# TABLE OF CONTENTS

- 1. INTRODUCTION
  - 1.1 Weka Introduction
  - 1.2KDD Process
  - 1.3Installation of Weka

# 2. LAUNCHING WEKA EXPLORER

- 2.1 Starting with Weka
- 2.2Pre-processing
- 2.3Loading the Data
- 2.4Setting Filters
- 3. <u>CLASSIFIER</u>
  - 3.1Building classifiers
  - 3.2Setting Test Options

# 4. <u>CLUSTERING</u>

- 4.1Clustering Data
- 4.2Choosing Clustering Scheme
- **4.3Setting Test Options**
- 4.4 Visualization of Results

# 5. <u>ASSOCIATIONS</u>

- 5.1 Finding Associations
- 5.2Setting Test Options
- 6. <u>ATTRIBUTE SELECTION</u>
  - 6.1 Introduction
  - **6.2Selecting Options**
- 7. DATA VISUALIZATION
  - 7.1 Introduction

- 7.2Changing the view
- 7.3Selecting instances
- 8. <u>CONCLUSION</u>
- 9. <u>REFERENCES</u>

# CHAPTER 1: INTRODUCTION TO WEKA

## **1.1 Introduction:**

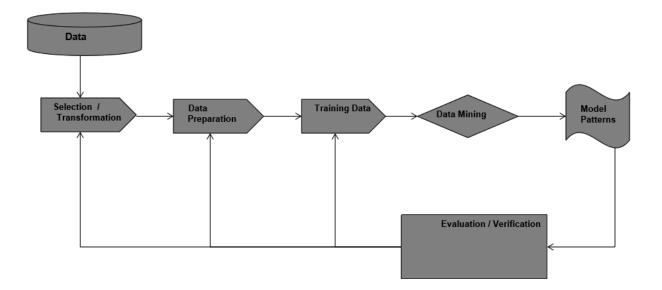
### What is WEKA?

WEKA, formally called **Waikato Environment for Knowledge Learning**, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from different domains.

WEKA supports many different standard data mining tasks such as data pre-processing, classification, clustering, regression, visualization and feature selection. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns.

WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation.

WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces.



### **1.2 KDD Process:**

### **1.3 Installation of Weka:**

The weka can be explored from the different sites, one of the sites is http://www.cs.waikato.ac.nz/m1/weka/downloading.html



# Downloading and installing Weka

☆

There are two versions of Weka: Weka 3.8 is the latest stable version, and Weka 3.9 is the development version. For the bleeding edge, it is also possible to download nightly snapshots.

Stable versions receive only bug fixes, while the development version receives new features. Weka 3.8 and 3.9 feature a package management system that makes it easy for the Weka community to add new functionality to Weka. The package management system requires an internet connection in order to download and install packages.

Note (1) for users upgrading from Weka 3.7 to Weka 3.8 or later: if the Weka 3.8 package manager does not start up, please delete the file installedPackageCache.ser in the packages folder that resides in the wekafiles folder in your user home.

Note (2) for users upgrading from Weka 3.7 to Weka 3.8 or later: serialized models created in 3.7 are not compatible with 3.8. We have a **model migrator** tool that can migrate some models to be compatible with 3.8.0. One exception is RandomForest, which can be migrated up to 3.7.13 but no further. Usage is as follows:

#### There are different options to launch weka depending the operating systems

Windows

Click here to download a self-extracting executable for 64-bit Windows that includes Oracle's 64-bit Java VM 1.8 (weka-3-8-0jre-x64.exe; 105.5 MB)

Click **here** to download a self-extracting executable for 64-bit Windows without a Java VM (weka-3-8-0-x64.exe; 50.2 MB)

Click **here** to download a self-extracting executable for 32-bit Windows that includes Oracle's 32-bit Java VM 1.8 (weka-3-8-0jre.exe; 100.8 MB)

Click **here** to download a self-extracting executable for 32-bit Windows without a Java VM (weka-3-8-0.exe; 50.2 MB)

These executables will install Weka in your Program Menu. Download the version without the Java VM if you already have Java 1.7 (or later) on your system.

• Mac OS X

Click here to download a disk image for OS X that contains a Mac application including Oracle's Java 1.8 JVM (weka-3-8-0-oracle-jvm.dmg; 125.8 MB)

#### Other platforms (Linux, etc.)

Click **here** to download a zip archive containing Weka (weka-3-8-0.zip; 50.6 MB)

First unzip the zip file. This will create a new directory called weka-3-8-0. To run Weka, change into that directory and type

java -jar weka.jar

Note that Java needs to be installed on your system for this to work. Also note, that using -jar will override your current CLASSPATH variable and only use the weka.jar.

#### • Windows x86

Click here to download a self-extracting executable that includes Java VM 1.8 (weka-3-9-0jre.exe; 100.7 MB)

Click **here** to download a self-extracting executable without the Java VM (weka-3-9-0.exe; 50.1 MB)

These executables will install Weka in your Program Menu. Download the second version if you already have Java 1.7 (or later) on your system.

• Windows x64

Click here to download a self-extracting executable that includes 64 bit Java VM 1.8 (weka-3-9-0jre-x64.exe; 105.4 MB)

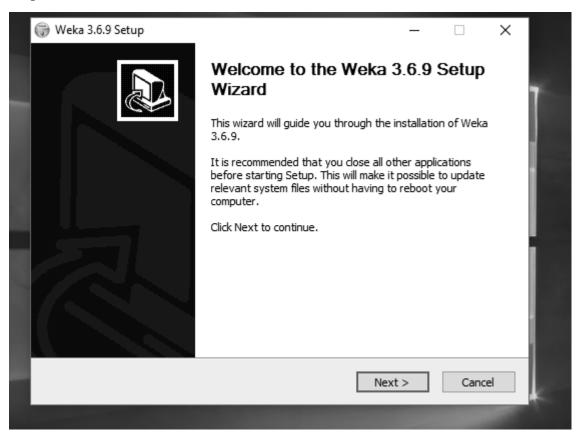
Click **here** to download a self-extracting executable without the Java VM (weka-3-9-0-x64.exe; 50.1 MB)

These executables will install Weka in your Program Menu. Download the second version if you already have Java 1.7 (or later) on your system.

Depending on the version click on the down load option. When we click on the download option setup of weka gets downloaded.

Click on setup and follow the below steps

#### Step 1:



#### Click on Next button

Step 2:

🗑 Weka 3.6.9 Setup			_		×	
Weko	License Agreem					
	Please review the	license terms befo	re installing We	eka 3.6.9.		
Press Page Down to see	the rest of the agreem	ent.				
G	NU GENERAL PUBLIC L	ICENSE		_	~	
Version 2, June 1991						
Copyright (C) 1989, 1991 Free Software Foundation, Inc.						
59 Temple Place, Suite 330, Boston, MA 02111-1307, USA.						
Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.						
	Preamble					
					<b>~</b>	
If you accept the terms agreement to install Wek		I Agree to continue	e. You must ac	cept the		
Jullsoft Install System v2,4	5					
vanoore 11150an by 500111 V21-11						
		< Back	I Agree	Cance	3	

click on I Agree option

Step 3:

Weka	Choose Components Choose which features of Weka 3	3.6.9 you want to install.
Check the components you install. Click Next to continu	want to install and uncheck the com e.	ponents you don't want to
Select the type of install:	Full 🗸	
Or, select the optional components you wish to install:	Associate Files	Description Position your mouse over a component to see its description.
Space required: 63.2MB		
Nullsoft Install System v2.46 –		
	< Back	Next > Cancel

# Click on Next Option

IIII - I	Choose Install Location
Weka	Choose the folder in which to install Weka 3.6.9.
	.6.9 in the following folder. To install in a different folder, click Browse
and select another folde	er. Click Next to continue.
Destination Folder	
Destination Folder	86)\Weka-3-6 Browse
C:\Program Files (x8	3
Space required: 63.2MB	3B

# Click on Next option

Weko	Choose Start M			
ovenue	Choose a Start	Menu folder for t	ne Weka 3.6.9 sl	nortcuts.
Select the Start Menu fo can also enter a name to			he program's sho	ortcuts. You
Weka 3.6.9				
Accessibility				^
Accessories Administrative Tools				
Bloodshed Dev-C++				
Broadcom Wireless CutePDF				
Energy Star				
Games HP				
Intel				
Java				*
Do not create shortc				
llsoft Install System v2.4	·6			
				Cancel

Click on install button. It extracts all the packages

Weka	Setup was completed successfully.	6	
Completed			
Create shortcut: C: \Pro Create shortcut: C: \Pro Created uninstaller: C: Output folder: C: \User Create shortcut: C: \Use Create shortcut: C: \Use Create shortcut: C: \Use Create shortcut: C: \Use	ram Files (x86)\Weka-3-6 ogram Files (x86)\Weka-3-6\Weka 3.6.9 ogram Files (x86)\Weka-3-6\Weka 3.6.9 \Program Files (x86)\Weka-3-6\uninstall s\Dhruva\AppData\Roaming\Microsoft\V ers\Dhruva\AppData\Roaming\Microsoft ers\Dhruva\AppData\Roaming\Microsoft ers\Dhruva\AppData\Roaming\Microsoft ers\Dhruva\AppData\Roaming\Microsoft	9 (with console).lnk .exe Windows\Start Menu t\Windows\Start Me t\Windows\Start Me t\Windows\Start Me	nu nu

### Click on next button



Click on finish button

### **CHAPTER 2: LAUNCHING WEKA EXPLORER**

#### 2.1 Starting with Weka

Once the program has been loaded on the user's machine it is opened by navigating to the programs start option and that will depend on the user's operating system.



