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	

Name: Zeenat **Class: TYIT** **Roll no: 578**

Practical 1

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

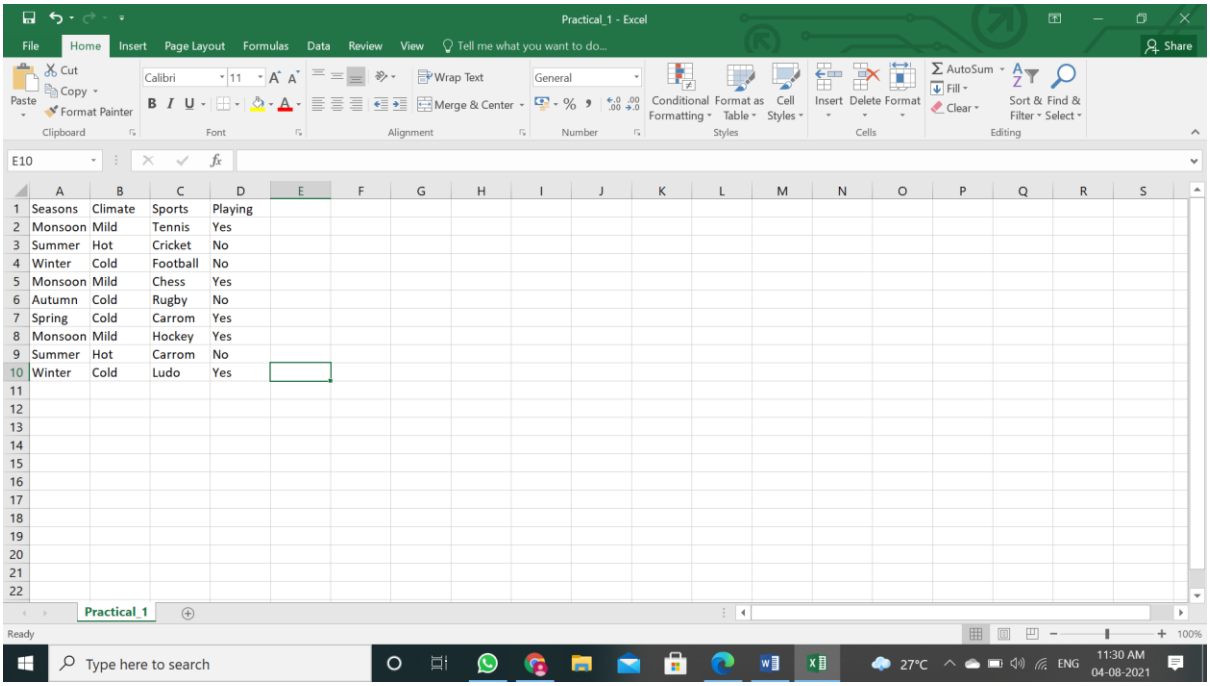
Theory:

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.

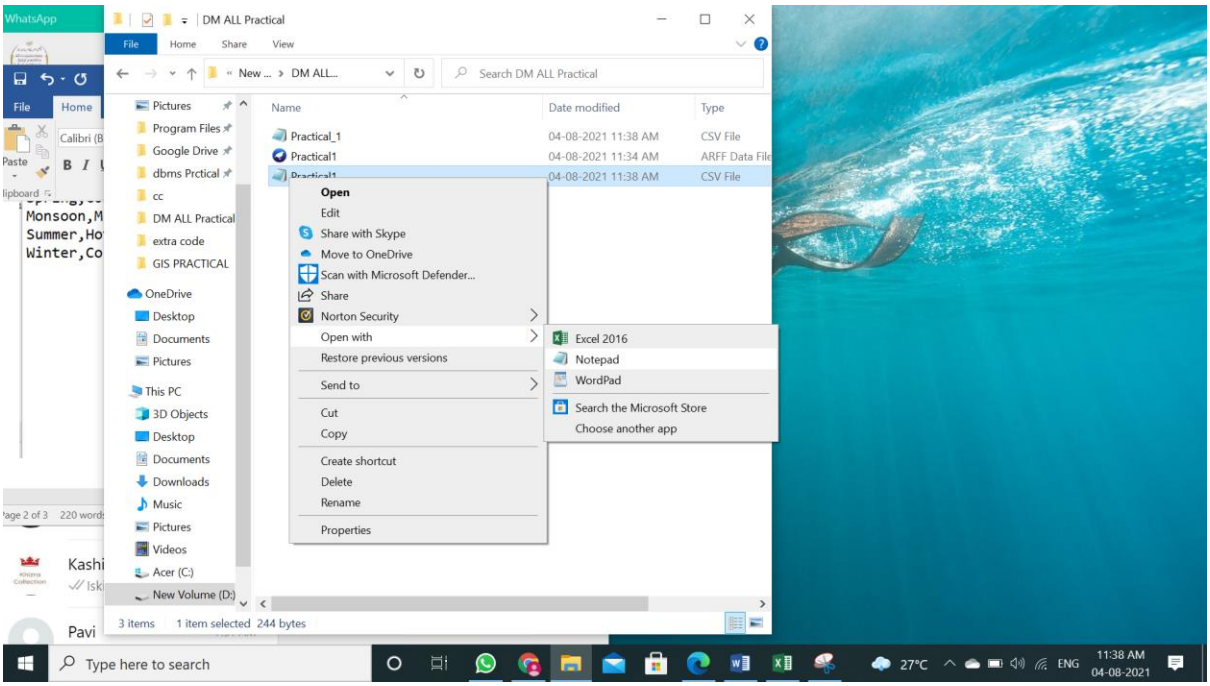
Market Based Analysis is one of the key techniques used by large relations to show associations between items.It allows retailers to identify relationships between the items that people buy together frequently.

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

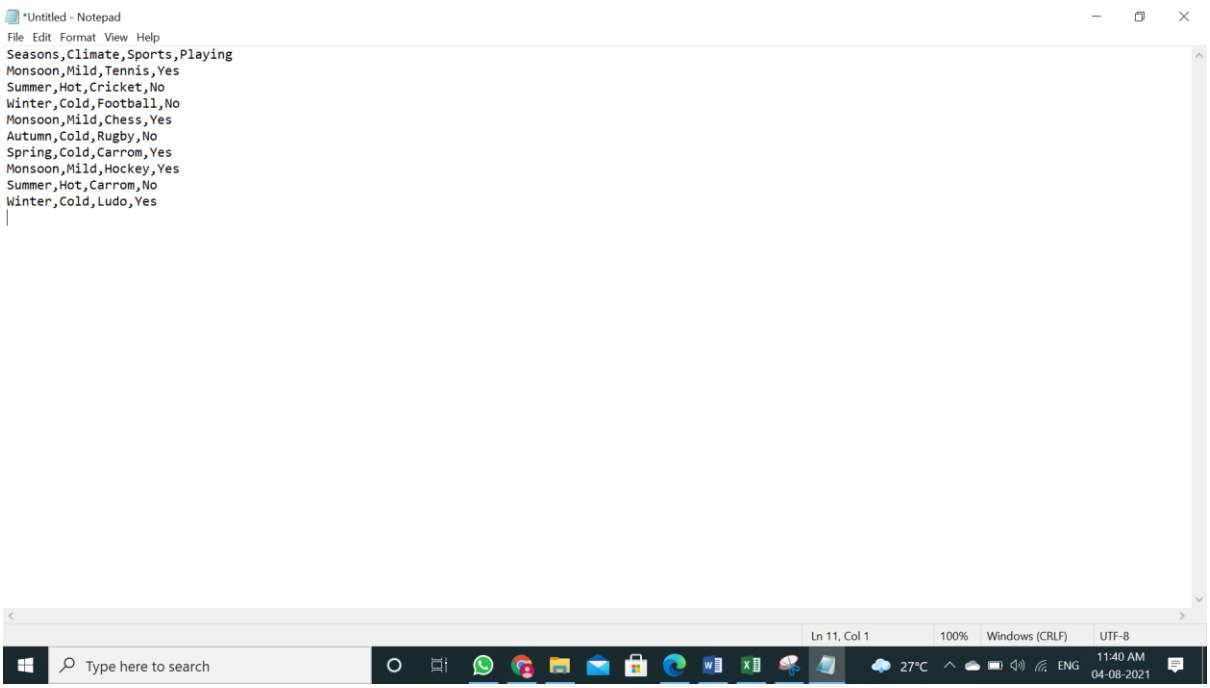


Step 2 : Open the same Excel File with the Open with Option and click on Notepad.

Name: Zeenat **Class: TYIT** **Roll no: 578**

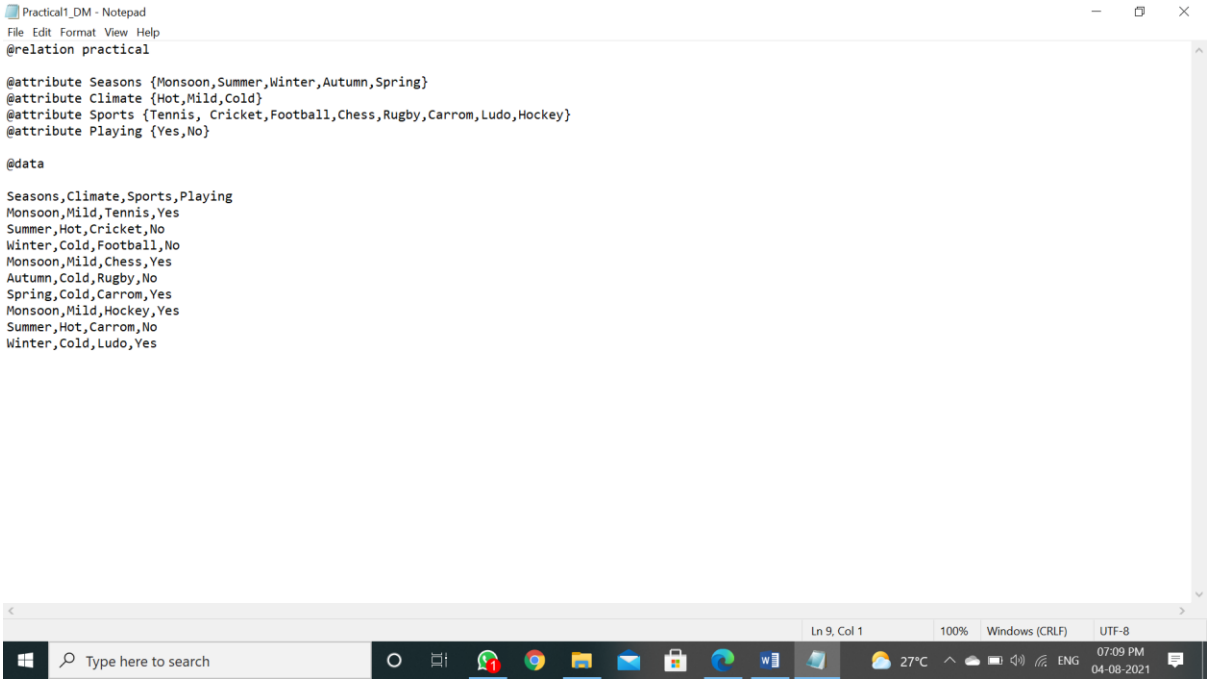


Step 3 : Copy and Insert one Notepad File Content into another Notepad.

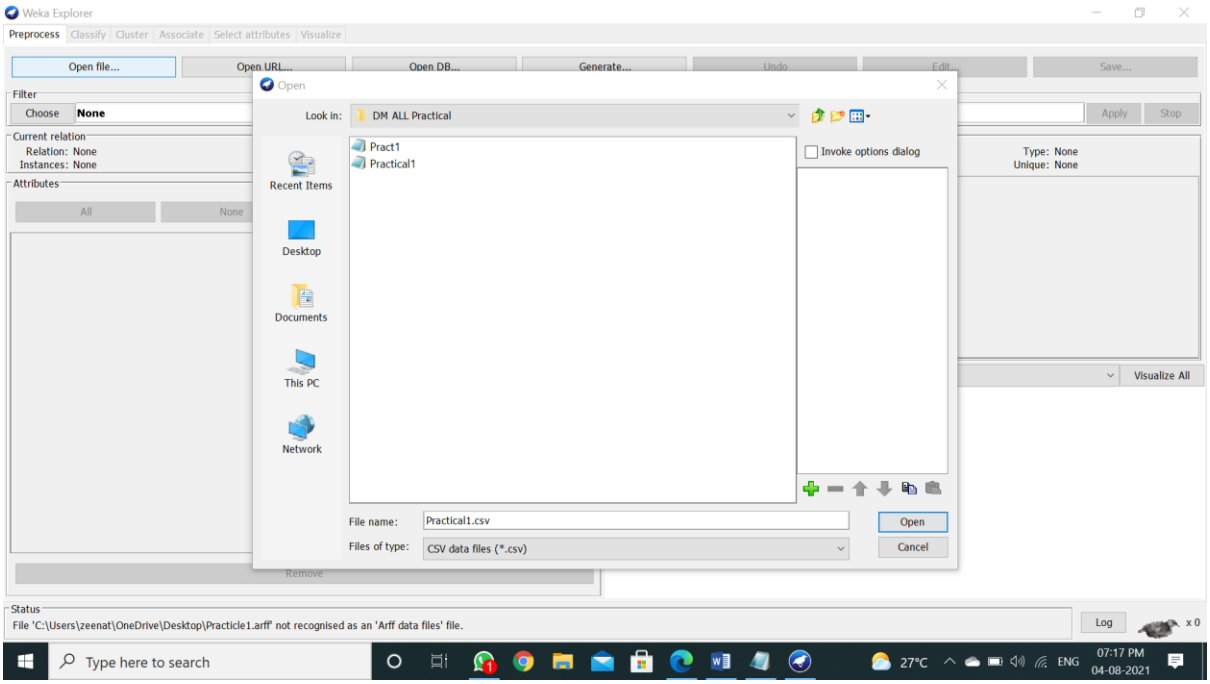


Name: Zeenat **Class: TYIT** **Roll no: 578**

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

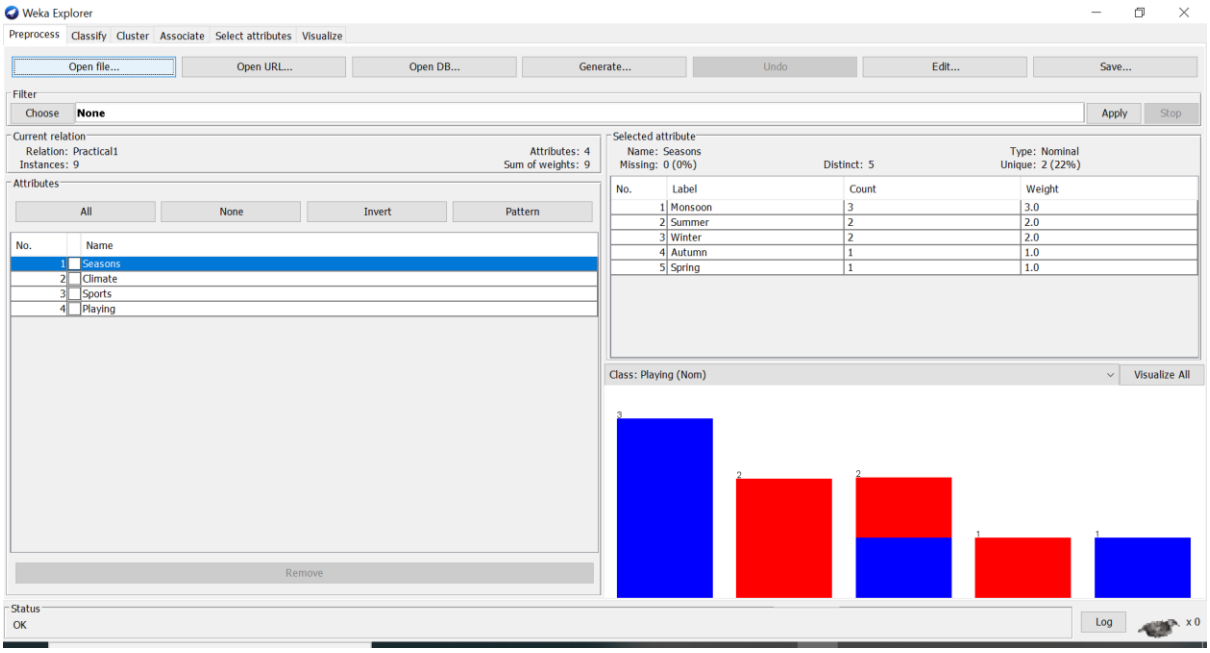


Step 5 : Open Weka Explorer Software and there is an Open File Option and browse for the file saved in folder.



