

EXERCISE -7

- (i) Rank the performance of j48, PART and oneR Algorithms on 'Weather' dataset using Experimenter.
- (ii) Build a Knowledge Flow model using Weka tool.

7.1.1 Problem Statement:

Rank the performance of j48, PART and oneR Algorithms on 'Weather' dataset using Experimenter.

7.1.2 Description:

About Dataset used

The weather data is a small [open data set](#) with only 14 examples.

In RapidMiner it is named [Golf Dataset](#), whereas Weka has two data set: weather.nominal.arff and weather.numeric.arff

The dataset contains data about weather conditions are suitable for playing a game of [golf](#). the original dataset that only has 5 variables:

- 1.outlook
- 2.temperature
- 3.humidity
- 4.windy
- 5.play

About Arff:

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes.

ARFF files have two distinct sections. The first section is the **Header** information, which is followed the **Data** information.

The **Header** of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types.

Lines that begin with a % are comments. The **@RELATION**, **@ATTRIBUTE** and **@DATA** declarations are case insensitive.

About CSV:

Files with .csv (Comma Separated Values) extension represent plain text files that contain records of data with comma separated values. Each line in a CSV file is a new record from the set of records contained in the file. Such files are generated when data transfer is intended from one storage system to another. Since all applications can recognize records separated by comma, import of such data files to database is done very conveniently. Almost all spreadsheet applications such as Microsoft Excel or OpenOffice Calc can import CSV without much effort. Data imported from such files is arranged in cells of a spreadsheet for representation to user.

Datatypes that are supported by Weka:

- numeric
- integer is treated as numeric
- real is treated as numeric
- [nominal-specification]
- string
- date [date-format]
- relational for multi-instance data (for future use)

where [nominal-specification] and [date-format] are defined below. The keywords **numeric**, **real**, **integer**, **string** and **date** are case insensitive.

Numeric attributes

Numeric attributes can be real or integer numbers.

Nominal attributes

Nominal values are defined by providing an [nominal-specification] listing the possible values: {[nominal-name1], [nominal-name2], [nominal-name3], ...}

