



MALAD KANDIVALI EDUCATION SOCIETY'S
NAGINDAS KHANDWALA COLLEGE OF COMMERCE, ARTS &
MANAGEMENT STUDIES & SHANTABEN NAGINDAS KHANDWALA
COLLEGE OF SCIENCE
MALAD [W], MUMBAI – 64
AUTONOMOUS INSTITUTION
(Affiliated To University Of Mumbai)
Reaccredited 'A' Grade by NAAC | ISO 9001:2015 Certified

CERTIFICATE

Name: Ms. Zeenat Hafeez Fazale Rab

Roll No: 578

Programme: BSc IT

Semester: V

This is certified to be a bonafide record of practical works done by the above student in the college laboratory for the course **DATA MINING AND WAREHOUSING (Course Code:2152UITDW)** for the partial fulfilment of Fifth Semester of BSc IT during the academic year 2021-22.

The journal work is the original study work that has been duly approved in the year 2021-22 by the undersigned.

External Examiner

Ms. Roshni Singh
(Subject-In-Charge)

Date of Examination:

(College Stamp)

Name: Zeenat Hafeez Fazale Rab

Class: T.Y. B.Sc. IT Sem- V

Roll No: 578

Subject: DATA MINING AND WAREHOUSING

[Course Code: 2152UITDW]

INDEX

Sr No	Name	Date	Sign
1	Perform data preprocessing tasks and demonstrate performing association rule mining on data sets.	23/07/2021	
2	Write ETL scripts and implement using data warehouse tools.	06/08/2021	
3	Perform Various OLAP operations such slice, dice, roll up, drill up and pivot.	13/08/2021	
4	Demonstrate performing KNN classification on data sets.	20/08/2021	
5	Demonstrate performing clustering on data sets.	17/09/2021	
6	Demonstrate performing Regression on data sets.	17/09/2021	
7	Implement a decision tree classifier.	20/08/2021	
8	Implement a Naive Bayes classifier.	27/08/2021	
9	Utilize data visualization techniques to analyse the given dataset.	17/09/2021	

Practical 1

Aim: Perform data preprocessing tasks and Demonstrate performing association rule mining on data sets.

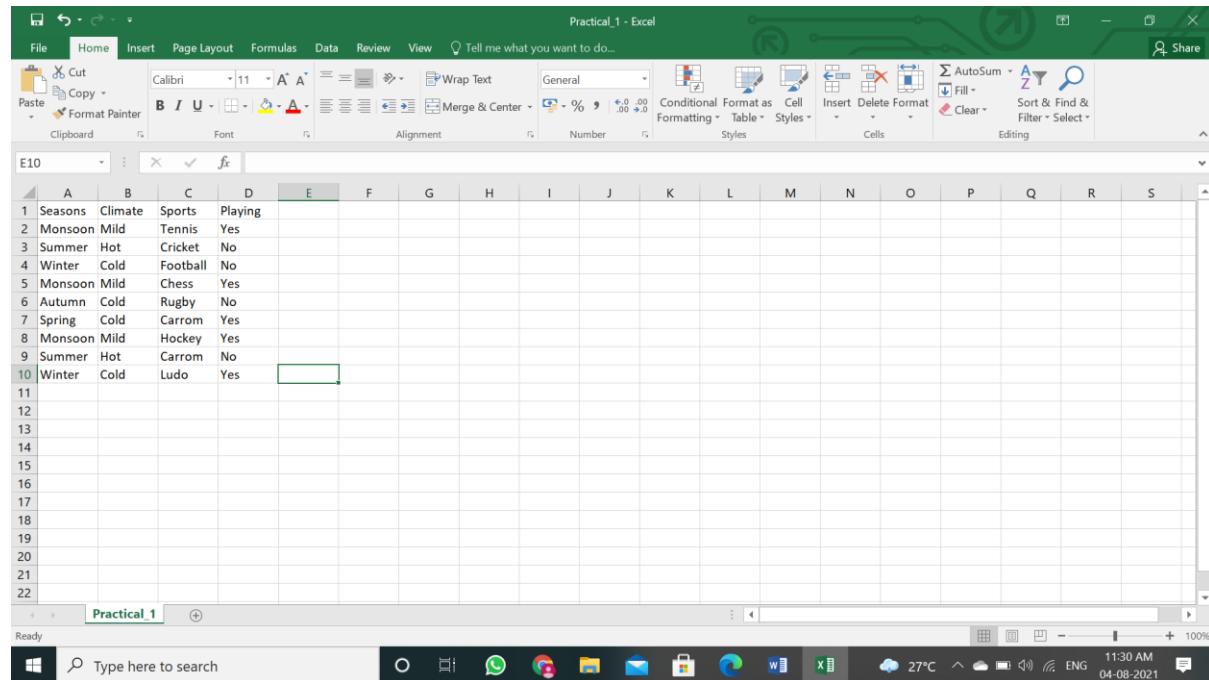
Theory:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is Market Based Analysis.

Market Based Analysis is one of the key techniques used by large relations to show associations between items. It allows retailers to identify relationships between the items that people buy together frequently.

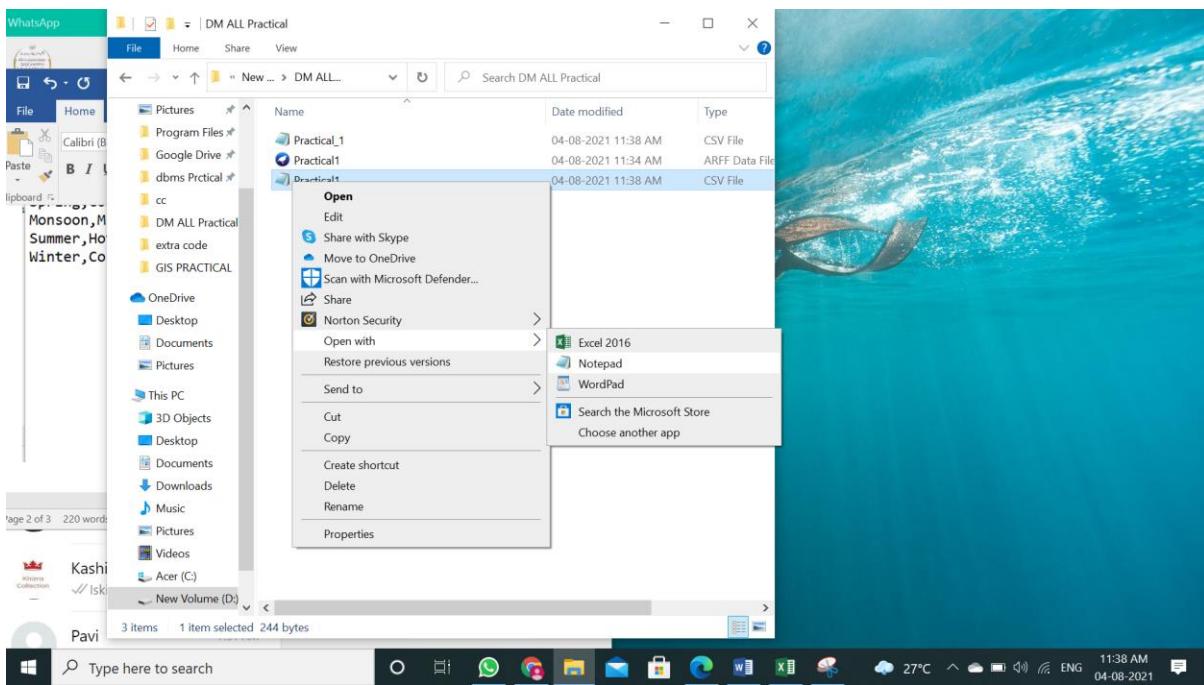
Step 1 : Put some Data in a Excel File and save it in CSV Format.



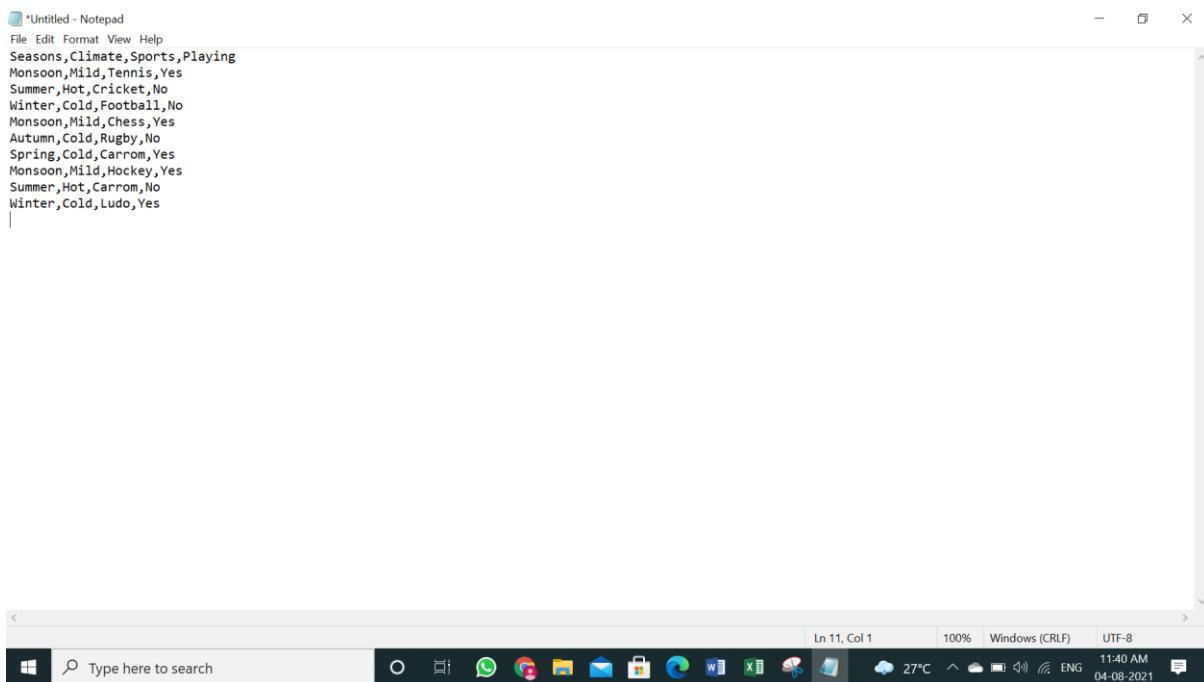
A screenshot of Microsoft Excel showing a data table in a CSV file named "Practical_1 - Excel". The table consists of 10 rows and 4 columns, with headers "Seasons", "Climate", "Sports", and "Playing". The data includes various weather conditions and sports activities. The Excel interface shows the Home tab selected, with the ribbon menu at the top and a toolbar below it. The status bar at the bottom right shows the date and time as 04-08-2021 11:30 AM.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Seasons	Climate	Sports	Playing															
2	Monsoon	Mild	Tennis	Yes															
3	Summer	Hot	Cricket	No															
4	Winter	Cold	Football	No															
5	Monsoon	Mild	Chess	Yes															
6	Autumn	Cold	Rugby	No															
7	Spring	Cold	Carrom	Yes															
8	Monsoon	Mild	Hockey	Yes															
9	Summer	Hot	Carrom	No															
10	Winter	Cold	Ludo	Yes															
11																			
12																			
13																			
14																			
15																			
16																			
17																			
18																			
19																			
20																			
21																			
22																			

Step 2 : Open the same Excel File with the Open with Option and click on Notepad.



Step 3 : Copy and Insert one Notepad File Content into another Notepad.



Step 4 : Always start a new file with @relation and name and enter @attribute column names with values in curly braces separated with commas and @data with all Excel values separated with commas and save this file with .arff extension i.e. for Weka Software



```
@Practical1_DM - Notepad
File Edit Format View Help
@relation practical

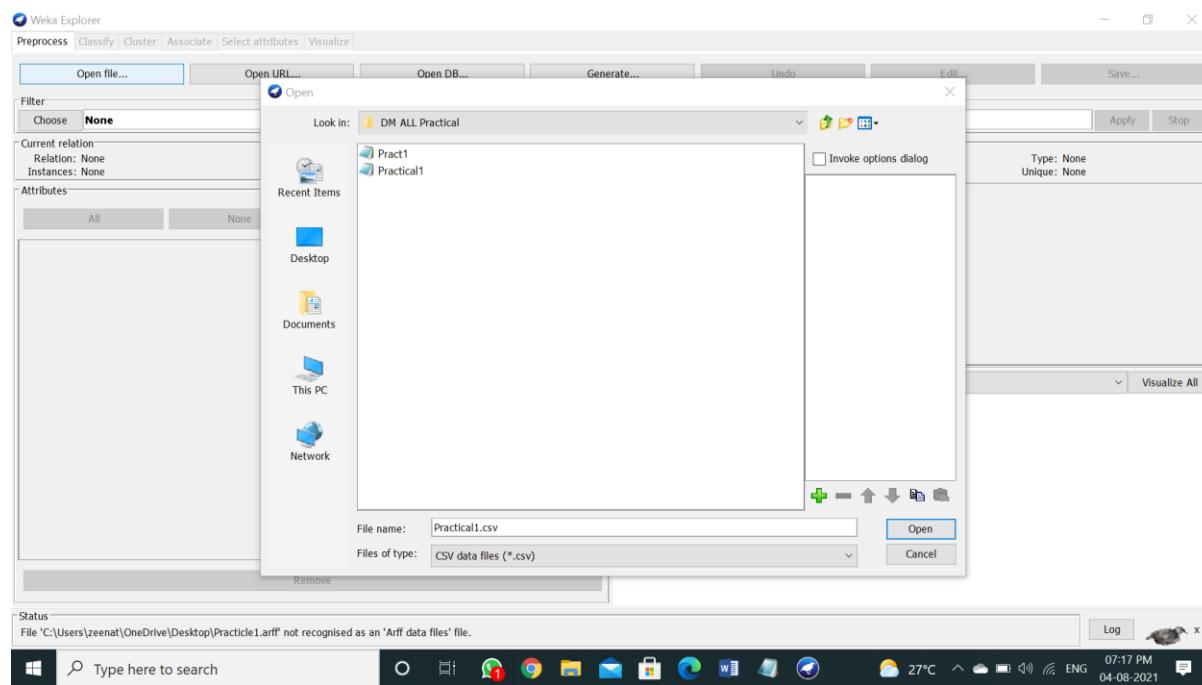
@attribute Seasons {Monsoon,Summer,Winter,Autumn,Spring}
@attribute Climate {Hot,Mild,Cold}
@attribute Sports {Tennis, Cricket,Football,Chess,Rugby,Carrom,Ludo,Hockey}
@attribute Playing {Yes,No}

@data

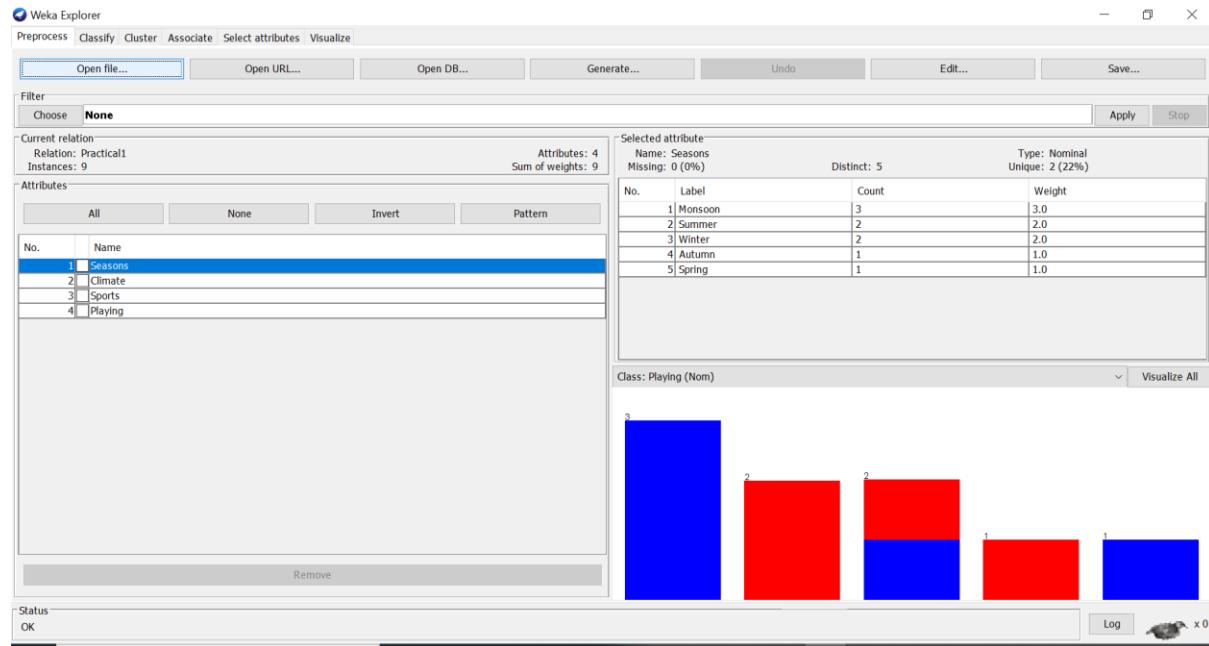
Seasons,Climate,Sports,Playing
Monsoon,Mild,Tennis,Yes
Summer,Hot,Cricket,No
Winter,Cold,Football,No
Monsoon,Mild,Chess,Yes
Autumn,Cold,Rugby,No
Spring,Cold,Carrom,Yes
Monsoon,Mild,Hockey,Yes
Summer,Hot,Carrom,No
Winter,Cold,Ludo,Yes
```



Step 5 : Open Weka Explorer Software and there is an Open File Option and browse for the file saved in folder.

