



## Experiment-2

Aim: Explore weka Data Mining/Machine Learning Toolkit.

- I. Downloading and/or installation of WEKA data mining toolkit.
- II. Understanding features of WEKA toolkit such as explorer, knowledge flow interface, Experimenter, command-line interface.
- III. Navigate the options available in the WEKA (ex: select attributes panel, pre process panel, classify panel, associate panel and visualize panel).
- IV. Study the arff file format.
- V. Explorer the available data sets in WEKA.
- VI. Load a data set (ex: weather dataset, iris dataset, etc).
- VII. Load each dataset and observe the following:
  - i. List the attributes names and they types.
  - ii. Number of records in each dataset.
  - iii. identify the class attribute.
  - iv. plot histogram.
  - v. Determine the number of records for each class.
  - vi. Visualize the data in various dimensions.

Objective:

To understand the features of weka and to know about navigate options available in weka. And to load a data set and experiment the results of it.

Preprocessors:

The data that is collected from the field contains many unwanted things that lead to wrong details. Thus, the data must be preprocessed to meet the requirements of the type of analysis you are seeking. This is done in the preprocessing module.

Classifiers:

Classifiers in WEKA are the models for predicting nominal or numeric quantities. The learning schemes available in WEKA include decision trees and lists, instance-based classifiers, classifiers include bagging, boosting, stacking. error-correcting output codes are locally weighted learning

## WEKA: waikato Environment for Knowledge Analysis.

The WEKA GUI chooser provides a starting point for launching WEKA's main GUI applications and supporting tools. The GUI chooser consists of four buttons-one for each of the four major WEKA applications-and four menus.

Explorer: It is an environment for exploring data explorer consists of several tabs. They are

- preprocess: It is the first step in machine learning is to process the data. It is used to select data file, process it and make it fit for applying the various machine learning algorithms.
- classify: The classify tab provides you several machine learning algorithms for the classification of your data such as linear Regression, logistic Regression.
- cluster: Under the cluster tab there are several clustering algorithms provided. Such as simplex means, Filtered cluster, hierarchical cluster.
- Associate: Under the Associate tab you would find Apriori filtered Associate and FD Growth.
- Select Attributes: Select Attributes allows you feature selection based on several algorithms such as classifier, Subset eval, principal components.
- Visualize tab: The visualize option allows you to visualize your processed data for analysis.

## Simple CLI:

It provides a simple command line interface and allows direct execution of WEKA commands.

## Experimentor:

It is an environment for performing experiment and conduction statistical tests between learning schemes.

## knowledge Flow:

It is a java-beans based interface for setting up and running machine learning experiments.

## trees J48 classifier:

It is an algorithm to generate a decision tree that is generated by C4.5. It is also known as statistical classifier. For decision tree classification, we need a database.

## weather-Nominal:

In weka, attributes can be nominal or numeric. The value of a nominal attribute is represented by a word: sunny, overcast and rainy for the outlook attribute; yes and no for play attribute.

## arff file format:

An ARFF (=Attribute-Relation File Format) file is an ASCII text file that describes a list of sharing a set of attributes. ARFF files are not the only format one can load, but all files that can be converted with WEKA's "core converters".

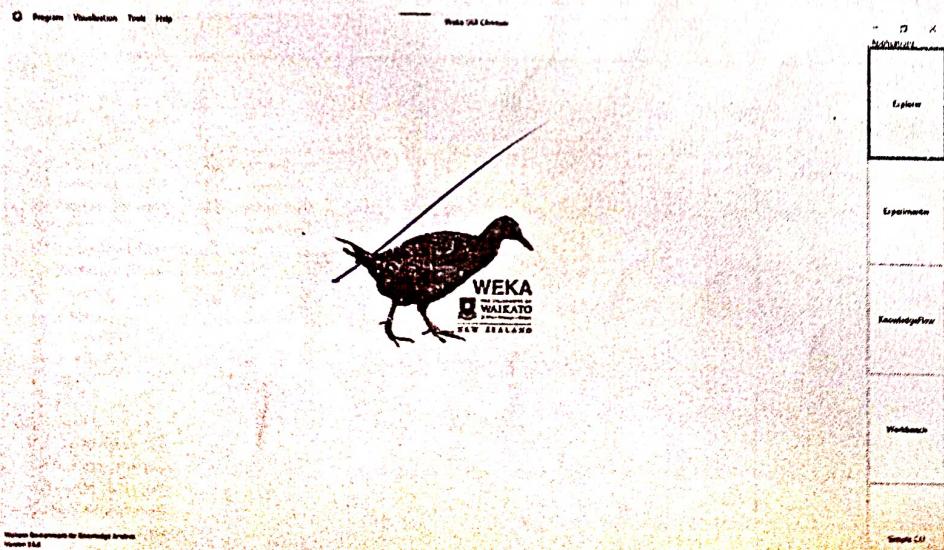


### Procedure:

- 1) Open WEKA. You can see 5 tabs on the right side of the application, those are: Explorer, Experimentor, Knowledge Flow, Workbench, Simple CLI.