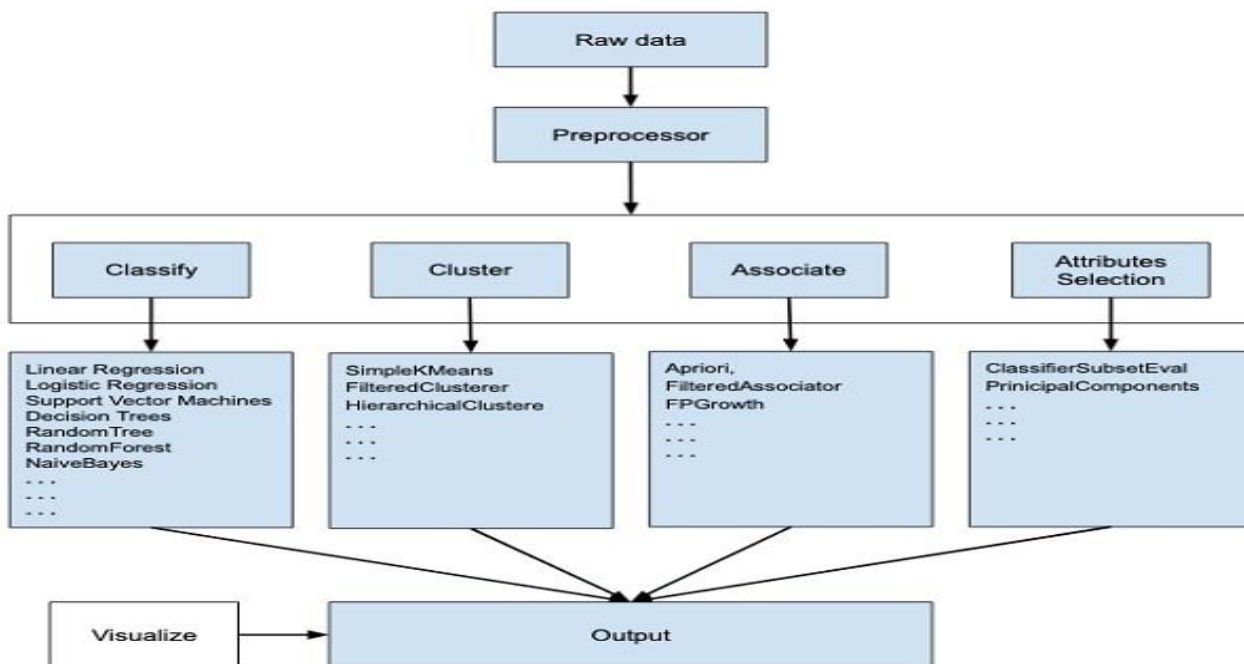
For example, the class value of the Iris dataset can be defined as follows:

```
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Values that contain spaces must be quoted.

About WEKA Software:

WEKA - an opensource software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. What WEKA offers is summarized in the following diagram –



If you observe the beginning of the flow of the image, you will understand that there are many stages in dealing with Big Data to make it suitable for machine learning

First, you will start with the raw data collected from the field. This data may contain several null values and irrelevant fields. You use the data preprocessing tools provided in WEKA to cleanse the data.

Then, you would save the preprocessed data in your local storage for applying ML algorithms.

Next, depending on the kind of ML model that you are trying to develop you would select one of the options such as **Classify**, **Cluster**, or **Associate**. The **Attributes Selection** allows the automatic selection of features to create a reduced dataset.

Note that under each category, WEKA provides the implementation of several algorithms. You would select an algorithm of your choice, set the desired parameters and run it on the dataset.

Then, WEKA would give you the statistical output of the model processing. It provides you a visualization tool to inspect the data.