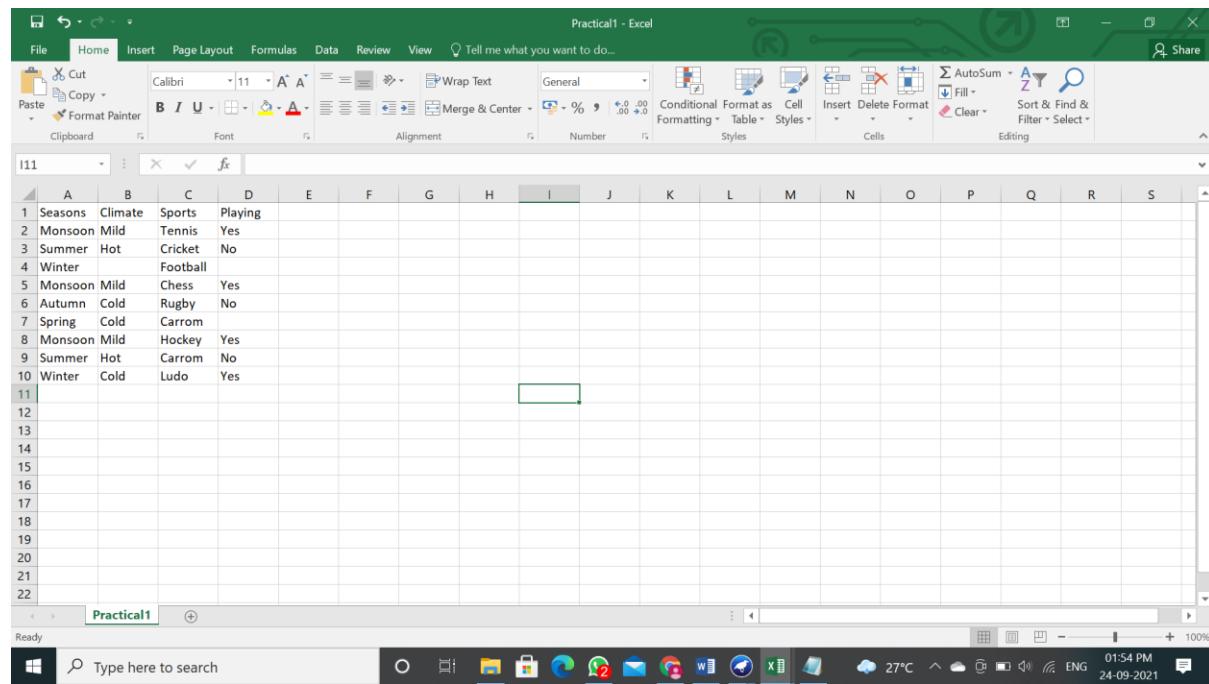


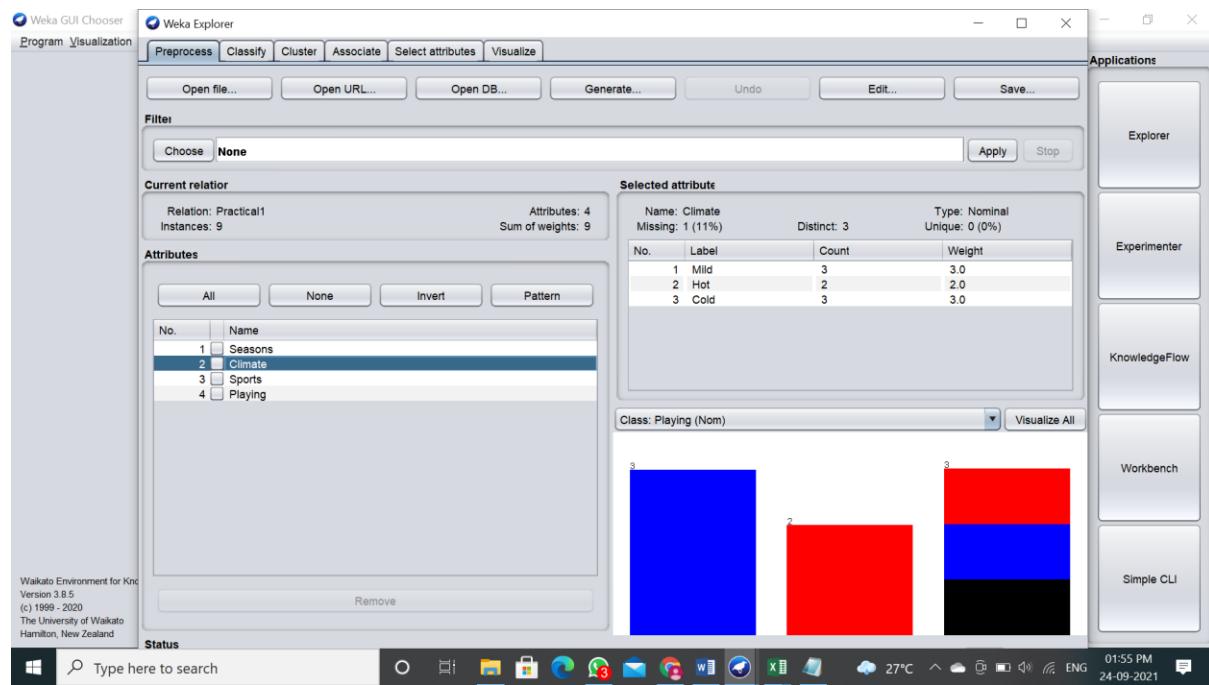
Step 6 : When you click on Open it will look like this.



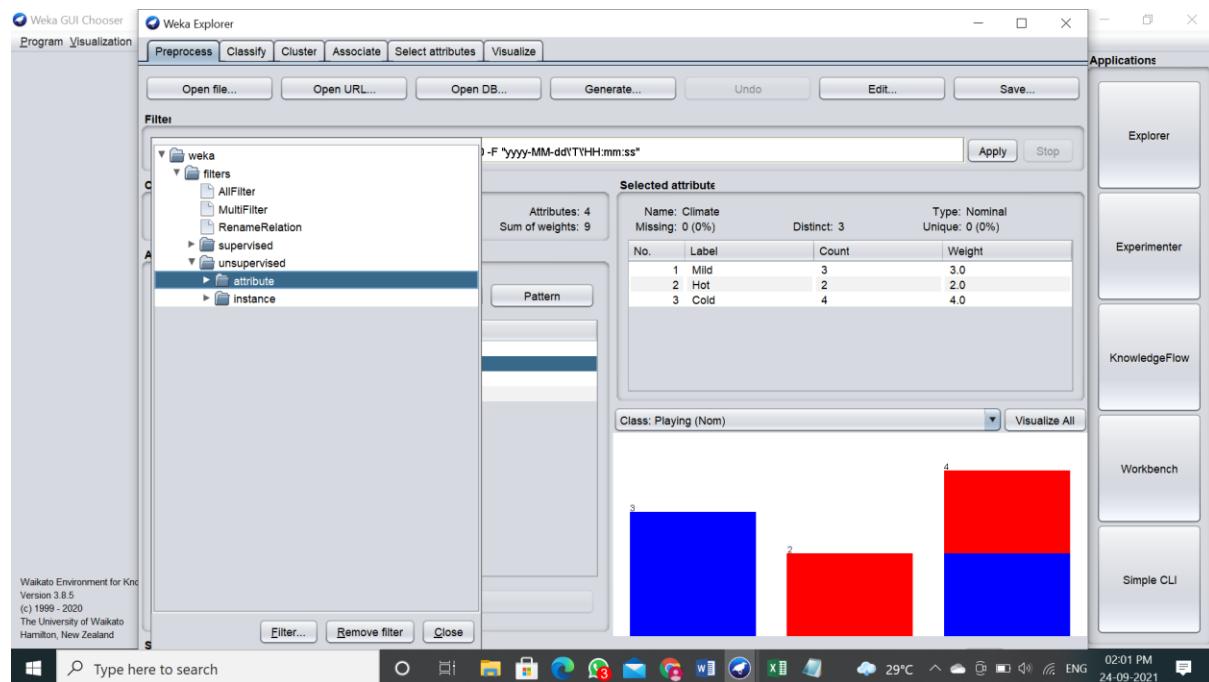
This is the Output for the Excel Data File.

This is file for remove data





Step 7 : If you want to do Filtration i.e. remove missing values from table go to Choose then Filter then Unsupervised then attribute and finally with ReplaceMissingWithUserConstant.

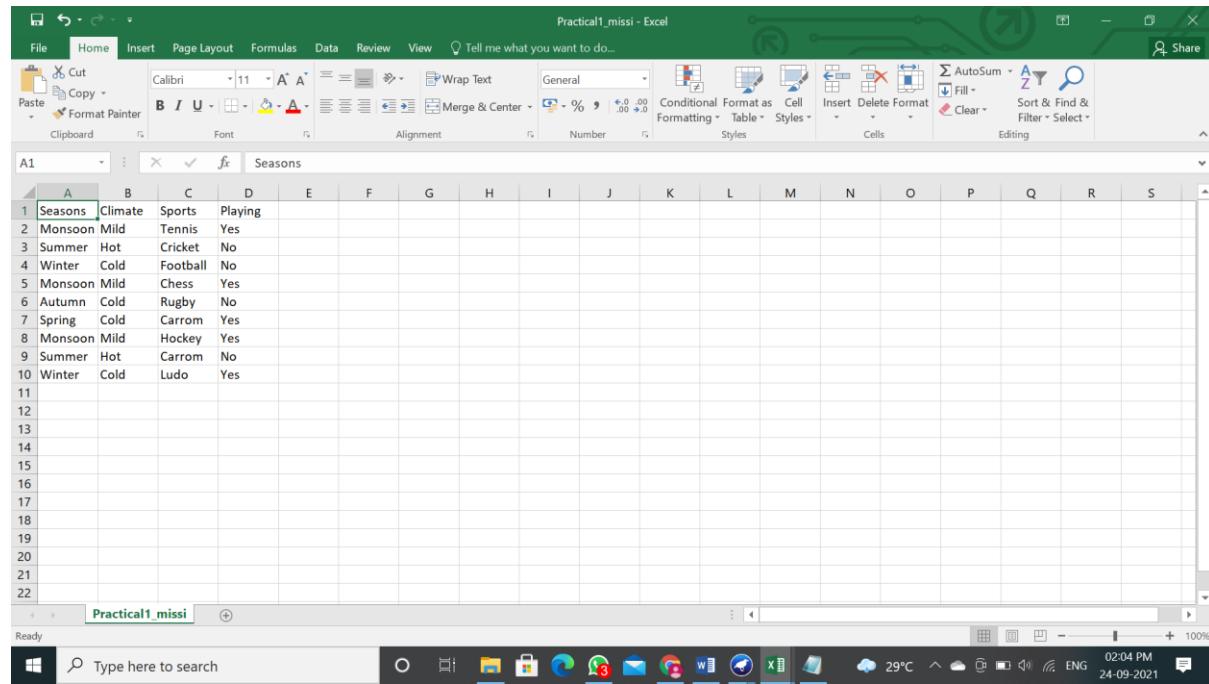
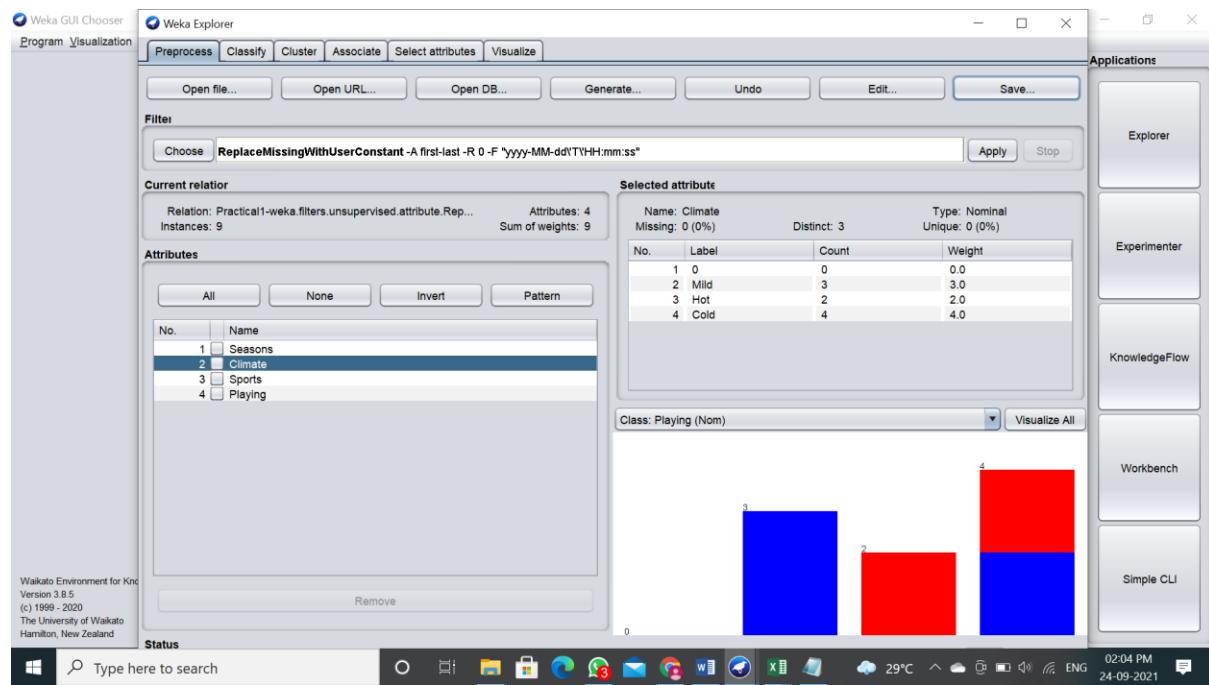


Step 8: Apply and Save file.

Name: Zeenat

Class: TYIT

Roll no: 578

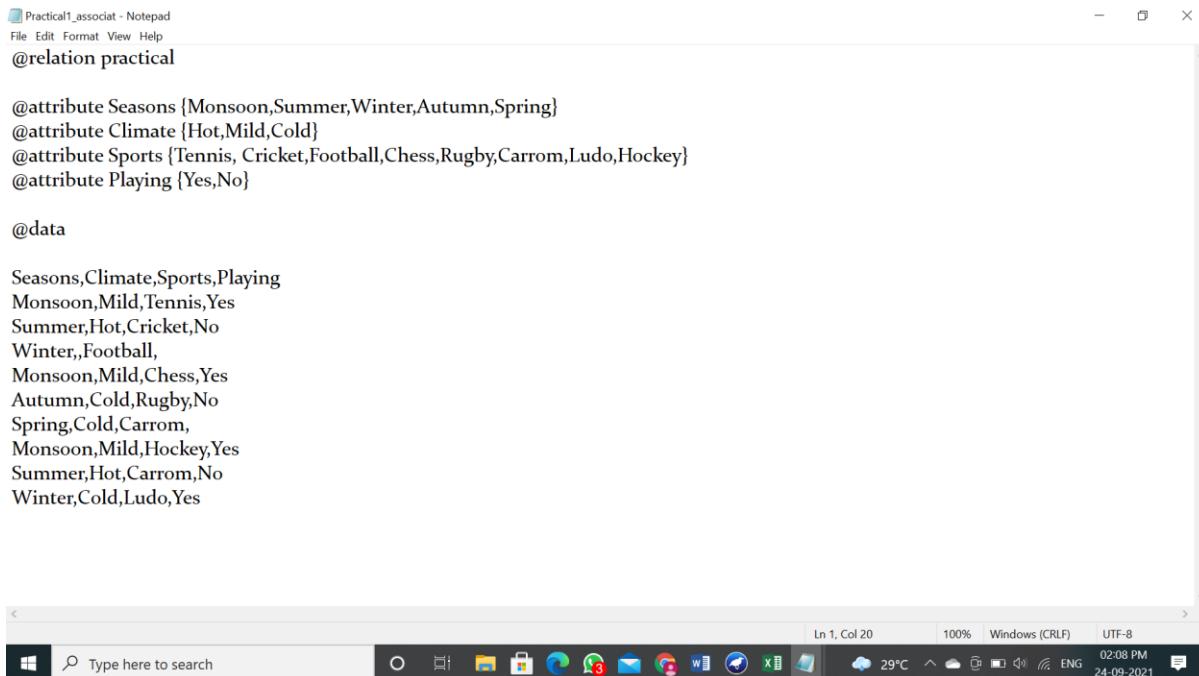


Step 9: Open Notepad and write the Relation, Attribute and Data for Dataset and save it with .arff extension.

Name: Zeenat

Class: TYIT

Roll no: 578

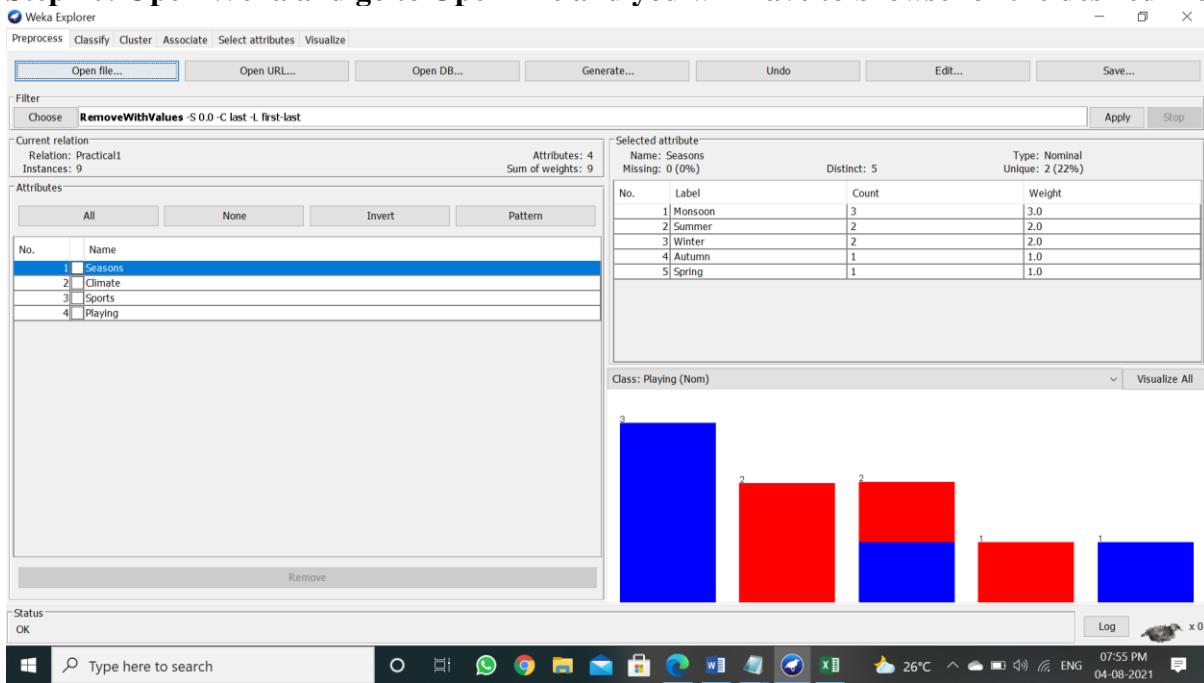


```
@attribute Seasons {Monsoon,Summer,Winter,Autumn,Spring}
@attribute Climate {Hot,Mild,Cold}
@attribute Sports {Tennis, Cricket,Football,Chess,Rugby,Carrom,Ludo,Hockey}
@attribute Playing [Yes,No]

@data

Seasons,Climate,Sports,Playing
Monsoon,Mild,Tennis,Yes
Summer,Hot,Cricket,No
Winter,Football,
Monsoon,Mild,Chess,Yes
Autumn,Cold,Rugby,No
Spring,Cold,Carrom,
Monsoon,Mild,Hockey,Yes
Summer,Hot,Carrom,No
Winter,Cold,Ludo,Yes
```

Step 10: Open Weka and go to Open File and you will have to browse for the desired file.



Step 11: Go to Associate Button click Apriori and click on Start you will get the output.