There are four options available on this initial screen.

1. Explorer- the graphical interface used to conduct experimentation on raw data

2. **Simple CLI-** provides users without a graphic interface option the ability to execute commands from a terminal window.

3. **Experimenter-** this option allows users to conduct different experimental variations on data sets and perform statistical manipulation

4. **Knowledge Flow-**basically the same functionality as Explorer with drag and drop functionality. The advantage of this option is that it supports incremental learning from previous results.

After selecting the Explorer option the program starts and provides the user with a separate graphical interface.

Preprocess         Classify         Cluster         Associate         Select attribution           Open         Open         Open         Open         Get	ener Undo Edit Save
Filter Choose None	Apply
Current relation Relation: None Instances: None Attributes: None	Selected attribute Name: None Type: None Missing: None Distinct: None Unique: None
Attributes           All         None         Invert         Pat	
	Visualize All
Remove	
Status Welcome to the Weka Explorer	Log x 0



Figure 2 shows the opening screen with the available options. At first there is only the option to select the Pre-process tab in the top left corner. This is due to the necessity to present the data set to the application so it can be manipulated. After the data has been pre-processed the other tabs become active for use.

There are six tabs:

1. Pre-process- used to choose the data file to be used by the application

2. **Classify-** used to test and train different learning schemes on the pre-processed data file under experimentation

3. Cluster- used to apply different tools that identify clusters within the data file

4. **Association-** used to apply different rules to the data file that identify association within the data

5. **Select attributes-**used to apply different rules to reveal changes based on selected attributes inclusion or exclusion from the experiment

6. **Visualize-** used to see what the various manipulation produced on the data set in a 2D format, in scatter plot and bar graph output.

### 2.2 Pre-processing:

In order to experiment with the application the data set needs to be presented to WEKA in a format that the program understands. There are rules for the type of data that WEKA will accept. There are three options for presenting data into the program.

• Open File- allows for the user to select files residing on the local machine or recorded medium.

• **Open URL-** provides a mechanism to locate a file or data source from a different location specified by the user.

• **Open Database-** allows the user to retrieve files or data from a database source provided by the user.

There are restrictions on the type of data that can be accepted into the program. Originally the software was designed to import only ARFF files, other versions allow different file types such as CSV, C4.5 and serialized instance formats. The extensions for these files include .csv, .arff, .names, .bsi and .data.

🖉 Weka Explorer			_		×
Preprocess Classify Cluster Associate Select attributes Visualize					
Open file Open URL Open DB Gen	erate Undo	Edit		Save	
Filter					
Choose None				Ap	ply
Current relation Relation: None Instances: None Attributes: None	Selected attribute Name: None Missing: None	Distinct: None	Type: Unique:		
Attributes					
All None Invert Pattern					
			~	Visualiz	e All
Remove	1				
Remove					
Status Welcome to the Weka Explorer			Log		r x 0

At the bottom of the window there is 'Status' box. The 'Status' box displays messages that keep you informed about what is going on. For example, when you first opened the 'Explorer', the message says, "Welcome to the Weka Explorer". When you loading "weather.arff" file, the 'Status' box displays the message "Reading from file...". Once the file is loaded, the message in the 'Status' box changes to say "OK". Right-click anywhere in 'Status box', it brings up a menu with two options:

1. **Available Memory** that displays in the log and in 'Status' box the amount of memory available to WEKA in bytes.

2. **Run garbage collector** that forces Java garbage collector to search for memory that is no longer used, free this memory up and to allow this memory for new tasks.

To the right of 'Status box' there is a 'Log' button that opens up the log. The log records every action in WEKA and keeps a record of what has happened. Each line of text in the log contains time of entry. For example, if the file you tried to open is not loaded, the log will have record of the problem that occurred during opening.

To the right of the 'Log' button there is an image of a bird. The bird is WEKA status icon. The number next to 'X' symbol indicates a number of concurrently running processes. When you loading a file, the bird sits down that means that there are no processes running. The number of processes besides symbol 'X' is zero that means that the system is idle. Later, in classification problem, when generating result look at the bird, it gets up and start moving that indicates that a process started. The number next to 'X' becomes 1 that means that there is one process running, in this case calculation.



### 2.3 Loading data:

The most common and easiest way of loading data into WEKA is from ARFF file, using 'Open file...' button . Click on 'Open file...' button and choose "project details " file from your local filesystem. Note, the data can be loaded from CSV file as well because some databases have the ability to **convert data only into CSV format**.

🖉 Weka Explorer	- 🗆 X
Preprocess Classify Cluster Associate Select attributes Visualize	
Open file Open URL Open DB Gene	rate Undo Edit Save
Filter	
Choose None	Apply
Current relation Relation: 6 sem stu project list1.arff Instances: 6 Attributes: 7	Selected attribute Name: Enroll No Type: Numeric Missing: 0 (0%) Distinct: 6 Unique: 6 (100%)
Attributes	Statistic Value
All None Invert Pattern	Minimum 1
	Maximum 42
No. Name	Mean 20
1 Enroll No	StdDev 15.786
	Class: Platform (Nom) Visualize All
3 Company Name	
4 project guide	
5 Phone No	8
6 E-mail id	
7 Platform	
Remove	
	1 21.5 42
- Status OK	Log x 0

Once the data is loaded, WEKA recognizes attributes that are shown in the 'Attribute' window. Left panel of 'Preprocess' window shows the list of recognized attributes:

No: is a number that identifies the order of the attribute as they are in data file.

Selection tick boxes: allow you to select the attributes for working relation.

Name: is a name of an attribute as it was declared in the data file.

The 'Current relation' box above 'Attribute' box displays the base relation (table) name and the current working relation - "project details", the number of instances - 6 and the number of attributes -7.

During the scan of the data, WEKA computes some basic statistics on each attribute. The following statistics are shown in 'Selected attribute' box on the right panel of 'Preprocess' window:

Name is the name of an attribute.

Type is most commonly Nominal or Numeric.

**Missing** is the number (percentage) of instances in the data for which this attribute is unspecified.

Distinct is the number of different values that the data contains for this attribute.

**Unique** is the number (percentage) of instances in the data having a value for this attribute that no other instances have.

Once the data is loaded into weka changes can be made to the attributes by clicking edit button shown above.

😨 Weka E	xplorer					- 🗆	$\times$		
Preprocess	Classify Cluster Associate	Sele	ect attribute	es Visual	ize				
Open	Open Open Open Gene Undo Edit Save								
Filter	Filter								
Choose	None						Apply		
Current rela	ation	4	Viewer						×
Relation Instances	: 6 sem stu project list1.arff : 6 Attributes: 7	Relati	on: 6 sem s	tu project	t list1.arff				
Attributes -	: 6 Attributes: 7	No.	Enroll No Numeric	name Nominal	Company Name Nominal	project guide Numeric	Phone No Numeric	E-mail id Nominal	Platform Nominal
All	None In	1			Headerlabs		9.71731		Ruby on
All	None In	2		F./	Aithent techol		9.71660		java
No.	Name	3		karan chand	Dattodils		9.81872		
	Enroll No	5			Leeway Hertz		9.09100	chanur	ios
2	<u> </u>	6		seshm			8.80204	shesh8	
3									
4	project guide	1							
5	Phone No	1							
6	E-mail id	]							
7	Platform								
	Remove								
•	rs \Dhruva \Desktop \mani new I								
ninar sub	sam mission						Undo	ОК	Cancel

To make the changes double click on the attribute value and update the details as user required .

Different operations can be performed through edit are as follows:

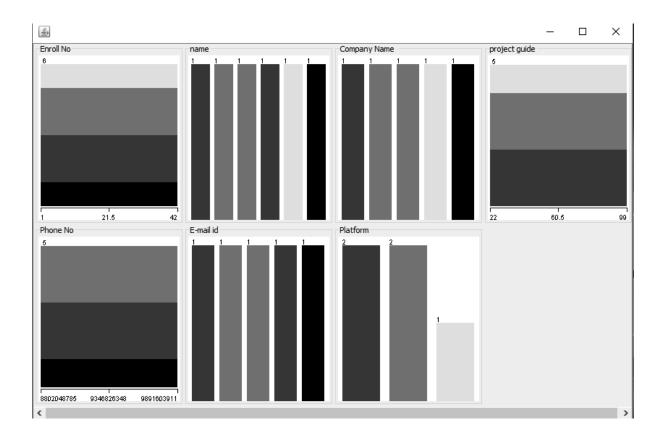
- 1) delete the attribute
- 2) Replace the attribute value
- 3) Set all values
- 4) Set missing values etc.