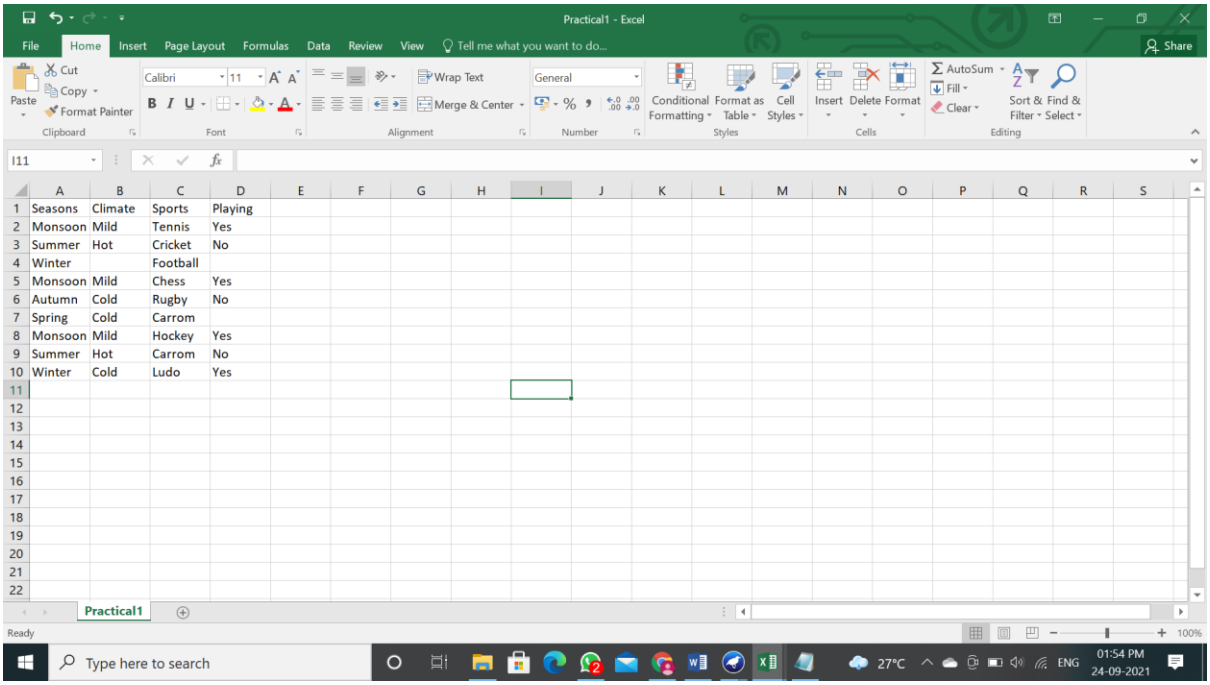
Name: Zeenat **Class: TYIT** **Roll no: 578**

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

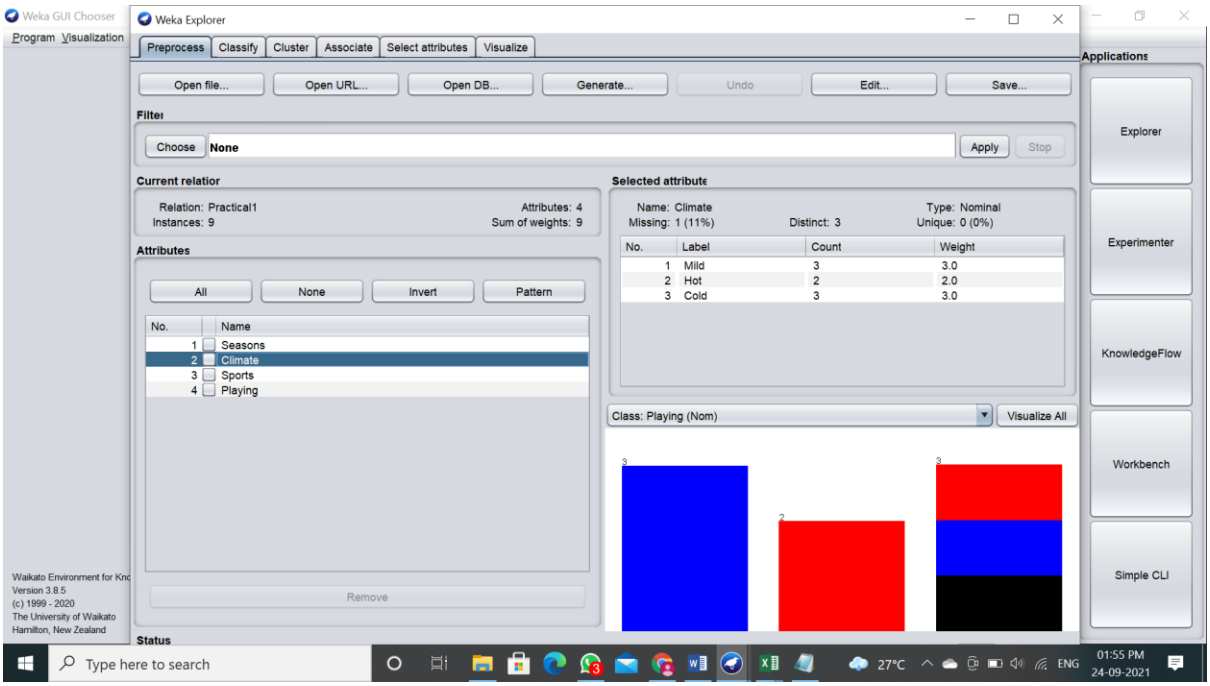


This is the Output for the Excel Data File.

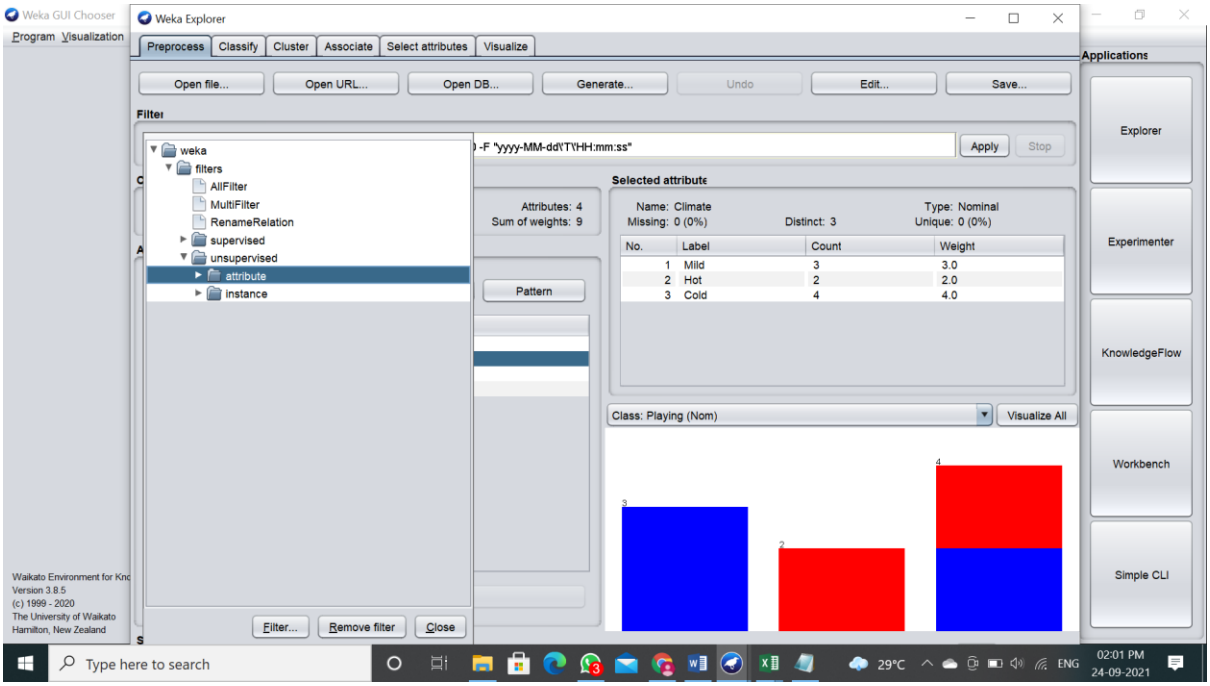
This is file for remove data



Name: Zeenat **Class: TYIT** **Roll no: 578**

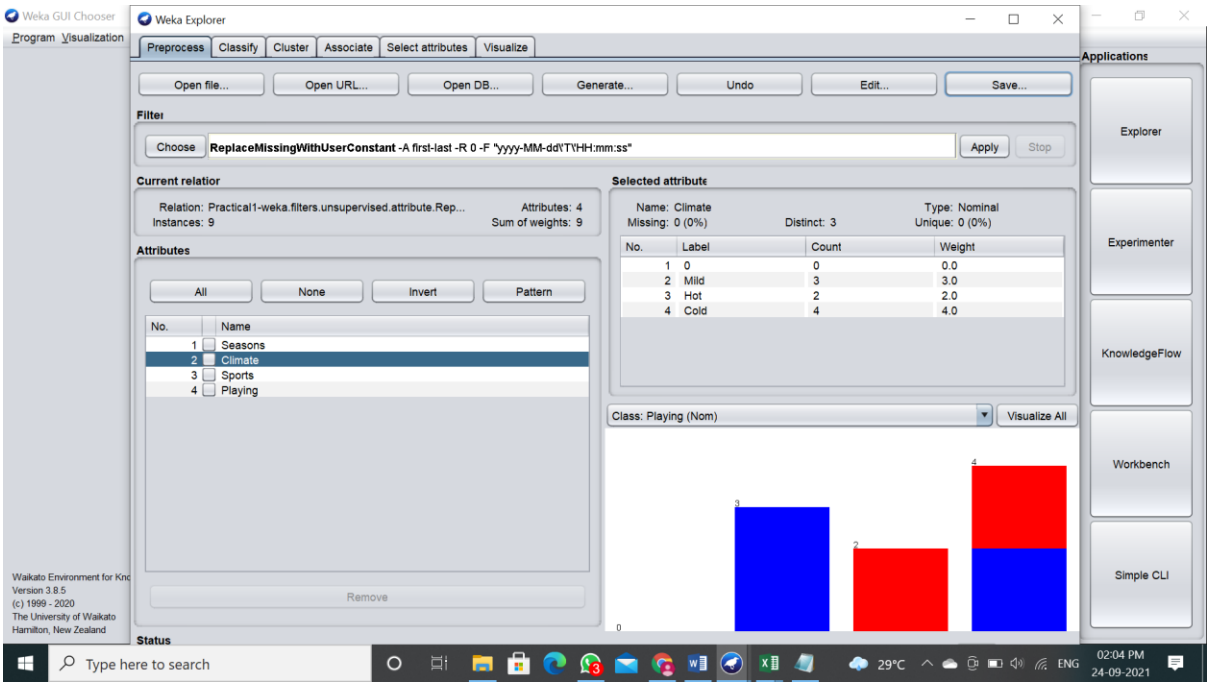


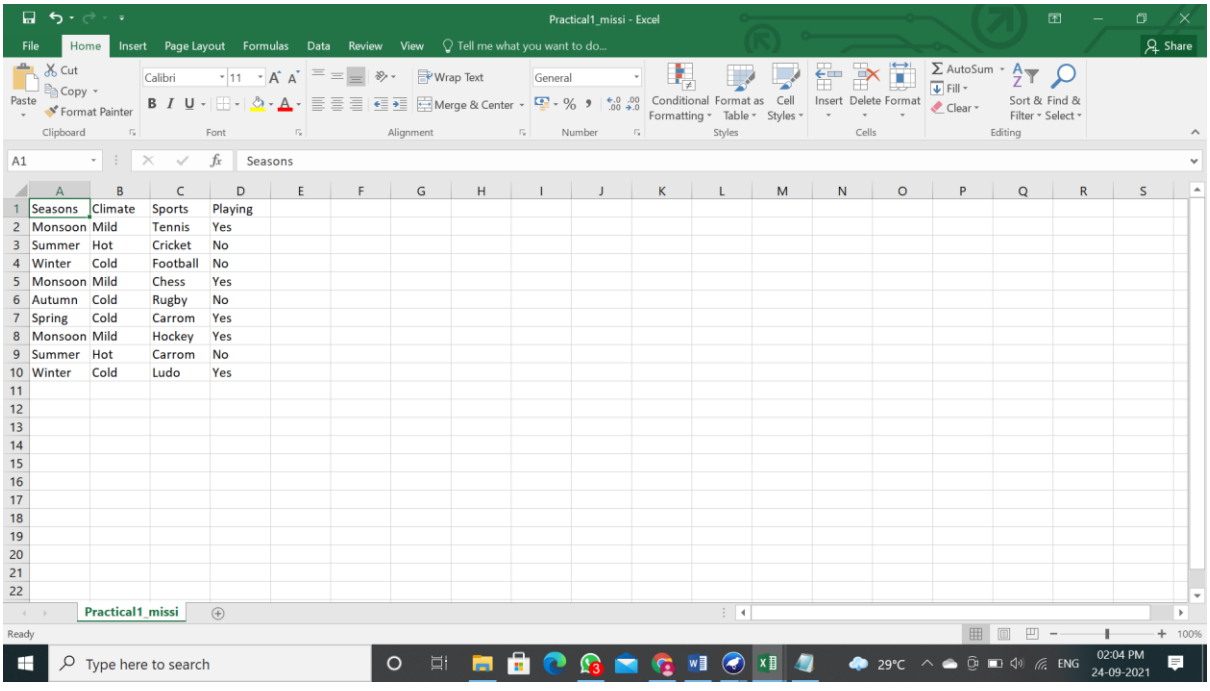
Step 7 : If you want to do Filtration i.e. remove missing values from table go to Choose then Filter then Unsupervised then attribute and finally with ReplaceMissingWithUserConstant.



Step 8: Apply and Save file.

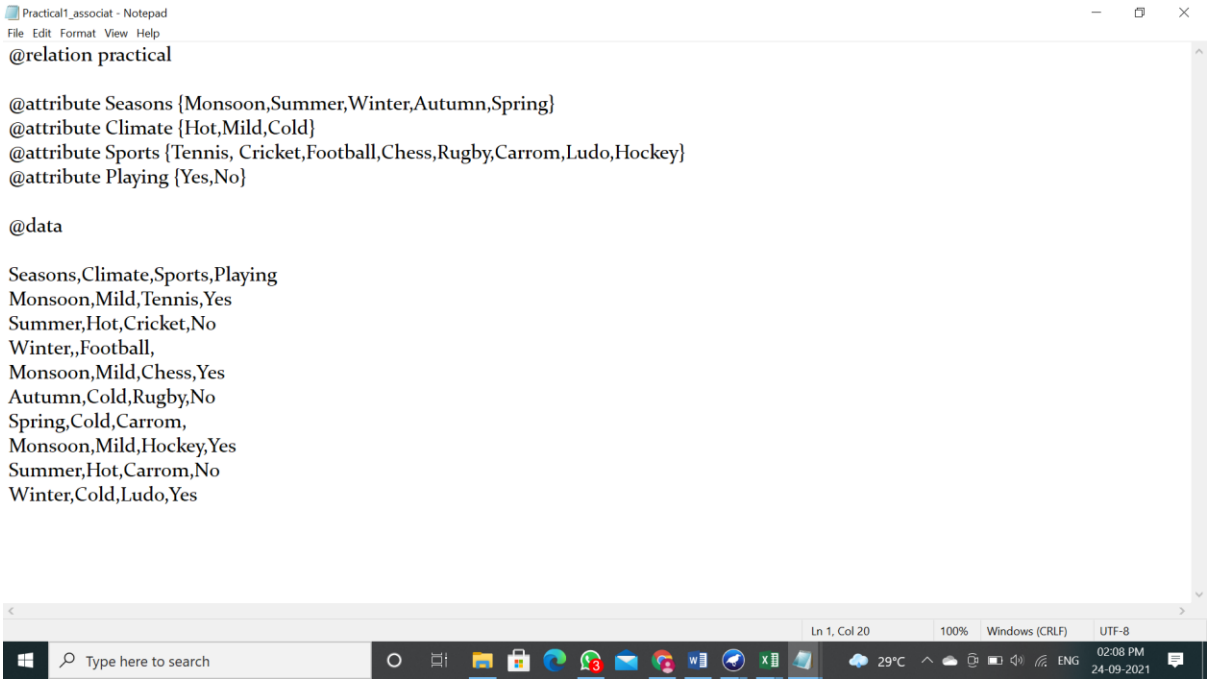
Name: Zeenat **Class: TYIT** **Roll no: 578**



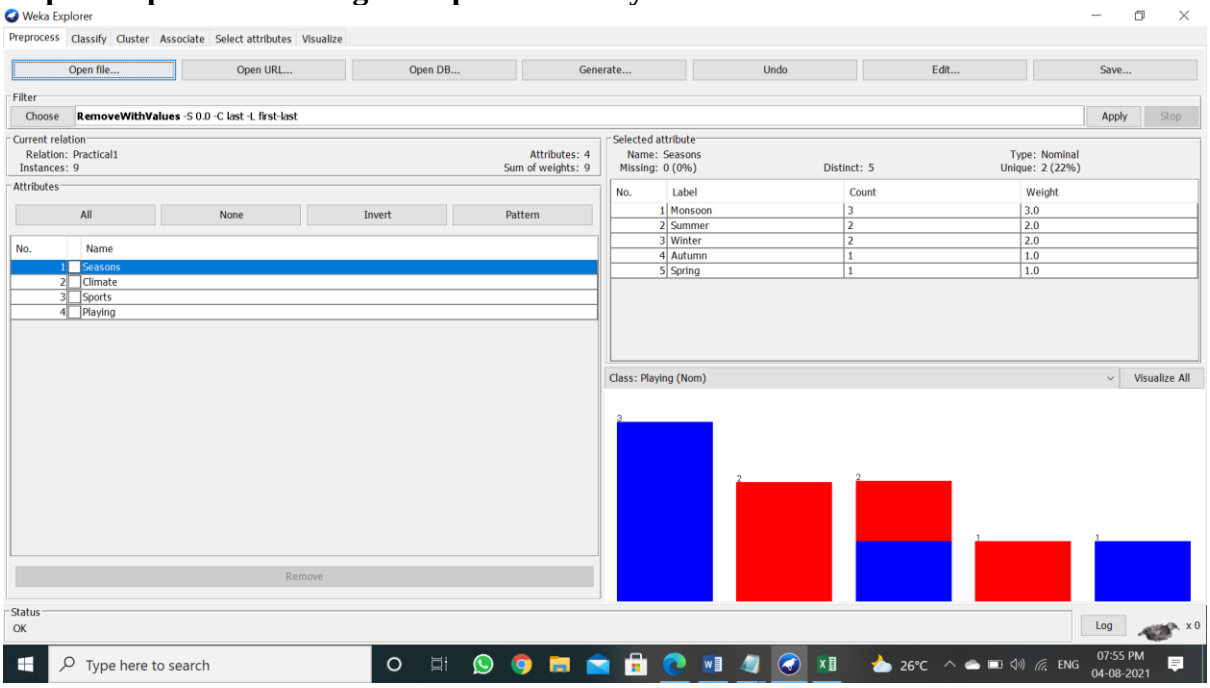


Step 9: Open Notepad and write the Relation, Attribute and Data for Dataset and save it with .arff extension.