The screenshot shows the Weka Explorer interface running on a Windows operating system. The main window displays the output of an Apriori association rule mining process. The command used was `Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1`. The output details the execution time (19:48:00 - 19:56:35), minimum support (0.25), minimum metric (confidence: 0.9), and number of cycles performed (15). It lists generated itemsets L(1) through L(3) and the best rules found, such as "Climate=Mild 3 => Seasons=Monsoon 3" and "Seasons=Monsoon Playing=Yes 3 => Climate=Mild 3". The Windows taskbar at the bottom shows various open applications and the date/time (04-08-2021, 08:01 PM).

```
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Associate
Choose Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Start Stop
Result list (right-click)
19:48:00 - Apriori
19:56:35 - Apriori

Apriori
=====
Minimum support: 0.25 (2 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9
Size of set of large itemsets L(2): 9
Size of set of large itemsets L(3): 2

Best rules found:

1. Climate=Mild 3 => Seasons=Monsoon 3    <conf:(1)> lift:(3) lev:(0.22) [2] conv:(2)
2. Seasons=Monsoon 3 => Climate=Mild 3    <conf:(1)> lift:(3) lev:(0.22) [2] conv:(2)
3. Seasons=Monsoon 3 => Playing=Yes 3    <conf:(1)> lift:(1.8) lev:(0.15) [1] conv:(1.33)
4. Climate=Mild 3 => Playing=Yes 3    <conf:(1)> lift:(1.8) lev:(0.15) [1] conv:(1.33)
5. Climate=Mild Playing=Yes 3 => Seasons=Monsoon 3    <conf:(1)> lift:(3) lev:(0.22) [2] conv:(2)
6. Seasons=Monsoon Playing=Yes 3 => Climate=Mild 3    <conf:(1)> lift:(3) lev:(0.22) [2] conv:(2)
7. Seasons=Monsoon Climate=Mild 3 => Playing=Yes 3    <conf:(1)> lift:(1.8) lev:(0.15) [1] conv:(1.33)
8. Climate=Mild 3 => Seasons=Monsoon Playing=Yes 3    <conf:(1)> lift:(3) lev:(0.22) [2] conv:(2)
9. Seasons=Monsoon 3 => Climate=Mild Playing=Yes 3    <conf:(1)> lift:(3) lev:(0.22) [2] conv:(2)
10. Climate=Hot 2 => Seasons=Summer 2    <conf:(1)> lift:(4.5) lev:(0.17) [1] conv:(1.56)
```

Practical 2

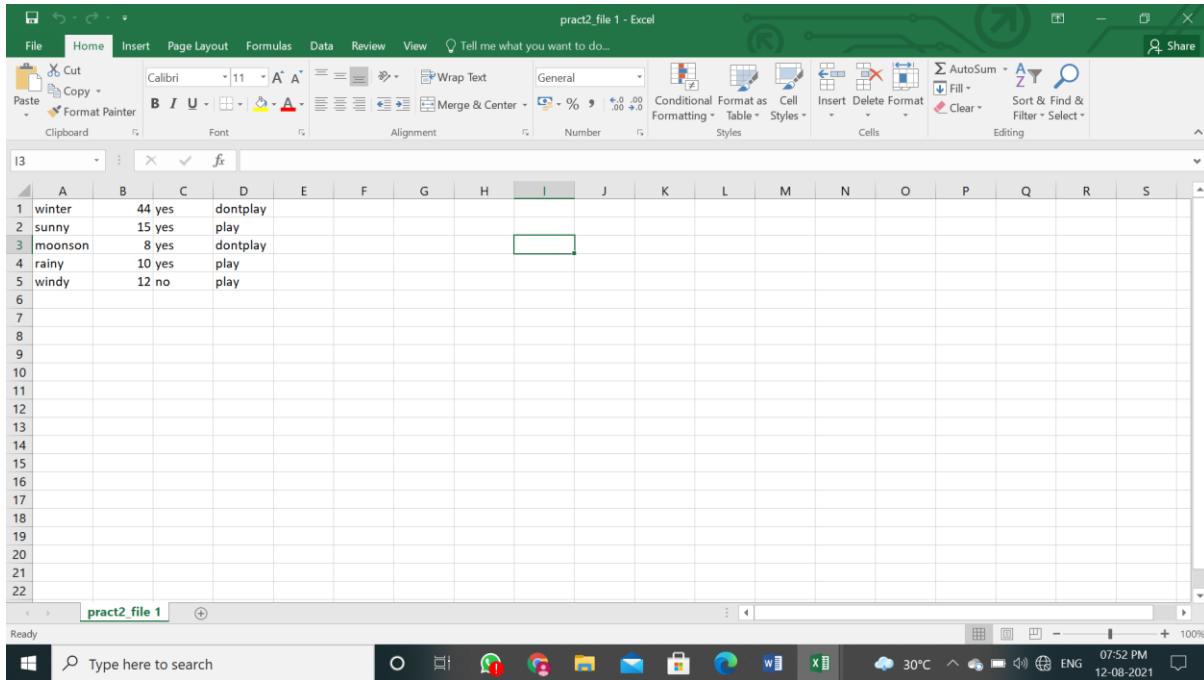
Aim: Write ETL scripts and implement using data warehouse tools.

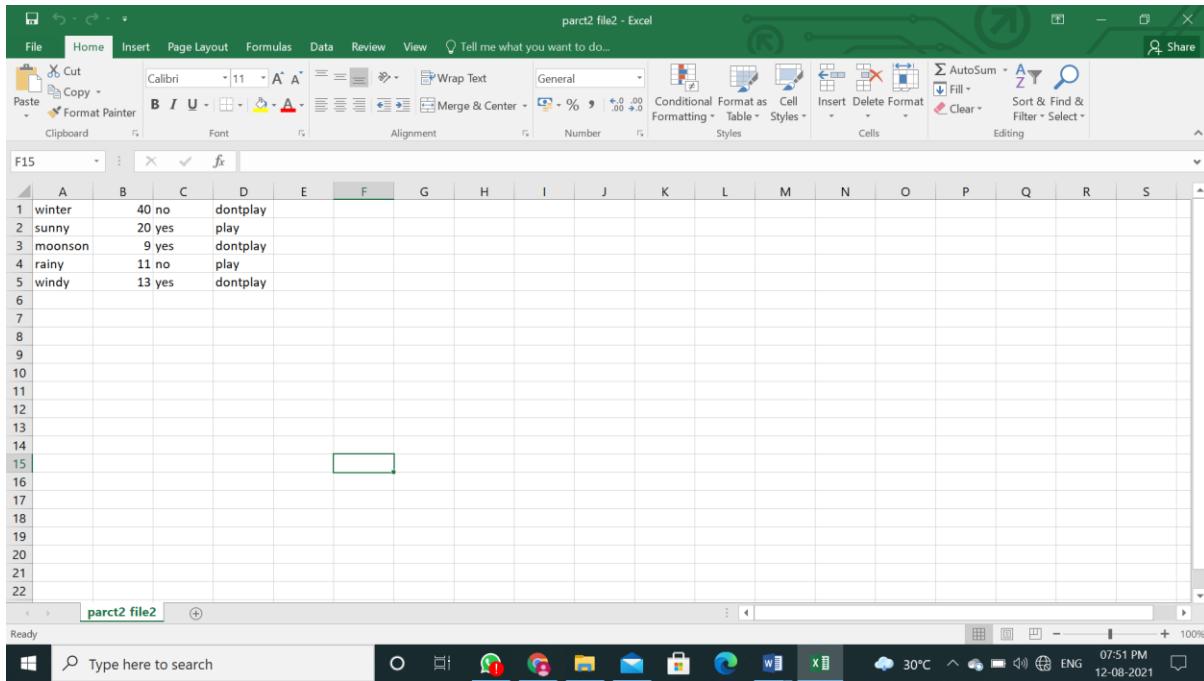
Theory:

The process of extracting data from source systems and bringing it into the data warehouse is commonly called ETL, which stands for extraction, transformation, and loading. ETL (Extract, Transform, Load) is an automated process which takes raw data, extracts the information required for analysis, transforms it into a format that can serve business needs, and loads it to a data warehouse. ETL typically summarizes data to reduce its size and improve performance for specific types of analysis.

The most common example of ETL is ETL is used in Data warehousing. User needs to fetch the historical data as well as current data for developing data warehouse. ... As The ETL definition suggests that ETL is nothing but Extract,Transform and loading of the data;This process needs to be used in data warehousing widely.

Step 1 : Open Excel File and fill in the Data and save it in .csv Format and create two Excel Files.

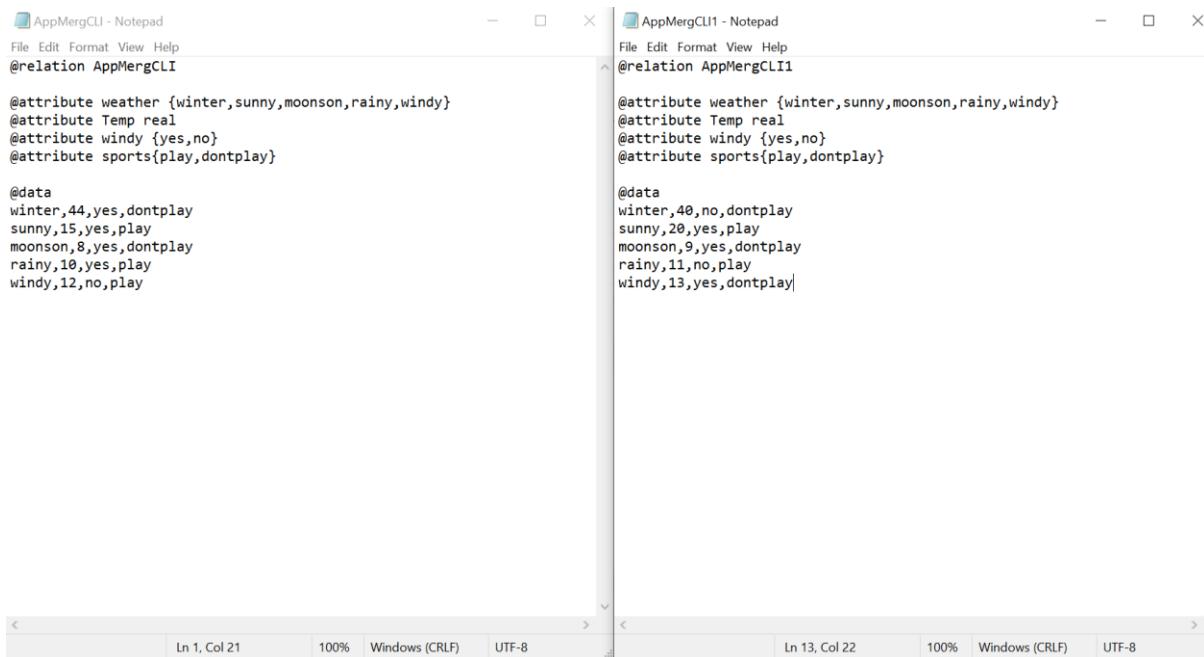




The screenshot shows an Excel spreadsheet titled "parct2 file2 - Excel". The data is organized into columns A through S. Column A contains weather conditions: winter, sunny, moonson, rainy, and windy. Columns B and C show numerical values (e.g., 40, 20, 9, 11, 13) and categorical responses (e.g., no, yes, play, dontplay). The formula bar at the top shows "F15". The status bar at the bottom indicates the date and time as 12-08-2021 07:51 PM.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	winter	40	no	dontplay															
2	sunny	20	yes	play															
3	moonson	9	yes	dontplay															
4	rainy	11	no	play															
5	windy	13	yes	dontplay															
6																			
7																			
8																			
9																			
10																			
11																			
12																			
13																			
14																			
15																			
16																			
17																			
18																			
19																			
20																			
21																			
22																			

Step 2 : On Excel File right click and open with Notepad and copy the content and paste it into new Notepad and write @relation and name, @attribute and column names and @data which contains all Excel Data with comma separated values.



The screenshot shows two Notepad windows. The left window, titled "AppMergCLI - Notepad", contains the following ARFF code:

```

@relation AppMergCLI
@attribute weather {winter,sunny,moonson,rainy,windy}
@attribute Temp real
@attribute windy {yes,no}
@attribute sports{play,dontplay}

@data
winter,44,yes,dontplay
sunny,15,yes,play
moonson,8,yes,dontplay
rainy,10,yes,play
windy,12,no,play

```

The right window, titled "AppMergCLI1 - Notepad", contains the following ARFF code:

```

@relation AppMergCLI1
@attribute weather {winter,sunny,moonson,rainy,windy}
@attribute Temp real
@attribute windy {yes,no}
@attribute sports{play,dontplay}

@data
winter,40,no,dontplay
sunny,20,yes,play
moonson,9,yes,dontplay
rainy,11,no,play
windy,13,yes,dontplay

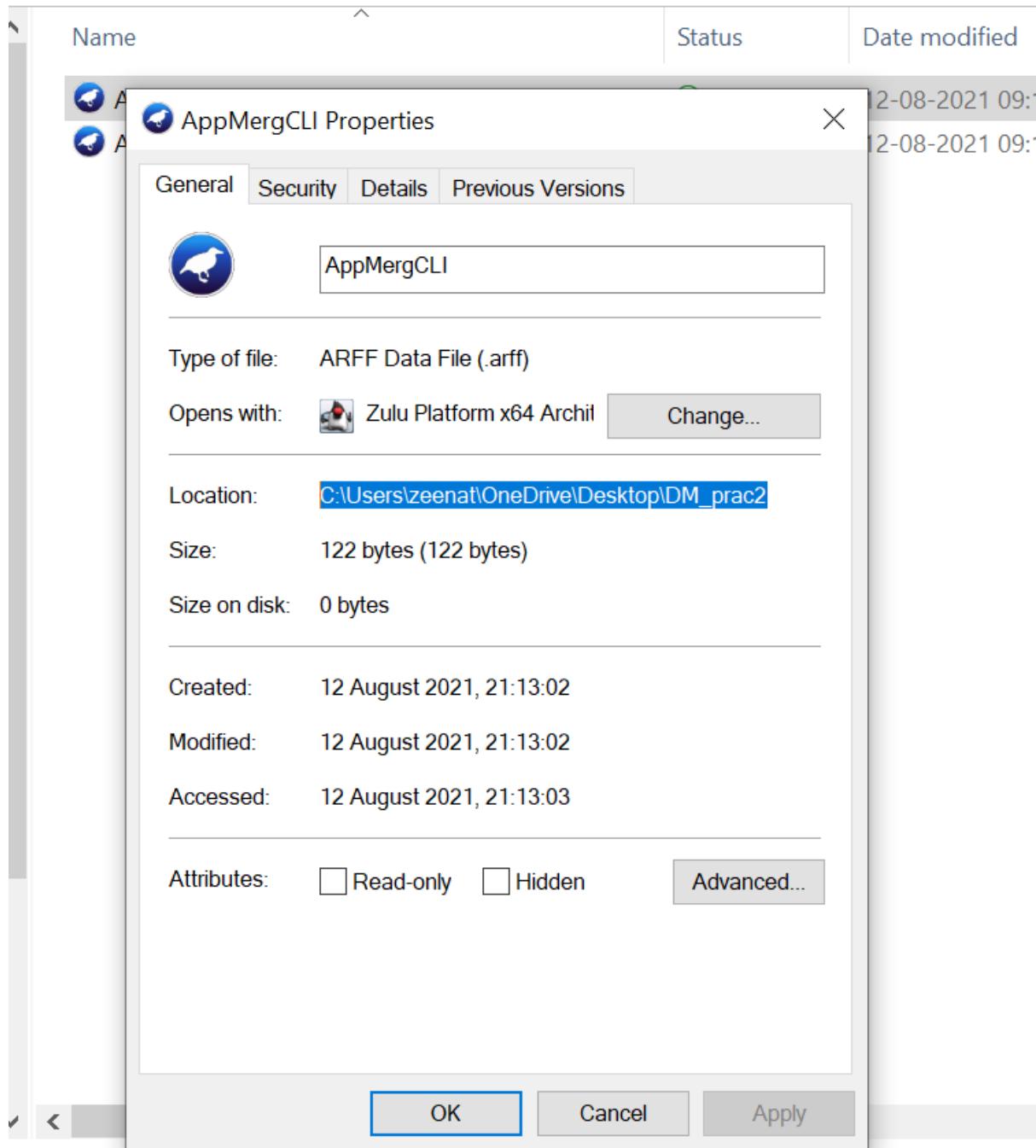
```

Step 3: Open the two arff Files created and right click and select on Properties and copy the path and paste it in the command given below.

Name: Zeenat

Class: TYIT

Roll no: 578



Name: Zeenat

Class: TYIT

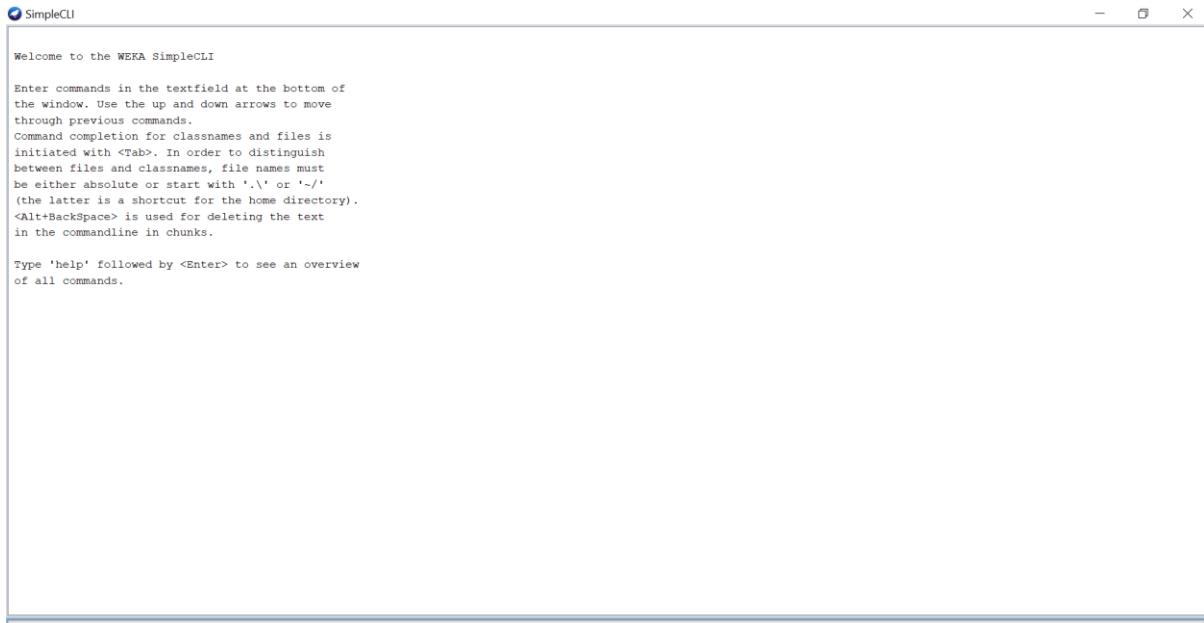
Roll no: 578

Name	Status	Date mod
AppMergCLI	✓	12-08-2021
AppMergCLI1	✓	12-08-2021
AppMergCLI1 Properties	X	
General	Security	Details
Previous Versions		
	AppMergCLI1	
Type of file:	ARFF Data File (.arff)	
Opens with:		Zulu Platform x64 Archiver
Change...		
Location:	C:\Users\zeenat\OneDrive\Desktop\DM_prac2	
Size:	277 bytes (277 bytes)	
Size on disk:	0 bytes	
Created:	12 August 2021, 21:12:37	
Modified:	12 August 2021, 21:32:32	
Accessed:	12 August 2021, 21:32:32	
Attributes:	<input type="checkbox"/> Read-only	<input type="checkbox"/> Hidden
Advanced...		
OK		Cancel
Apply		

Step 4: Open Weka Software and go to SimpleCLI Mode and type the following command in the box given below:

```
java weka.core.Instances append C:\Users\Public\Prac\fruits1.arff  
C:\Users\Public\Prac\fruits2.arff > C:\Users\Public\Prac\integrate.arff
```

*Note : There should be no spaces in File or Folder name or else can be replaced by underscore.