🛛 🖉 Weka B	Explorer				- 0	×							
v Preprocess	Classify Clust	ter Associat	te Select attr	ributes Visualize									
Open	Open	Open	1 Ge	ne Undo	Edit Save								
Filter						4	Viewer						X
Choose	None				Apply		tion: 6 sem	stu proiec	t list1.arff				147
Current rela Relation	ation 1: 6 sem stu proje	ect list1.arff		Selected attribute Name: Enroll No	Type: Numeri	No	1	name Nominal	Company Name Nominal	project guide Numeric	Phone No Numeric	E-mail id Nomina	I
Instances		Attributes:		Missi 0 ( Dis	tinct: Unique: 6 (10				Headerlabs		9.71731		Get mean
Attributes				Statistic	Value	2			Aithent techol		9.71660		Set all values to
				Minimum	1	3		karan chand	Daffodils		9.81872		
All	None	In	Pa	Maximum	42	- 4			Leeway Hertz		9.89160	chandr.	Set missing values to
				Mean	20	- 6		seshm			8.80204	shesh8	Replace values with
No.	Name			StdDev	15.786		1 10	- acommun	Judean		0.0020 111	ancarios	
	1 🔤 Enroll No		^		1								Rename attribute
2.7	2 name		100	Class: Platform (Nom)	✓ Visualize #	11							Attribute as class
25	3 Company Na		ų.										Delete attribute
	1 project guide	e		6									Delete attributes
	5 E-mail id			_		- 8 -							
	7 Platform		~	_									Sort data (ascending)
	Remo	nve											Optimal column width (curren
				1	21.5	42							Optimal column width (all)
Status OK					Log 💞								
al qp	10. 1			submission	discourse in the								

After update of values the minimum ,maximum , mean and standard deviation values gets changed.

🕝 Weka Explorer		— C	x c					
Preprocess Classify Cluster Associate Select	attr	ributes Visualize						
Open Open Open Gene Undo Edit Save								
Filter								
Choose None Apply								
Current relation Relation: 6 sem stu project list1.arff Instances: 6 Attributes: 7		Selected attribute Name: Enroll No Type: Numeric Missi 0 ( Distinct: Unique: 6 (10						
Attributes		Statistic Value						
All None In Pa		Minimum 1	1					
All None In Pa		Maximum 42						
		Mean 20						
No. Name		StdDev 15.786						
1 Enroll No	$\mathbf{A}$							
1 Enroll No 2 name 3 Company Name 4 project guide		Class: Platform (Nom) Visualize All						
3 Company Name								
4 project guide		6						
5 Phone No		·						
6 E-mail id								
7 Platform	<b>Y</b>							
Remove								
		1 21.5	42					
Status OK		Log	×0					

Click on visualize all



Attribute selection:

Weka Explorer	– 🗆 X
Preprocess Classify Cluster Associate Select att	ributes Visualize
Open Open Ge	ene Undo Edit Save
Filter	
Choose None	Apply
Current relation Relation: 6 sem stu project list1.arff Instances: 6 Attributes: 7	Selected attribute Name: name Type: Nominal Missi 0 ( Distinct: Unique: 6 (10
Attributes	No. Label Count
All None In Pa	1 Akanksha 1
	2 piyush 1
No. Name	3 karan jaryal 1 4 chandrakant 1
1 Enroll No	4 chandrakant 1 5 Abhishek chawla 1
2 🔽 name	6 seshmanipandey 1
3 Company Name	
4 project guide	Class: Platform (Nom) Visualize All
5 Phone No 6 E-mail id	
7 Platform	1 1 1 1 1 1
Remove	
Status OK	Log 💉 V O

🖉 Weka Explorer	– 🗆 X
Preprocess Classify Cluster Associate Sele	ct attributes Visualize
Op Op Op Ge	en Undo Edit Sav
Filter	
Choose None	Apply
Current relation Relation: 6 sem stu project list1.arff Instances: 6 Attributes: 7	Selected attribute Name: Company Name Type: No Missi 1 ( Distinct: . Unique: 5 (
Attributes	No. Label Count
All N I P	1 Headerlabs 1
No. Name	3 Daffodils 1
	4 Leeway Hertz 1
2 name	5 ducat 1 🗸
3 🗹 Company Name	Class: Platform (Nom) 🗸 Visualize All
4 project guide	L
5 Phone No	1 1 1 1 1
6 E-mail id	
7 Platform Remove	
Status OK	Log 💉 x 0

## 2.4 Setting Filters

Pre-processing tools in WEKA are called "filters". WEKA contains filters for discretization, normalization, resampling, attribute selection, transformation and combination of attributes. Some techniques, such as association rule mining, can only be performed on categorical data. This requires performing discretization on numeric or continuous attributes.

Using filters you can replace the discrete values to nominal values.

🖉 Weka Explorer	– 🗆 X
Preprocess Classify Cluster Associate Select attributes Visualize	
Open file Open URL Open DB Ge	nerate Undo Edit Save
Filter	
Choose None	Apply
-Current relation Relation: 6 sem stu project list1.arff Instances: 6 Attributes: 7	Selected attribute Name: Enroll No Type: Numeric Missing: 0 (0%) Distinct: 6 Unique: 6 (100%)
Attributes	Statistic Value
All None Invert Pattern	Minimum 1
	Maximum 42
No. Name	Mean 20
1 Enroll No	StdDev 15.786
	Class: Platform (Nom) Visualize All
3 Company Name	
4 project guide	
5 Phone No	8
6 E-mail id	
7 Platform	
Remove	
Status	1 21.5 42
OK	Log ×0

This will show pull-down menu with a list of available filters. Select Supervised Æ Attribute Æ Discretize and click on 'Apply' button. The filter will convert Numeric values into Nominal

Weka Explorer					-	
Preprocess Classify Cluster Associate Select attributes	Visualize					
Open file Open URL Open DB	Gene	erate	Undo	Edit		Save
Filter	1					
weka						Apply
AllFilter     MultiFilter		Selected at Name: f Missing: (	Enroll No	Distinct: 6	Type: N Unique: 6	
terren en e		Statistic		Value	2	
	Pattern	Minimum		1		
	ruccini	Maximum		42		
		Mean StdDev		20	26	
		Judev		15.70	30	
		Class: Platfo	orm (Nom)		~	Visualize All
		6				
			_	_	_	
-		1		21.5		42
				2110		_
					Log	×°
1						
Filter Remove filter Close		1		21.5		42
OK	]				Log	

When filter is chosen, the fields in the window changes to reflect available options.

🗿 Weka E	xplorer										_		×
Preprocess	Classify	Cluster	Associate	Select attributes	Visualize								
Open f	ìle	Ope	n URL	Open DB		Genera	ate	Undo	E	dit		Save	
Filter		_											_
Choose	AddCla	assificati	on -W "weł	a.classifiers.rules.	ZeroR "							App	ly
Current rela Relation Instances	: 6 sem st	u project l		tributes: 7			Selected a Name: Missing:	name	Distinct: 6	ι	Type: No Jnique: 6 (		
Attributes							No.	Label		Count			
All		None	•	Invert	Pattern			1 Akanksha		1			^
			·	2.172.12	, accent			2 piyush		1			
No.	Name							3 karan jaryal 4 chandrakant		1			- 11
t I	Enroll	No							•	1			×
	l 🗸 name						Class: Platf	form (Nom)			~	Visualize	All
	<u> </u>	any Name											
	projec	t guide				-1	1	1	1 1		1	1	
						-1							
	Platfo												
			Remove										
Status OK											Log		×0

As you can see, there is no change in the value Outlook. Select value Temperature, look at the 'Selected attribute' box, the 'Type' field shows that the attribute type has changed from Numeric to Nominal. The list has changed as well: instead of statistical values there is count of instances.

#### **CHAPTER 3: CLASSIFIERS**

#### 3.1 Building "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, support vector machines, multi-layer perceptrons, logistic regression, and bayes' nets. "Meta" classifiers include bagging, boosting, stacking, error-correcting output codes, and locally weighted learning.

Once you have your data set loaded, all the tabs are available to you. Click on the 'Classify' tab.

'Classify' window comes up on the screen.

🚺 Weka E	xplorer							_	$\times$
Preprocess	Classify	Cluster	Associate	Select attribute	s Visualize				
Classifier									
Choose	Zero	۱							
Test options	5			Cla	ssifier output				
🖲 Use tra	iining set								
🔿 Supplie	d test set		Set						
O Cross-	alidation/	Folds	10						
O Percen	tage split	%	66						
	Мо	re option	s						
(Nom) Platfo	orm			~					
St	art		Stop						
Result list (r	ight-click f	for option	ns)						

Now you can start analyzing the data using the provided algorithms. In this exercise you will analyze the data.

🗿 Weka E	xplorer								_		×
Preprocess	Classify	Cluster	Associate	Select attributes	Vi	alize					
Classifier											
Choose	Zero	ર									
Test option:	s		C	Classifier output							
🔿 Use tra	aining set										^
🔿 Supplie	ed test set	Set									
Cross-	validation	Folds	10								
O Percen	itage split	%	66								
1	More optio	ns									
(Nom) Platfo	orm		~								
Start		Stop									
Result list (r	right-click	for option	s)								
											~
Status										1 .	
OK									Log	-	r, x 0

### 3.2 Setting Test Options:

Before you run the classification algorithm, you need to 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 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 the '%' field.

In the 'Classifier evaluation options' make sure that the following options are checked

1. **Output model.** The output is the classification model on the full training set, so that it can be viewed, visualized, etc.