The various models can be applied on the same dataset. You can then compare the outputs of different models and select the best that meets your purpose.

Thus, the use of WEKA results in a quicker development of machine learning models on the whole.

Now that we have seen what WEKA is and what it does, in the next chapter let us learn how to install WEKA on your local computer

7.1.3 Steps to implement the classification:

1.To Rank the performance of different algorithm on Weather dataset using Experiment.

So first open the file by using the **Open file ...** option and select the **weather-numeric.arff** file.

The screenshot displays the WEKA 'Select attributes' window. The 'Current relation' is 'weather' with 14 instances and 5 attributes. The 'Attributes' list on the left includes outlook, temperature, humidity, windy, and play. The 'Selected attribute' panel on the right shows details for 'outlook', including a table of counts for sunny (5), overcast (4), and rainy (5). Below this, a bar chart visualizes the distribution of the 'play' class (Nom) for each outlook category, with red bars for 'yes' and blue bars for 'no'.

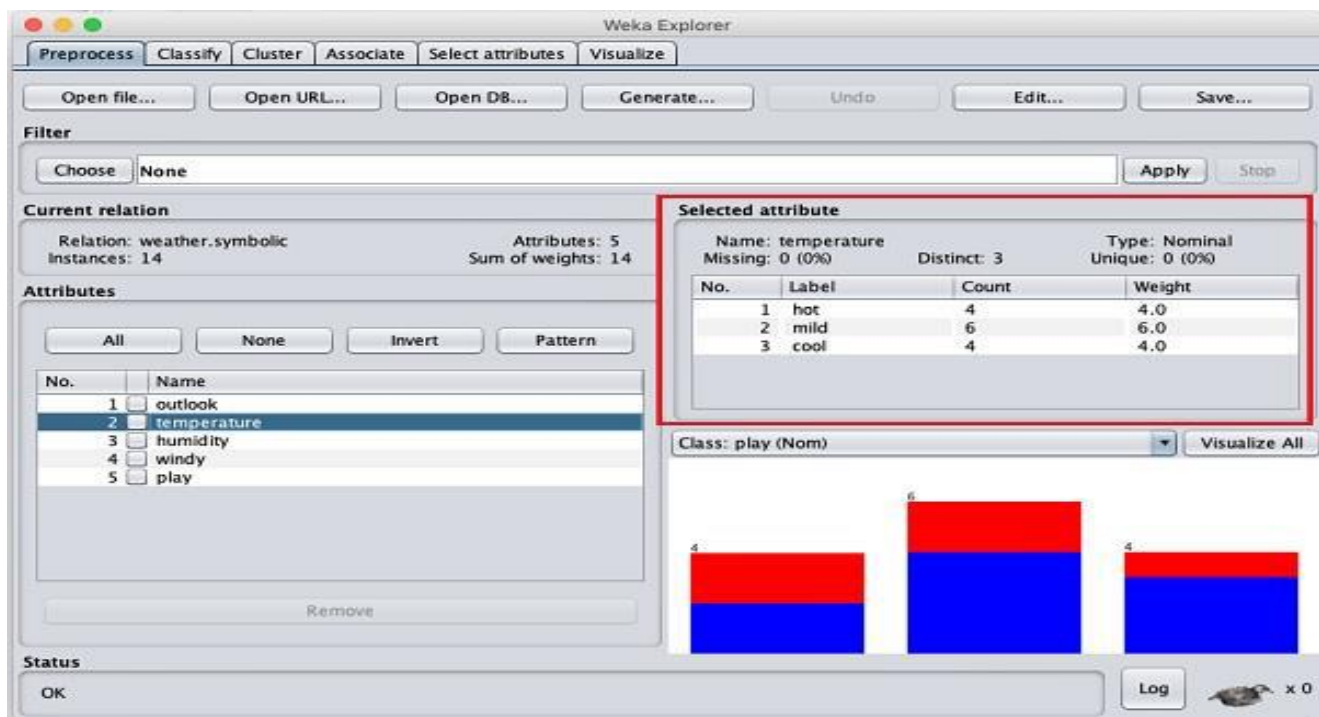
No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

Class: play (Nom)

Visualize All

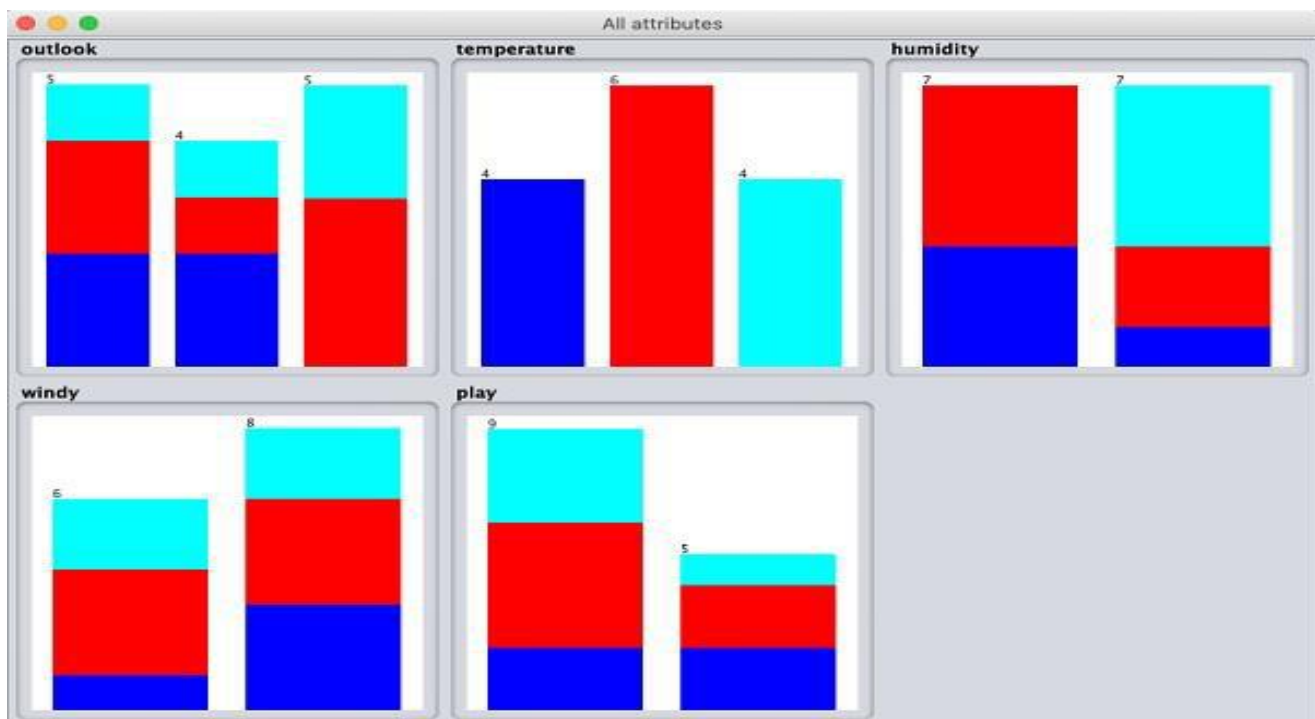
2. The **weather** database contains five fields - outlook, temperature, humidity, windy and play. When you select an attribute from this list by clicking on it, further details on the attribute itself are displayed on the right hand side.