The screenshot shows the WEKA SimpleCLI interface. The title bar says "SimpleCLI". The main area displays the following text:

```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.' or '-'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

Type 'help' followed by <Enter> to see an overview
of all commands.
```

In the bottom command-line window, the following command is visible:

```
zeenat/OneDrive/Desktop/DM_prac2/AppMergCLI.arff C:/Users/zeenat/OneDrive/Desktop/DM_prac2/AppMergCLI1.arff > C:/Users/zeenat/OneDrive/Desktop/DM_prac2/AppMergCLlappend.arff
```

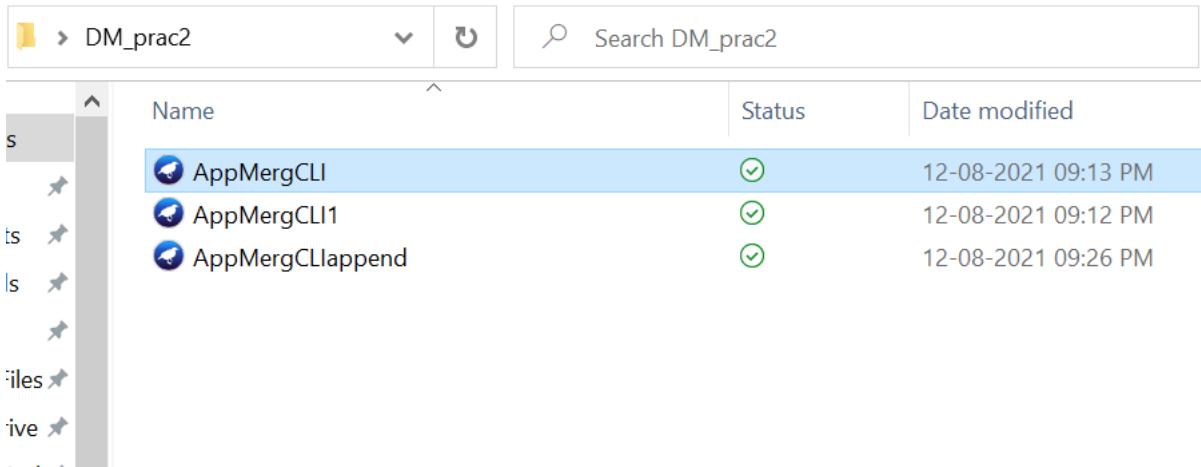
```
java weka.core Instances append
C:/Users/zeenat/OneDrive/Desktop/DM_prac2/AppMergCLI.arff
C:/Users/zeenat/OneDrive/Desktop/DM_prac2/AppMergCLI1.arff >
C:/Users/zeenat/OneDrive/Desktop/DM_prac2/AppMergCLlappend.arff
```

Step 5: If a new File is created with content of two arff Files that means we have done with Integration of two Files and open new File with Notepad to see the Output.

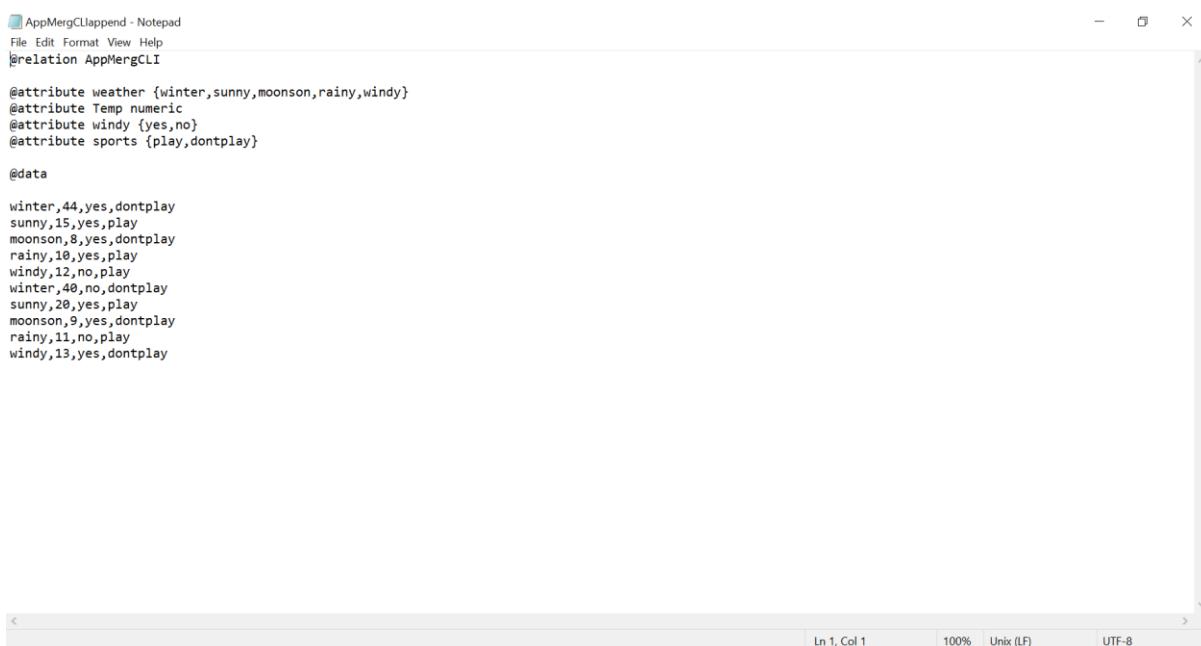
Name: Zeenat

Class: TYIT

Roll no: 578



Name	Status	Date modified
AppMergCLI	✓	12-08-2021 09:13 PM
AppMergCLI1	✓	12-08-2021 09:12 PM
AppMergCLIappend	✓	12-08-2021 09:26 PM



```
AppMergCLIappend - Notepad
File Edit Format View Help
@relation AppMergCLI

@attribute weather {winter,sunny,moonson,rainy,windy}
@attribute Temp numeric
@attribute windy {yes,no}
@attribute sports {play,dontplay}

@data

winter,44,yes,dontplay
sunny,15,yes,play
moonson,8,yes,dontplay
rainy,10,yes,play
windy,12,no,play
winter,40,no,dontplay
sunny,28,yes,play
moonson,9,yes,dontplay
rainy,11,no,play
windy,13,yes,dontplay
```

Ln 1, Col 1 100% Unix (LF) UTF-8

Practical 3**Aim: To perform OLAP Operations in Data Mining****Theory:**

Online Analytical Processing (OLAP) is a category of software that allows users to analyze information from multiple database systems at the same time. It is a technology that enables analysts to extract and view business data from different points of view. It is based on multidimensional data model and allows the user to query on multi-dimensional data.

Analysts frequently need to group, aggregate and join data. These OLAP operations in data mining are resource intensive. With OLAP data can be pre-calculated and pre-aggregated, making analysis faster.

OLAP databases are divided into one or more cubes. The cubes are designed in such a way that creating and viewing reports become easy.

OLAP cube:

The OLAP cube is a data structure optimized for very quick data analysis.

The OLAP Cube consists of numeric facts called measures which are categorized by dimensions. OLAP Cube is also called the hypercube.

Four types of analytical OLAP operations are:

1. Roll-up
2. Drill-down
3. Slice and dice
4. Pivot

Step 1: Open Oracle Live and create a Table with some Records.

SQL Worksheet

```
1 create table salestable(product varchar(10) not null,quarter varchar(10) not null, region varchar(20)
2 not null, sales int not null);
3
4 insert into salestable values('A','Q1','EUROPE',10);
5 insert into salestable values('A','Q1','AMERICA',20);
6 insert into salestable values('A','Q2','EUROPE',20);
7 insert into salestable values('A','Q2','AMERICA',50);
8 insert into salestable values('A','Q3','AMERICA',20);
9 insert into salestable values('A','Q4','EUROPE',10);
10 insert into salestable values('A','Q4','AMERICA',30);
11 insert into salestable values('B','Q1','EUROPE',40);
12 insert into salestable values('B','Q1','AMERICA',60);
13 insert into salestable values('B','Q2','EUROPE',20);
14 insert into salestable values('B','Q2','AMERICA',10);
15 insert into salestable values('B','Q3','AMERICA',20);
16 insert into salestable values('B','Q4','EUROPE',10);
17 insert into salestable values('B','Q4','AMERICA',40);
18
19 |
```

Table created.

1 row(s) inserted.

Step 2: Fire Queries on data by using GROUP BY Clauses and use Aggregate Functions like Sum, Count, Average, Max and Min. Also use Roll Up and Cube Clauses.

Name: Zeenat

Class: TYIT

Roll no: 578

The screenshot shows a "Live SQL" interface with a "SQL Worksheet". The query entered is:

```
1 SELECT QUARTER, REGION, SUM(SALES)
2 FROM SALESTABLE
3 GROUP BY CUBE ((QUARTER, REGION))
```

The results are displayed in two tables. The first table shows data for the first three quarters:

QUARTER	REGION	SUM(SALES)
-	-	360
-	EUROPE	110
-	AMERICA	250
Q1	-	130
Q1	EUROPE	50
Q1	AMERICA	80
Q2	-	100
Q2	EUROPE	40

The second table shows data for the last four quarters:

Q2	-	100
Q2	EUROPE	40
Q2	AMERICA	60
Q3	-	40
Q3	AMERICA	40
Q4	-	90
Q4	EUROPE	20
Q4	AMERICA	70

[Download CSV](#)
14 rows selected.

Live SQL

Feedback Help zeenatrab@gmail.com

SQL Worksheet

Clear Find Actions Save Run

```
1 SELECT CASE WHEN grouping(QUARTER) = 1 THEN 'All' ELSE QUARTER END AS QUARTER, CASE
2 WHEN grouping(REGION) = 1 THEN 'All' ELSE REGION END AS REGION, SUM(SALES)
3 FROM SALESTABLE
4 GROUP BY CUBE (QUARTER, REGION)
```

QUARTER	REGION	SUM(SALES)
All	All	360
All	EUROPE	110
All	AMERICA	250
Q1	All	130
Q1	EUROPE	50

```
1 SELECT CASE WHEN grouping(QUARTER) = 1 THEN 'All' ELSE QUARTER END AS QUARTER, CASE
2 WHEN grouping(REGION) = 1 THEN 'All' ELSE REGION END AS REGION, SUM(SALES)
3 FROM SALESTABLE
4 GROUP BY CUBE (QUARTER, REGION)
```

QUARTER	REGION	SUM(SALES)
Q1	AMERICA	80
Q2	All	100
Q2	EUROPE	40
Q2	AMERICA	60
Q3	All	40
Q3	AMERICA	40
Q4	All	90
Q4	EUROPE	20
Q4	AMERICA	70

Download CSV
14 rows selected.

Result from SQL query with Cube operator

2: ROLLUP

```
1 SELECT QUARTER, REGION, SUM(SALES)
2 FROM SALESTABLE
3 GROUP BY ROLLUP (QUARTER, REGION)
```

QUARTER	REGION	SUM(SALES)
Q1	EUROPE	50
Q1	AMERICA	80
Q1	-	130
Q2	EUROPE	40
Q2	AMERICA	60
Q2	-	100
Q3	AMERICA	40
Q3	-	40
Q4	EUROPE	20

Q4	AMERICA	70
Q4	-	90
-	-	360

[Download CSV](#)

12 rows selected.

Result from SQL query with ROLLUP operator

GROUPING SETS

This query is equivalent to

```
1 SELECT QUARTER, REGION, SUM(SALES)
2 FROM SALESTABLE
3 GROUP BY GROUPING SETS ((QUARTER), (REGION))
```

QUARTER	REGION	SUM(SALES)
Q4	-	90
Q1	-	130
Q2	-	100
Q3	-	40
-	EUROPE	110
-	AMERICA	250

[Download CSV](#)

6 rows selected.

SQL Worksheet

 Clear Find

```
1 SELECT QUARTER, NULL, SUM(SALES)
2 FROM SALESTABLE
3 GROUP BY QUARTER
4 UNION ALL
5 SELECT NULL, REGION, SUM(SALES)
6 FROM SALESTABLE
7 GROUP BY REGION
```

QUARTER	NULL	SUM(SALES)
Q4	-	90
Q1	-	130
Q2	-	100
Q3	-	40
-	EUROPE	110
-	AMERICA	250

[Download CSV](#)

6 rows selected.

Multiple CUBE, ROLLUP and GROUPING SETS statements can be used in a single SQL query. Different combinations of CUBE, ROLLUP and GROUPING SETS can generate equivalent result sets. Consider the following qu

Name: Zeenat

Class: TYIT

Roll no: 578

```
1 SELECT QUARTER, REGION, SUM(SALES)
2 FROM SALESTABLE
3 GROUP BY CUBE (QUARTER, REGION)
```

QUARTER	REGION	SUM(SALES)
-	-	360
-	EUROPE	110
-	AMERICA	250
Q1	-	130
Q1	EUROPE	50
Q1	AMERICA	80

Q2	-	100
Q2	EUROPE	40
Q2	AMERICA	60
Q3	-	40
Q3	AMERICA	40
Q4	-	90
Q4	EUROPE	20
Q4	AMERICA	70

[Download CSV](#)

14 rows selected.

Name: Zeenat

Class: TYIT

Roll no: 578

```
1 SELECT QUARTER, REGION, SUM(SALES)
2 FROM SALESTABLE
3 GROUP BY GROUPING SETS ((QUARTER, REGION), (QUARTER), (REGION), ())
```

QUARTER	REGION	SUM(SALES)
Q1	EUROPE	50
Q1	AMERICA	80
Q2	EUROPE	40
Q2	AMERICA	60
Q3	AMERICA	40
Q4	EUROPE	20

Q4	AMERICA	70
-	AMERICA	250
-	EUROPE	110
-	-	360
Q1	-	130
Q2	-	100
Q3	-	40
Q4	-	90

[Download CSV](#)

14 rows selected.

Name: Zeenat

Class: TYIT

Roll no: 578

```
1 SELECT QUARTER, REGION, SUM(SALES)
2 FROM SALESTABLE
3 GROUP BY ROLLUP (QUARTER, REGION)
```

QUARTER	REGION	SUM(SALES)
Q1	EUROPE	50
Q1	AMERICA	80
Q1	-	130
Q2	EUROPE	40
Q2	AMERICA	60
Q2	-	100
Q3	AMERICA	40

Q3	-	40
Q4	EUROPE	20
Q4	AMERICA	70
Q4	-	90
-	-	360

[Download CSV](#)

12 rows selected.

Name: Zeenat

Class: TYIT

Roll no: 578

```
1 SELECT QUARTER, REGION, SUM(SALES)
2 FROM SALESTABLE
3 GROUP BY GROUPING SETS ((QUARTER, REGION), (QUARTER),())
```

QUARTER	REGION	SUM(SALES)
Q1	EUROPE	50
Q1	AMERICA	80
Q1	-	130
Q2	EUROPE	40
Q2	AMERICA	60
Q2	-	100
Q3	AMERICA	40

Q3	-	40
Q4	EUROPE	20
Q4	AMERICA	70
Q4	-	90
-	-	360

[Download CSV](#)

12 rows selected.

Practical 4

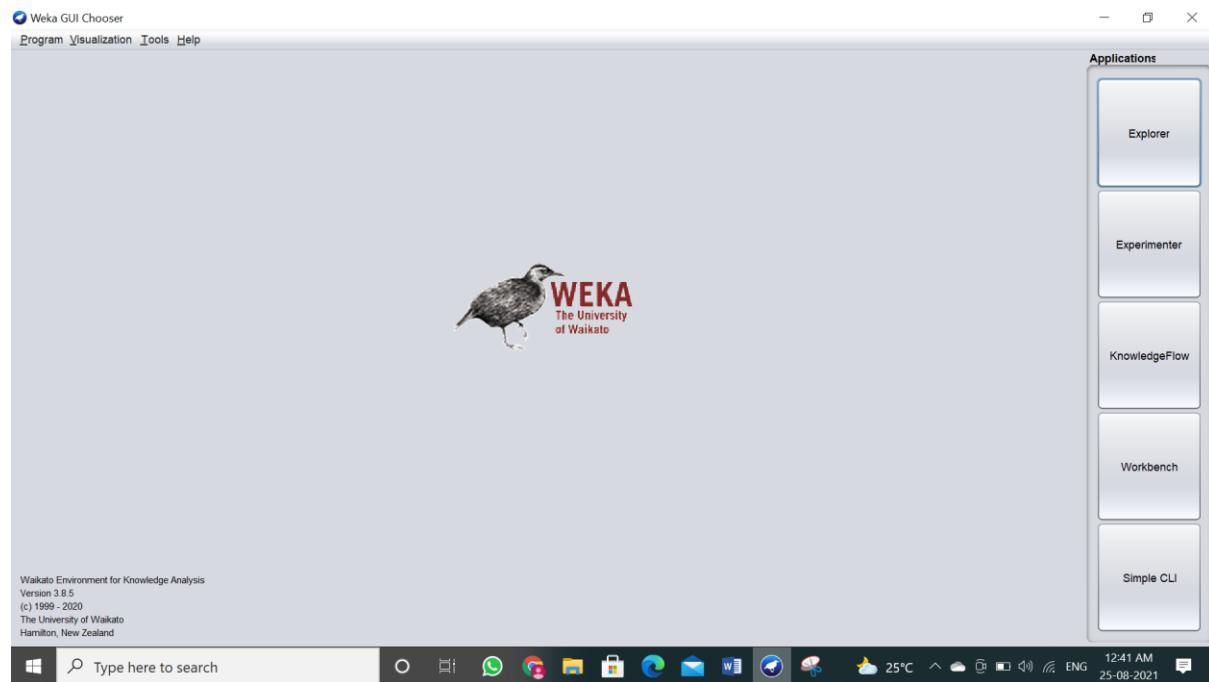
Aim: To perform kth nearest Algorithm.

Theory:

KNN (K — Nearest Neighbors) is one of many (supervised learning) algorithms used in data mining and machine learning, it's a classifier algorithm where the learning is based "how similar" is a data (a vector) from other. A k-nearest-neighbor algorithm, often abbreviated k-nn, is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in. The k-nearest-neighbor is an example of a "lazy learner" algorithm, meaning that it does not build a model using the training set until a query of the data set is performed.

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression)

Step 1: Open Weka Software and go to Explorer Option.