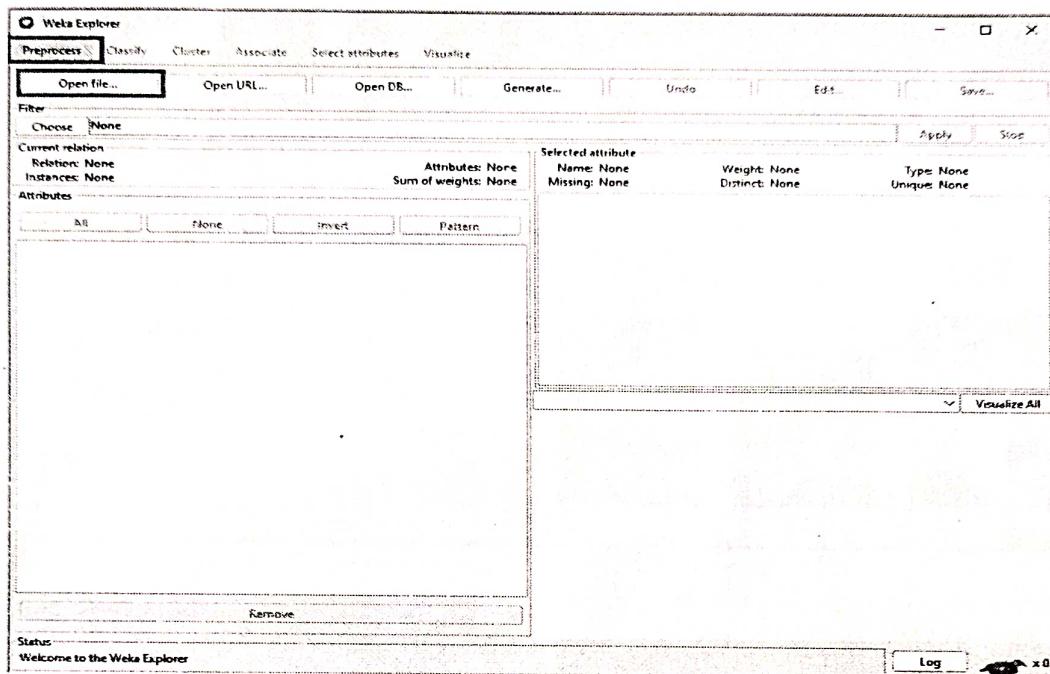


- 2) Click on Explorer.