Name: Zeenat **Class: TYIT** **Roll no: 578**

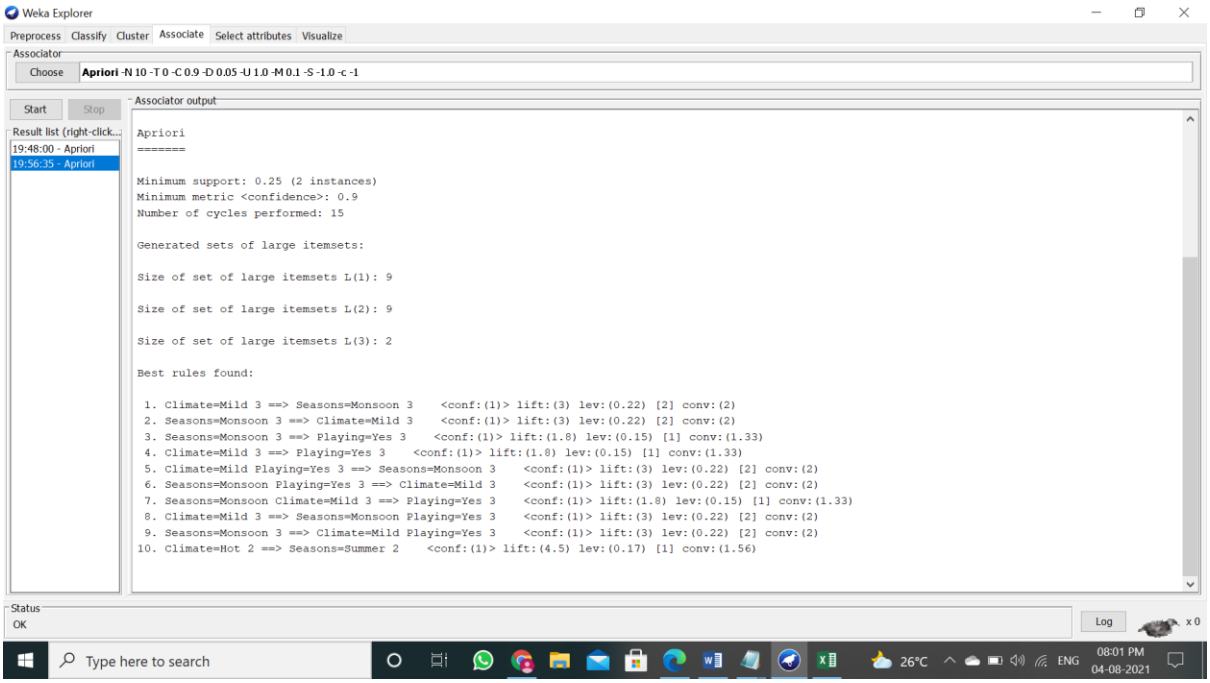


Step 10: Open Weka and go to Open File and you will have to browse for the desired file.



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

Name: Zeenat **Class: TYIT** **Roll no: 578**



Name: Zeenat **Class: TYIT** **Roll no: 578**

Practical 2

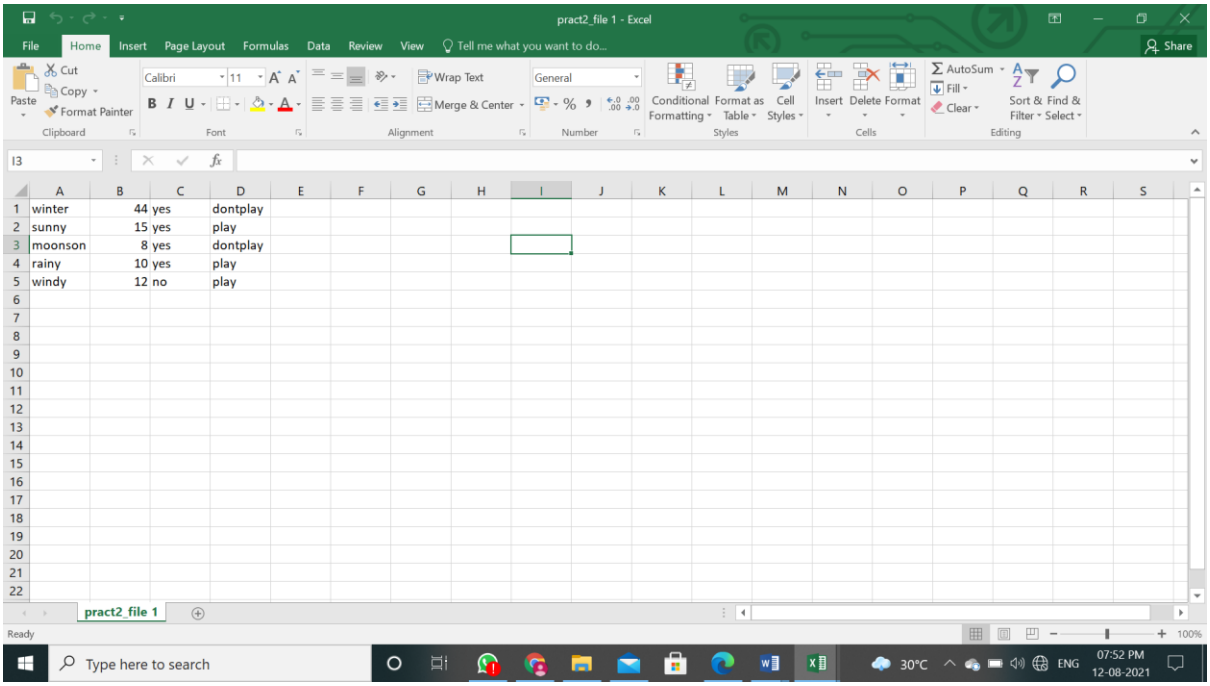
Aim: Write ETL scripts and implement using data warehouse tools.

Theory:

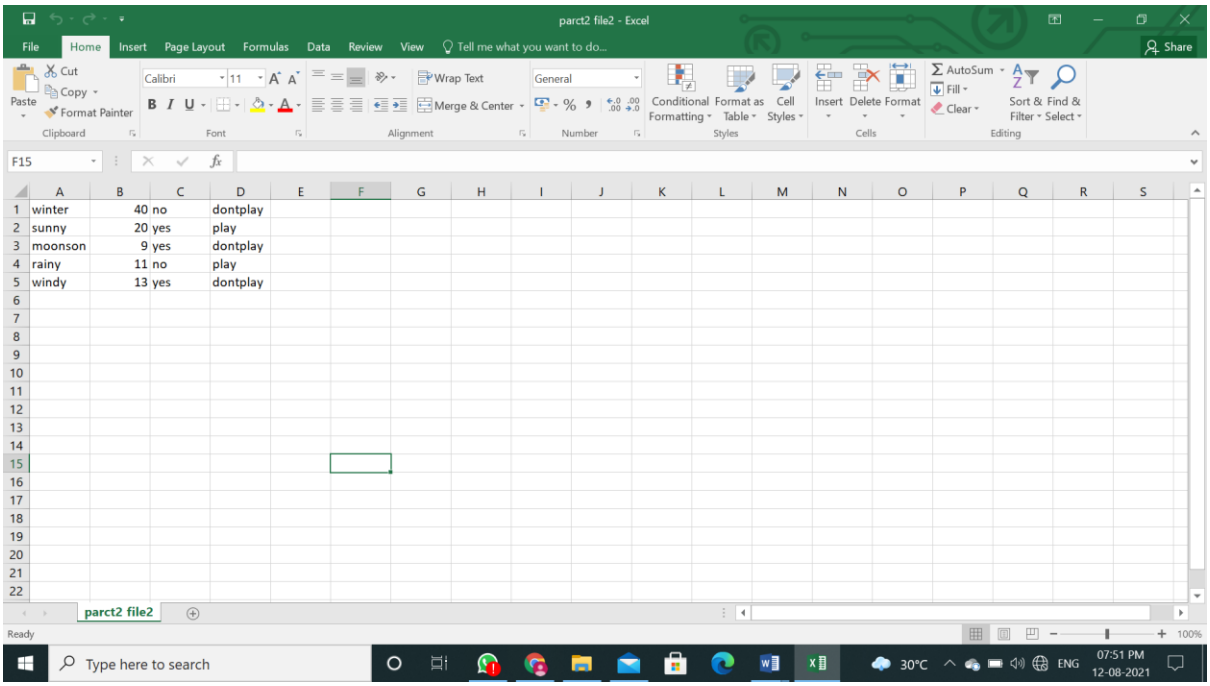
The process of extracting data from source systems and bringing it into the data warehouse is commonly called ETL, which stands for extraction, transformation, and loading. ETL (Extract, Transform, Load) is an automated process which takes raw data, extracts the information required for analysis, transforms 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.

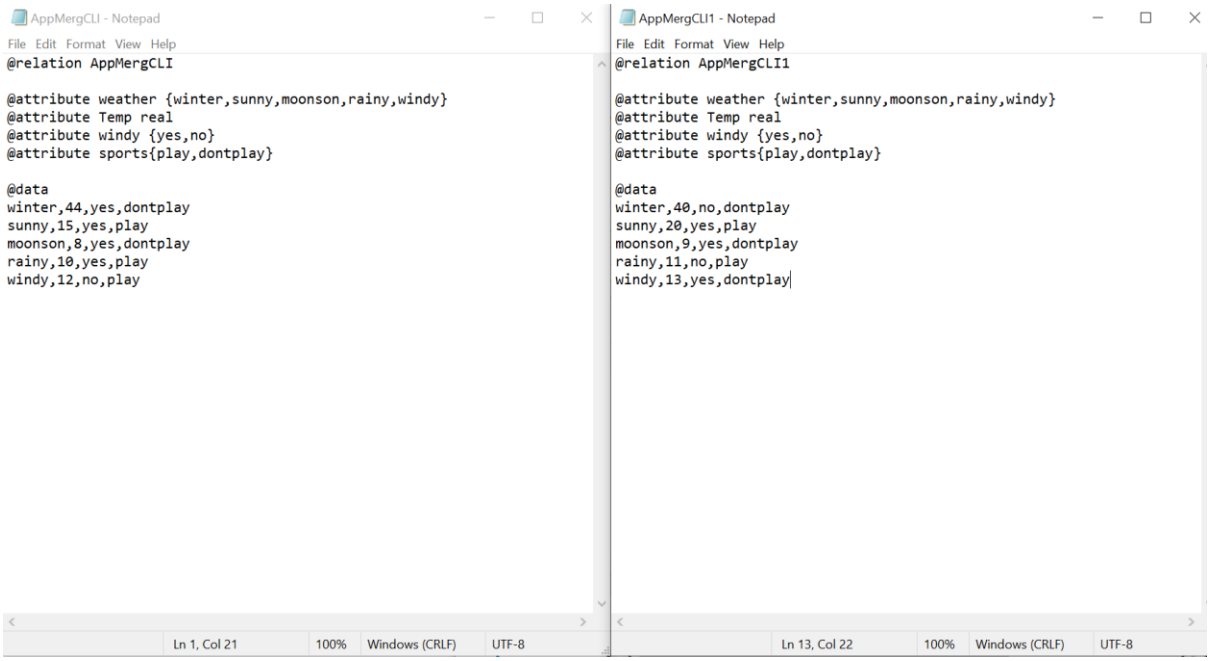
Step 1 : Open Excel File and fill in the Data and save it in .csv Format and create two Excel Files.



Name: Zeenat **Class: TYIT** **Roll no: 578**

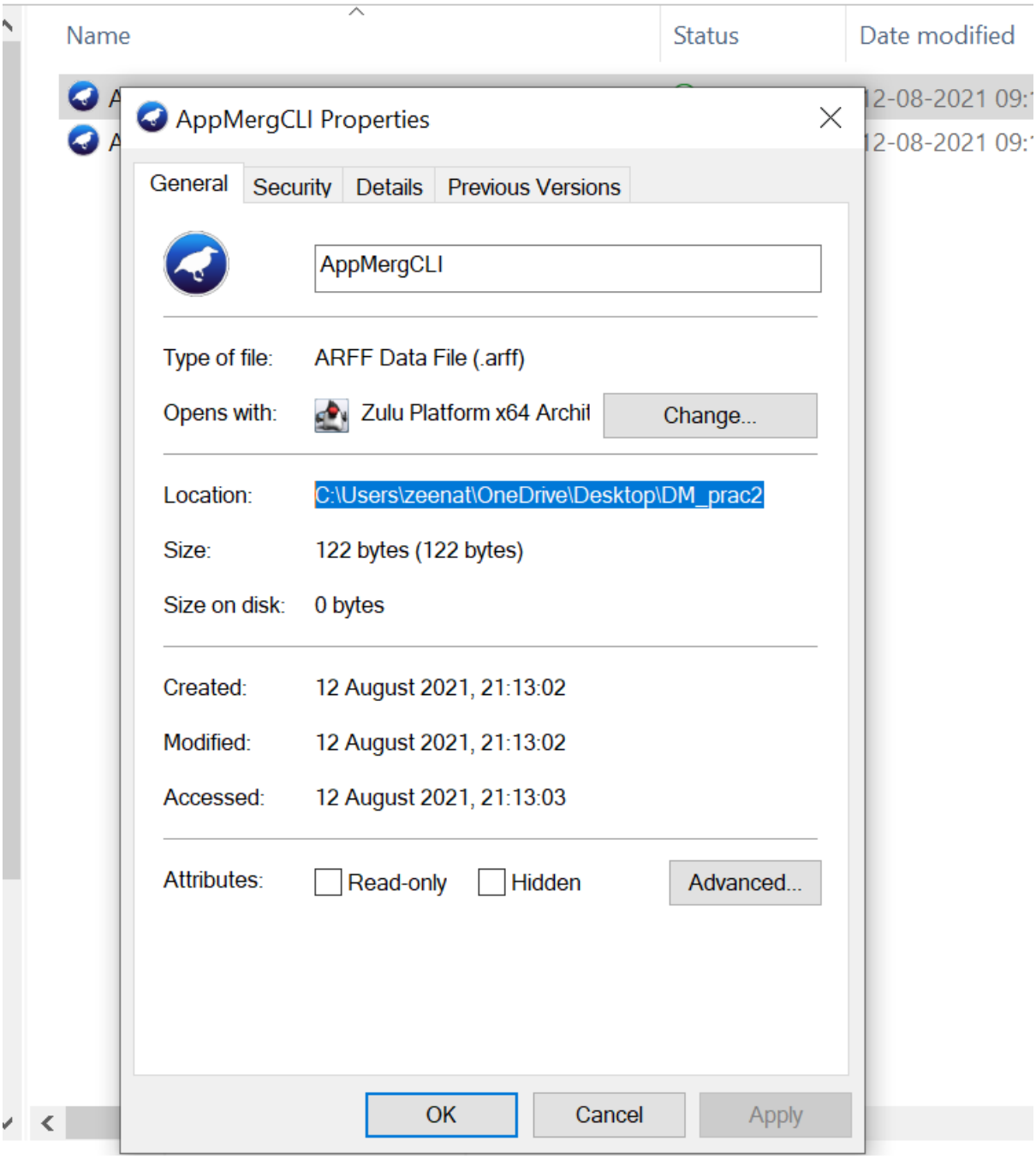


Step 2 : On Excel File right click and open with Notepad and copy the content and paste it into new Notepad and write @relation and name, @attribute and column names and @data which contains all Excel Data with comma separated values.