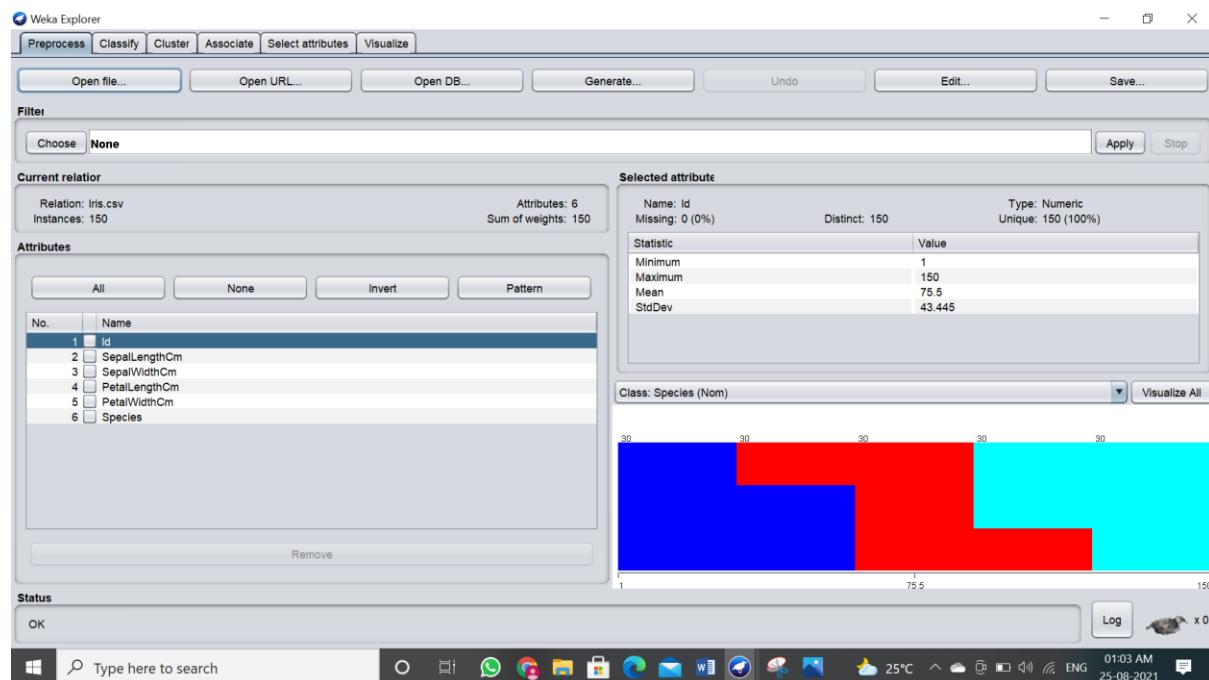
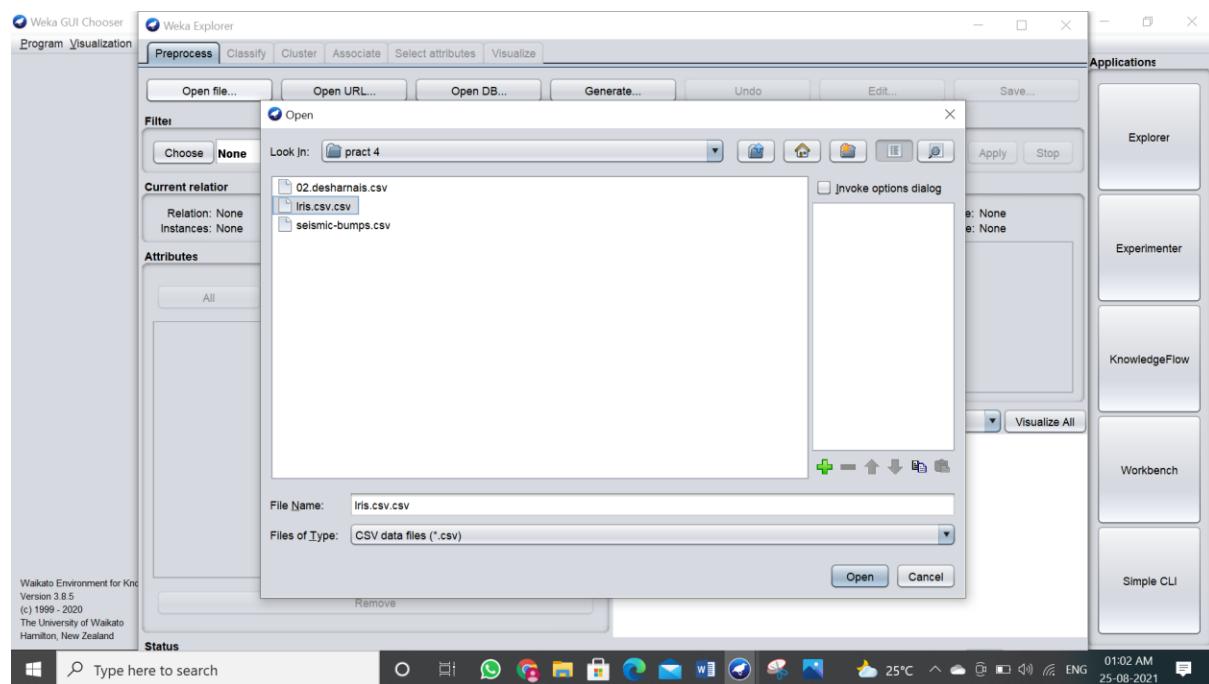


Step 2: Go to Open File option and choose a CSV File and click on Open.

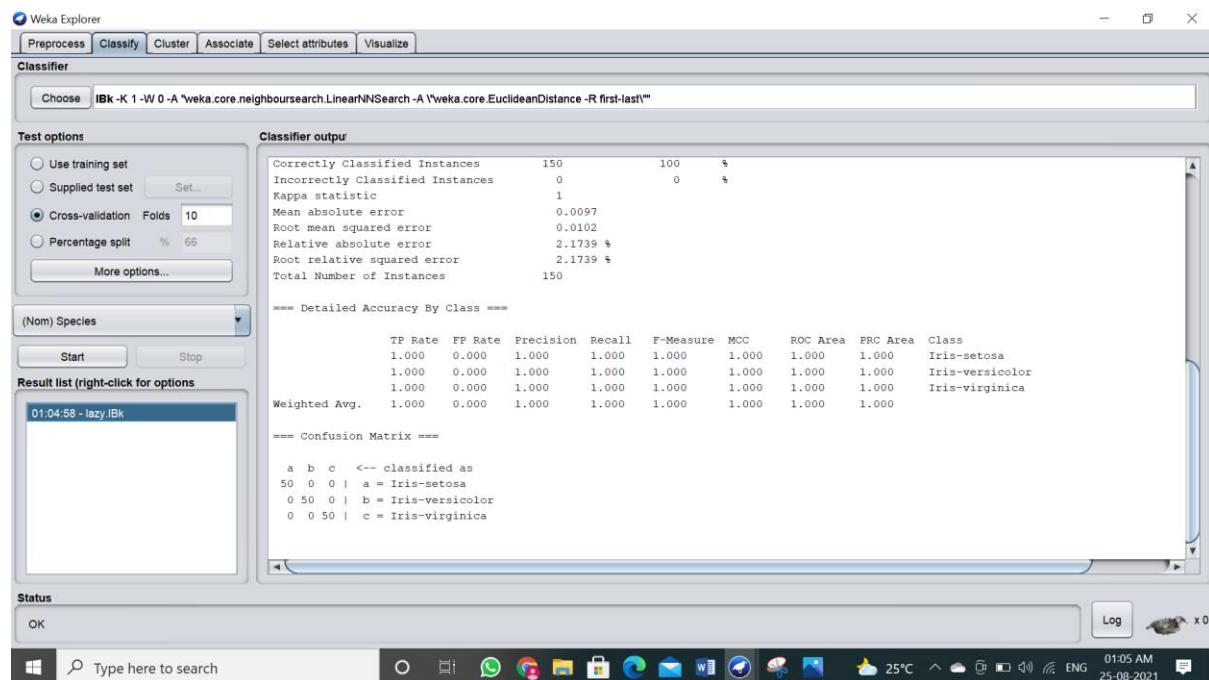
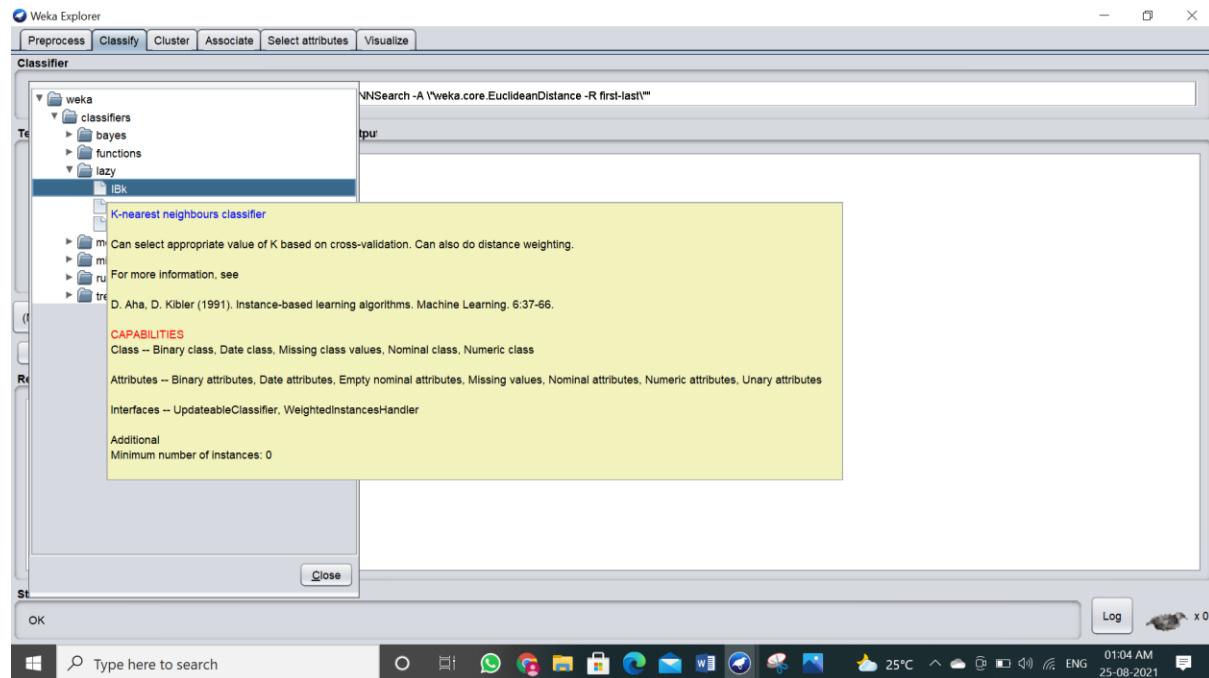
Name: Zeenat

Class: TYIT

Roll no: 578



Step 3: Go to Classify, click on Choose > Lazy > Ibk File and click on Start Button and we will get the Output.



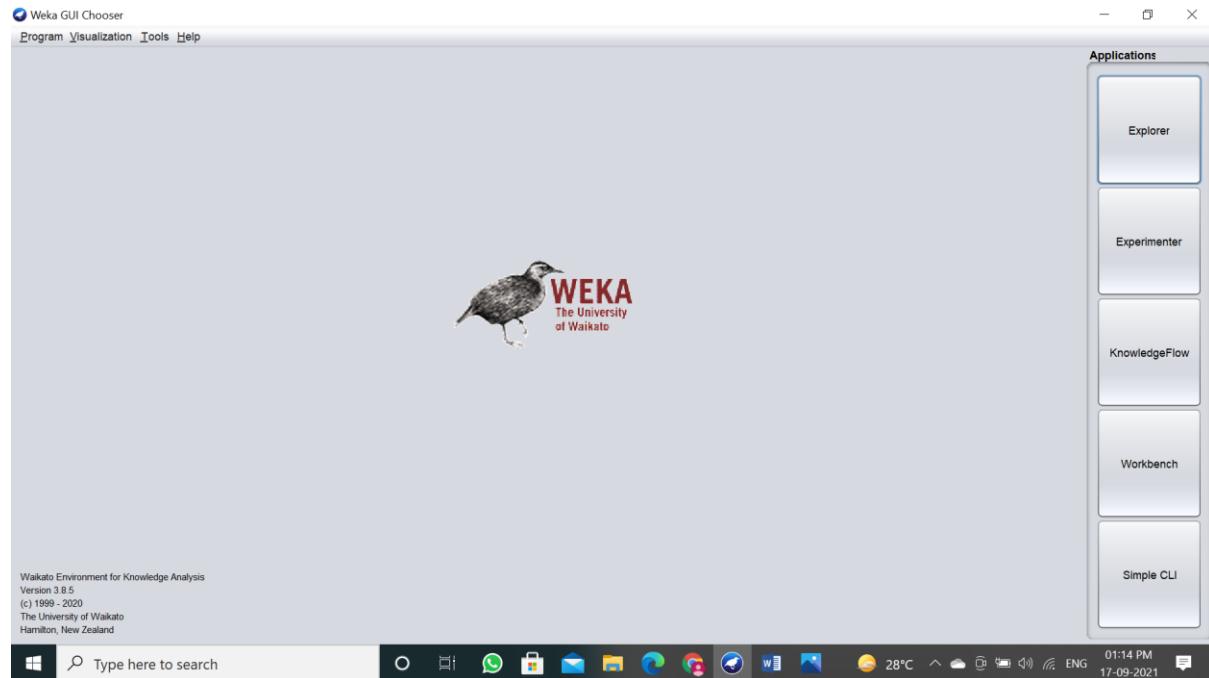
Practical 5

Aim: Demonstrate performing clustering on data sets.

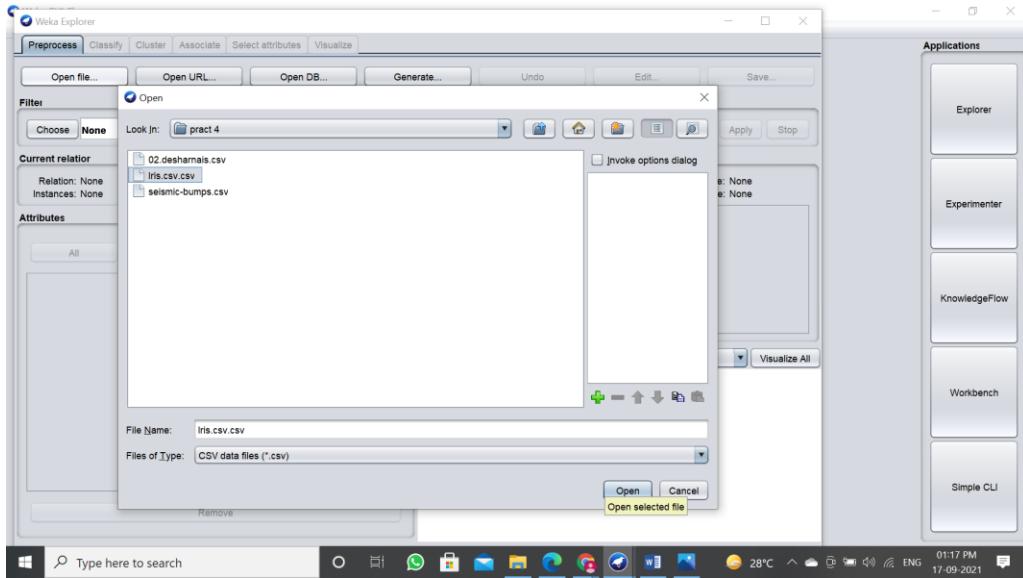
Theory:

Clustering only utilizes input data, to determine patterns, anomalies, or similarities in its input data. A good clustering algorithm aims to obtain clusters whose: The intracluster similarities are high, It implies that the data present inside the cluster is similar to one another. The process of making a group of abstract objects into classes of similar objects is known as clustering. In the process of cluster analysis, the first step is to partition the set of data into groups with the help of data similarity, and then groups are assigned to their respective labels.

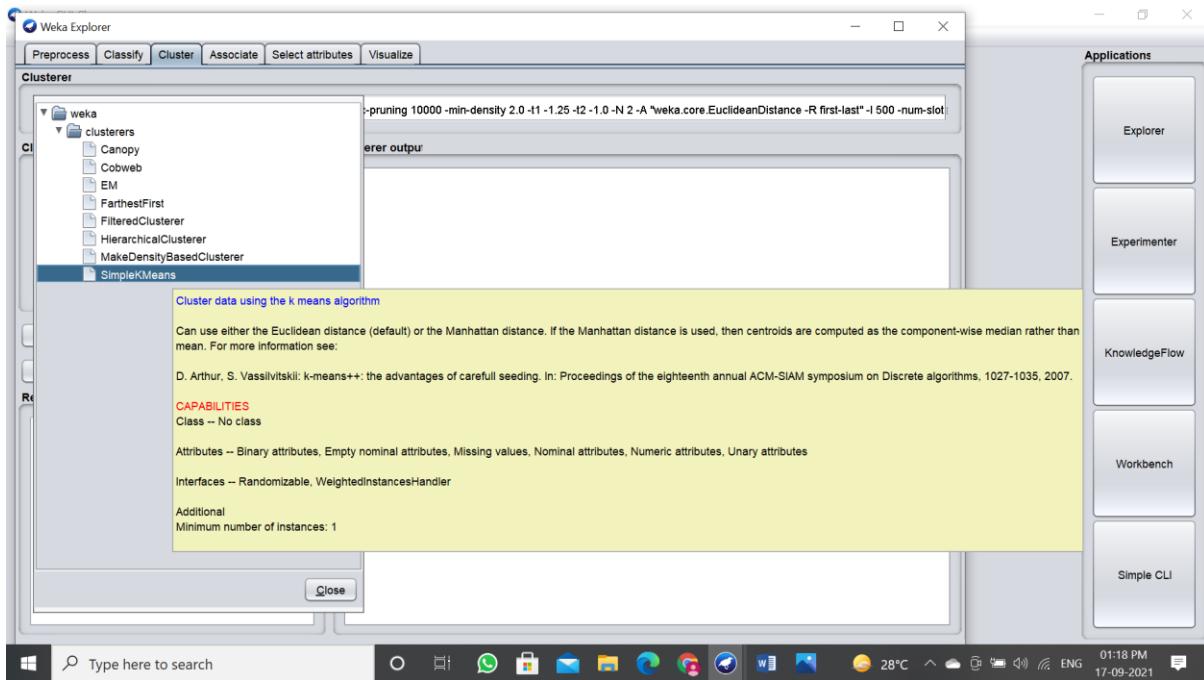
Step 1: Open Weka Software and go to Explorer Option.