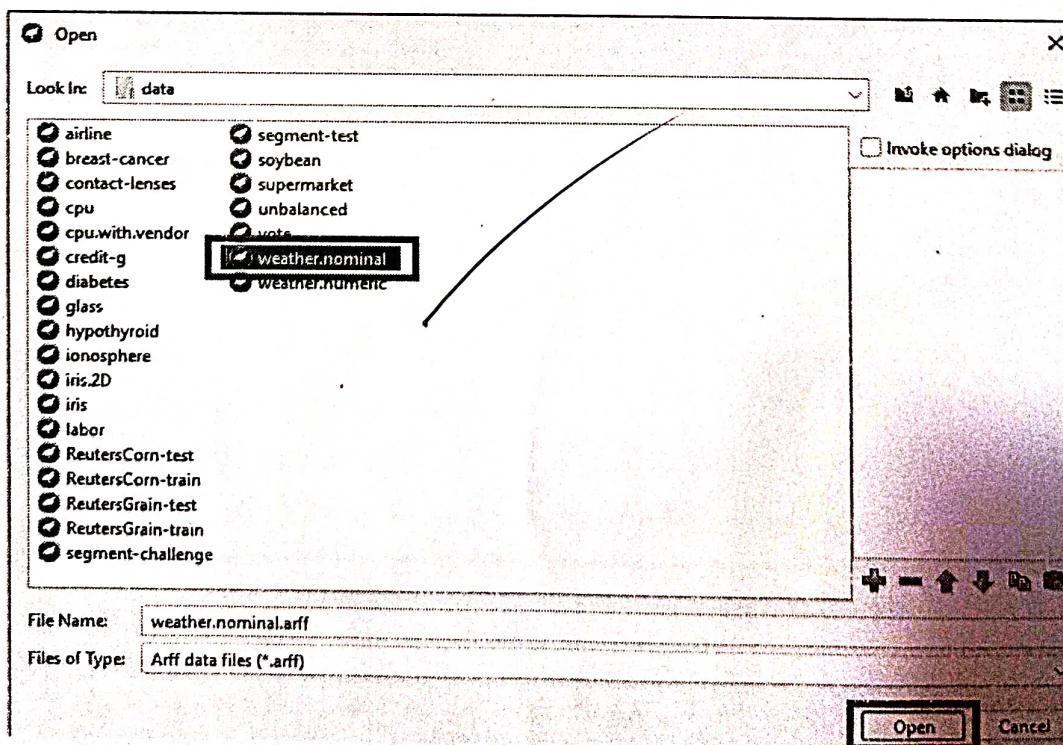


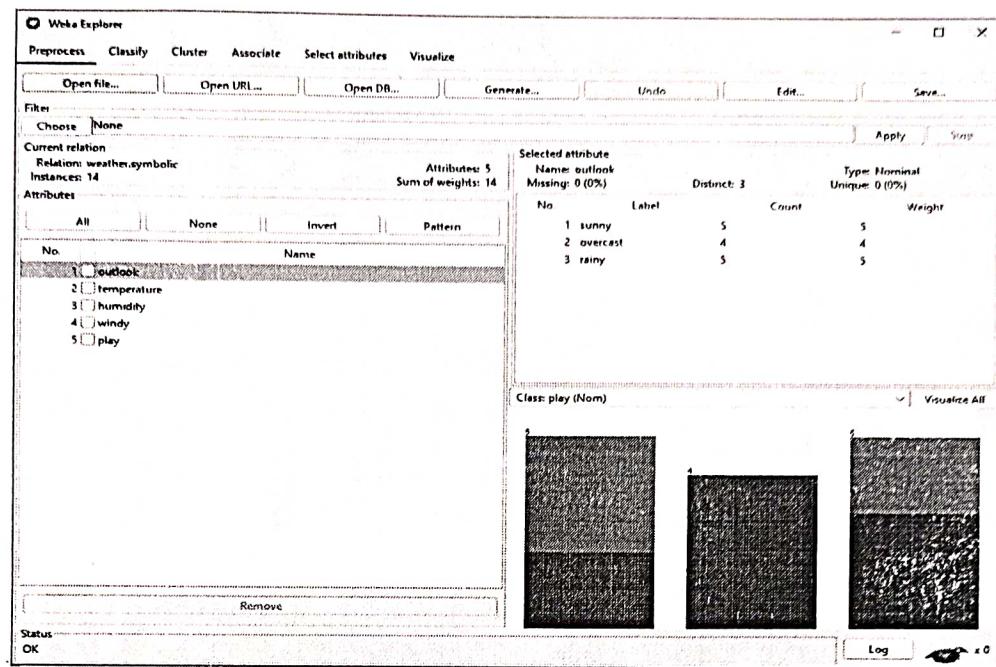


3) On preprocess, click on "Open file".

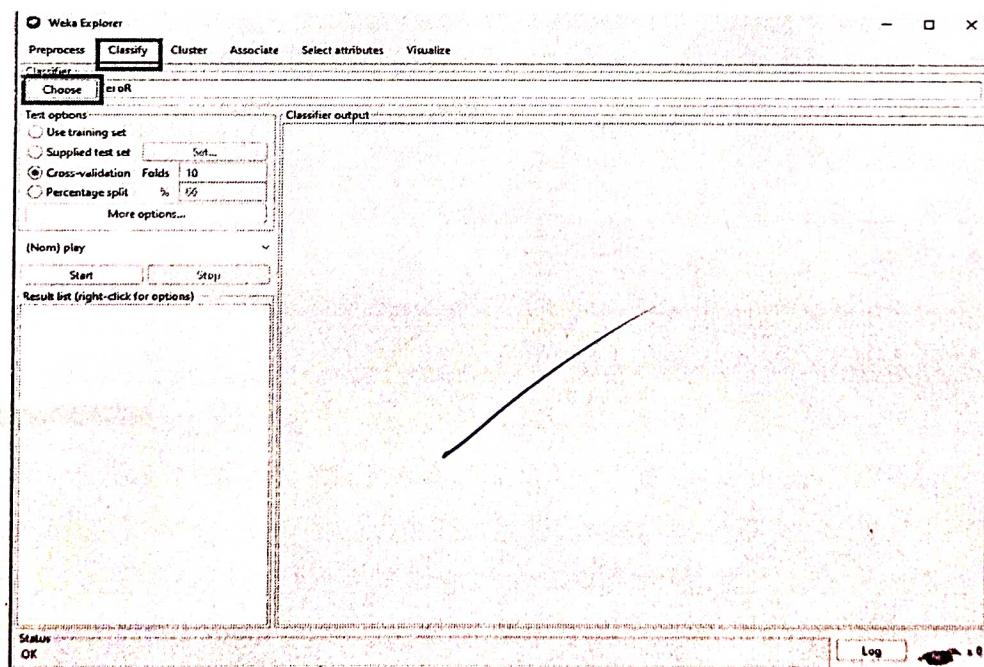


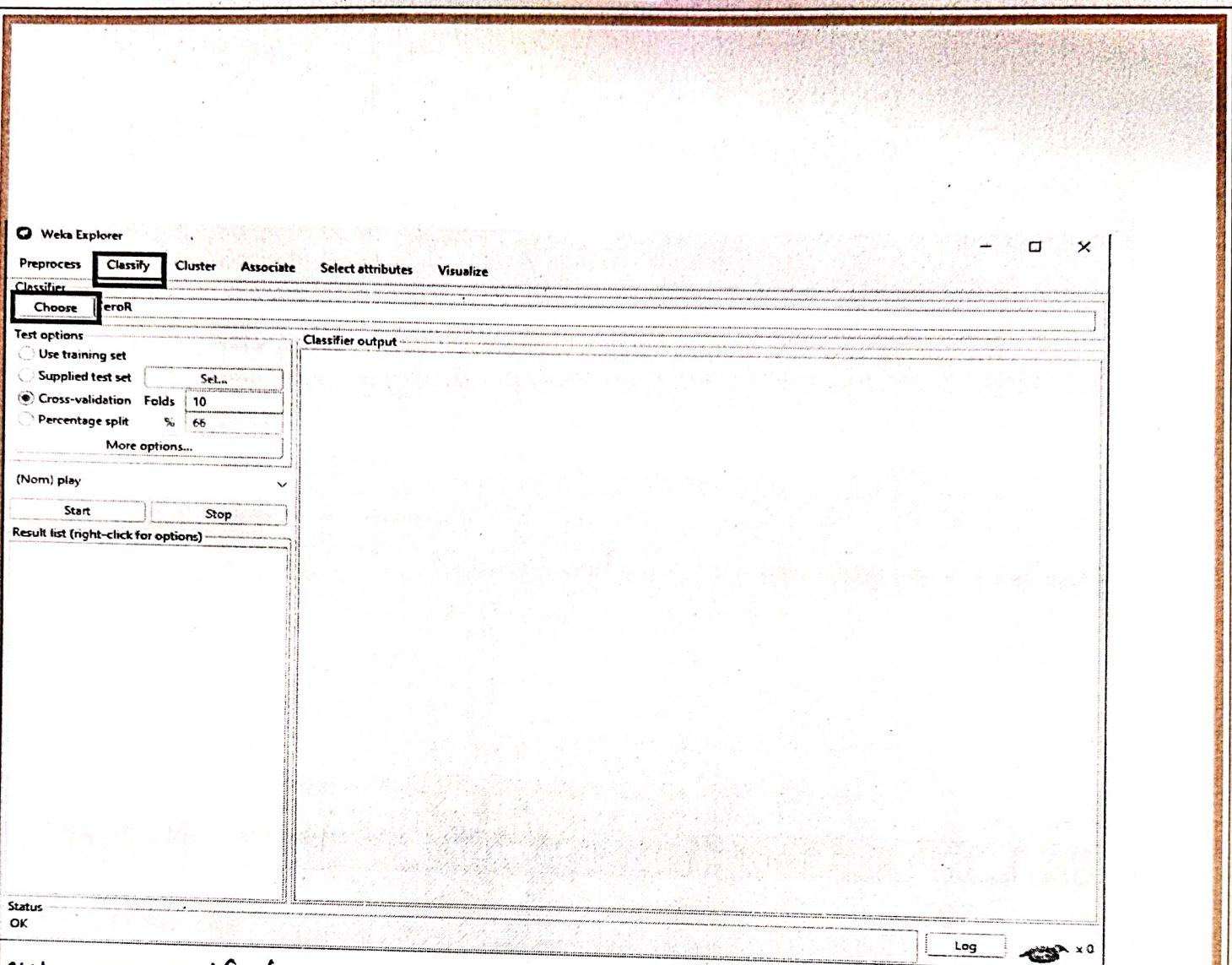
4) Go to "C:\Program Files\Weka-3-8-6\data". Select "weather.nominal.arff" and click on open.