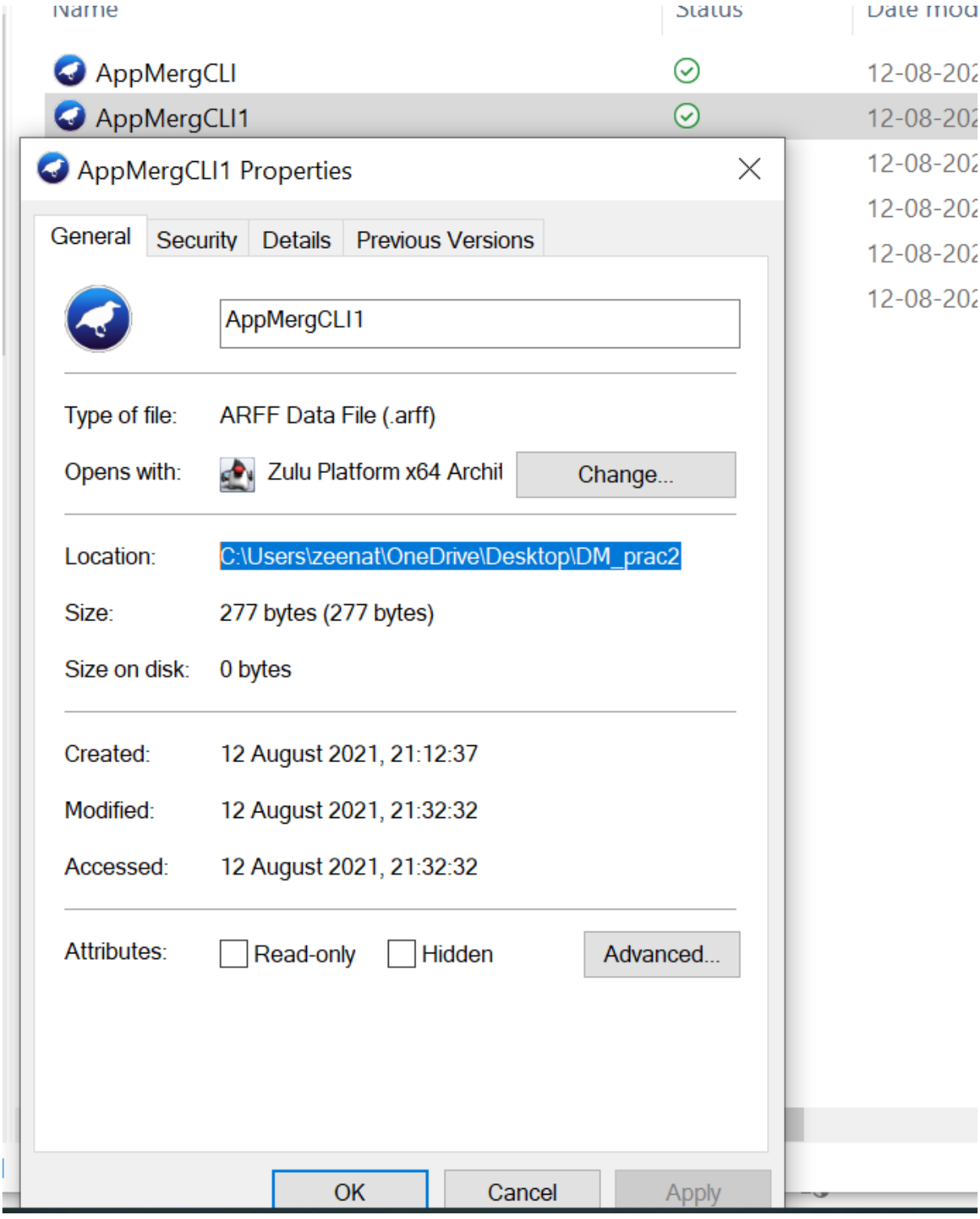


Step 3: Open the two arff Files created and right click and select on Properties and copy the path and paste it in the command given below.

Name: Zeenat **Class: TYIT** **Roll no: 578**



Name: Zeenat **Class: TYIT** **Roll no: 578**

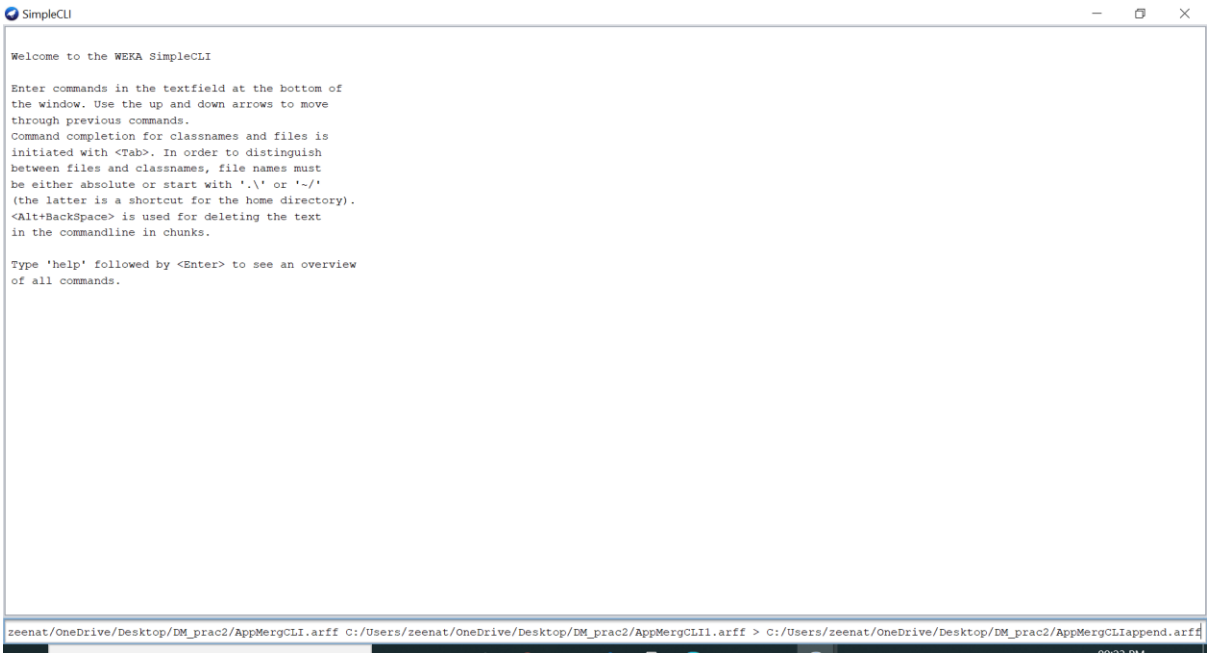


Step 4: Open Weka Software and go to SimpleCLI Mode and type the following command in the box given below:

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

Name: Zeenat **Class: TYIT** **Roll no: 578**

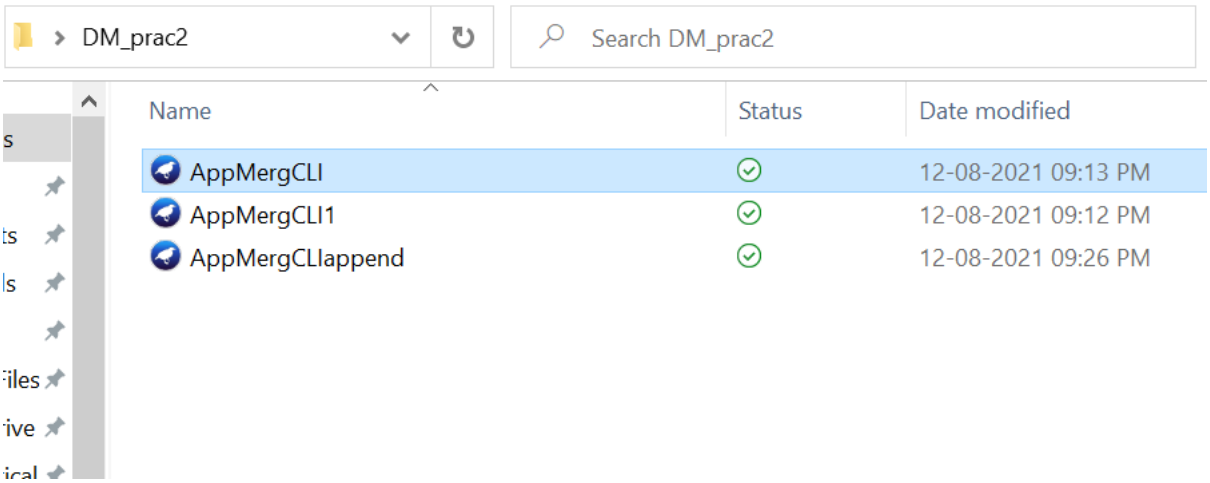
***Note : There should be no spaces in File or Folder name or else can be replaced by underscore.**

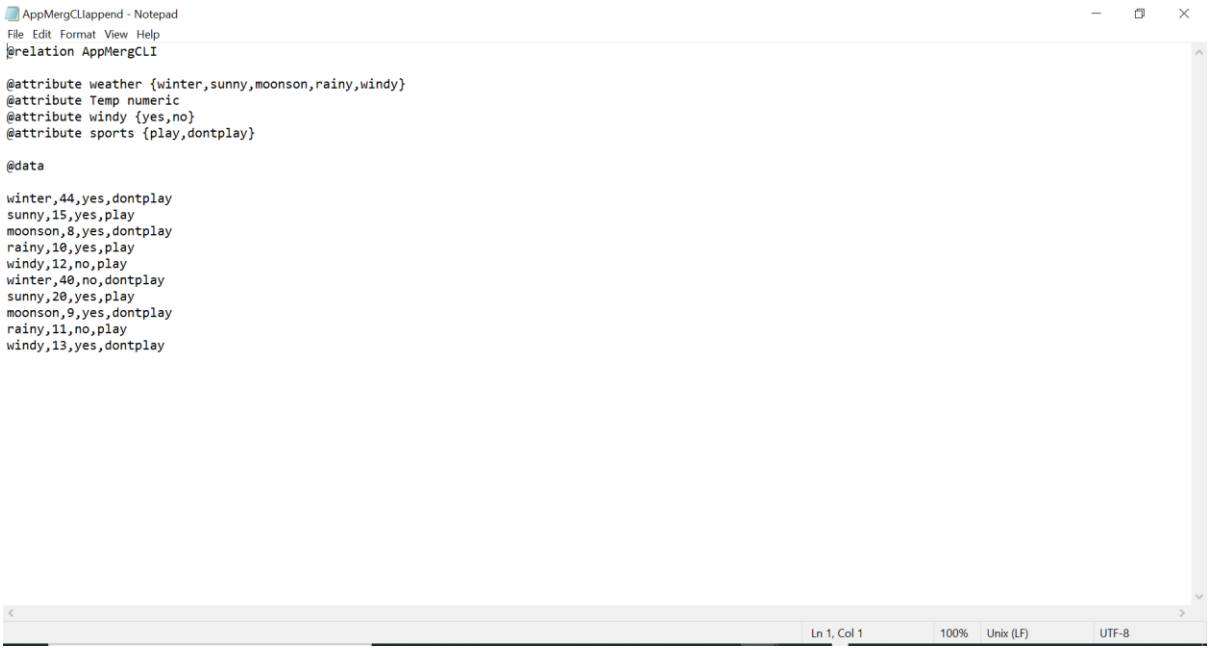


java weka.core.Instances append C:/Users/zeenat/OneDrive/Desktop/DM_prac2/AppMergCLI.arff C:/Users/zeenat/OneDrive/Desktop/DM_prac2/AppMergCLI1.arff > C:/Users/zeenat/OneDrive/Desktop/DM_prac2/AppMergCLIappend.arff

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: Zeenat **Class: TYIT** **Roll no: 578**

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.