Let us select the temperature attribute first. When you click on it, you would see the following screen –



At the bottom of the window, you see the visual representation of the **class** values.

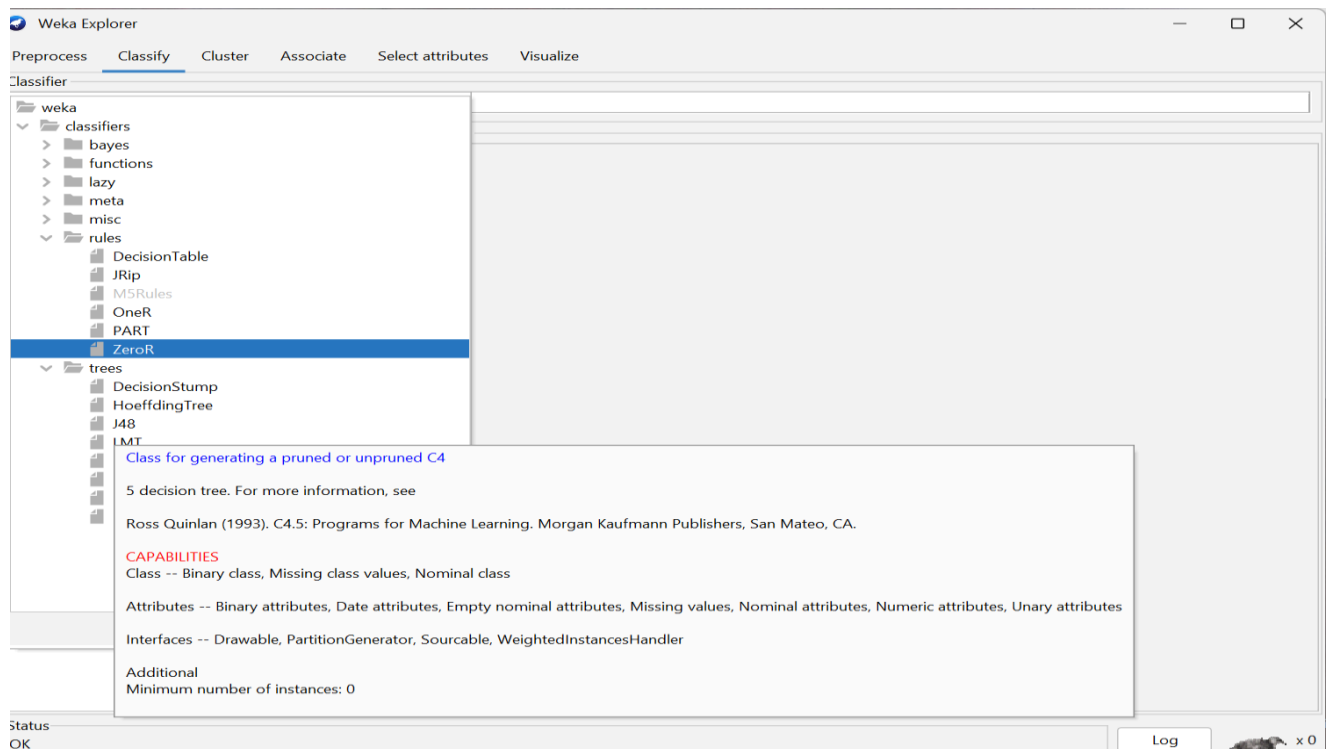
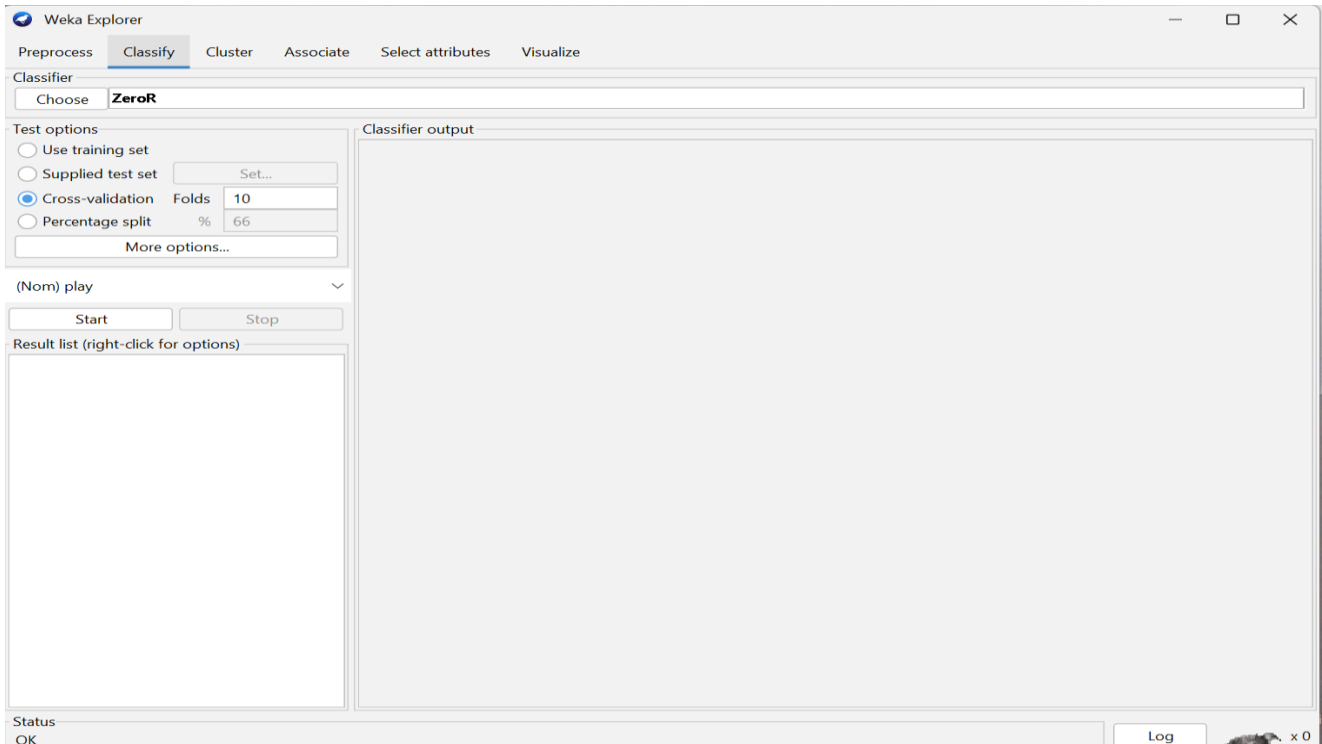
If you click on the **Visualize All** button, you will be able to see all features in one single window as shown here



3. Now to Rank the performance of j48, PART and oneR Algorithms on 'Weather' dataset first we have to classify the dataset using these algorithms.

4. First we will classify the dataset using j48 algorithm.

To do that click on classify and choose j48 Tree-based algorithm and click on start to get the classifier output.



Weka Explorer

Preprocess **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 % **66**

More options...

(Norm) play

Start Stop

Result list (click to select one or more)

23:08:41 Starts the classification

Classifier output

Size of the tree : 8

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	0.186	
Mean absolute error	0.2857	
Root mean squared error	0.4818	
Relative absolute error	60 %	
Root relative squared error	97.6586 %	
Total Number of Instances	14	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.778	0.600	0.700	0.778	0.737	0.189	0.789	0.847	yes
	0.400	0.222	0.500	0.400	0.444	0.189	0.789	0.738	no