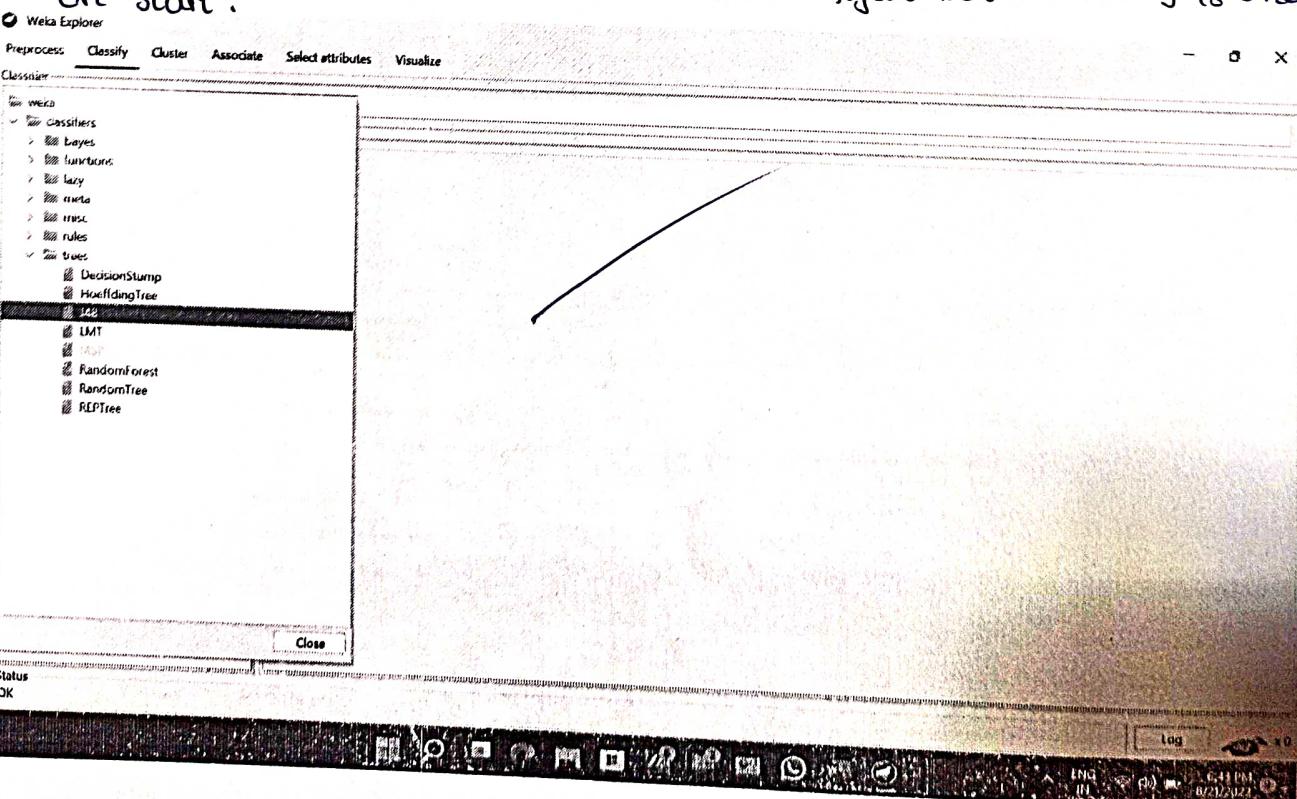


5) click on "classify" as shown in below figure and then click on choose.





6) You can see the following options as shown in figure below select j48 and click on "Start".





7) You will see the following

**WEKA Explorer**

Progress Classify Cluster Associate Select attributes Visualize

Classifier Choose J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set Set
- Cross-validation Folds 10
- Percentage split % 65
- More options...