Name: Zeenat **Class: TYIT** **Roll no: 578**

SQL Worksheet

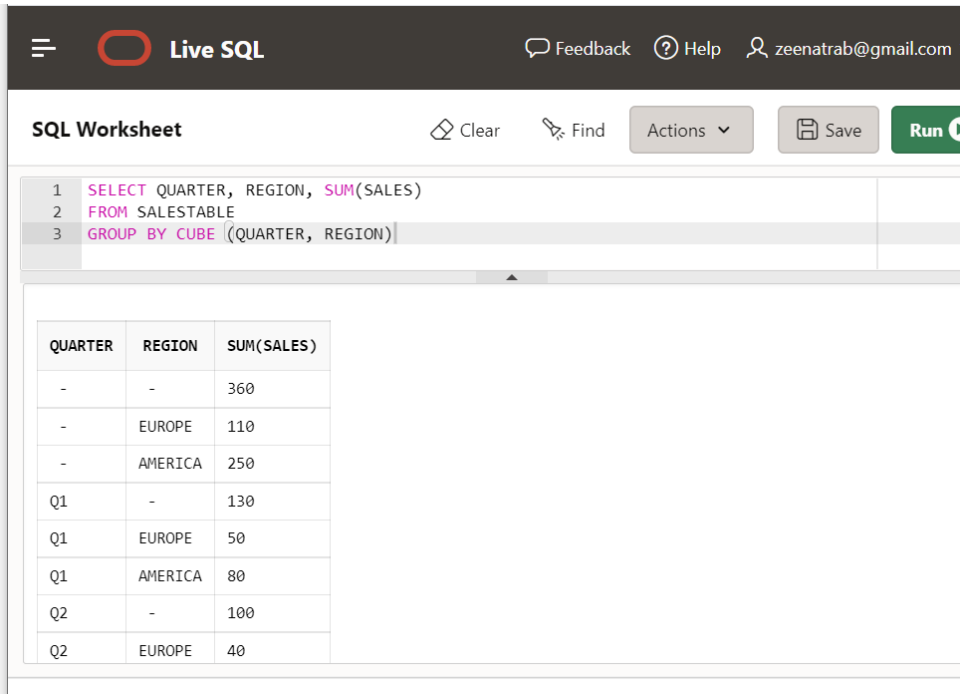
```
1  create table salestable(product varchar(10) not null,quarter varchar(10) not null, region varchar(20)
2  not null, sales int not null);
3
4  insert into salestable values('A','Q1','EUROPE',10);
5  insert into salestable values('A','Q1','AMERICA',20);
6  insert into salestable values('A','Q2','EUROPE',20);
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.

1 row(s) inserted.

1 row(s) inserted.

1 row(s) inserted.

1 row(s) inserted.

1 row(s) inserted.

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**



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

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**

Result from SQL query with Cube operator

2: ROLLUP

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.

Result from SQL query with ROLLUP operator

GROUPING SETS

This query is equivalent to

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	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.

Name: Zeenat **Class: TYIT** **Roll no: 578**

Practical 4

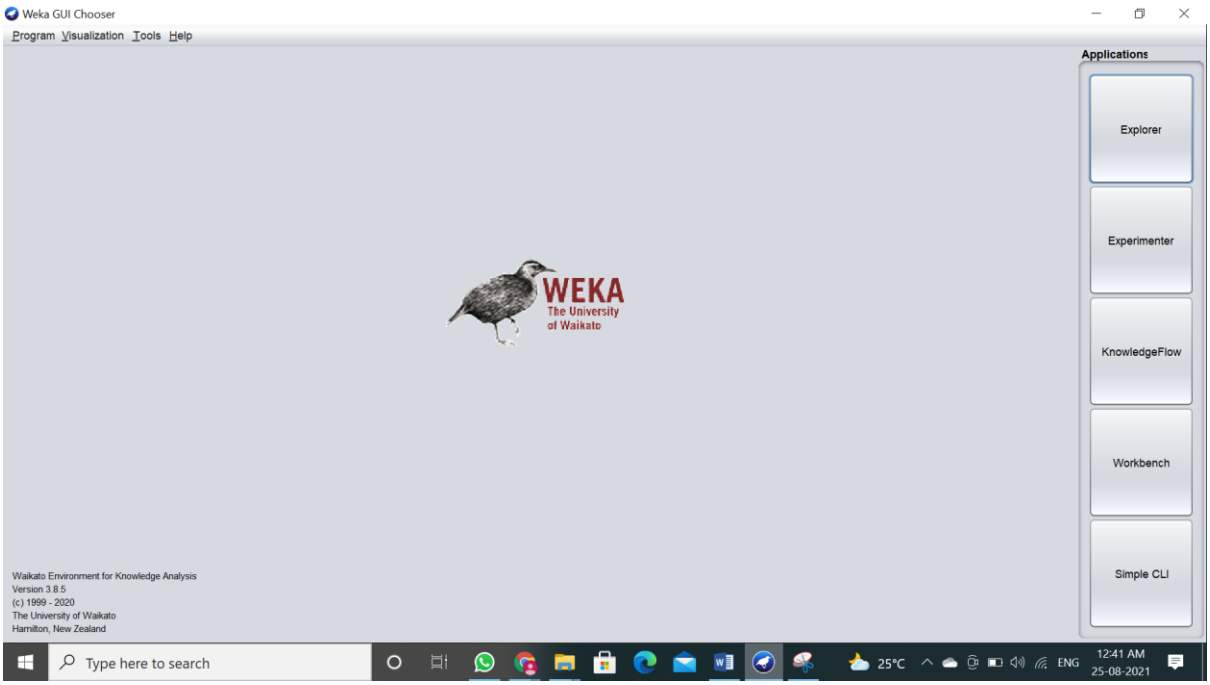
Aim: To perform kth nearest Algorithm.

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.

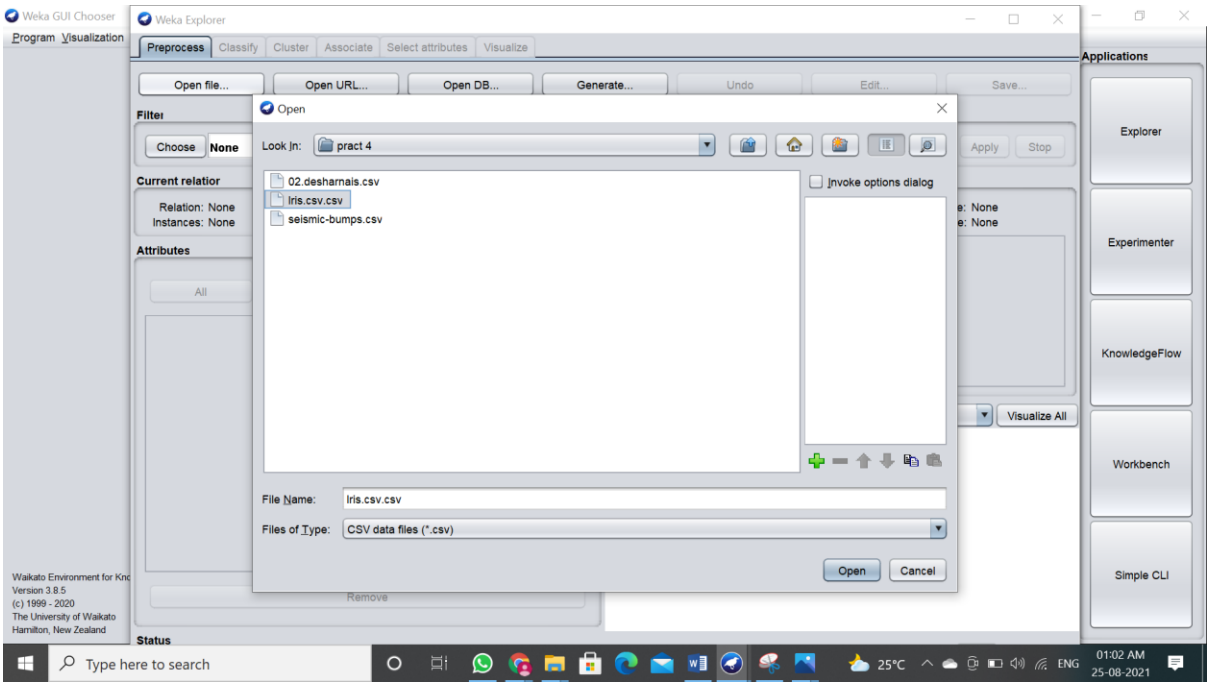
KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression)

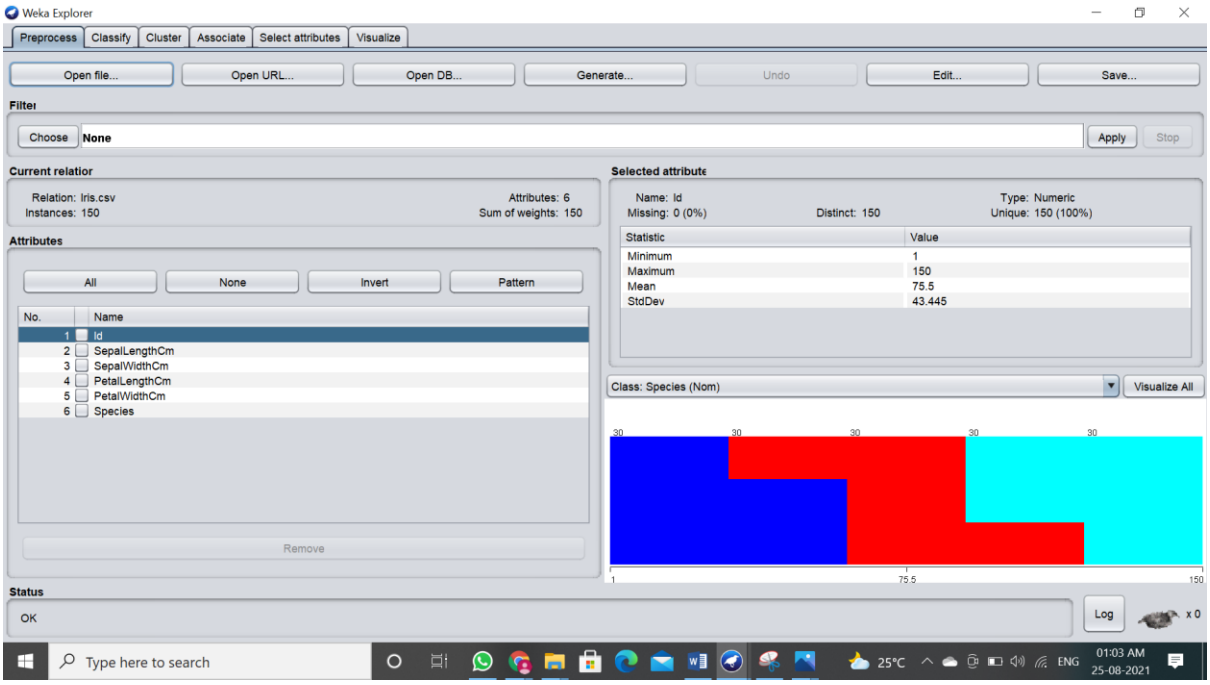
Step 1: Open Weka Software and go to Explorer Option.