Weighted Avg.	0.643	0.465	0.629	0.643	0.632	0.189	0.789	0.808	

=== Confusion Matrix ===

a b <-- classified as

7 2 | a = yes

3 2 | b = no

Status OK

Log x0

5. Now we will classify the dataset using PART algorithm.

Under classify choose PART rule-based algorithm and click on start to get the classifier output.

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose **PART**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds **10**

☐ Percentage split % **66**

More options...

(Norm) play

Start Stop

Result list (click to select one or more)

23:08:41 Starts the classification

Classifier output

Size of the tree : 8

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	0.186	
Mean absolute error	0.2857	
Root mean squared error	0.4818	
Relative absolute error	60 %	
Root relative squared error	97.6586 %	
Total Number of Instances	14	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.778	0.600	0.700	0.778	0.737	0.189	0.789	0.847	yes
	0.400	0.222	0.500	0.400	0.444	0.189	0.789	0.738	no
Weighted Avg.	0.643	0.465	0.629	0.643	0.632	0.189	0.789	0.808	

=== Confusion Matrix ===

a b <-- classified as

7 2 | a = yes

3 2 | b = no

Status OK

Log x0

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose **PART -C 0.25 -M 2**

Test options:

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds %
- ☐ Percentage split

(Norm) play

Result list (right-click to start the classification)

- 23:08:48 - trees.J48
- 23:10:16 - rules.PART

Classifier output