Normal display

Start Stop

Result list (right-click for options) 165433 -trees.J48

**Classifier output**

Run information

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2  
 Relation: weather.symbolic  
 Instances: 14  
 Attributes: 5  
  
 outlook  
 temperature  
 humidity  
 windy  
 play  
 Test mode: 10-fold cross-validation

Classifier model (full training set)

J48 pruned tree

```

outlook = sunny
| humidity = high: no (3.0)
| humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
| windy = TRUE: no (2.0)
| windy = FALSE: yes (3.0)

Number of Leaves : 5
Size of the tree : 8
  
```

Status OK

8) Click on the resulted list to see the visual.

**WEKA Explorer**

Progress Classify Cluster Associate Select attributes Visualize

Classifier Choose J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set Set
- Cross-validation Folds 10
- Percentage split % 65
- More options...

Normal display

Start Stop

Result list (right-click for options) 165433 -trees.J48

**Classifier output**

Time taken to build model: 0 seconds.

Stratified cross-validation

Summary

	Correctly Classified Instances	7	50	%
Incorrectly Classified Instances	7	50	%	
Kappa statistic	-0.0426			
Mean absolute error	0.4167			
Root mean squared error	0.5064			
Relative absolute error	51.5 %			
Root relative squared error	121.2587 %			
Total Number of Instances	14			

Detailed Accuracy By class

	TP Rate	FP Rate	Precision	Recall	F-Measure	NCC	ROC Area	PRC Area	CLASS
0   1	0.556	0.800	0.625	0.556	0.588	-0.041	0.633	0.759	yes
1   0	0.400	0.444	0.333	0.400	0.364	-0.043	0.633	0.457	no
Weighted Avg.	0.500	0.544	0.521	0.500	0.508	-0.043	0.633	0.650	

Confusion Matrix

0 | 1 --> classified as  
 5 | 4 | 1 | a = yes  
 3 | 2 | b = no

*2 | 4 | 8 | 2*

Status OK

## Output:

By the end of the experiment, we learned the installation of WEKA tool and we have understood the features of WEKA tool. and we have learned how to load a dataset in WEKA.

## Experiment-4

Aim: Demonstrate performing classification on data.

- Load each dataset into weka and run ID3,J48 classification algorithm. Study the classified output. Compute the classifier output. Compute entropy values, kappa statistic.
- Extract if-then rules from the decision tree generated by the classifier, observe the confusion matrix.
- Load each dataset into weka and perform Naive-bays classification and K-Nearest Neighbour classification. Interpret the results obtained.
- Plot ROC Curves.
- Compare classification result of ID3,J48, Naive-Bayes and K-NN classifiers for each data set, and reduce which classifier is performing best and poor for each dataset and justifies.

Objective:

The ultimate objective of classification is to relate a variable of interest with observed variables. The actual variables of interest is meant to be of "Qualitative" type. The algorithm required for performing the classification is known as the classifier.

Zero-R:

- ZeroR is the simplest classification method which relies on the target and ignores all predictors.
- ZeroR classifier simply predicts the majority category.
- Although there is no predictability power in zeroR it is useful for determining a baseline performance as a benchmark for other classification methods.

One-R:

- This method is used in the sequential learning algorithm for learning the values.
- It returns a single rule that covers atleast some examples.
- However, what makes it really powerful, is its ability to create relations among the attributes given. Hence covering a larger hypothesis space.

Explorer: It is an environment for exploring data.

Simple CLI: It provides a simple command-line interface and allows direct execution of WEKA commands.

Experimentor: It is an environment for performing experiment and conducting statistical tests between learning schemes.

Knowledge Flow: It is an Java-Beans based interface to setting up and running machine learning experiments.



**Preprocess:** It is the first step in machine learning to preprocess the data. It is used to select the data file preprocessing and make it fit for applying the various machine learning algorithms.

**Classify:** The classify tab provides you several machine learning algorithms for the classification of your data. Such as Linear Regression, Logistic Regression.

**Test options:** Before you run the classification algorithm, you need to set test options. Set test options. Set test options in the Test options box. The test options that available are.

1. Use training set:

Evaluates the classifier on how well it predicts the class of the instances it was trained on.

2. Supplied test set:

Evaluates the classifier on how well it predicts the class of a set of instances it was trained on. Loaded from a file.

Clicking on the 'Set...' button brings up a dialog allowing you to choose the file to test on.

3. Cross Validation:

Evaluates the classifier by cross-validation, using the number of folds that are entered in the 'Folds' text field.