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

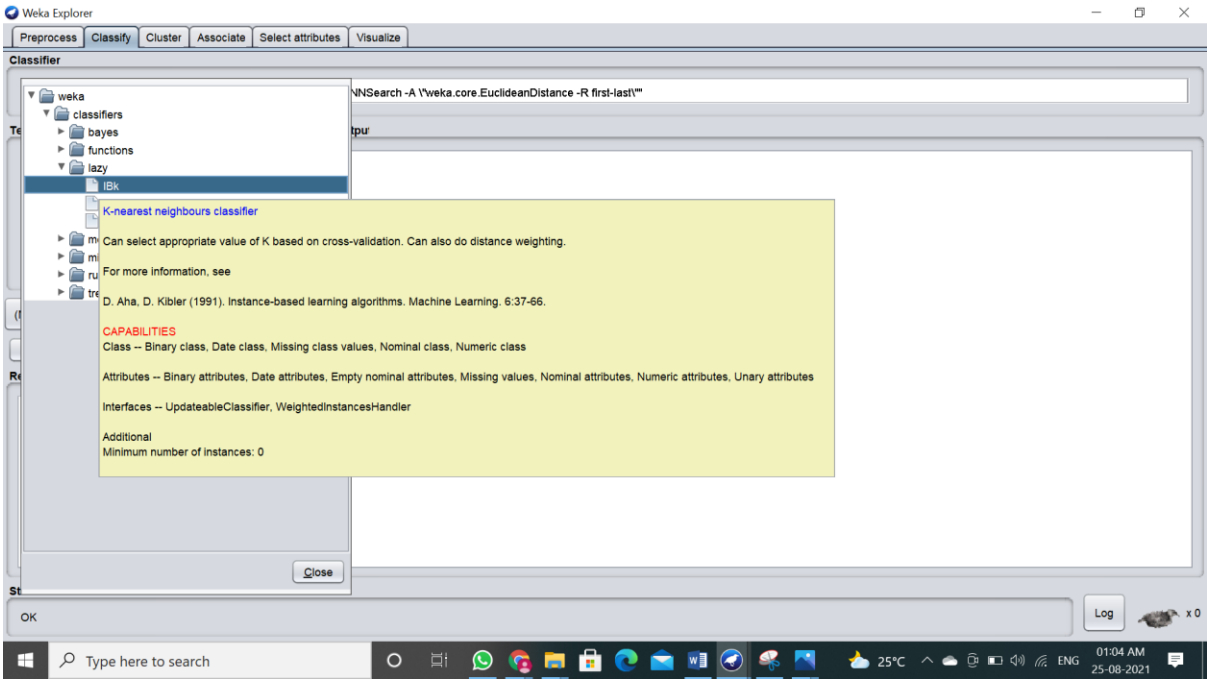
Name: Zeenat **Class: TYIT** **Roll no: 578**

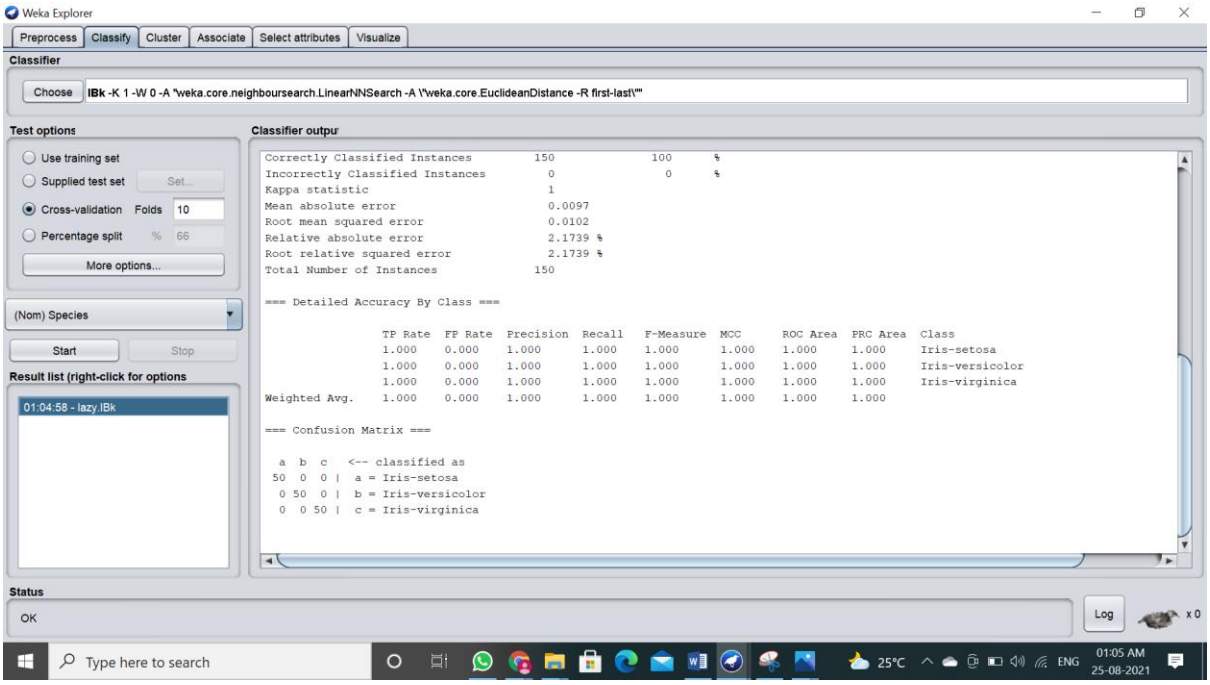




Step 3: Go to Classify, click on Choose > Lazy > Ibk File and click on Start Button and we will get the Output.

Name: Zeenat **Class: TYIT** **Roll no: 578**





Name: Zeenat **Class: TYIT** **Roll no: 578**

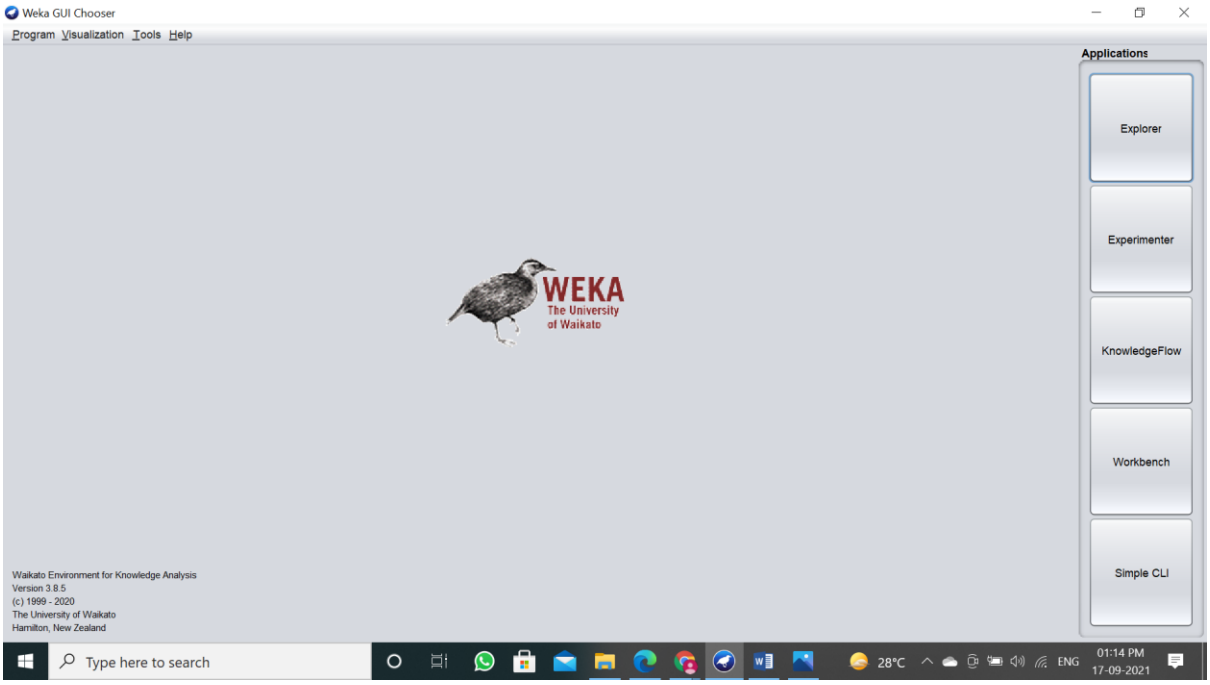
Practical 5

Aim: Demonstrate performing clustering on data sets.

Theory:

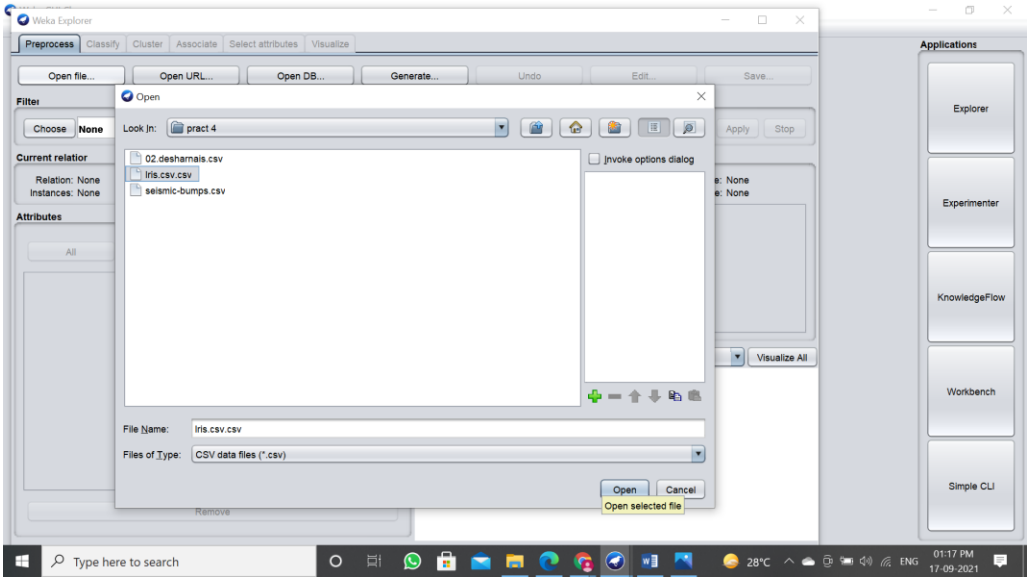
Clustering only utilizes input data, to determine patterns, anomalies, or similarities in its input data. A good clustering algorithm aims to obtain clusters whose: The intracluster similarities are high, It implies that the data present inside the cluster is similar to one another. The process of making 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.

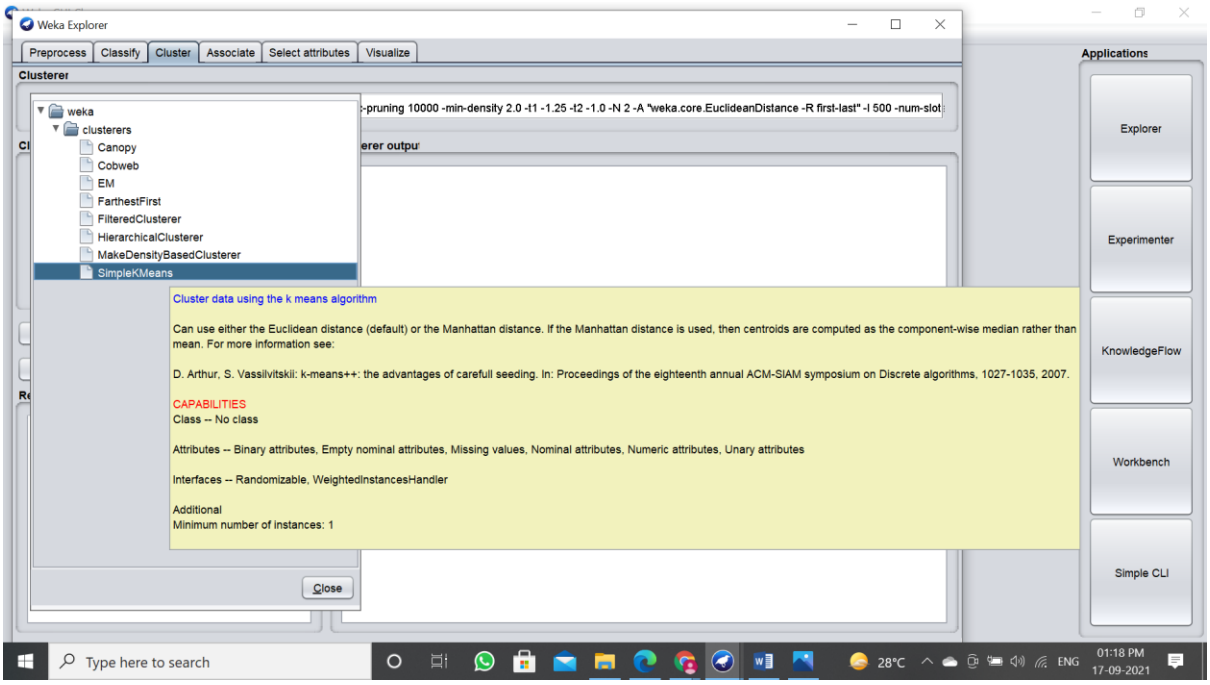


Step 2: Go to Open File option and choose a CSV File an click on Open.

Name: Zeenat **Class: TYIT** **Roll no: 578**

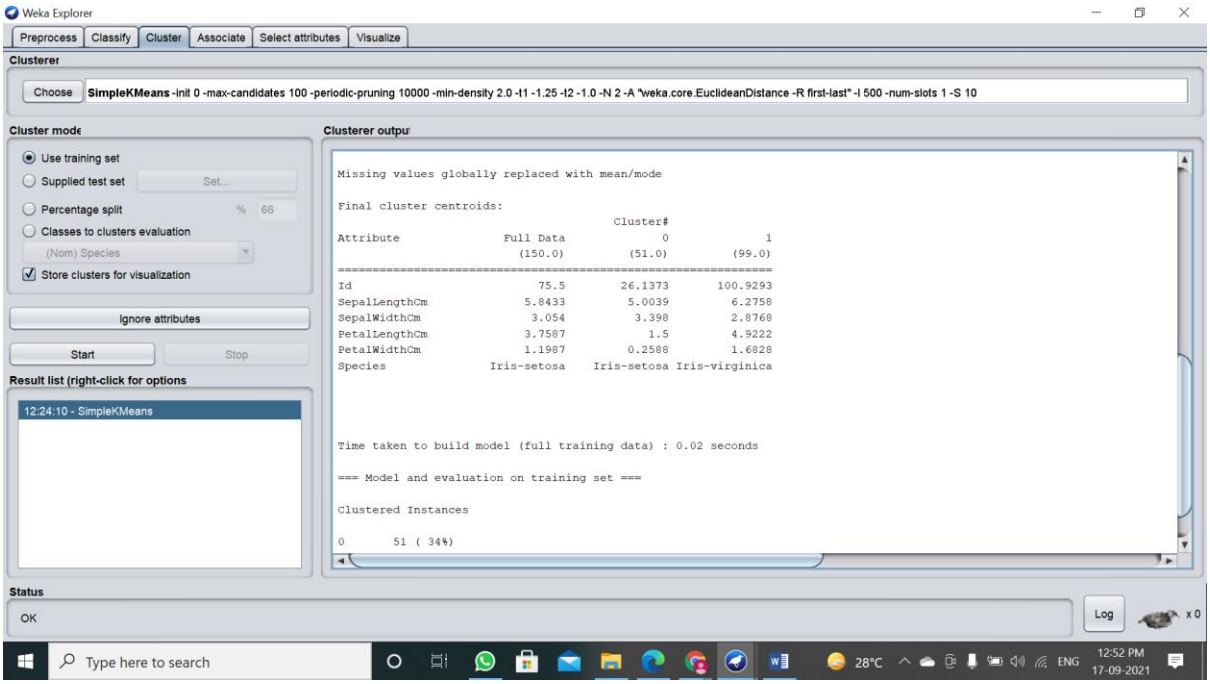


Step 3: Go to Cluster and Choose SimpleKMeans option.

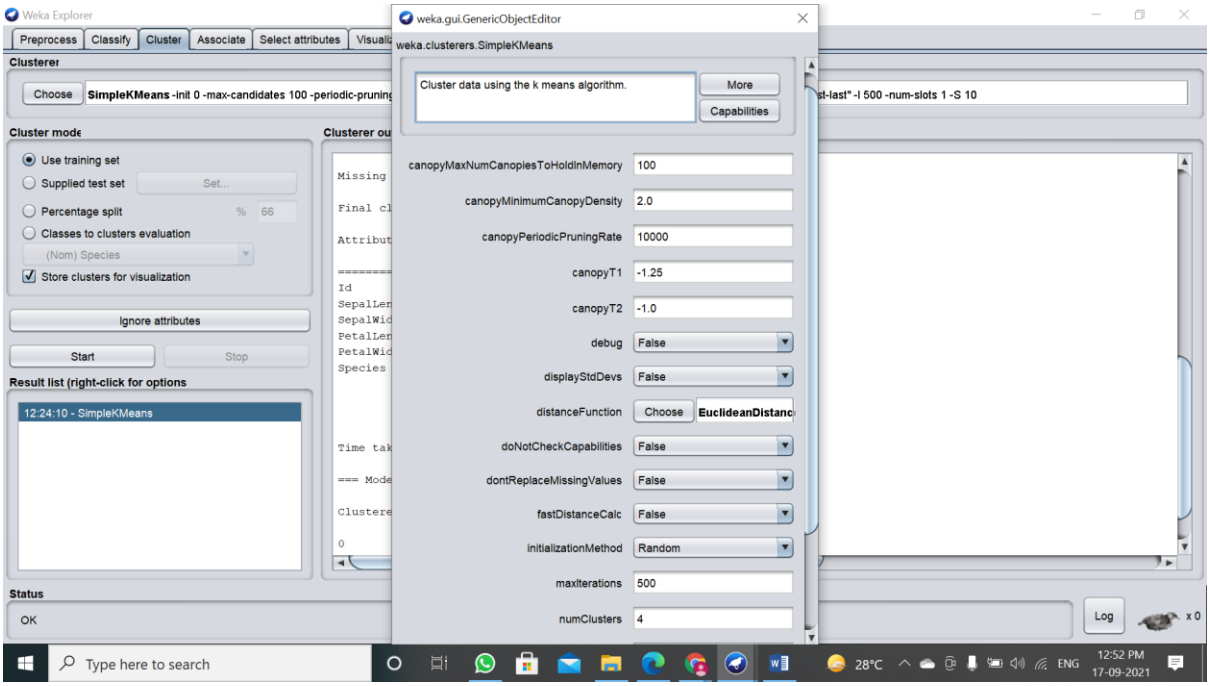


Step 4: Click on Start.

Name: Zeenat **Class: TYIT** **Roll no: 578**

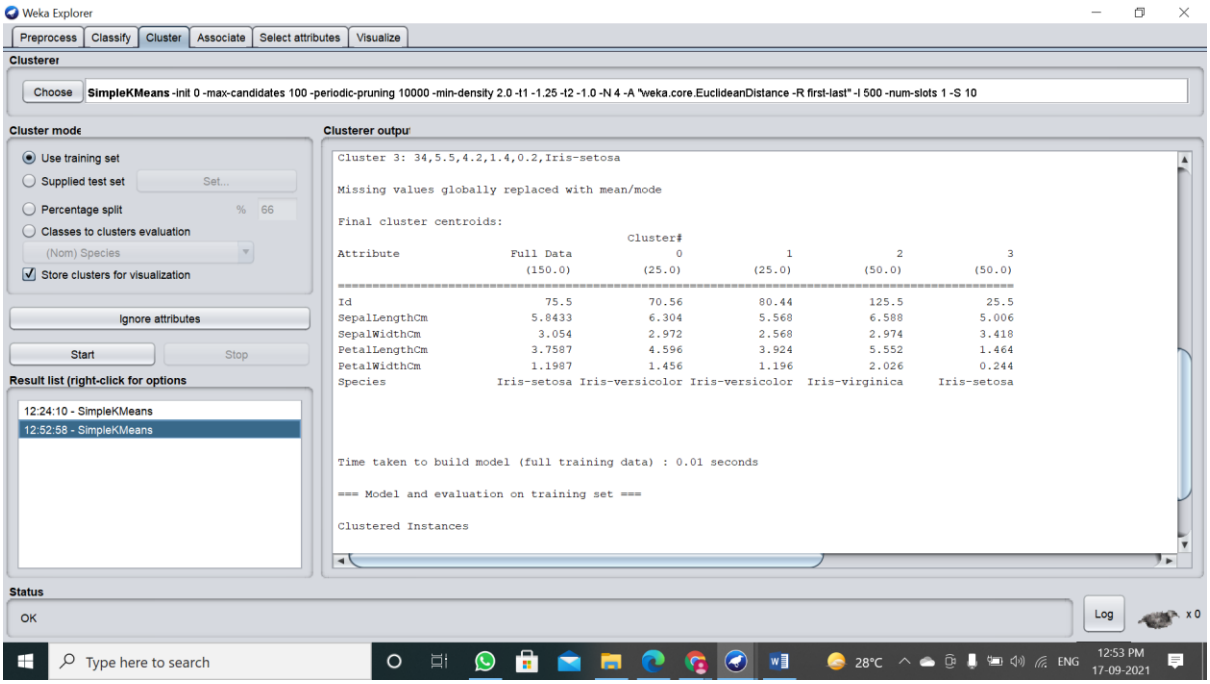


Step 5: Open GenericObjectEditor and changes value numClusters to 4.



Output:

Name: Zeenat **Class: TYIT** **Roll no: 578**