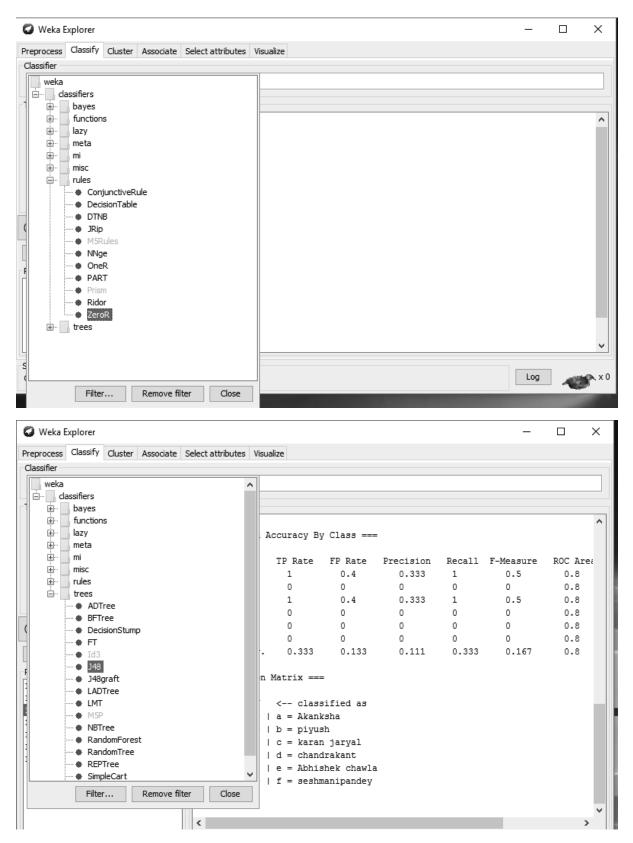
2. Output per-class stats. The precision/recall and true/false statistics for each class output.

3. **Output confusion matrix.** The confusion matrix of the classifier's predictions is included in the output.

4. **Store predictions for visualization.** The classifier's predictions are remembered so that they can be visualized.

5. Set 'Random seed for Xval / % Split' to 1. This specifies the random seed used when randomizing the data before it is divided up for evaluation purposes

🗿 Weka E	xplorer								_	×
Preprocess	Classify	Cluster	Associate	Select attri	ibutes	Visualize				
Classifier										
±	bayes function lazy meta mis misc rules trees		Remove fil		Close	routput				
	Filter		Keniove II		JUSE					

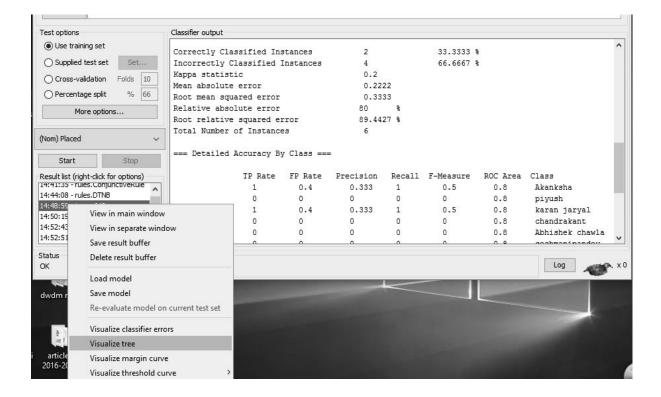


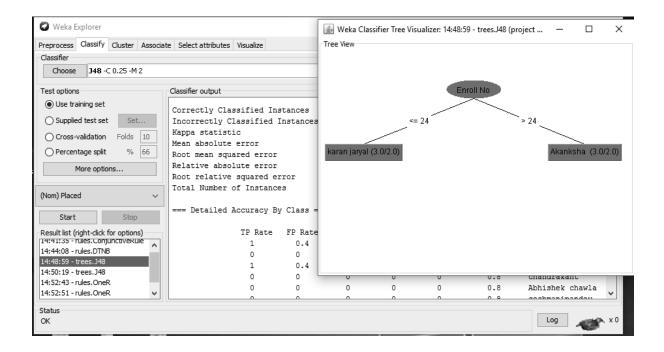
Once the options have been specified, you can run the classification algorithm. Click on 'Start' button to start the learning process. You can stop learning process at any time by clicking on 'Stop' button

When training set is complete, the 'Classifier' output area on the right panel of 'Classify' window is filled with text describing the results of training and testing. A new entry appears in the 'Result list' box on the left panel of 'Classify' window.

🗿 Weka E	xplorer								_		×
Preprocess	Classify	Cluster	Associate	Select attributes	Visualize						
Classifier											
Choose	Conju	inctiveR	<b>ule</b> -N 3 -M	2.0 -P -1 -S 1							
Test option:	S			Classifier output	t						
🖲 Use tra	aining set			Mean absolu	ute error		0.44	44			~
🔿 Supplie	d test set		Set	Root mean	squared er	ror	0.52	7			
0				Relative al	bsolute er	ror	100	÷.			
O Cross-	validation	Folds	10		-	d error	111.80	34 %			
O Percen	tage split	%	66	Total Numb			3				
				Ignored Cla	ass Unknow	n Instances		3			
	More op	tions									
				==== Detail	ed Accuracy	y By Class =					
(Nom) Platfo	orm		~		TP Rat	te FP Rate	Precision	Recall	F-Measure	ROC Ar	re;
Star	+		Stop		0	0	0	0	0	0.5	
					1	1	0.333	1	0.5	0.5	
Result list (r	_		-		0	0	0	0	0	0.5	
14:41:35 - r	ules.Conji	unctiveRu	ıle	Weighted A	vg. 0.3	33 0.333	0.111	0.333	0.167	0.5	
											1
				=== Confus:	ion Matrix	===					
					classif:	ind an					
					= Ruby on						
				01010	-	Nalls					
				01010	2						
				01010	100						
				<							

Preprocess Classify Cluster Associate	Select attributes Visua	lize						
Classifier								
Choose DTNB -X 1								
Test options	Classifier output							
Use training set	Mean absolute e	error		0.44	44			T
O Supplied test set Set	Root mean squar	red error		0.50	16			
O Supplied test set	Relative absolu	ite error		100	*			
O Cross-validation Folds 10	Root relative a	squared ex	rror	106.40	64 %			
O Percentage split % 66	Total Number of	E Instance	es	3				
	Ignored Class (	Jnknown I:	nstances		3			
More options								
	=== Detailed Ad	ccuracy By	y Class ===	=				
(Nom) Platform 🗸 🗸								
	1	TP Rate	FP Rate	Precision		F-Measure		ŝŝ
Start Stop		0	0	0	0	0	0.5	
Result list (right-dick for options)		1	1	0.333	-	0.5	0.5	
14:41:35 - rules.ConjunctiveRule		0	0	0	0	0	0.5	
14:44:08 - rules.DTNB	Weighted Avg.	0.333	0.333	0.111	0.333	0.167	0.5	
	=== Confusion N	atrix ==:	=					
	abc < c							
	$010   a = R_1$							
	0 1 0   a = R(a)	-	115					
	0 1 0   b = ja							
	01010-10							
								1





🖉 Weka Explorer							8_8		×	
Preprocess Classify Cluster Association	ate Select attributes Visualize								_	
									_	
Choose 348 -C 0.25 -M 2		Woka Ch	continue Viewa	Grou 14:49:50	troop 1/10 (pr	oject details.arff)		_		×
Test options	Classifier output	I VVEKa Cia	ssirier visua	1/26, 14,40,33 - 1	riees:540 (pi	oject detalis.am)				~
Our Use training set	Correctly Classified Ins	X: name (Nom	)		~	Y: predictedname	(Nom)			~
O Supplied test set Set	Incorrectly Classified 1	Colour: name	(Nom)			Select Instance				~
O Cross-validation Folds 10	Kappa statistic Mean absolute error	Reset	Clear	Open	Save	Jitter				
O Percentage split % 66	Root mean squared error	500		/ L / L_		Jitter				
More options	Relative absolute error	Plot:project de	tails.arff_pre	edicted			Г			11.20
(Nom) Placed ~ Start Stop	Root relative squared er Total Number of Instance === Detailed Accuracy By	eA_ b c_						Y : X · ·		
Result list (right-click for options) 14:41:35 - rules.ComunctiveRule 14:44:08 - rules.DTNB 14:48:59 - trees.J48 14:50:19 - trees.J48	TP Rate 1 0 1	pi A Akanksh	a piyush	karan jar	chandrak	Q Abhishek a seshi	nanip			• •
14:50:19 - trees.148 14:52:43 - rules.OneR 14:52:51 - rules.OneR  ✓	0	Class colour								13035
Status OK		Akanksha chaufrakau		piy Abh	ush ishek cha	wle	karan seshma	jaryal nipandey		

#### **CHAPTER 4: CLUSTERING**

#### 4.1 Clustering Data:

WEKA contains "clusterers" for finding groups of similar instances in a dataset. The clustering schemes available in WEKA are k-Means, EM, Cobweb, X-means, Farthest First. Clusters can be visualized and compared to "true" clusters (if given). Evaluation is based on log likelihood if clustering scheme produces a probability distribution.