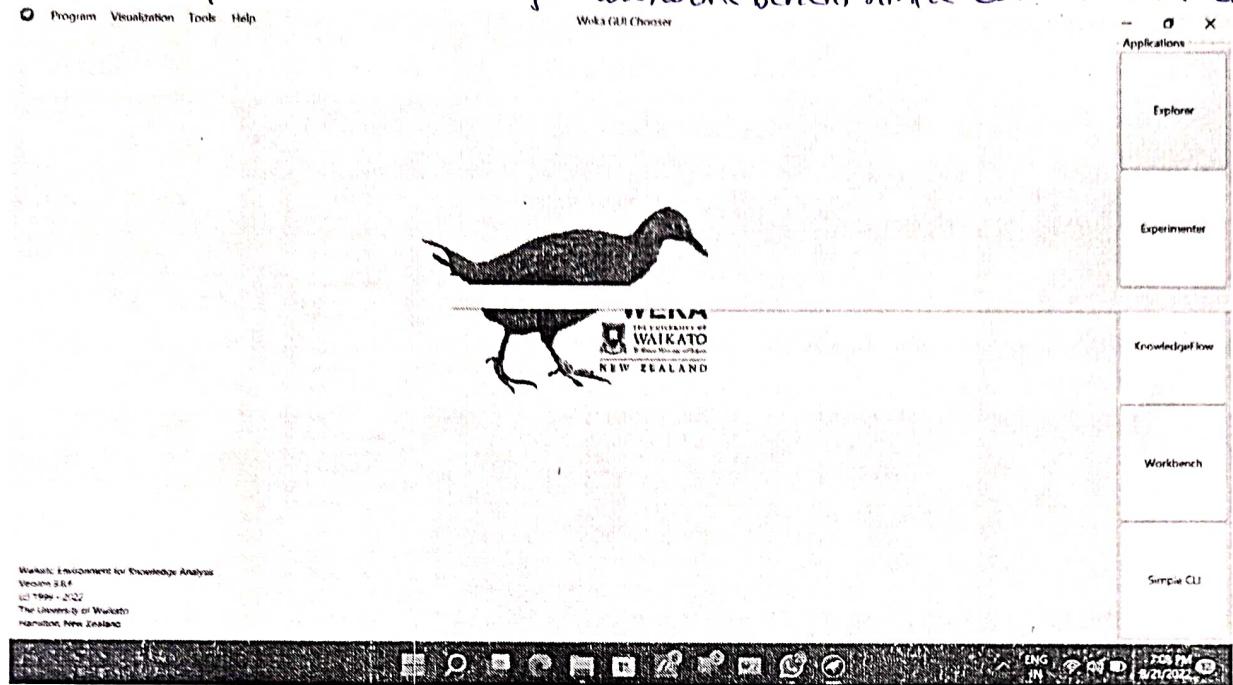
4. Percentage split:

Evaluates the classifier on how well it predicts a certain percentage of the data, which is held out for testing. The amount of data held out depends on the value entered in '%' field.

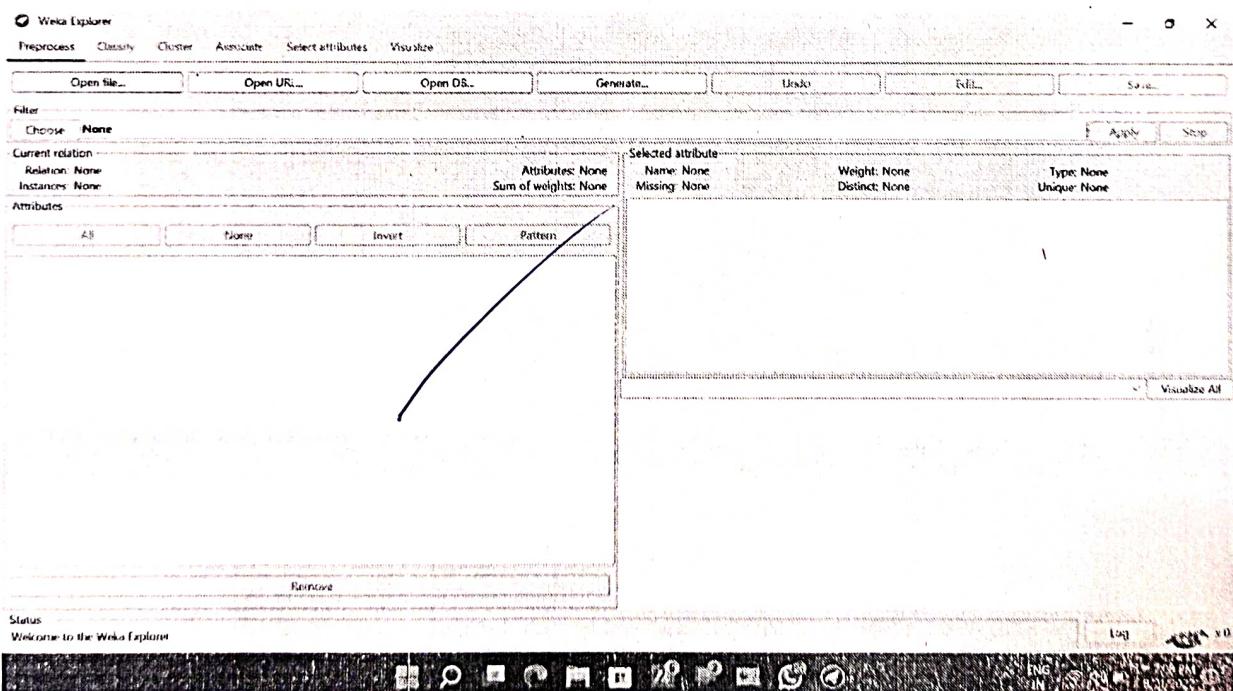


## procedure

- 1) Open WEKA You can see 5 tabs on the right side of the application. Those are explorer, experimentor, Knowledge Flow, work bench, simple CLI. Click on Explorer.