Name: Zeenat **Class: TYIT** **Roll no: 578**

Practical 6

Aim: Demonstrate performing Regression on data sets.

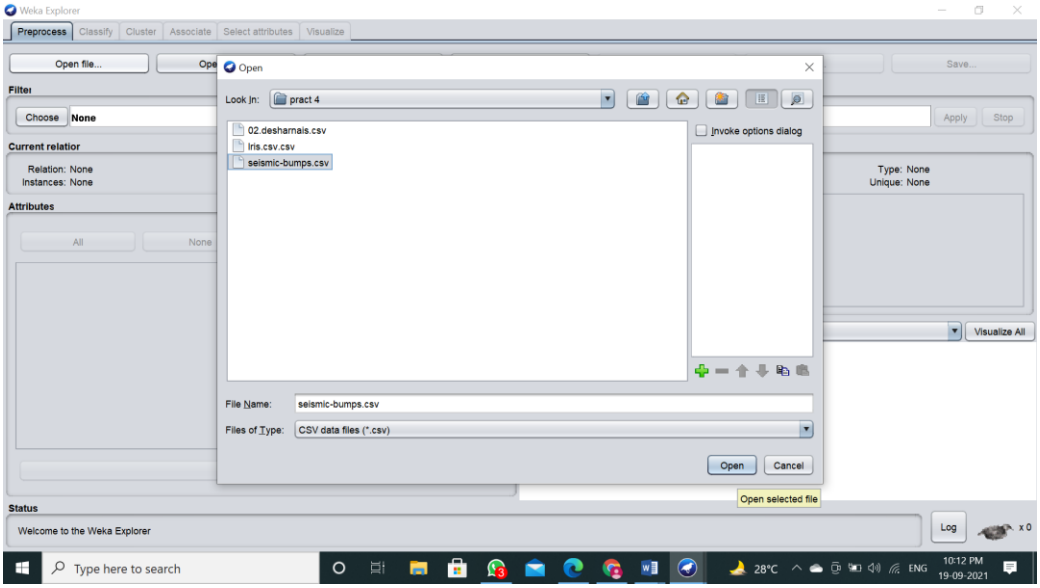
Theory:

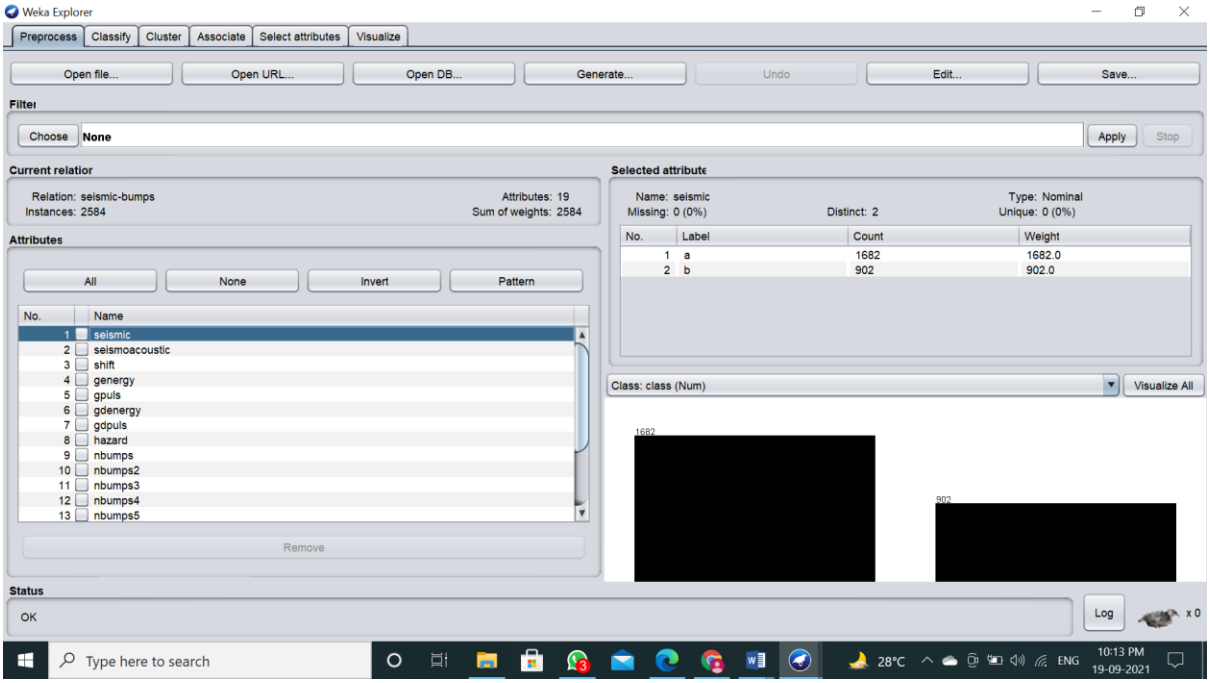
Regression algorithms predict the output values based on input features from the data fed in the system. The go-to methodology is the algorithm builds a model on the features of training data and using the model to predict value for new data. According to Oracle, here's a great definition of Regression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

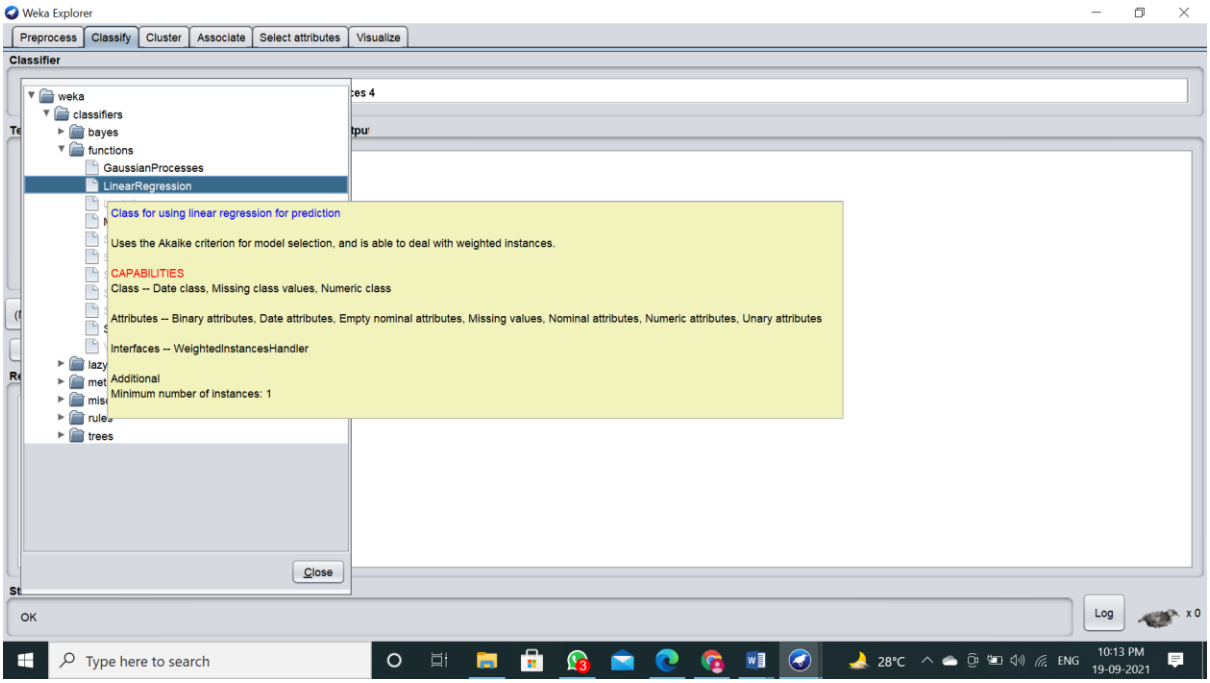
Name: Zeenat **Class: TYIT** **Roll no: 578**



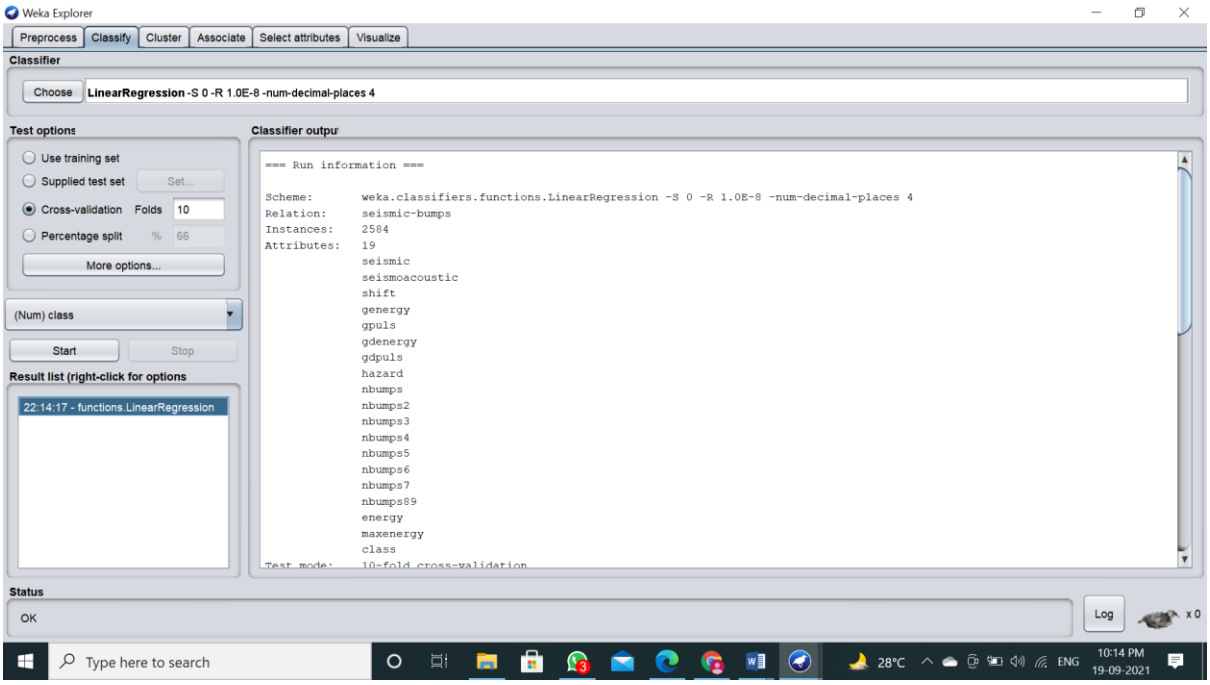


Step 3: Go to Classify > Choose > Functions > Linear Regression

Name: Zeenat **Class: TYIT** **Roll no: 578**



Step 4: Click on Start for the Output.



Name: Zeenat **Class: TYIT** **Roll no: 578**

Practical 7

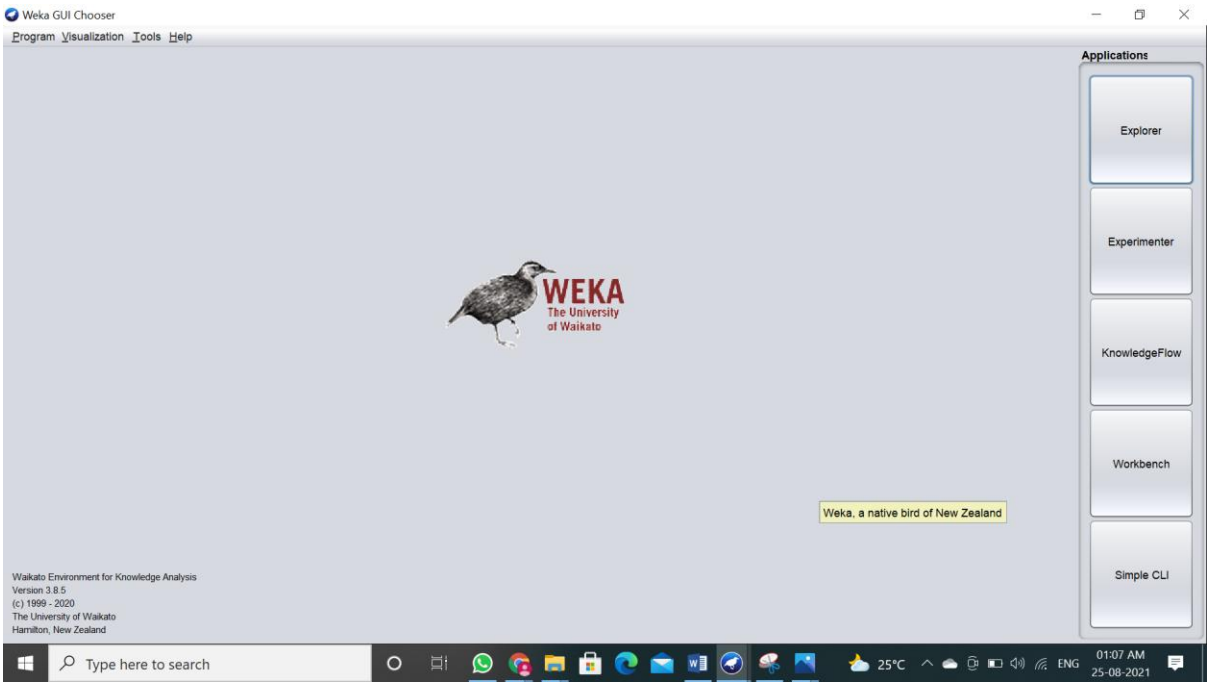
Aim: To implement Decision Tree

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.

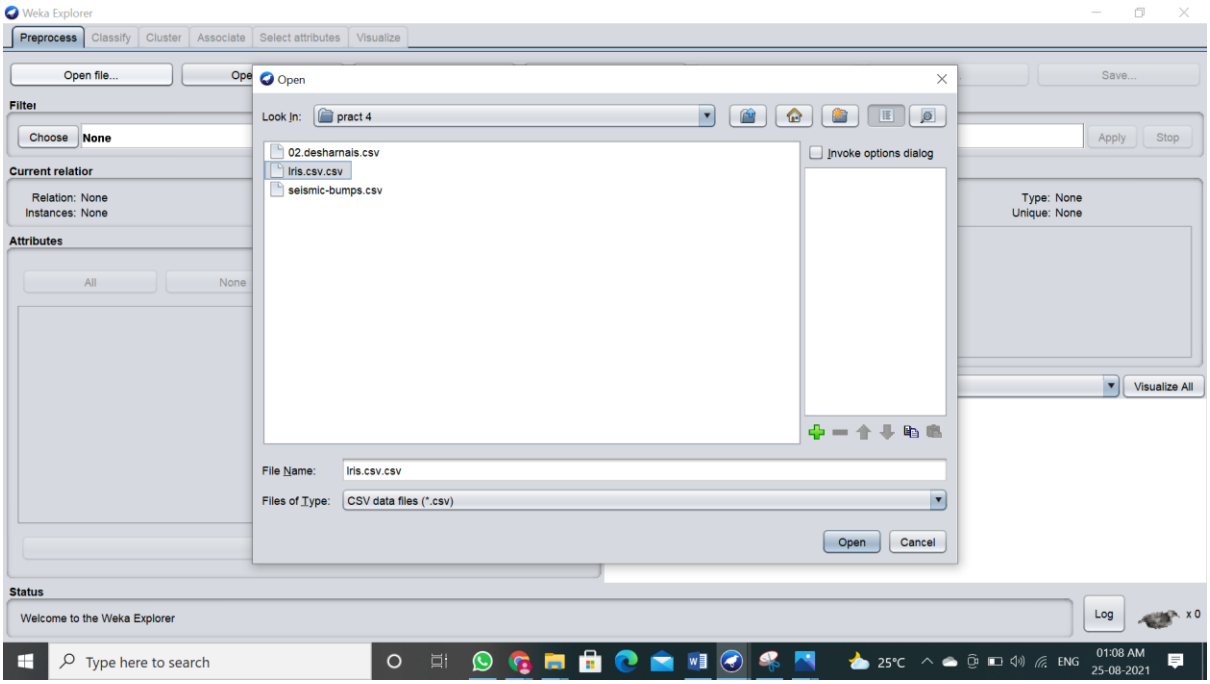
The Decision Tree algorithm was released as ID3 (Iterative Dichotomiser) by machine researcher J. Ross Quinlan.

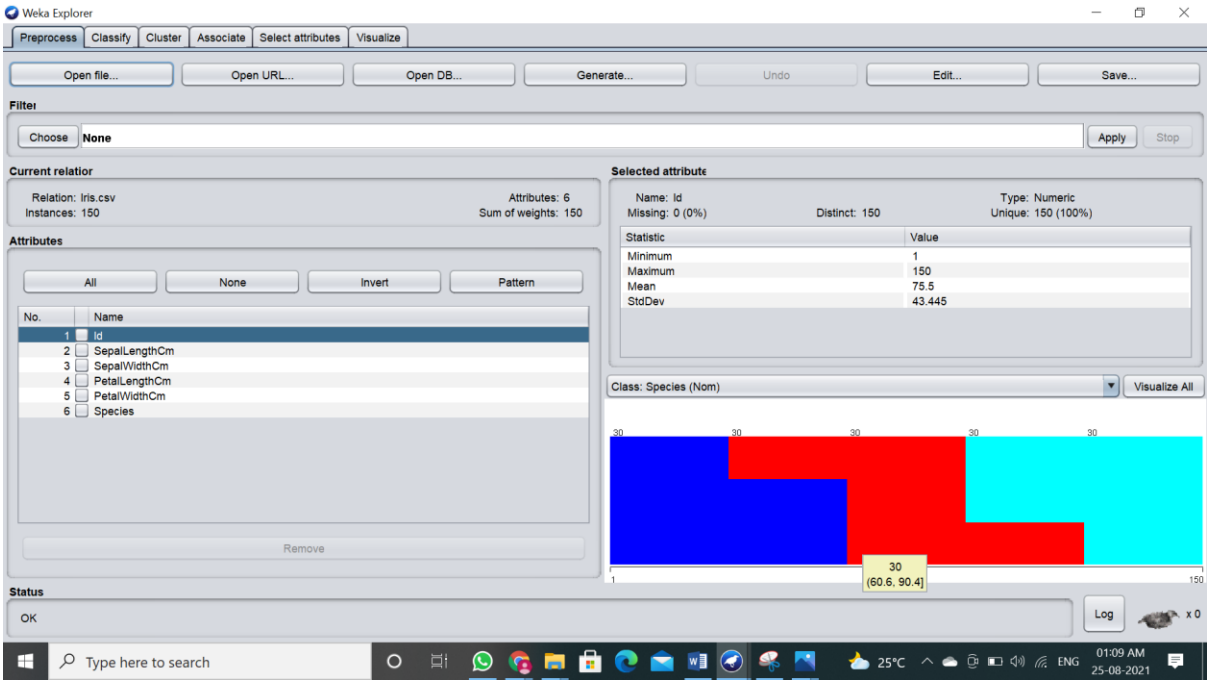
Step 1: Open Weka Software and go to Explorer.



Name: Zeenat **Class: TYIT** **Roll no: 578**

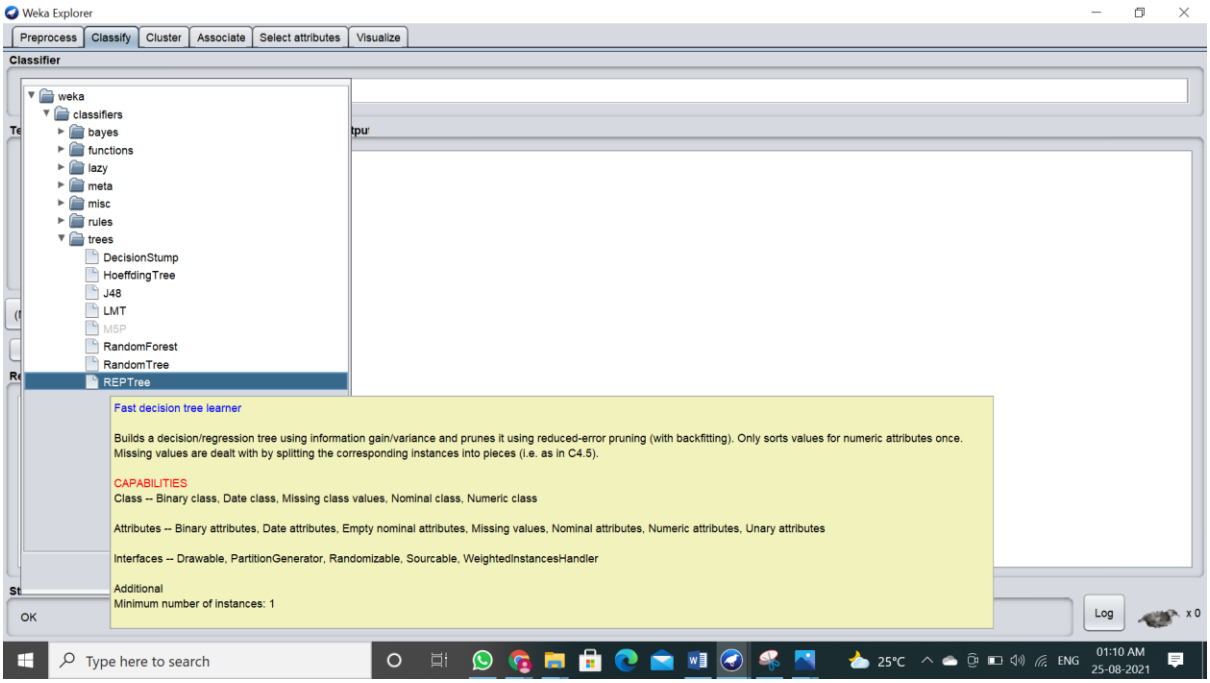
Step 2: Go to Open File and open CSV File.

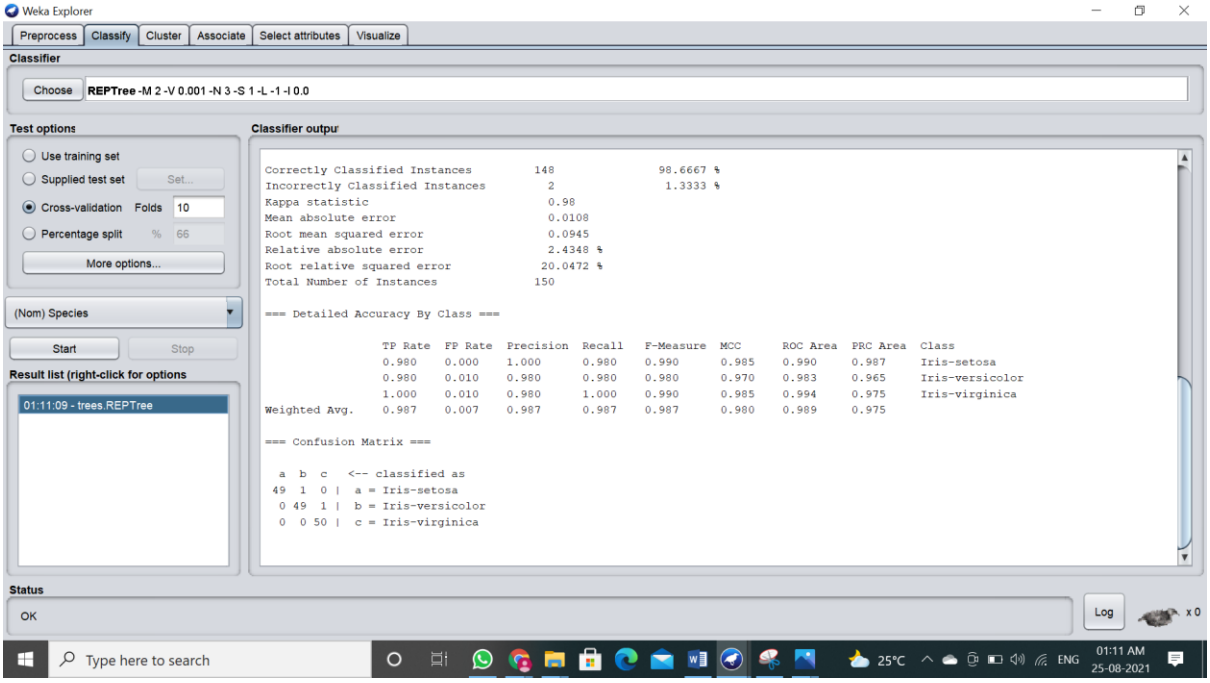