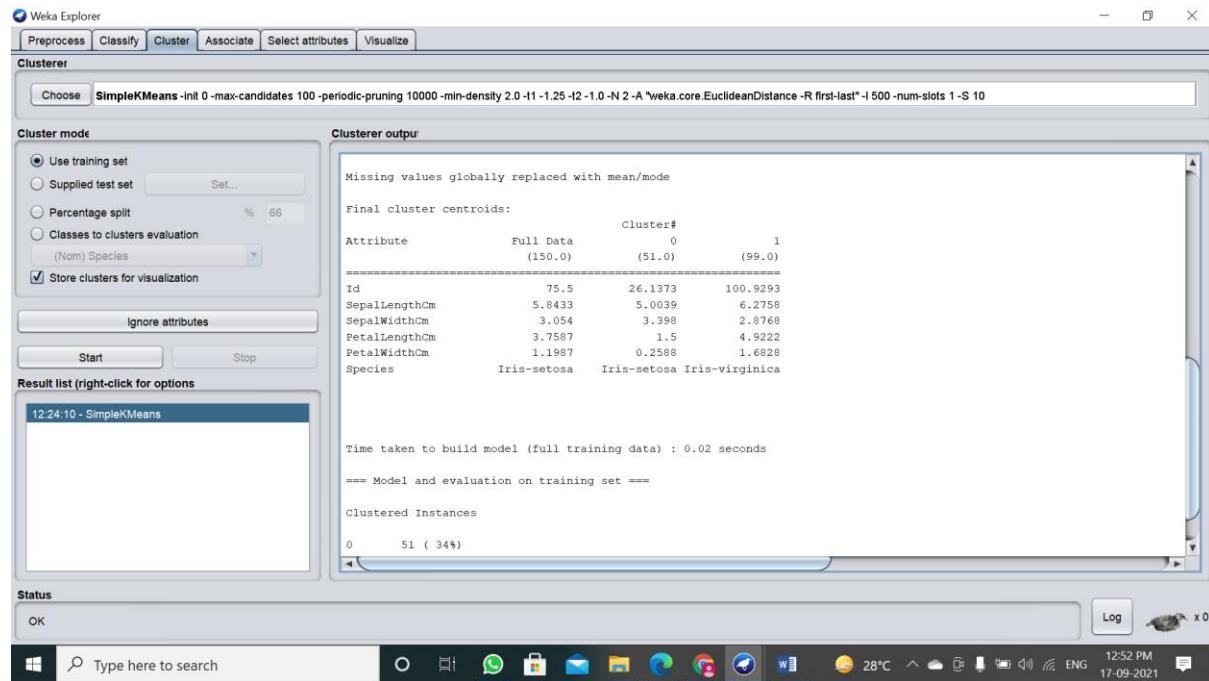
Step 2: Go to Open File option and choose a CSV File and click on Open.



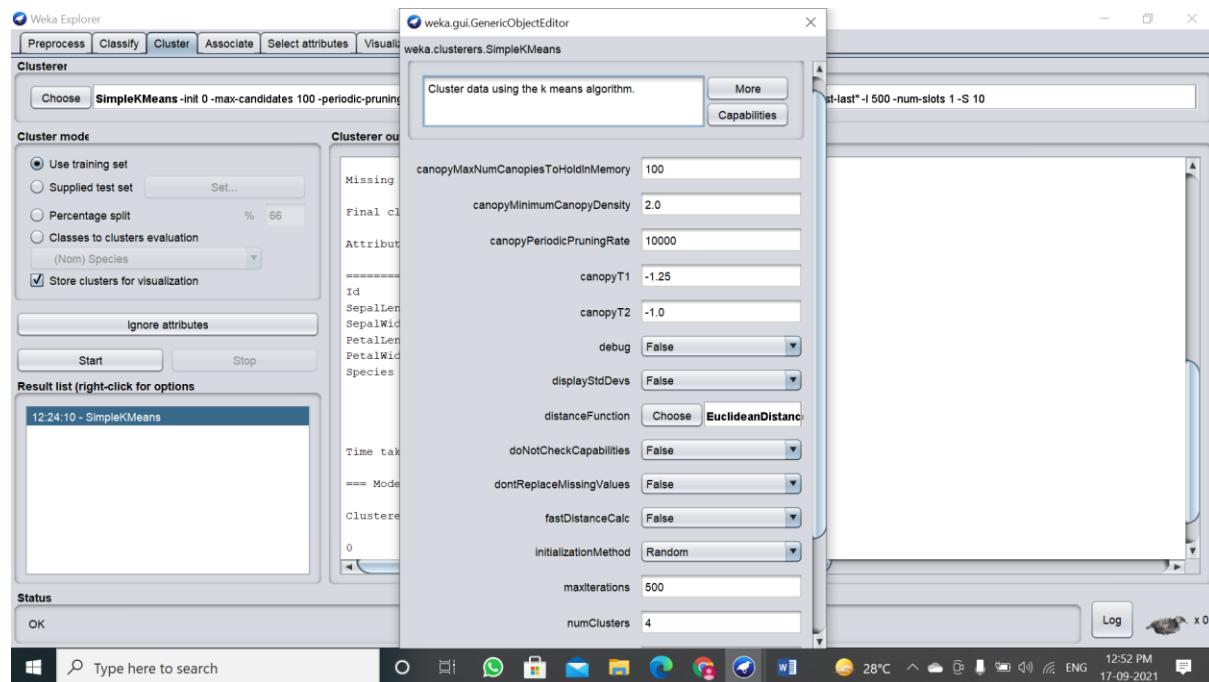
Step 3: Go to Cluster and Choose SimpleKMeans option.



Step 4: Click on Start.



Step 5: Open GenericObjectEditor and changes value numClusters to 4.

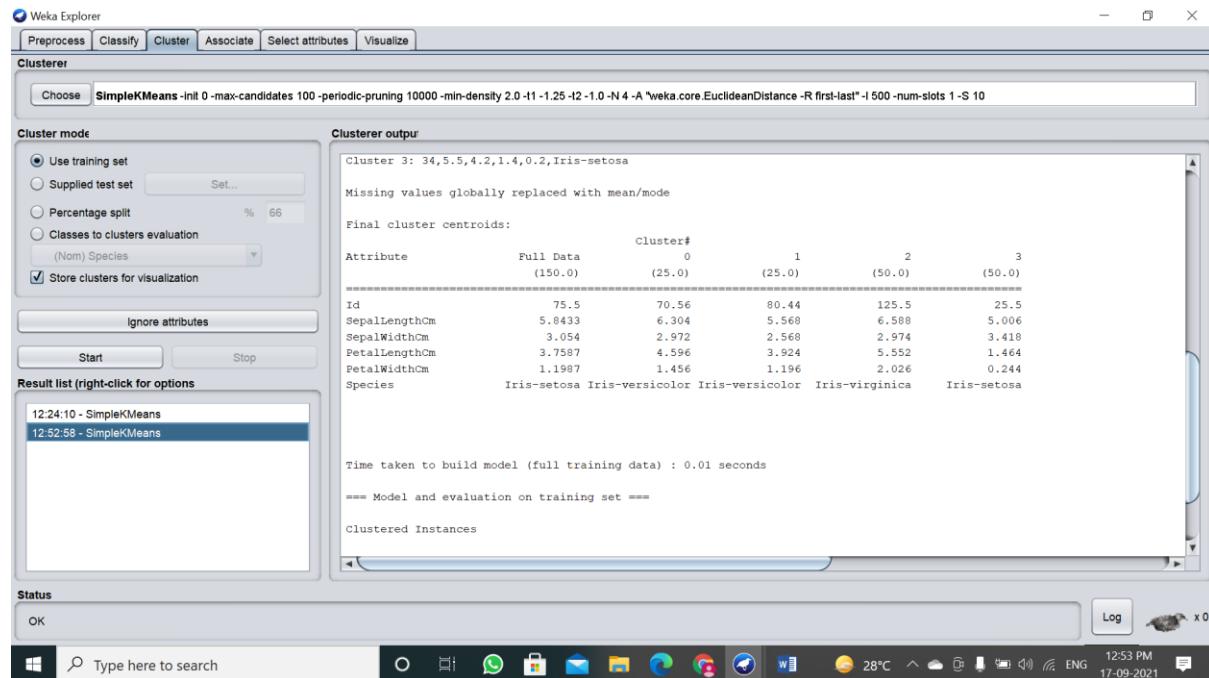


Output:

Name: Zeenat

Class: TYIT

Roll no: 578



Name: Zeenat

Class: TYIT

Roll no: 578

Practical 6

Aim: Demonstrate performing Regression on data sets.

Theory:

Regression algorithms predict the output values based on input features from the data fed in the system. The go-to methodology is the algorithm builds a model on the features of training data and using the model to predict value for new data. According to Oracle, here's a great definition of Regression – a data mining function to predict a number.

Step 1: Open Weka Software and go to Explorer Option.

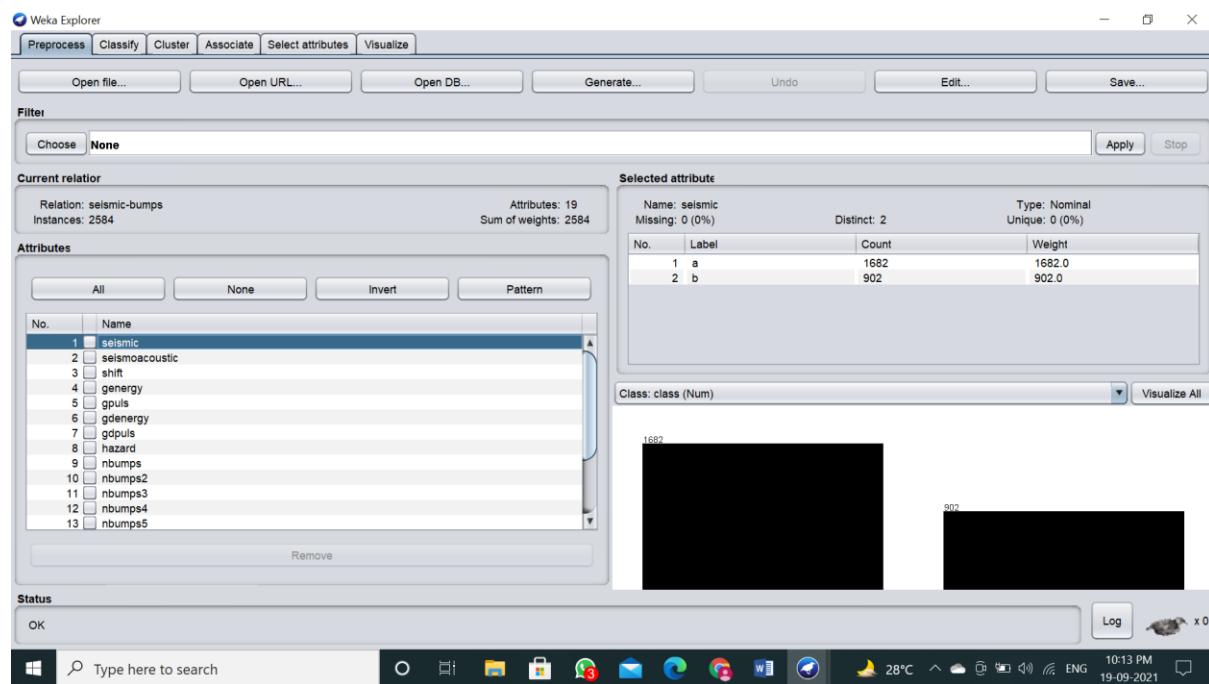
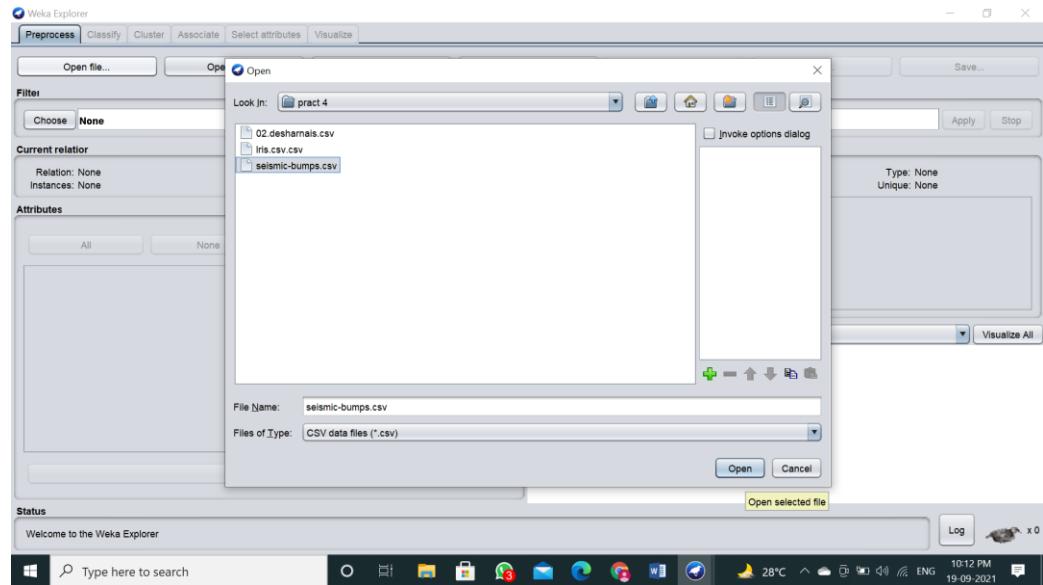


Step 2: Go to Open File option and choose a CSV File and click on Open.

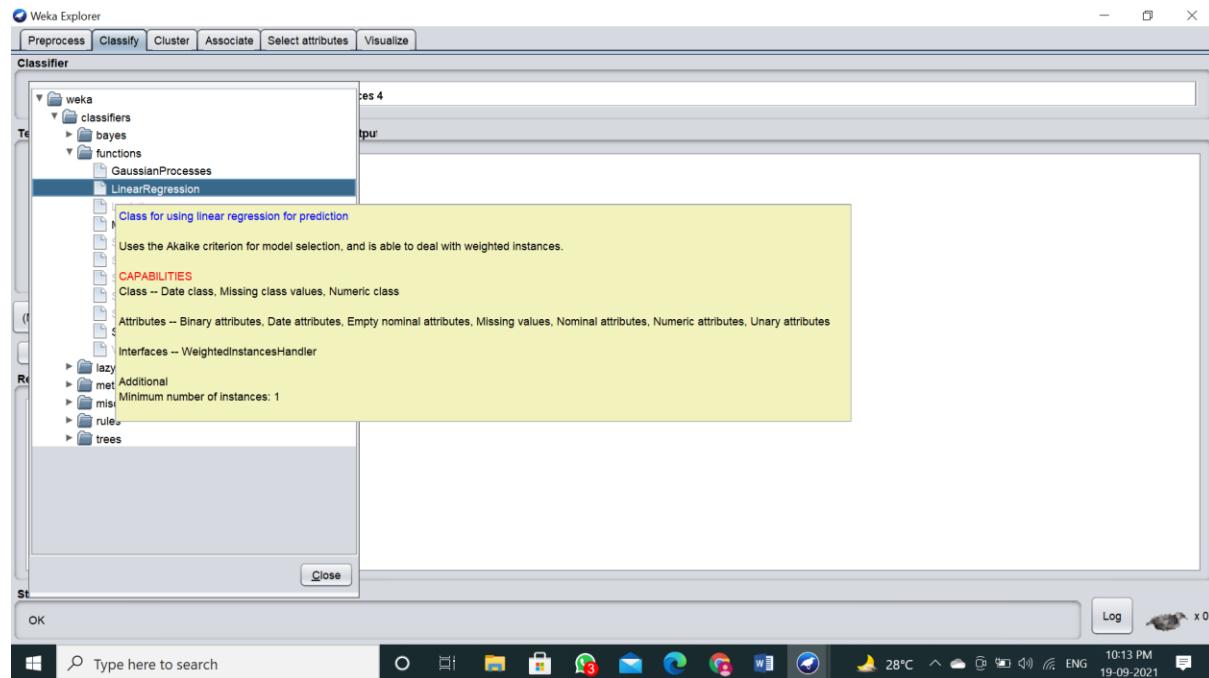
Name: Zeenat

Class: TYIT

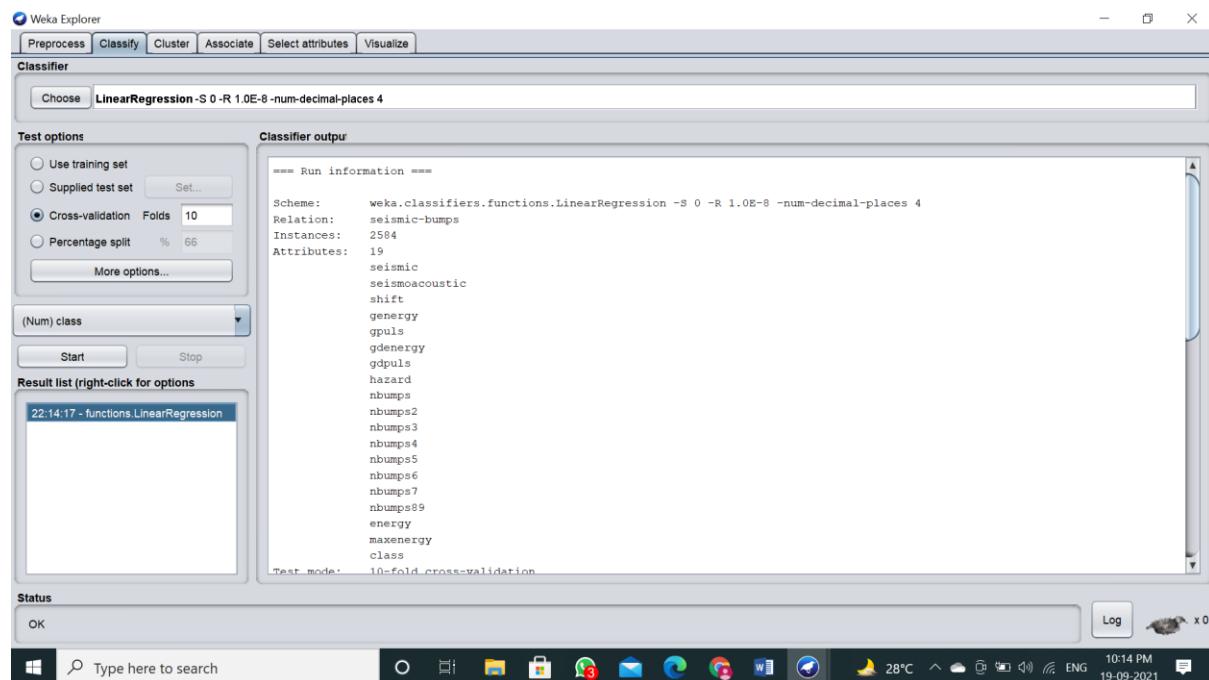
Roll no: 578



Step 3: Go to Classify > Choose > Functions > Linear Regression



Step 4: Click on Start for the Output.