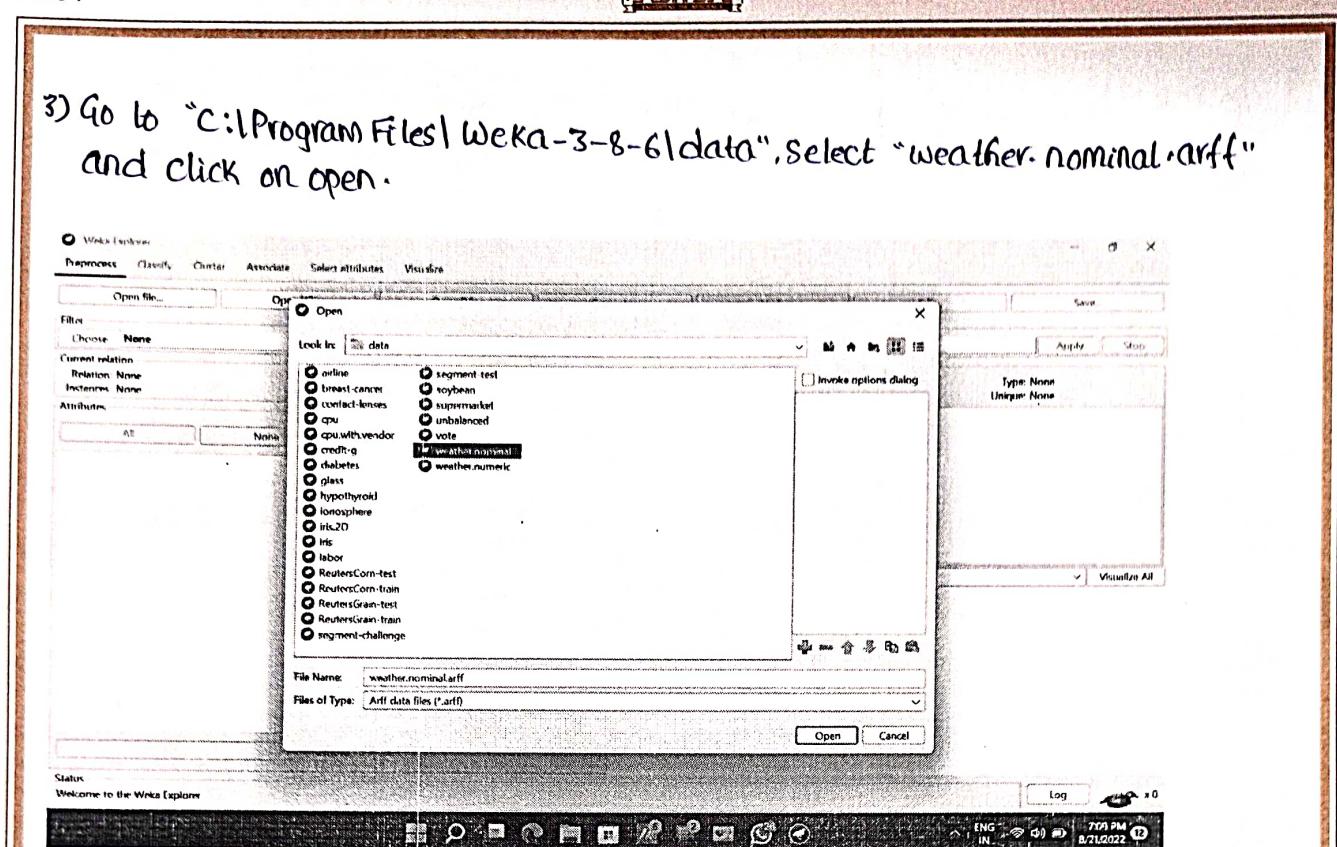
- 2) On preprocess click on "open file".



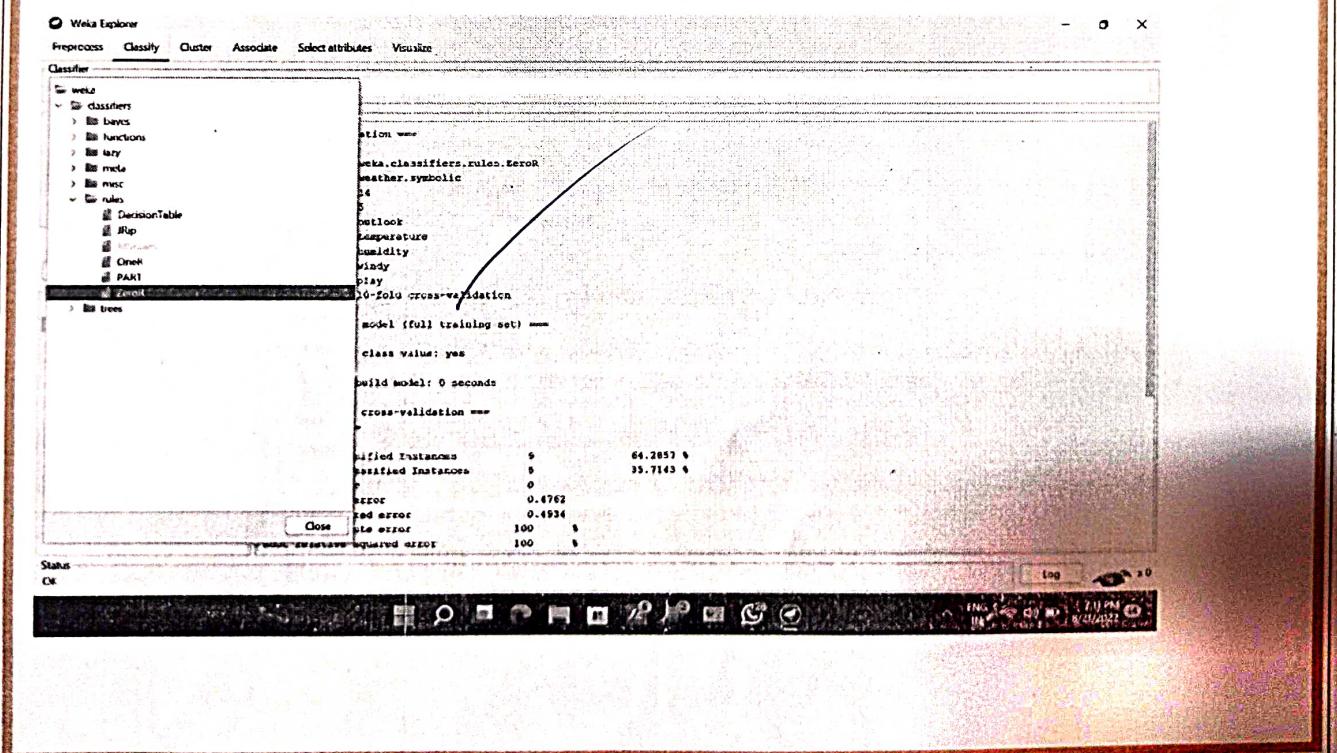
Date :