x∎	5	~ @ - <del>.</del>					samp	le-sales-da	ata - Excel					? 📧	- 🗆	×
FI	LE HO	DME IN	SERT PA	AGE LAYOU	FORM	ULAS D.	ATA RE	VIEW	VIEW PE	DF				<u>A</u> M	urali N 👻	0
Pas	D 🚌 -	Calibri B I <u>U</u>		- A A		≡ ॐ∙ ≣ ∉ Æ	e" L	General \$ - % .₀ .₀0		ditional Form nat as Table ។ Styles <del>-</del>	-	Hinsert ▼ E× Delete ▼ Hint Format ▼	∑ • <sup>A</sup> Z▼• ↓ • <i>M</i> h•• ≪ •	Sign and Encrypt •		
Clip	board 🗔		Font		G /	lignment	G	Number	G.	Styles		Cells	Editing	Privacy		^
A1		• : )	X 🗸	<i>fx</i> ⊂u	stId											~
	А	В	С	D	E	F	G	Н	I	J	К	L	М	Ν	0	
1	custId	custName	custCount	productS	salesChan	unitsSold	dateSold									
2	23262	Candice L	Congo	SUPA101	Retail	117	8/9/2012									
3	23263	Xerxes Sn	Panama	DETA200	Retail	73	7/6/2012									
4			Tanzania,				*****									
5			South Afri			14	8/5/2012									
6		Celeste P		PURA200		170	*****									
7			Syrian Ara			129	****									
8			Guadelou		Retail		*****									
9			Macedoni				6/3/2012									
10			Kyrgyzsta				6/7/2012									
11			Reunion				*****									$\square$
12			Turks and				*****		_							$\square$
13	23273	Isaac Coo	Netherlar	SUPA104	Online	22	*****									
-	• •	sample	e-sales-dat	a (	Ð					÷ •						Þ
REA	DY											III II	<b>—</b> —		+ 10	00%

An international online catalog company wishes to group its customers based on common features. Company management does not have any predefined labels for these groups. Based on the outcome of the grouping, they will target marketing and advertising campaigns to the different groups. The information they have about the customers includes customer ID ,Customer Name , customer count,ProductSold, Sales Channel,Units Sold,Date Sold.

For our exercise we will use a part of the database for customers in US. Depending on the type of products sold, not all attributes are important. For example, suppose the to know the det

In 'Preprocess' window click on 'Open file...' button and select "customers.csv" file. Click 'Cluster' tab at the top of WEKA Explorer window.

Weka Explorer	- 🗆 X
Preprocess Classifier Cluster Associate Select attributes Visualize	
Open file Open URL Open DB	Generate Undo Edit Save
Filter	
Choose None	Apply
Current relation	Selected attribute
Relation: None Instances: None Attributes: None	Name: None Type: None Missing: None Distinct: None Unique: None
Attributes	
All None Invert Pattern	
Aii None Invert Pattern	
	Visualize All
December 201	
Remove	-
Status	
Welcome to the Weka Explorer	Log x 0

🖉 Weka Explorer		- 🗆 X
Preprocess Classify Cluster Associate Select attributes Visualize		
Open file Open URL Open DB Gene	erate Undo	Edit Save
Filter		
Choose None		Apply
Current relation Relation: sample-sales-data Instances: 998 Attributes: 7	Selected attribute Name: custId Missing: 0 (0%) Distir	Type: Numeric nct: 998 Unique: 998 (100%)
Attributes	Statistic	Value
All None Invert Pattern	Minimum	23262
All None Invert Pattern	Maximum	24259
No. Name	Mean	23760.5
	StdDev	288.242
2 custName		
3 custCountry		
4 productSold		
5 salesChannel	Class: dateSold (Nom)	✓ Visualize All
6 unitsSold	. ,	
7 dateSold		
Remove		100 <u>100 99</u> 100 100 100
	23262	23760.5 24259

# 4.2 Choosing Clustering Scheme:

In the 'Clusterer' box click on 'Choose' button. In pull-down menu select WEKA Æ Clusterers, and select the cluster scheme 'SimpleKMeans'. Some implementations of K-means only allow numerical values for attributes; therefore, we do not need to use a filter.

🖉 Weka Explorer		– 🗆 X
Preprocess Classify Cluster Associate Select attributes Visualize		
	erate Undo	Edit Save
Filter Shoose None		Apply
Current relation Relation: sample-sales-data Instances: 998 Attributes: 7	Selected attribute Name: custId Missing: 0 (0%) Distinct: 99	Type: Numeric 98 Unique: 998 (100%)
Attributes All None Invert Pattern	Statistic Minimum Maximum	Value 23262 24259
No.     Name       1     custId       2     custName       3     custCountry	Mean StdDev	23760.5 288.242
4 productSold 5 salesChannel 6 unitsSold 7 dateSold	Class: dateSold (Nom)	Visualize All
		<u>100 99 100 100 100</u>

🖉 Weka Explorer		_		×
Preprocess Classify Cluster Associate Select attributes	Visualize			
Clusterer				
weka	stance -R first-last" -I 500 -S 10			
ClOPE Cobweb DBSCAN EM FarthestFirst	Clusterer output Number of iterations: 15 Within cluster sum of squared errors: 3941.62 Missing values globally replaced with mean/mod Cluster centroids:		91428	^
HierarchicalClusterer     MakeDensityBasedClusterer     OPTICS     SIB	Attribute		Full Dat (99)	
SimpleKMeans XMeans	custId custName custCountry productSold salesChannel unitsSold dateSold		23760 ndice Lev Denma: PURA1( Onlin 108.256 2012-05-	vy rk 00 ne 65
Close	Time taken to build model (full training data)	) : 0.	06 secon	is v
Status OK		Log		р, х(

the clustering algorithm is chosen, right-click the algorithm, Once on "weak.gui.GenericObjectEditor" comes up to the screen. Set the value in "numClusters" box to 5 (instead of default 2) because you have five clusters in your .arff file. Leave the value of 'seed' as is. The seed value is used in generating a random number, which is used for making the initial assignment of instances to clusters. Note that, in general, K-means is quite sensitive to how clusters are initially assigned. Thus, it is often necessary to try different values and evaluate the results.

🖉 Weka Explorer	G weka.gui.GenericObjectEditor	××
Preprocess Classify Cluster Associate Select attribut Clusterer Choose SimpleKMeans -N 5 - A "weka.core.Euclid	About	
Cluster mode	Capabilities	
O Supplied test set Set	displayStdDevs False	~
O Percentage split % 66	distanceFunction Choose EuclideanDistance -R first-last	
Classes to dusters evaluation (Nom) dateSold	dontReplaceMissingValues False	- ;a
Store dusters for visualization	numClusters 5	3)
Ignore attributes		.5 7Y
Start Stop	seed 10	00
Result list (right-dick for options) 12:45:27 - FilteredClusterer		1e
12:57:40 - SimpleKMeans	Open Save OK Cancel	.5
	Time taken to build model (full training data) : 0.06 sec === Model and evaluation on training set === <	conds
OK	Log	. × 0

# 4.3 Setting Test Options:

Before you run the clustering algorithm, you need to choose 'Cluster mode'. Click on 'Classes to cluster evaluation' radio-button in 'Cluster mode' box and select in the pull-down box below.

Once the options have been specified, you can run the clustering algorithm. Click on the 'Start' button to execute the algorithm.

🗿 Weka E	xplorer						-		×
Preprocess	Classify	Cluster	Associate	Select attributes	Visualize				
Clusterer									
Choose	Simp	leKMear	<b>ns</b> -N 5 -A "v	veka.core.Euclidea	Distance -R first-last" -I 500 -S 10				
Cluster mod					Clusterer output	19			^
Obeur	aning set				Within cluster sum of	squared errors: 39	941.626631269	91428	
🔘 Supplie	ed test set	:	Sei	t	Missing values global	ly replaced with me	an/mode		
Percen     Olasses	itage split s to cluste		ation	% 66	Cluster centroids:				
	dateSold			~	Attribute			Full Da (99	
Store of	dusters fo	r visualiza	ation					(95	
					custId			23760	.5
		Ignore a	attributes		custName		Car	ndice Le	vy
					custCountry			Denma	rk
	Start			Stop	productSold			PURA1	.00
Result list (r	right-click	for optior	ns)		salesChannel			Onli	
12:45:27 - 6	FilteredClu	isterer			unitsSold			108.25	
12:57:40 - 5	SimpleKMe	ans			dateSold		2	2012-05-	.15
					Time taken to build m	odel (full training	g data) : 0.0	)6 secon	lds
					=== Model and evaluat	ion on training set	; ===		
					<				>
Status								_	
OK							Log	-	, x
	c unc an	yonann	1.						

#### 🗿 Weka Explorer

Preprocess Classify Cluster Associate Select attributes	(feedback							
Ousterer	VBUAIZE							
Choose SimpleKMeans -N 5 -A 'weka.core.Eudidean0	Astance -R. Hrst-last" -1 500 -5 10							
Duster mode	Clusterer output							
O Use training set	NUMBER OF ICERCIONS: 15							
· · · · · ·	Within cluster sum of squared errors: 3941.626 Missing values globally replaced with mean/mode							
	Missing values globally replaced with mean/mode							
O Percentage split % 66	Cluster centroids:							
Classes to clusters evaluation			Cluster#					
(Nom) dateSold	Attribute	Full Data	0	1	2	3	4	
Store clusters for visualization		(998)	(264)	(176)	(248)	(160)	(150)	
Store custors for vacanzation								
Ignore attributes	custId	23760.5	23744.5189	23999.6932	23811.3831	23630,5875	23562.42	
Ignore actributes	custName				Anjolie Hicks			
Start Stoo	custCountry productSold	PURA100	Bouvet Island SUPA101	Anguilla DETA100	Swaziland PURA200	Denmark SUPA103	Panama PURAS00	
	salesChannel	Online	Online	Retail		Retail	Retail	
Result list (right-dick for options) 12:45:27 - FilteredClusterer	unitsSold	108.2565	146.7652	119.0966		73.0938	145.9867	
12:45:27 - Hiteredulusterer 12:57:40 - SimpleKMeans	dateSold	2012-05-15	2012-05-15	2011-07-27		2012-08-11	2011-11-11	
	Time taken to build model (full training data) Model and evaluation on training set Clustered Instances 0 264 (26%) 1 176 (36%) 2 248 (25%) 3 180 (15%) 4 150 (15%)	: 0.06 seconds						
Status								
DK							Log	-08
		ø 📀					e <sup>w</sup> へ (1) 阳 <sup>1:</sup> 12/	