Practical 7

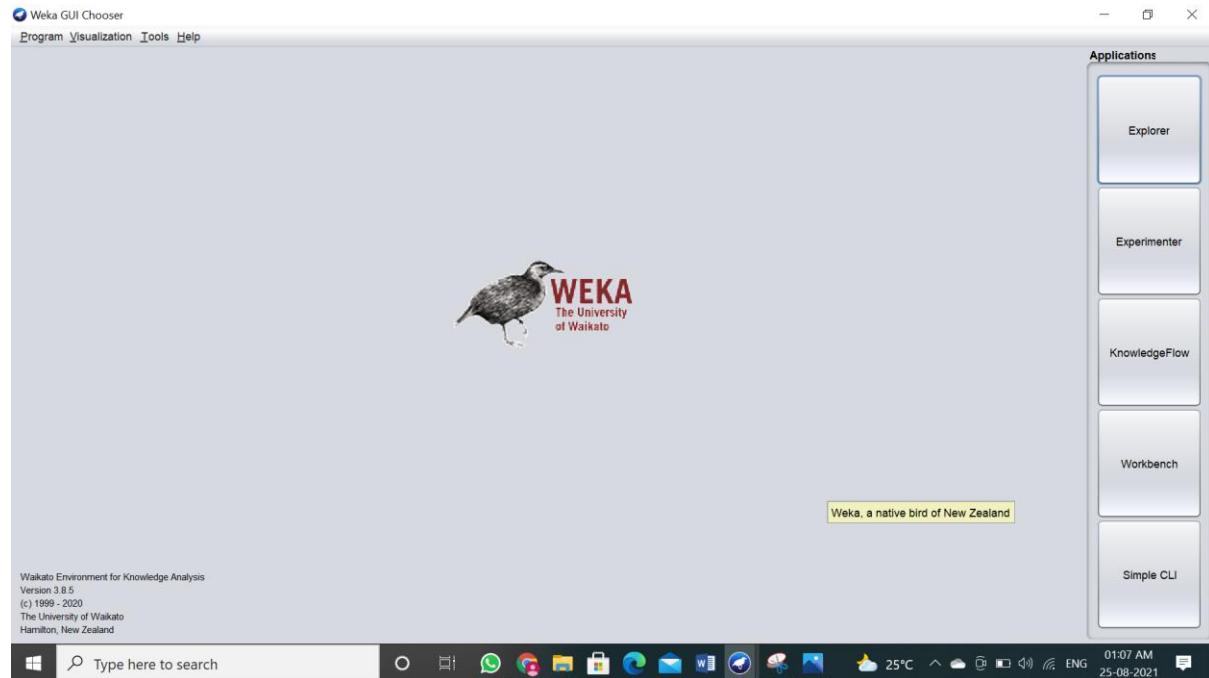
Aim: To implement Decision Tree

Theory:

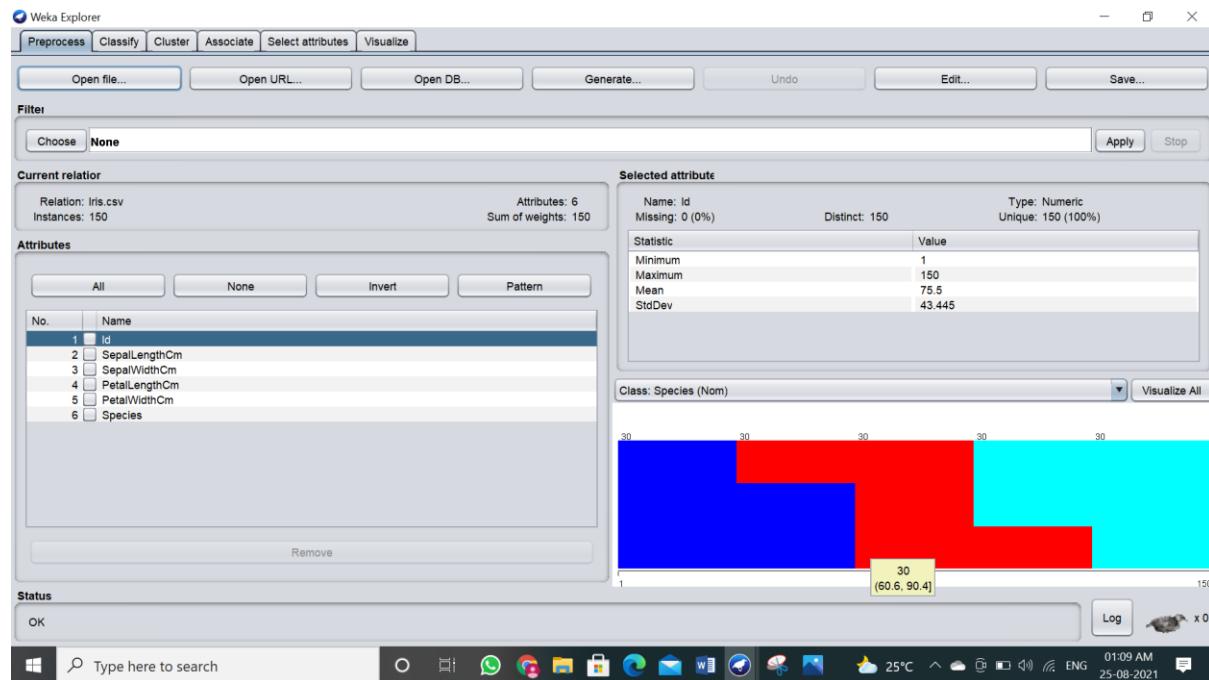
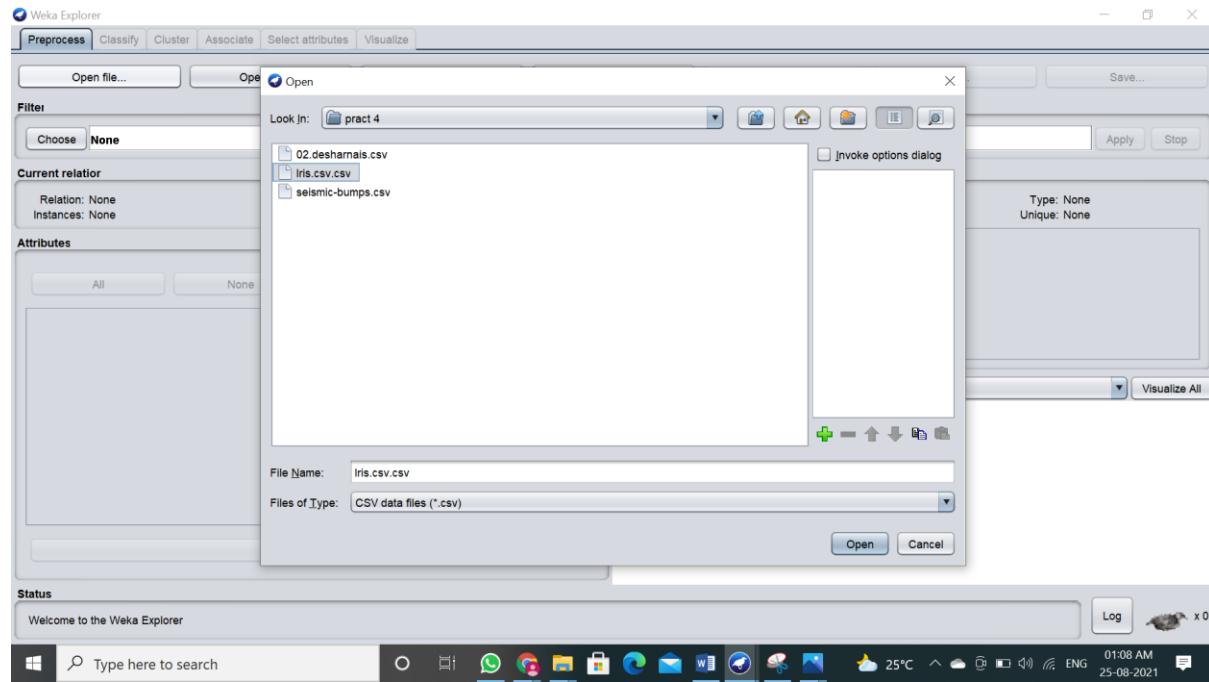
A decision tree is a supervised learning approach wherein we train the data present knowing the target variable. As the name suggests, this algorithm has a tree type of structure. Let us first look into the decision tree's theoretical aspect and then look into the same graphical approach. In Decision Tree, the algorithm splits the dataset into subsets based on the most important or significant attribute. The most significant attribute is designated in the root node, and that is where the splitting takes the place of the entire dataset present in the root node. This splitting done is known as decision nodes. In case no more split is possible, that node is termed as a leaf node. To stop the algorithm from reaching an overwhelming stage, a stop criterion is employed. One of the stop criteria is the minimum number of observations in the node before the split happens. While applying the decision tree in splitting the dataset, one must be careful that many nodes might have noisy data. To cater to an outlier or noisy data problems, we employ techniques known as Data Pruning. Data pruning is nothing but an algorithm to classify out data from the subset, making it difficult for learning from a given model.

The Decision Tree algorithm was released as ID3 (Iterative Dichotomiser) by machine researcher J. Ross Quinlan.

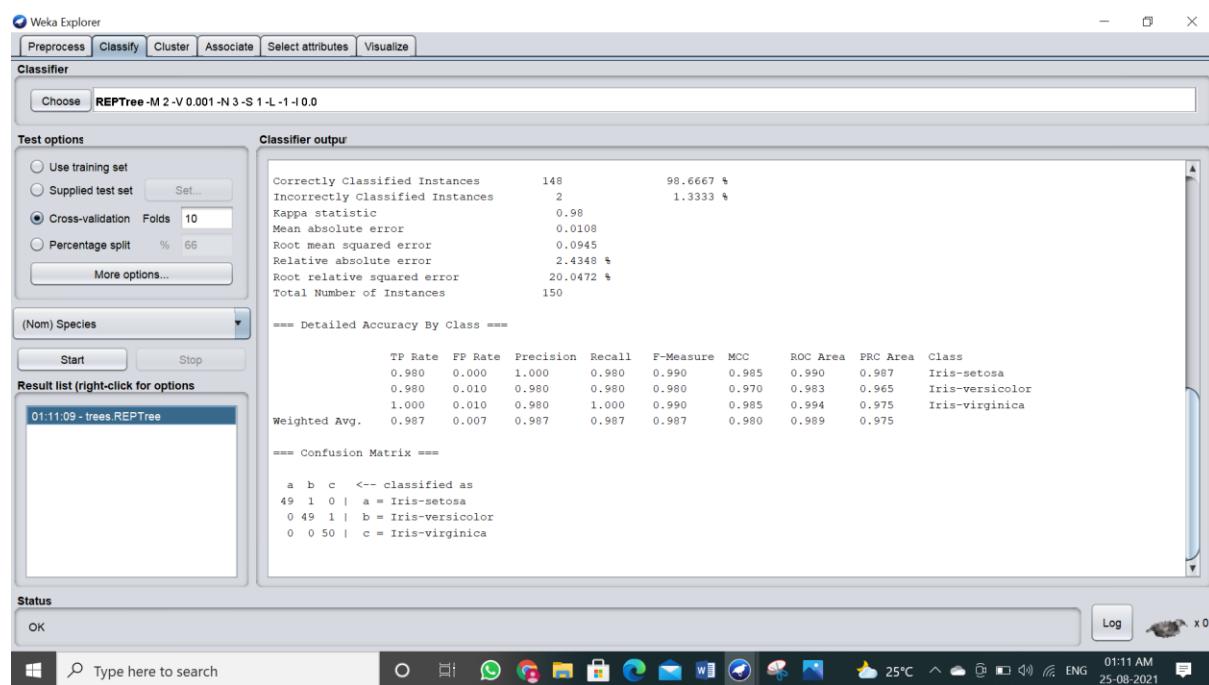
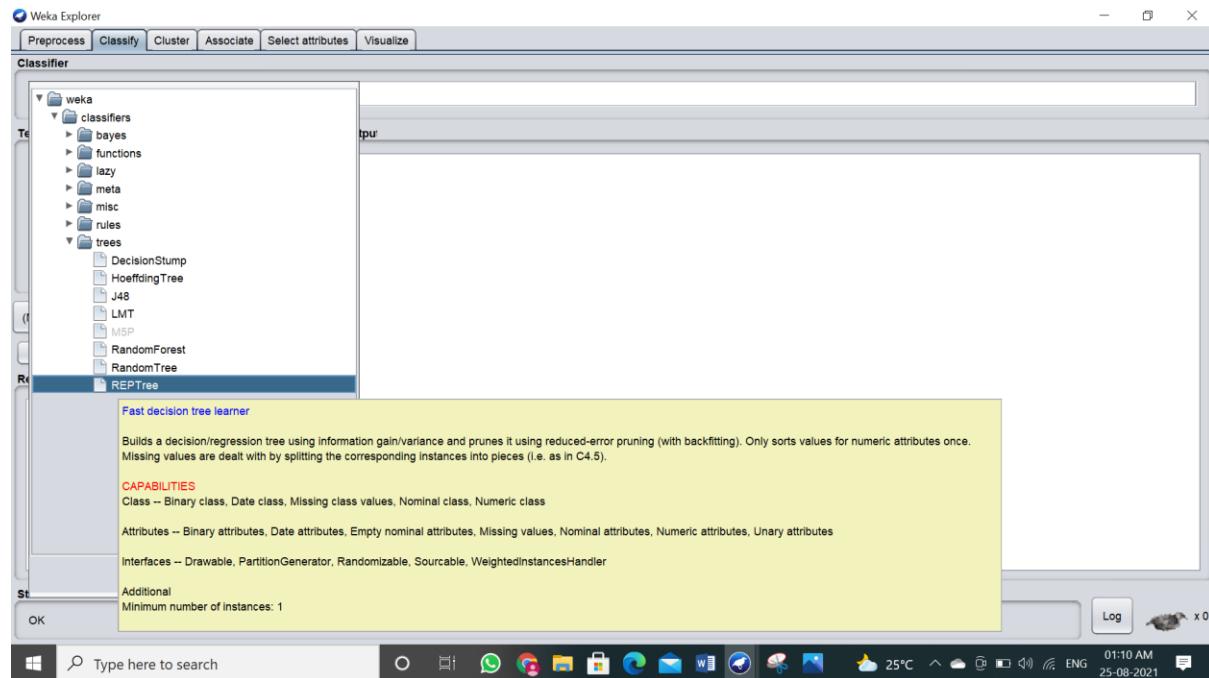
Step 1: Open Weka Software and go to Explorer.



Step 2: Go to Open File and open CSV File.



Step 3: Go to Classify > Choose > Trees > REPTree and click on Start Button to get Output



Step 4: To view the Tree right click on REPtree and click on Visualize Tree.

Name: Zeenat

Class: TYIT

Roll no: 578

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose REPTree -M 2 -V 0.001 -N 3 -S 1 -L 1 -I 0.0

Test options

Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...

(Nom) Species

Start Stop

Result list (right-click for options)

01:11:09 - trees.REPTree

- View in main window
- View in separate window
- Save result buffer
- Delete result buffer(s)
- Load model
- Save model
- Re-evaluate model on current test set
- Re-apply this model's configuration
- Visualize classifier errors

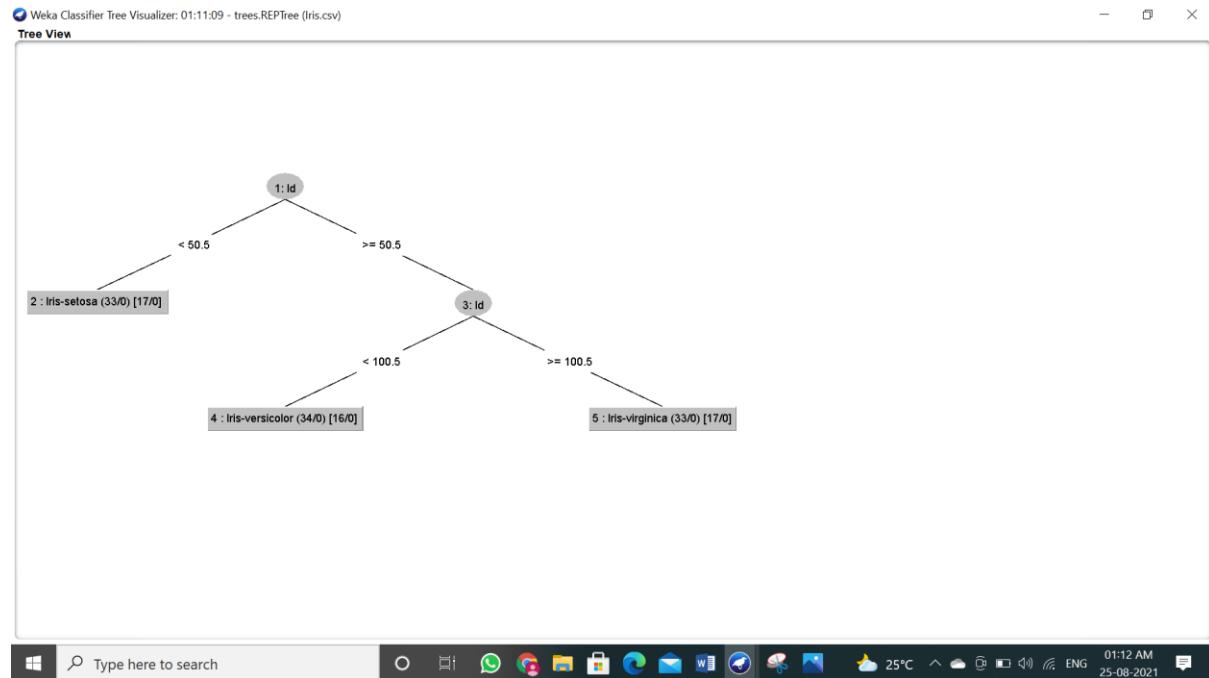
Status OK

Type here to see

Classifier output

	Correctly Classified Instances	148	98.6667 %
Incorrectly Classified Instances	2	1.3333 %	
Kappa statistic	0.98		
Mean absolute error	0.0108		
Root mean squared error	0.0945		
Relative absolute error	2.4348 %		
Root relative squared error	20.0472 %		
Total Number of Instances	150		

==== Detailed Accuracy By Class ====
Iris-setosa Iris-versicolor Iris-virginica
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.980 0.000 1.000 0.980 0.990 0.985 0.990 0.987 Iris-setosa
0.980 0.010 0.980 0.980 0.980 0.970 0.983 0.965 Iris-versicolor
1.000 0.010 0.980 1.000 0.990 0.985 0.994 0.975 Iris-virginica
0.007 0.987 0.987 0.987 0.980 0.989 0.989 0.975



Practical 8**Aim: Implement Naïve Bayes****Theory:**

The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions.

The independence assumptions often do not have an impact on reality. Therefore they are considered as naive. You can derive probability models by using Bayes' theorem (credited to Thomas Bayes). Depending on the nature of the probability model, you can train the Naive Bayes algorithm in a supervised learning setting.

Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.