- 3) Go to "C:\Program Files\Weka-3-8-6\data", Select "weather.nominal.arff" and click on open.



- 4) You can observe choose zeroR . Click on it.





5) In test option you can see cross-validation folds. set it as 10. and click on "Start" button.

The screenshot shows the Weka Explorer interface with the 'Classifier' tab selected. Under 'Test options', 'Cross-validation' is chosen with 'Folds' set to 10. The 'Start' button is highlighted. The 'Classifier output' pane displays the following information:

```

Classifier output
--- Run information ---
Scheme: weka.classifiers.rules.ZeroR
Relation: weather.symbolic
Instances: 14
Attributes: 4
    outlook
    temperature
    humidity
    windy
    play
Test mode: 10-fold cross-validation
--- Classifier model (full training set) ---
Error predicts class value: yes
Time taken to build model: 0 seconds
--- Stratified cross-validation ---
--- Summary ---
Correctly Classified Instances      9      64.2857 %
Incorrectly Classified Instances   5      35.7143 %
Kappa statistic                   0
Mean absolute error               0.4762
Root mean squared error           0.4934
Relative absolute error            100 %
Root relative squared error       100 %
Total Number of Instances         14

```

The status bar at the bottom shows the date and time: 8/21/2022 7:27 PM.

6) You can scroll down in the classifier output and see the mean absolute error, Root mean squared error etc.

The screenshot shows the Weka Explorer interface with the 'Classifier' tab selected. Under 'Test options', 'Cross-validation' is chosen with 'Folds' set to 10. The 'Start' button is highlighted. The 'Classifier output' pane displays the following information:

```

Classifier output
--- Run information ---
Time taken to build model: 0 seconds
--- Stratified cross-validation ---
--- Summary ---
Correctly Classified Instances      9      64.2857 %
Incorrectly Classified Instances   5      35.7143 %
Kappa statistic                   0
Mean absolute error               0.4762
Root mean squared error           0.4934
Relative absolute error            100 %
Root relative squared error       100 %
Total Number of Instances         14
--- Detailed Accuracy by Class ---
          TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PNC Area Class
        1.000  1.000  0.643  1.000  0.783  7     0.170  0.555  yes
        0.000  0.000  0.000  0.000  0.000  7     0.178  0.318  no
Weighted Avg.  0.643  0.643  0.643  0.643  0.643  7     0.178  0.470
--- Confusion Matrix ---
4 0 -- classified as
9 0 ; 1 -- yes
5 0 ; 1 -- no

```

The status bar at the bottom shows the date and time: 8/21/2022 7:27 PM.



7) Click on 'choose' and select 'OneR' and set cross Validation folds to 10 and click on "start" button."

**Weka Explorer**

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose **OneR - E6**

**Test options**

- Use training set
- Supplied test set
- Cross-validation Folds
- Percentage split
- More options...

(Normal play)

**Start Stop**

Result list (right-click for options)

191231 - rules.ZeroR  
191324 - rules.OneR

**Classifier output**

```
==> Run information ==>
Scheme: weka.classifiers.rules.OneR - E 6
Relation: weather.symbolic
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
Test mode: 10-fold cross-validation

==> Classifier model (full training set) ==>
outlook:
  sunny -> no
  overcast-> yes
  rainy -> yes
(10/14 instances correct)

Time taken to build model: 0.01 seconds

==> Stratified cross-validation ==>
==> Summary ==>
Correctly Classified Instances       6      42.8571 %
Incorrectly Classified Instances     8      57.1429 %

```

**Status**

OK

**Weka Explorer**

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose **OneR - E6**

**Test options**

- Use training set
- Supplied test set
- Cross-validation Folds
- Percentage split
- More options...

(Normal play)

**Start Stop**

Result list (right-click for options)

191231 - rules.ZeroR  
191324 - rules.OneR

**Classifier output**

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                 0.5714
Root mean squared error             0.7859
Relative absolute error              120 %
Root relative squared error         151.2194 %
Total Number of Instances           14

==> Detailed Accuracy By Class ==>

```

TP Rate	FP Rate	Precision	Recall	F-Measure	NCC	ROC Area	PRJ Area	Classid
0.444	0.600	0.371	0.444	0.500	-0.149	0.422	0.611	yes
0.400	0.556	0.286	0.400	0.333	-0.149	0.422	0.329	no
Weighted Avg.	0.429	0.584	0.409	0.429	0.449	0.422	0.510	

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

**Status**

OK

Output:

By the end of the experiment we have experimented with zeroR, OneR algorithm and we observed the difference between these two algorithms.