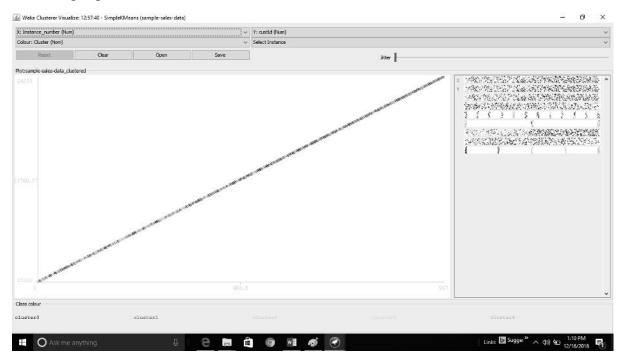
- a ×

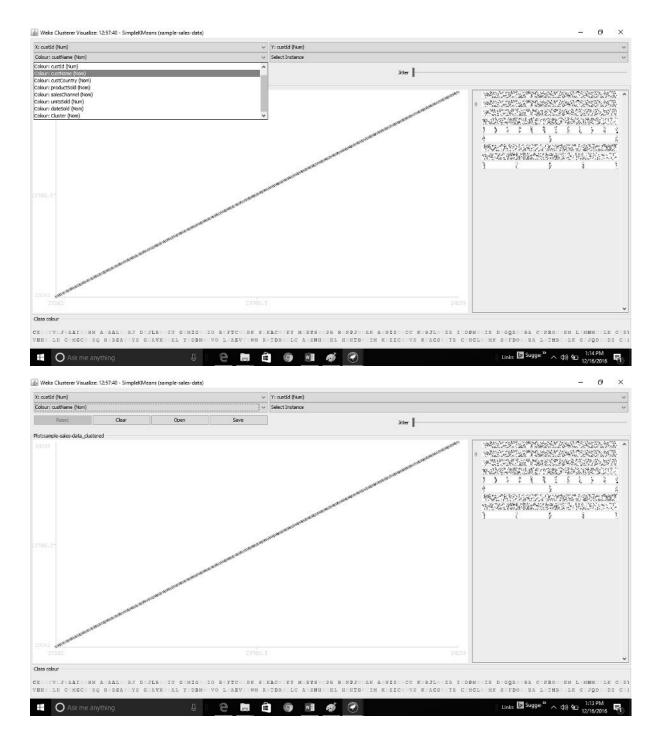
#### 4.4 Visualization of Results

Another way of representation of results of clustering is through visualization. Right-click on the entry in the 'Result list' and select 'Visualize cluster assignments' in the pull-down window.

reprocess Classify Cluster Associate Select a	attributes Visualize								
Choose SimpleKMeans -N 5 -A 'weka.com	a.EudideanDistance -R first-last" -I 500 -S	10							
		10							
Duster mode	Clusterer output	13: 13							
O Use training set		of squared errors: 3941.6260							
O Supplied test set Set	Missing values glo	ally replaced with mean/mode	2						
O Percentage split %	66 Cluster centroids:								
Classes to clusters evaluation	Cluster centrolus:			Cluster#					
(Nom) dateSold	Attribute		Full Data	0	1	2	3	4	
Store clusters for visualization			(998)	(264)	(176)	(248)	(160)	(150)	
Ignore attributes	custId		23760.5 Candi ca Tarri	23744.5189 Vance Campos	23999.6932	23811.3831	23630.5875 Candice Levy	23562.42	
	custCountry			Bouvet Island	Anguilla	Swaziland	Denmark	Panama	
Start Stop	product5old		PURA100	SUPA101	DETA100	PURA200	SUPA103	PURA500	
Result list (right-dick for options)	salesChannel		Online	Online	Retail	Online	Retail	Retail	
2:45:27 - FilteredClusterer	unitsSold		108.2565	146.7652	119.0966	59.4355	73.0938	145.9867	
12:57:40 - SimpleKMeans	dateSold	1	2012-05-15	2012-05-15	2011-07-27	2012-06-17	2012-08-11	2011-11-11	
	ew in main window								
	ew in separate window								
Si	ive result buffer								
D	elete result buffer	model (full training data)	: 0.06 seconds						
Le	ad model	ation on training set							
S	we model								
R	e-evaluate model on current test set								
	sualize cluster assignments								
Vi	sualize tree								
	3 160 (16%) 4 150 (15%)								
tatus XK								Log	

This brings up the 'Weka Clusterer Visualize' window.



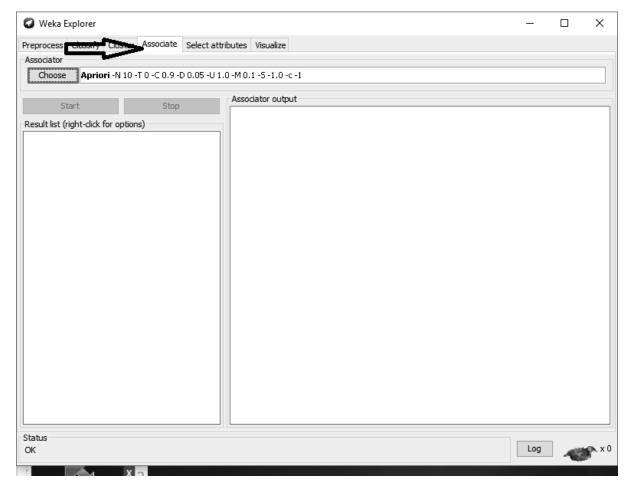


On the 'Weka Clusterer Visualize' window, beneath the X-axis selector there is a dropdown list, 'Colour', for choosing the color scheme. This allows you to choose the color of points based on the attribute selected. Below the plot area, there is a legend that describes what values the colors correspond to.

## **CHAPTER 5: ASSOCIATION**

#### **5.1 Finding Associations**

WEKA contains an implementation of the Apriori algorithm for learning association rules. This is the only currently available scheme for learning associations in WEKA. It works only with discrete data and will identify statistical dependencies between groups of attributes. Apriori can compute all rules that have a given minimum support and exceed a given confidence.

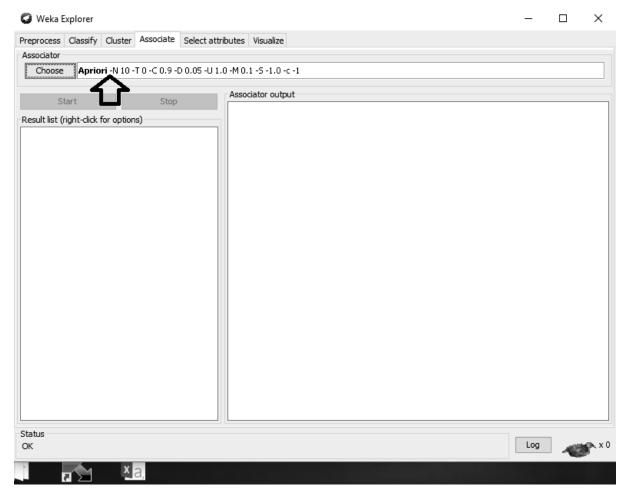


For this exercise you will use sales data from the "sales-sample-data.csv" file.

🖉 Weka Explorer			– 🗆 X		
Preprocess Classify Cluster Associate Select attributes Visualize					
Open file Open URL Open DB Gene	erate Un	do Edit	Save		
Filter					
Choose None			Apply		
Current relation Relation: sample-sales-data Instances: 998 Attributes: 7	Selected attribute Name: custId Missing: 0 (0%)	Distinct: 998	Type: Numeric Unique: 998 (100%)		
Attributes	Statistic	Value			
All None Invert Pattern	Minimum	23262			
All None Invert Pattern	Maximum	24259			
No. Name	Mean	23760.5			
	StdDev	288.242	<u>.</u>		
1 custId 2 custName					
3 custCountry					
4 productSold					
5 salesChannel	Class: dateSold (Nom) Visualize All				
6 unitsSold			- Hoddinee Fill		
7 dateSold					
	100 100 100	99 <u>100 100</u>	99 100 100 100		
Remove					
Kemove	23262	23760.5	24259		
Status	20202	23700.0	24208		
OK			Log x 0		

5.2 Setting Test Options

Check the text field in the 'Associator' box at the top of the window. As you can see, there are no other associators to choose and no extra options for testing the learning scheme



Right-click on the 'Associator' box, and click on show properties, 'GenericObjectEditor' appears on your screen. In the dialog box, change the value in 'minMetric' to 0.4 for confidence = 40%. Make sure that the default value of rules is set to 100. The upper bound for minimum support 'upperBoundMinSupport' should be set to 1.0 (100%) and 'lowerBoundMinSupport' to 0.1. Apriori in WEKA starts with the upper bound support and incrementally decreases support (by delta increments, which by default is set to 0.05 or 5%). The algorithm halts when either the specified number of rules is generated, or the lower bound for minimum support is reached. The 'significanceLevel' testing option is only applicable in the case of confidence and is (-1.0) by default (not used).