```

=== Run information ===

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

=== Classifier model (full training set) ===

PART decision list
-----

outlook = overcast: yes (4.0)

windy = TRUE: no (4.0/1.0)

outlook = sunny: no (3.0/1.0)


: yes (3.0)

Number of Rules :      4

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

```

Status: OK  x0

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose **PART -C 0.25 -M 2**

Test options:

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds %
- ☐ Percentage split

(Norm) play

Result list (right-click to start the classification)

- 23:08:48 - trees.J48
- 23:10:16 - rules.PART

Classifier output

```

Number of Rules :      4

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      5           35.7143 %
Incorrectly Classified Instances    9           64.2857 %
Kappa statistic                    -0.3404
Mean absolute error                 0.5518
Root mean squared error             0.6935
Relative absolute error             115.875 %
Root relative squared error         140.5649 %
Total Number of Instances          14


=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          0.444    0.800    0.500    0.444    0.471     -0.344   0.522    0.792    yes
          0.200    0.556    0.167    0.200    0.182     -0.344   0.522    0.385    no
Weighted Avg.   0.357    0.713    0.381    0.357    0.367     -0.344   0.522    0.647

=== Confusion Matrix ===

a b  <-- classified as
4 5 | a = yes
4 1 | b = no

```

Status: OK  x0

6. Now we will classify the dataset using oneR algorithm.

Under classify choose oneR rule-based algorithm and click on start to get the classifier output.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

weka

- classifiers
 - bayes
 - functions
 - lazy
 - meta
 - misc
 - rules
 - DecisionTable
 - JRip
 - MSRules
 - OneR**
 - PART
 - ZeroR
 - trees

build model: 0 seconds

cross-validation ==

	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error
5	35.7143 %	64.2857 %	-0.3404	0.5518	0.6935
9				115.875 %	140.5649 %

Accuracy By Class ==

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.444	0.800	0.500	0.444	0.471	-0.344	0.522	0.792	yes
0.200	0.556	0.167	0.200	0.182	-0.344	0.522	0.385	no
0.357	0.713	0.381	0.357	0.367	-0.344	0.522	0.647	

Matrix ==

classified as

4 1 | b = no

Status OK

Log x0

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **OneR -B 6**

Test options

☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds **10**

☐ Percentage split % **66**

More options...

(Nom) play

Start Stop

Result list (right-click)

- 23:08:48 - trees.J48
- 23:10:16 - rules.PART
- 23:11:16 - rules.OneR**

Starts the classification

Classifier output

=== Run information ===

Scheme: weka.classifiers.rules.OneR -B 6

Relation: weather

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 seconds

=== Stratified cross-validation ===

=== Summary ===

	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error
6	42.8571 %	57.1429 %	-0.2444	0.5714	0.7559

Status OK

Log x0

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose OneR - 8 6

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) play

Start Stop

Result list (right-click to start the classification)

- 23:08:48 - trees.J48
- 23:10:16 - rules.PART
- 23:11:16 - rules.OneR

Classifier output

rainy -> yes
(10/14 instances correct)

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	6	42.8571 %
Incorrectly Classified Instances	8	57.1429 %
Kappa statistic	-0.2444	
Mean absolute error	0.5714	
Root mean squared error	0.7559	
Relative absolute error	120 %	
Root relative squared error	153.2194 %	
Total Number of Instances	14	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.556	0.800	0.556	0.556	0.556	-0.244	0.378	0.594	yes
	0.200	0.444	0.200	0.200	0.200	-0.244	0.378	0.326	no
Weighted Avg.	0.429	0.673	0.429	0.429	0.429	-0.244	0.378	0.498	

=== Confusion Matrix ===

```

a b  <-- classified as
5 4 | a = yes
4 1 | b = no

```

Status OK

Log x0

7. We now completed to classify the data using given algorithm

Now to rank the performance of all algorithms we use classifier output and analysis the output.

7.1.4 Results and Discussion:

We performed the j48, PART and oneR algorithms successfully on the weather dataset. Based on the results, we can say that the j48 algorithm worked best among all three with 64.2% accuracy. The oneR algorithm ranked second with 42.8% accuracy and the PART algorithm is ranked last as it is the least performing among all three with 35.7% accuracy.