A Naive Bayes model consists of a large cube that includes the following dimensions:

- Input field name
- Input field value for discrete fields, or input field value range for continuous fields.
Continuous fields are divided into discrete bins by the Naive Bayes algorithm
- Target field value

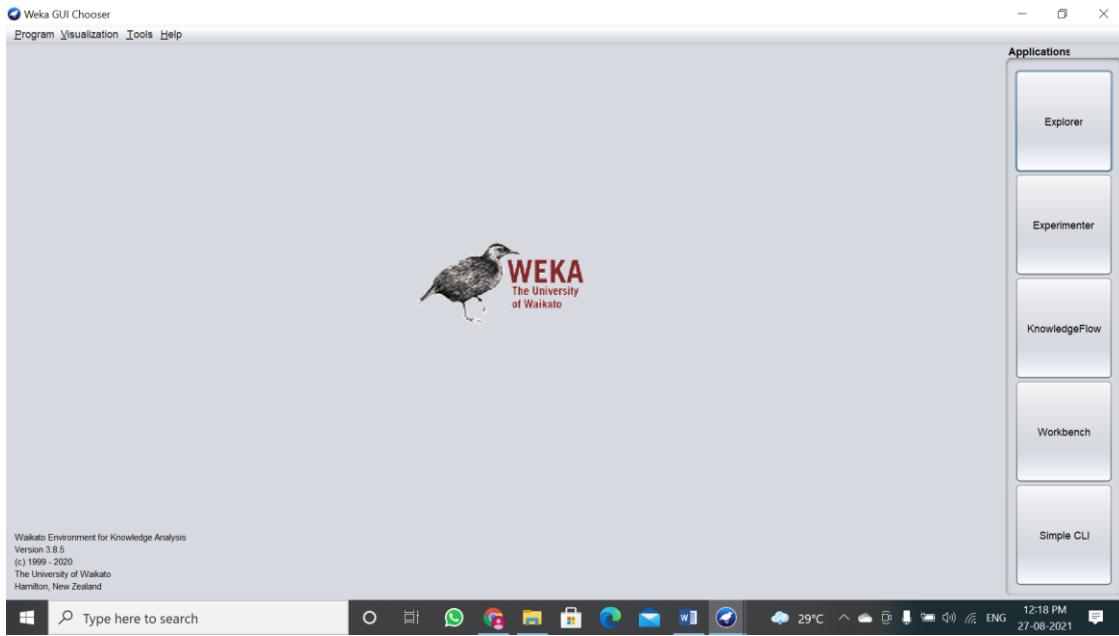
This means that a Naive Bayes model records how often a target field value appears together with a value of an input field.

Step 1: Open Weka Software and go to Explorer.

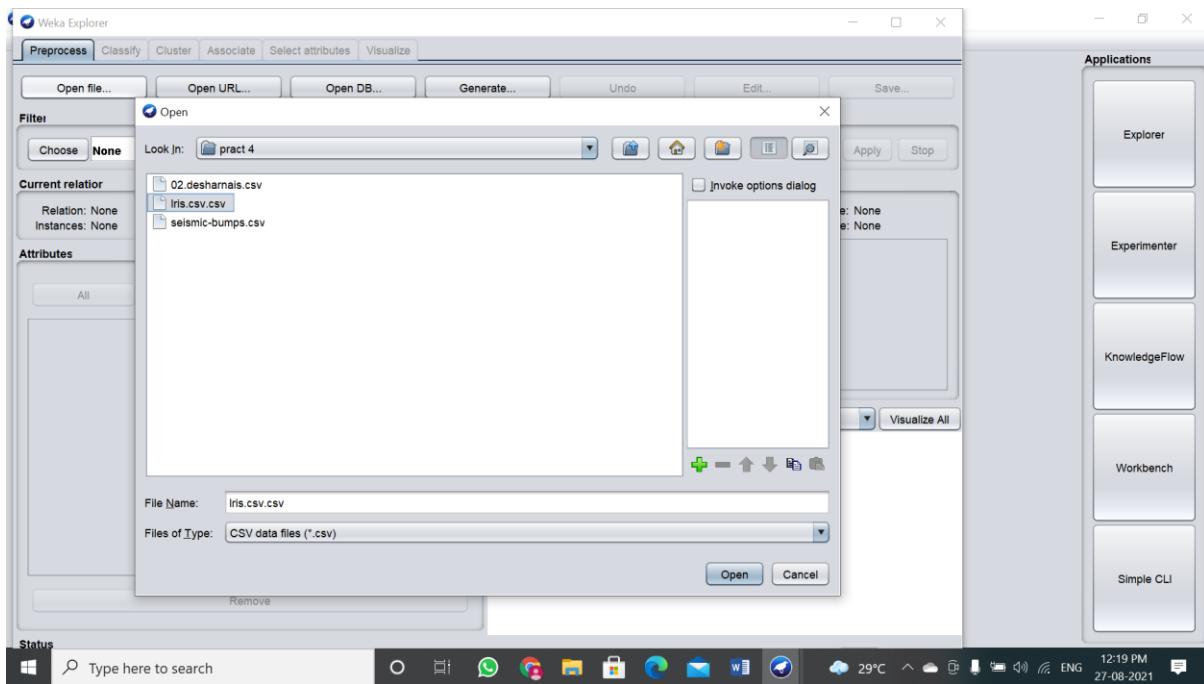
Name: Zeenat

Class: TYIT

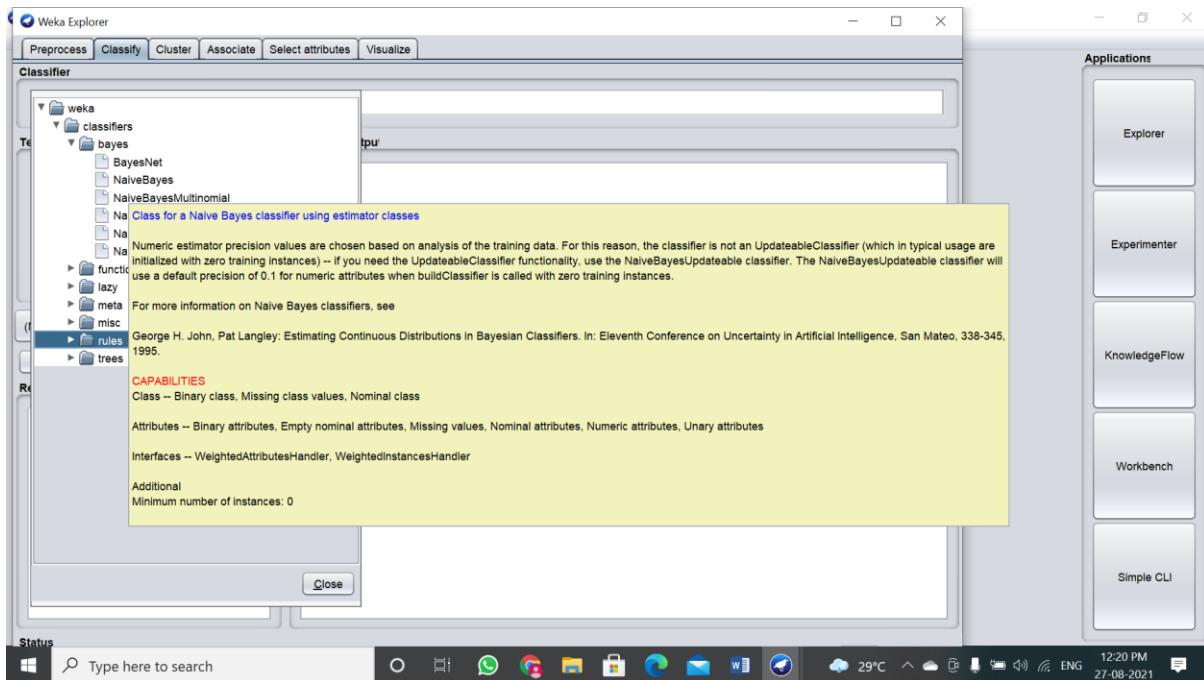
Roll no: 578



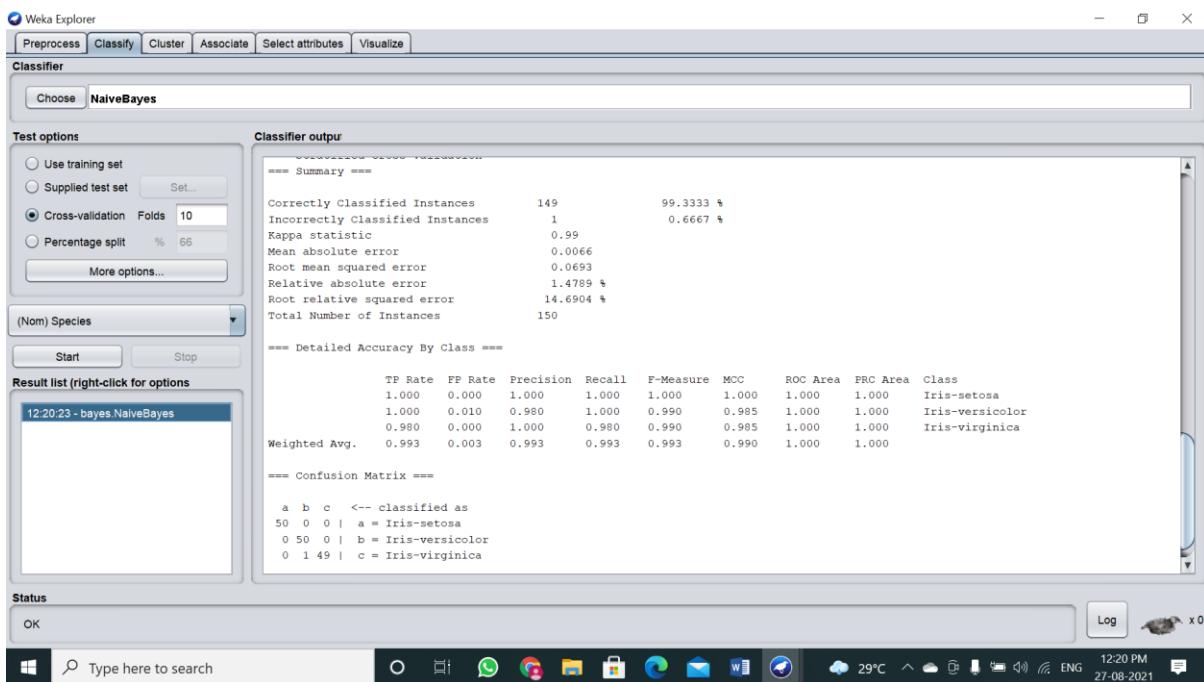
Step 2: Go to Open File and open CSV File.



Step 3: Go to Classify > Choose > bayes> NaiveBayes and click on Start Button.



Output:



Practical 9

Aim: Utilise data visualisation techniques to analyse the given dataset.

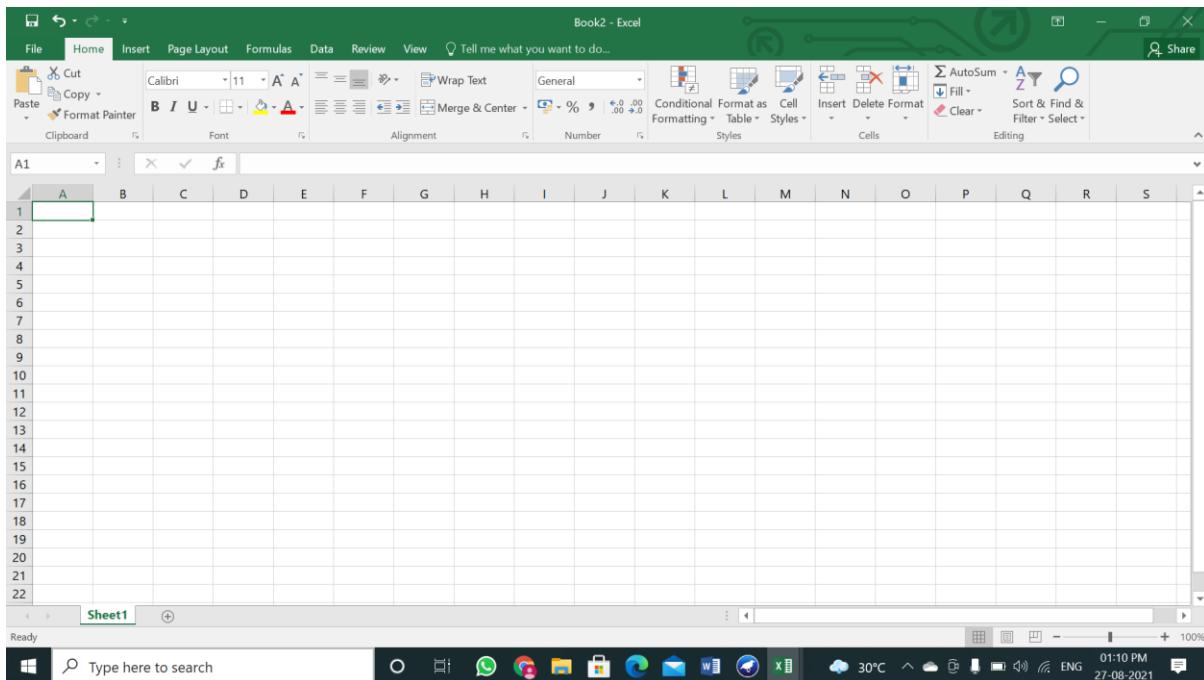
Theory:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Data visualization gives us a clear idea of what the information means by giving it visual context through maps or graphs. This makes the data more natural for the human mind to comprehend and therefore makes it easier to identify trends, patterns, and outliers within large data sets.

Data in a database or data warehouse can be viewed at different granularity or abstraction levels, or as different combinations of attributes or dimensions. Data can be presented in various visual forms, such as boxplots, 3-D cubes, data distribution charts, curves, surfaces, and link graphs, as shown in the data visualization section of Chapter 2. Figures 13.4 and 13.5 from StatSoft show data distributions in multidimensional space. Visual display can help give users a clear impression and overview of the data characteristics in a large data set.

Step 1: Open Excel Sheet

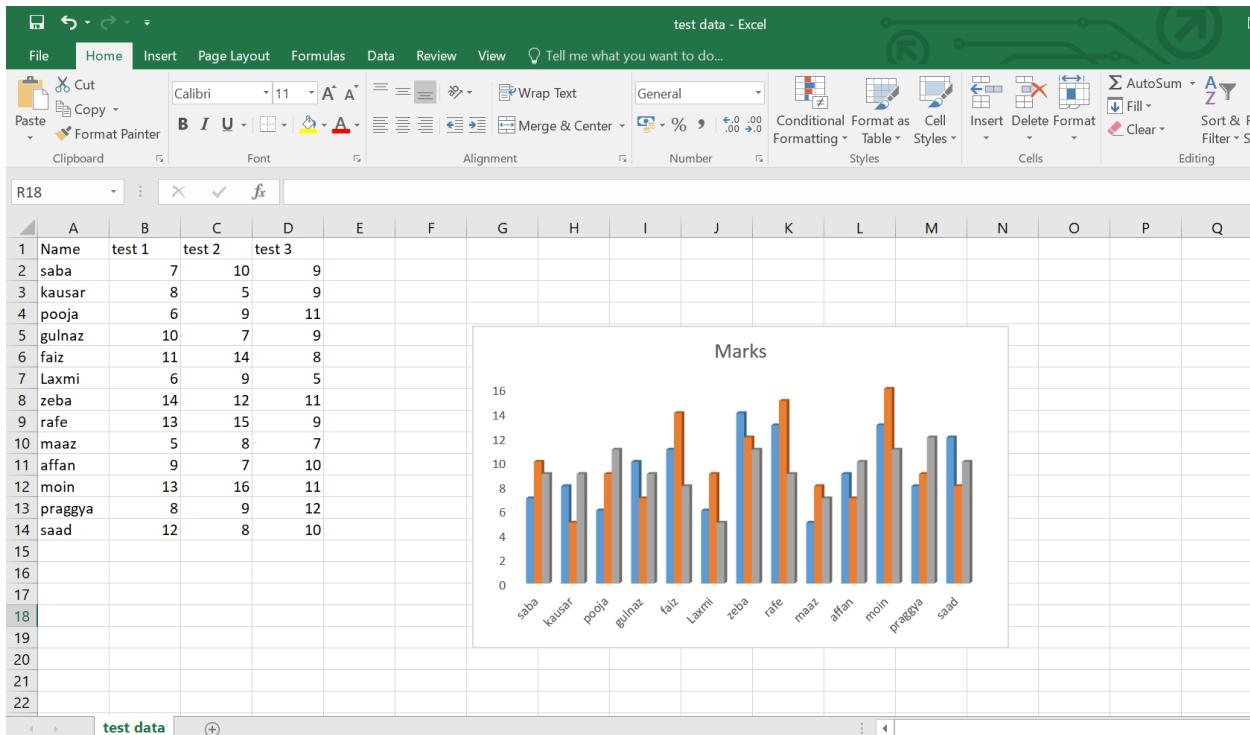


Step 2: Give some data

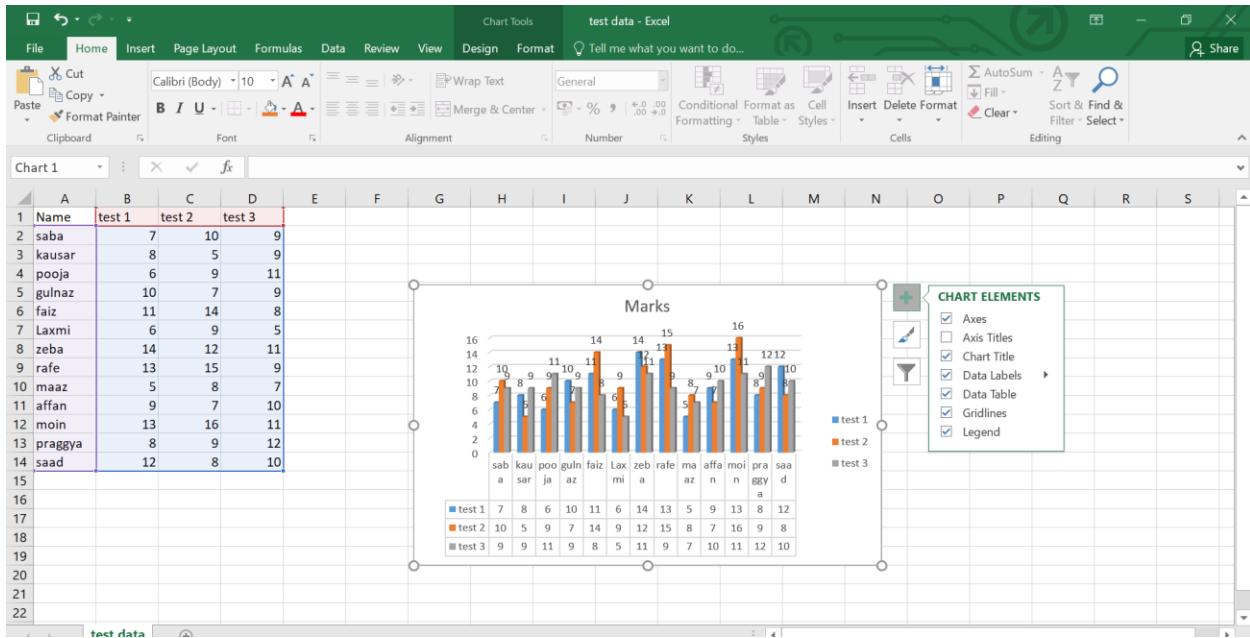
Screenshot of Microsoft Excel showing a table titled "test data". The table has columns A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S. Rows 1 through 14 contain data. Row 1 is the header with columns A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S. Rows 2 through 14 list names and their corresponding test scores. Row 15 is blank.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Name	test 1	test 2	test 3															
2	saba	7	10	9															
3	kausar	8	5	9															
4	pooja	6	9	11															
5	gulnaz	10	7	9															
6	faiz	11	14	8															
7	Laxmi	6	9	5															
8	zeba	14	12	11															
9	rafe	13	15	9															
10	maaz	5	8	7															
11	affan	9	7	10															
12	moin	13	16	11															
13	praggya	8	9	12															
14	saad	12	8	10															
15																			
16																			
17																			
18																			
19																			
20																			
21																			
22																			

Step 3: Make a Graph for the test data.



Step 4: Add Chart Element

**Output:**