🗿 Weka Explorer			_		$\times$
Preprocess Classify Cluster Associate Select attrib	utes Visualize				
Associator					
Choose Apriori -N 10 -T 0 -C 0.4 -D 0.05 -U 1.0	-M 0.1 -S -1.0 -c -1				
Start Stop	Associator output				
Result list (right-click for options)	🗿 weka.gui.GenericObject	Editor		$\times$	
	weka.associations.Apriori About				
	Class implementing an	Apriori-type algorithm.	More Capabilities	5	
	car Fal	-			
	car Fals	ie		~	
	delta 0.0	5			
	lowerBoundMinSupport 0.1				
	metricType Cor	fidence		$\sim$	
	minMetric 0.4				
	numRules 10			_	
	outputItemSets Fak			~	
	removeAllMissingCols Fals			~	
Status OK	significanceLevel -1.0				
2016 player project list1	verbose Fals	e		~	

Once the options have been specified, you can run Apriori algorithm. Click on the 'Start' button to execute the algorithm.

Weka Explorer	-		$\times$
Preprocess Classify Cluster Associate Select attributes Visualize			
Associator			
Choose Aprior MID TO COO DOOS UI O MOI S 1.0 -c -1			
Show properties			
Start Copy configuration to clipboard butput			
Result list (right-click f Enter configuration			
Status OK	Log	-	∧ × 0

# **CHAPTER 6: ATTRIBUTE SELECTION**

# 6.1 Introduction:

Attribute selection searches through all possible combinations of attributes in the data and finds which subset of attributes works best for prediction. Attribute selection methods contain two parts: a search method such as best-first, forward selection, random, exhaustive, genetic algorithm, ranking, and an evaluation method such as correlation-based, wrapper, information gain, chi-squared. Attribute selection mechanism is very flexible - WEKA allows (almost) arbitrary combinations of the two methods.

O Weka Explorer Preprocess Classify Cluster Cl		– 🗆 ×
Open file Open URL Open DB Gen	erate Undo	Edit Save
Choose None		Apply
Current relation Relation: sample-sales-data Instances: 998 Attributes: 7	Selected attribute Name: custId Missing: 0 (0%) Distinct	Type: Numeric :: 998 Unique: 998 (100%)
Attributes All None Invert Pattern	Statistic Minimum Maximum	Value 23262 24259
No. Name	Mean StdDev	23760.5 288.242
2 custName 3 custCountry 4 productSold		
5 salesChannel 6 unitsSold	Class: dateSold (Nom)	<ul> <li>✓ Visualize All</li> </ul>
7 dateSold		0 <u>100 99 100 100 100</u>
Remove		
Status DK	23262	23760.5 242

To begin an attribute selection, click 'Select attributes' tab.

# 6.2 Selecting Options

To search through all possible combinations of attributes in the data and find which subset of attributes works best for prediction, make sure that you set up attribute evaluator to 'CfsSubsetEval' and a search method to 'BestFirst'. The evaluator will determine what method to use to assign a worth to each subset of attributes. The search method will determine what style of search to perform. The options that you can set for selection in the 'Attribute Selection Mode' box are :

1. Use full training set. The worth of the attribute subset is determined using the full set of training data.

**2. Cross-validation.** The worth of the attribute subset is determined by a process of cross-validation. The 'Fold' and 'Seed' fields set the number of folds to use and the random seed used when shuffling the data.

Specify which attribute to treat as the class in the drop-down box below the test options. Once all the test options are set, you can start the attribute selection process by clicking on 'Start' button.

Preprocess Classify Cluster Associate Select attributes Visualize Attribute Evaluator Choose ChiSquaredAttributeEval architectod Choose Ranker -T -1.7976931348623157E308 -N -1 Attribute Selection Mode O Use full training set Cross-validation Folds 10 Seed 1 Evaluator: weka.attributeSelection.ChiSquaredAttributeEval Search:weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1 Relation: sample-sales-data	
Choose ChiSquaredAttributeEval	
arch Nathod         Choose       Ranker -T -1.7976931348623157E308 -N -1         Attribute Selection Mode       Attribute selection output         Image: Cross-validation       Folds         1       Seed         1       Evaluator:         Weka.attributeSelection.ChiSquaredAttributeEval         Search:weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1         Relation:       sample-sales-data	
Cho se Ranker -T -1.7976931348623157E308 -N -1 Attribute Selection Mode	
Cho se Ranker -T -1.7976931348623157E308 -N -1 Attribute Selection Mode	
Attribute Selection Mode       Attribute selection output         Image: Seed 1       Image: Seed 1         Image: Seed 1       Evaluator:         Image: Seed 1       Seed 1         Image: Seed 1       Seed 1	
<pre> O Use full training set C Cross-validation Folds 10 Seed 1 Evaluator: weka.attributeSelection.ChiSquaredAttributeEval Search:weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1 Relation: sample-sales-data</pre>	^
O Cross-validation       Folds       10         Seed       1       Evaluator:       weka.attributeSelection.ChiSquaredAttributeEval         Search:weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1         Relation:       sample-sales-data	^
Seed 1 Evaluator: weka.attributeSelection.ChiSquaredAttributeEval Search:weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1 Relation: sample-sales-data	
Seed 1 Search:weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1 Relation: sample-sales-data	
Search:weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1 Relation: sample-sales-data	
Relation: sample-sales-data	
(Nom) dateSold	
Attributes: 7	
Start Stop custId	
Result list (right-dick for options) cust Name	
10:53:02 - BestFirst + CfsSubsetEval custCountry	
11:05:48 - Ranker + ChiSquaredAttributeE productSold	
salesChannel	
unitsSold	
dateSold	
Evaluation mode:evaluate on all training data	
=== Attribute Selection on all input data ===	
< >> Search Method:	1
Status	
OK Log	к 0
arch 2016 (1) notes 2016-2016 player project list1	

	sociate Select attributes Visualize	
Attribute Evaluator		
Choose CfsSubsetEval		
Search Method		
Choose BestFirst -D 1 -N 5	i	
Attribute Selection Mode	Attribute selection output	
Use full training set		
Cross-validation Folds		
Seed	1 === Attribute Selection on all input data ===	
	Search Method:	
(Nom) dateSold	Attribute ranking.	
Start Stop		
	Attribute Evaluator (supervised, Class (nominal): 7 dateSold):	
Result list hight-click for options)	Chi-squared Ranking Filter	
	Ranked attributes:	
	462074.0000017349632 2 custName	
	106169.4096881520256 3 custCountry	
	5082.2492880671016 4 productSold	
	3992.000000000192 1 custId	
	875.0094324211856 5 salesChannel	
	0 6 unitsSold	
	Selected attributes: 2,3,4,1,5,6 : 6	
L		

### **CHAPTER 7: DATA VISUALIZATION**

#### 7.1 Introduction:

WEKA's visualization allows you to visualize a 2-D plot of the current working relation. Visualization is very useful in practice, it helps to determine difficulty of the learning problem. WEKA can visualize single attributes (1-d) and pairs of attributes (2-d), rotate 3-d visualizations (Xgobi-style). WEKA has "Jitter" option to deal with nominal attributes and to detect "hidden" data points.

🙆 Weka	Explorer								_		×
Preprocess	Classify Cluster	Associate	Select attributes	Visualize.							
Open	file Op	en URL	Open DB.	- <del>6</del> 7	erate	Undo		Edit		Save	
Filter											
Choose	None									Ap	ply
Current rela Relation Instances	: 6 sem stu projec		ttributes: 7		Selected at Name: E Missing: 0	Enroll No	Distinct: 6	;	Type: Nu Unique: 6 (		
Attributes					Statistic			Value			
All	Nor		Invert	Pattern	Minimum			1			
All	INOF		Invert	Pattern	Maximum			42			
No.	Name				Mean			20			
					StdDev			15.786			
	1 Enroll No 2 name										
	2name 3Company Nam	e			-						
	1 project guide										
	5 Phone No				Class: Platfo	rm (Nom)			~	Visualiz	o All
6					Class: Flatio	ann (nonn)			~	visualiz	e All
7	7 Platform				1						
					6						
					_						
		Remove									
					1		21				42
Status											
ОК									Log	-	r, x 0

🗿 Weka E	xplorer								_		×
Preprocess	Classify	Cluster	Associate	Select attributes	Visua	lize					
Plot Matrix		Enr	oll No	name		Company Name	project guide	Phone No		E-mail id	
			•			•					^
name											
			•	·		•					_
Enroll No			1	· · ·		· · · ·		:		· · .	
											. 🗸
		<									>
PlotSize: [10	0]	-									
PointSize: [1	]						Update				
Jitter:	÷.						Select Attributes				
								100			
Colour: Plat		m)				~	SubSample % :	100			
Class Color	ur —										
					Rub	y on Rails java	ios				
Status OK									Lo	g	P. x 0
1270			-								

Select a square that corresponds to the attributes you would like to visualize. For example, let's choose 'outlook' for X – axis and 'play' for Y – axis. Click anywhere inside the square that corresponds to 'play on the left and 'outlook' at the top.

🗿 Weka 🛙	xplorer										_		Х
Preprocess	Classify	Cluster	Associate	Select attributes	Visualiz	e.							
Plot Matrix	1	Enr	oll No	name		Company Nam	е	project guide	Phone No		E-r	nail id	
E-mail id										•			^
												•	
Phone No													
													<u> </u>
													U
		<			Ш			1	11				>
PlotSize: [1	00]												
PointSize: [	1]							Update					
Jitter:	- i -							Select Attributes					
									100				
Colour: Pla		m)				~		SubSample % :	100				
Class Colo	ur												
					Ruby	on Rails ja	ava						
Status													
OK											Log	160	P, x 0

# A 'Visualizing r' window appears on the screen

(: name (Nom)				✓ Y: Phone N	o (Num)			
Colour: Platform	(Nom)			✓ Select Inst	ance			
Reset	Clear	Open	Save		Jitter			
Plot:6 sem stu pr	oject list.arff				-			
9891603911	×	×	M	М		x	•	
						Y	:	
9346826348	_							
3802048785	Akanksha	1	ıran jarya		Abhishek ch	M		
		yush ka	iran Jarya		seshmani	pan		

7.2 Changing the View

In the visualization window, beneath the X-axis selector there is a drop-down list, 'Colour', for choosing the color scheme. This allows you to choose the color of points based on the attribute selected.

Below the plot area, there is a legend that describes what values the colors correspond to. In your example, red represents 'no', while blue represents 'yes'. For better visibility you should change the color of label 'yes'. Left-click on 'yes' in the 'Class colour' box and select lighter color from the color palette. To the right of the plot area there are series of horizontal strips.

Each strip represents an attribute, and the dots within it show the distribution values of the attribute. You can choose what axes are used in the main graph by clicking on these strips (left-click changes X-axis, rightclick changes Y-axis). The software sets X - axis to 'Outlook' attribute and Y - axis to 'Play'. The instances are spread out in the plot area and concentration points are not visible. Keep sliding 'Jitter', a random displacement given to all points in the plot, to the right, until you can spot concentration points.

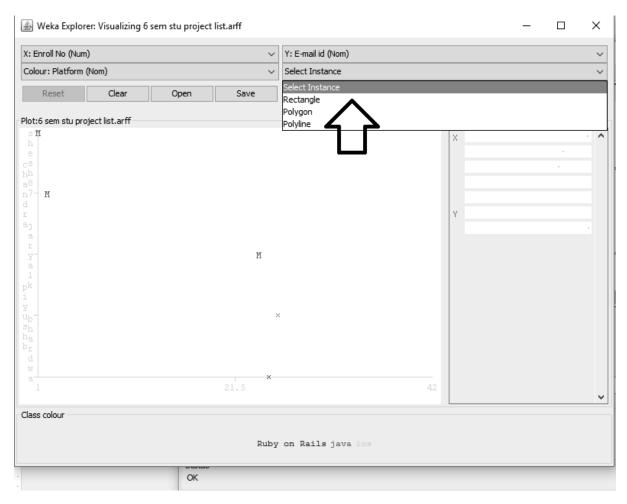
🛓 Weka Explo	rer: Visualizir	ng 6 sem stu p	oroject list.arff				-	×
X: name (Nom)				✓ Y: Phone No.	(Num)			~
Colour: project o	guide (Num)			✓ Select Insta	nce			~
Reset	Clear	Ope	n Save		Jitter			
Plot:6 sem stu pr	oject list.arff							
9891603911	-			М				• •
			Μ			×		_
	м	Μ						
						Y		
9346826348	_							
8802048785					1	M		
	Akanksha	piyush	karan jarya	chandrakan	Abhishek ch seshmani	.pan		~
Class colour								
0				o'				0
K I		8.6	13	- a S	×a,			

X: Enroll No (Num)   Colour: project guide (Num)   Colour: project guide (Num)   Select Instance     V   Plote6 sem stu procedulst.aff   9991603911     M   M   M   M   M   M   M   Y     Plote6 sem stu procedulst.aff     991603911     M   M   M   M   M   M   M   Y     Plote6 sem stu procedulst.aff     99346626348-     Plote6     Plote7     Plote7	📓 Weka Explorer: Visualizing 6 sem stu project list.arff		- C	×
Reset         Clear         Open         Save         Jitter           Plot6 sem stu protectist.arff         #         .         .         .           9891603911         #         .         .         .         .           9346826348         .         .         .         .         .           9346826348         .         .         .         .         .           88020048785         #         .         .         .         .           1         .21,5         .42         .         .	X: Enroll No (Num)	V: Phone No (Num)		~
Ploté sem stu profiliat.arff 9891603911 9346826348- 9346826348- 86002048785 <u>H</u> 21.5 42 Class colour	Colour: project guide (Num)	✓ Select Instance		~
9891603911 W 9346826348- 8802048785 M 1 21.5 42	Reset Clear Open Save	Jitter		
9346826348- 9346826348- 8802048785 И 1 21.5 42	12			
8802048785 M 1 21.5 42	9891603911	M MM		· · · · · · · · · · · · · · · · · · ·
Class colour	9346826348-			
	1 21.	5 42		~
				_

## 7.3 Selecting Instances:

Sometimes it is helpful to select a subset of the data using visualization tool. A special case is the 'User Classifier', which lets you to build your own classifier by interactively selecting instances. Below the Y - axis there is a drop-down list that allows you to choose a selection method. A group of points on the graph can be selected in four ways.

📓 Weka Explorer: Visualizing 6 sem stu project list.arff - 🗆 🗙						×	
X: Enroll No (Num)	~	Y: E-mail id (Nom)				~	
Colour: Platform (Nom)	~	Select Instance				~	
Reset Clear	Open Save						
Plot:6 sem stu project list.arff							
sta h e			×			•	
cs hh a8 n7- <b>M</b>							
n7- M d r aj			Y			-	
a r	М						
y- a l pk i							
y u <sub>b</sub> - s <sub>h</sub>	:	×					
ha br d							
a l	21.5 ×	42				~	
Class colour							
Ruby on Rails java ios							
						-	



# 1. Select Instance

Click on an individual data point. It brings up a window listing attributes of the point. If more than one point will appear at the same location, more than one set of attributes will be shown.

: Enroll No (Num)	$\sim$	Y: E-mail id (Nom)	
olour: Platform (Nom)	~	Select Instance	
Reset Clear Open Sav	e	Jitter	
<pre>Weka: Instance info Weka: Instance info Plot : Master Plot Instance: 2 Enroll No: 26.0 name: piyush Company Name : Aithent techologies project guide: Missing Phone No: 9.71660565E9 E-mail id: piyushbukaria026@gmail. Platform: java</pre>	×.	42	
ass o			
	Ruby	on Rails java ios	

2. Rectangle: You can create a rectangle by dragging it around the points.

our: Platform (Nom) Reset Clear Open :6 sem stu project list.arff	∽ Save	ctangle ect Instance :tangle	
:6 sem stu project list.arff	Save		_
M		/gon /line	
М	м	X Y	
	×		
1	21.5 ×	42	
is colour			
	Ruby	Rails java ios	

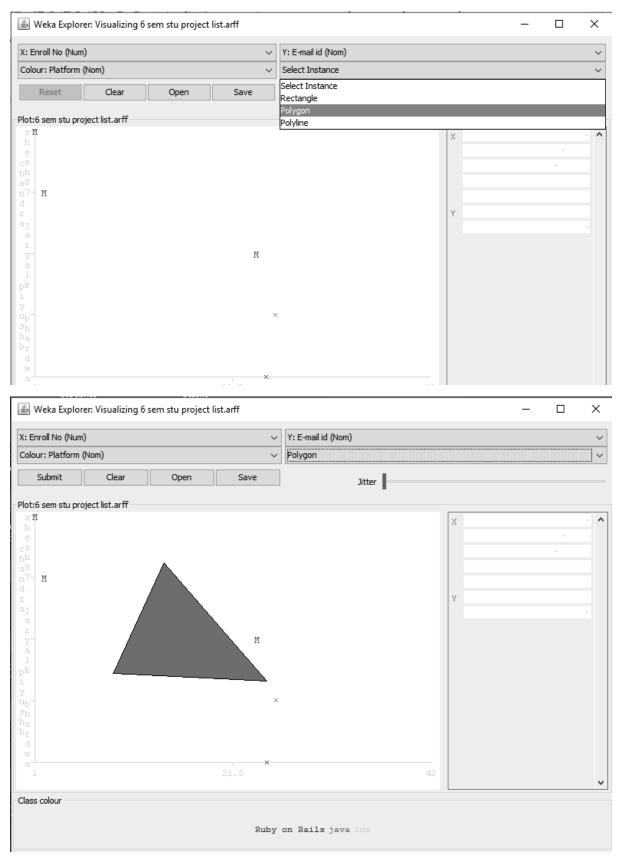
#### 🕌 Weka Explorer: Visualizing 6 sem stu project list.arff

X: Enroll No (Num)			~	Y: E-mail id (Nom) ~
Colour: Platform (Nom)			~	Rectangle
Submit	Clear	Open	Save	Jitter
Plot:6 sem stu proj	ject list.arff			
s M h c cs hh a <sup>8</sup> n <sup>7</sup> - M d r a r y a 1 pk i y u <sub>b</sub> sh ha br d v a a a a a a a a a a a a a			M 21.5	42
T			21.5	42
Class colour				

\_

×

**3. Polygon:** You can select several points by building a free-form polygon. Left-click on the graph to add vertices to the polygon and right-click to complete it.



4. Polyline: To distinguish the points on one side from the once on another, you can build a polyline. Left-click on the graph to add vertices to the polyline and right-click to finish.