Step 3: Go to Classify > Choose > Trees > REPTree and click on Start Button to get Output

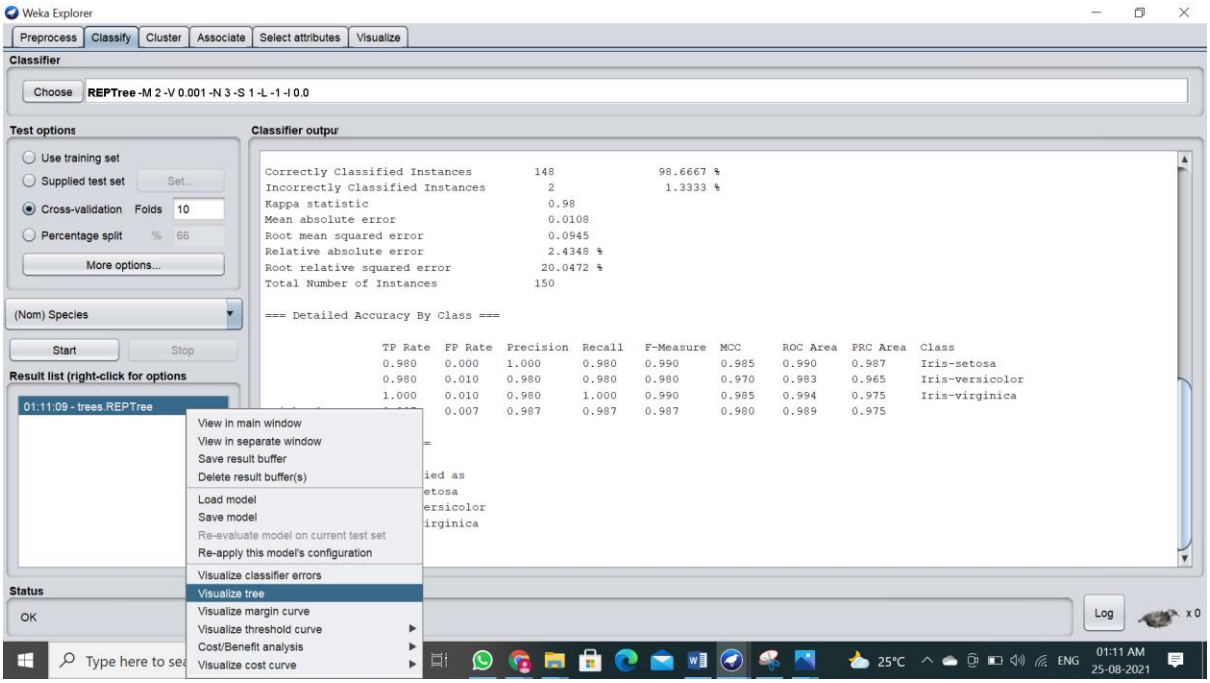
Name: Zeenat **Class: TYIT** **Roll no: 578**

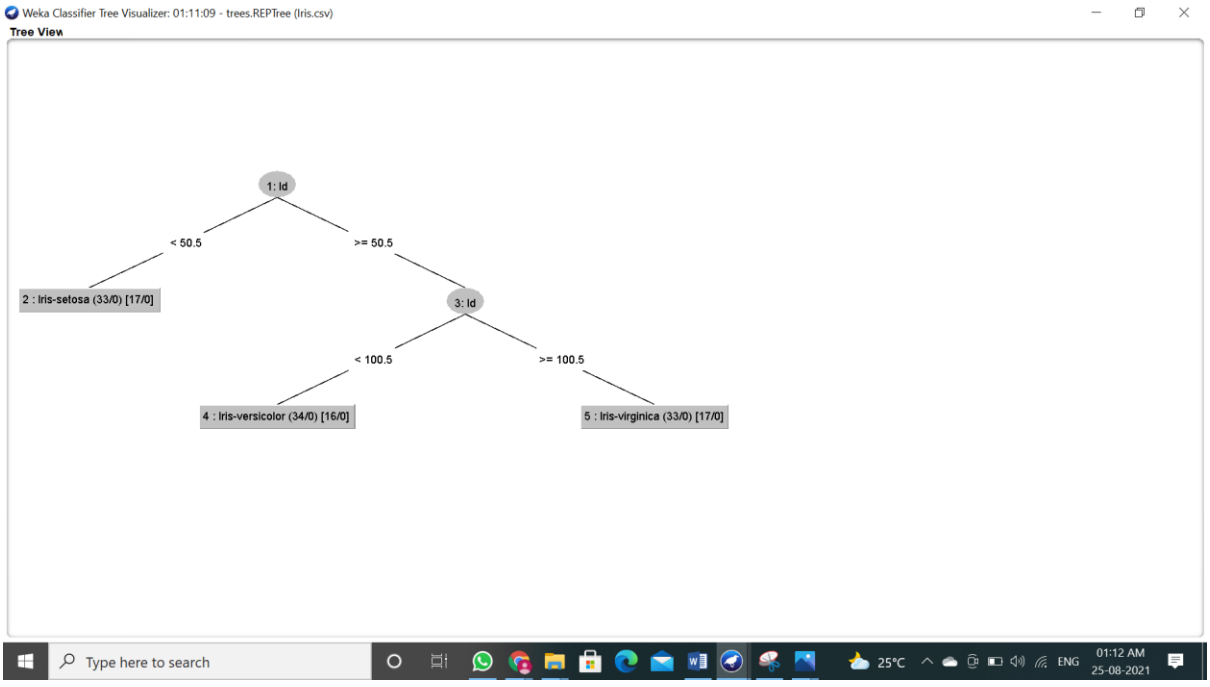




Step 4: To view the Tree right click on REPTree and click on Visualize Tree.

Name: Zeenat **Class: TYIT** **Roll no: 578**





Name: Zeenat **Class: TYIT** **Roll no: 578**

Practical 8

Aim: Implement Naìve 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.

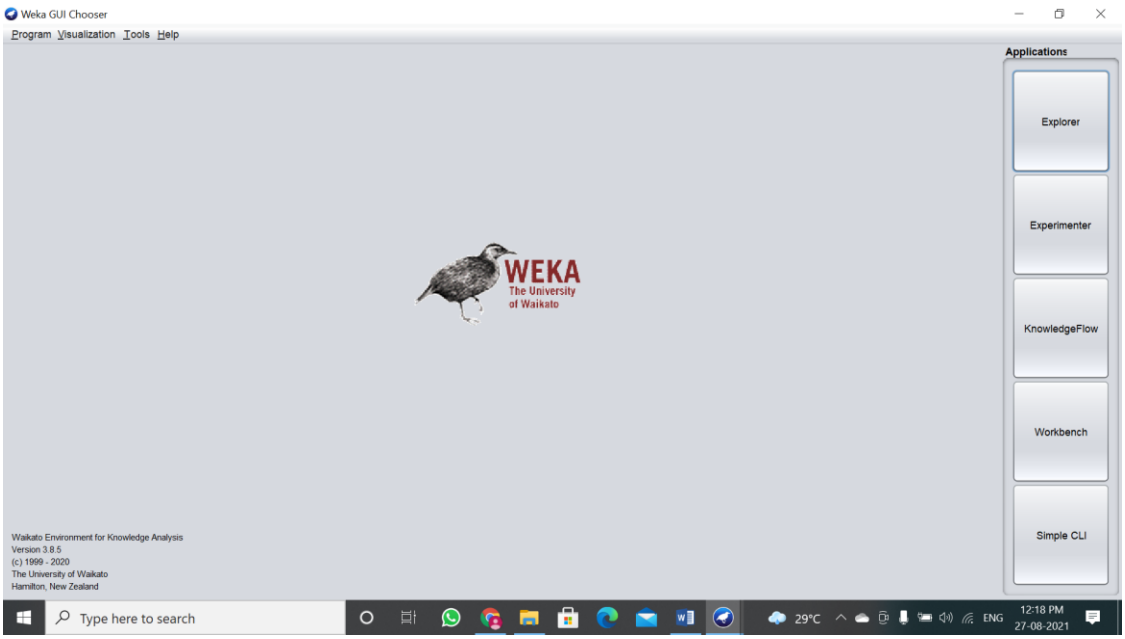
Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.
A Naive Bayes model consists of a large cube that includes the following dimensions:

- 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

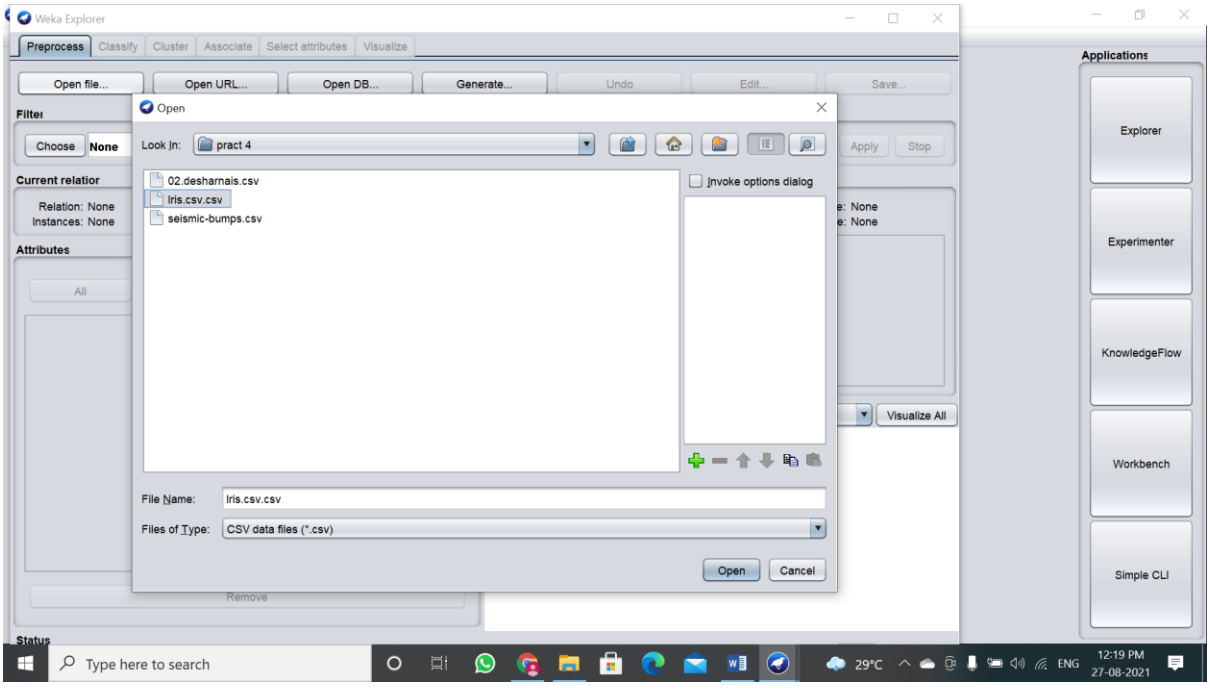
This means that a Naive Bayes model records how often a target field value appears together with a value of an input field.

Step 1: Open Weka Software and go to Explorer.

Name: Zeenat **Class: TYIT** **Roll no: 578**

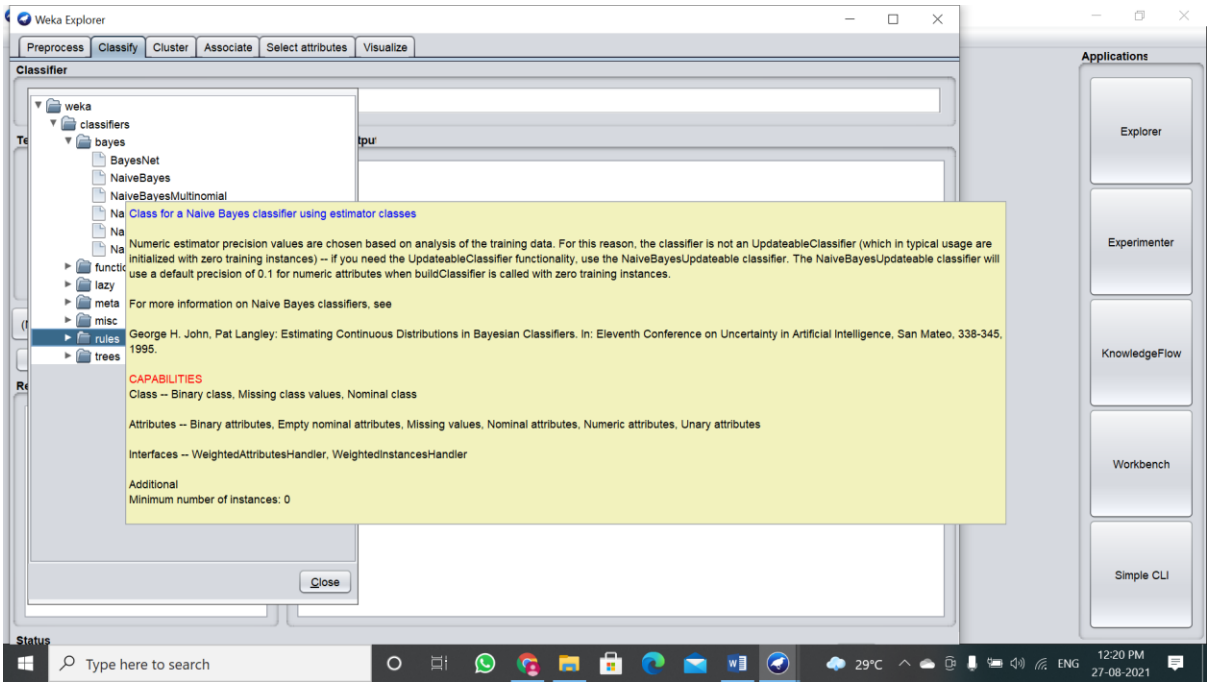


Step 2: Go to Open File and open CSV File.

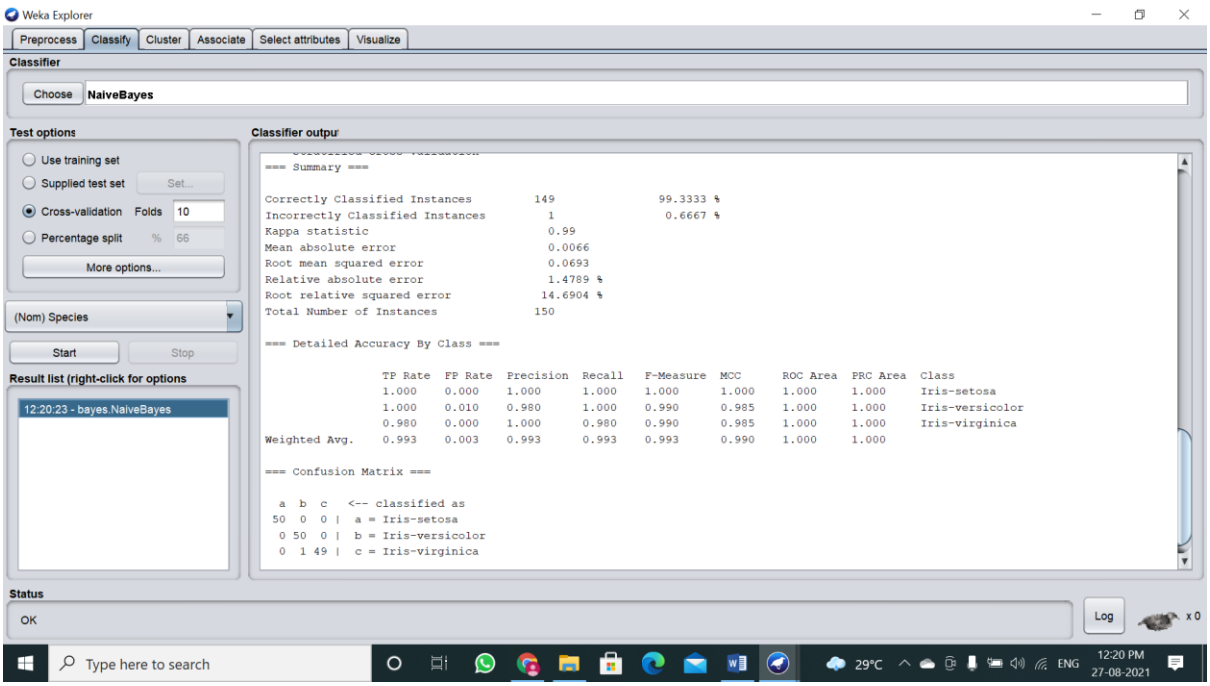


Step 3: Go to Classify > Choose > bayes> NaiveBayes and click on Start Button.

Name: Zeenat **Class: TYIT** **Roll no: 578**



Output:



Name: Zeenat **Class: TYIT** **Roll no: 578**

Practical 9

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

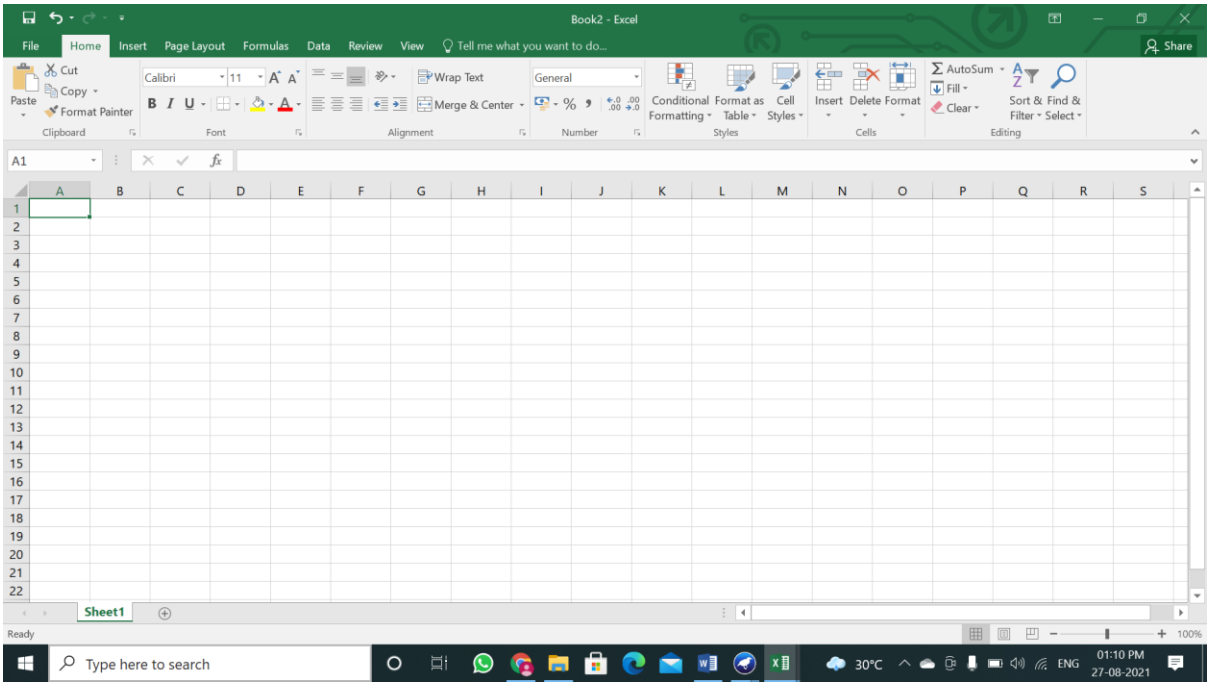
Theory:

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.

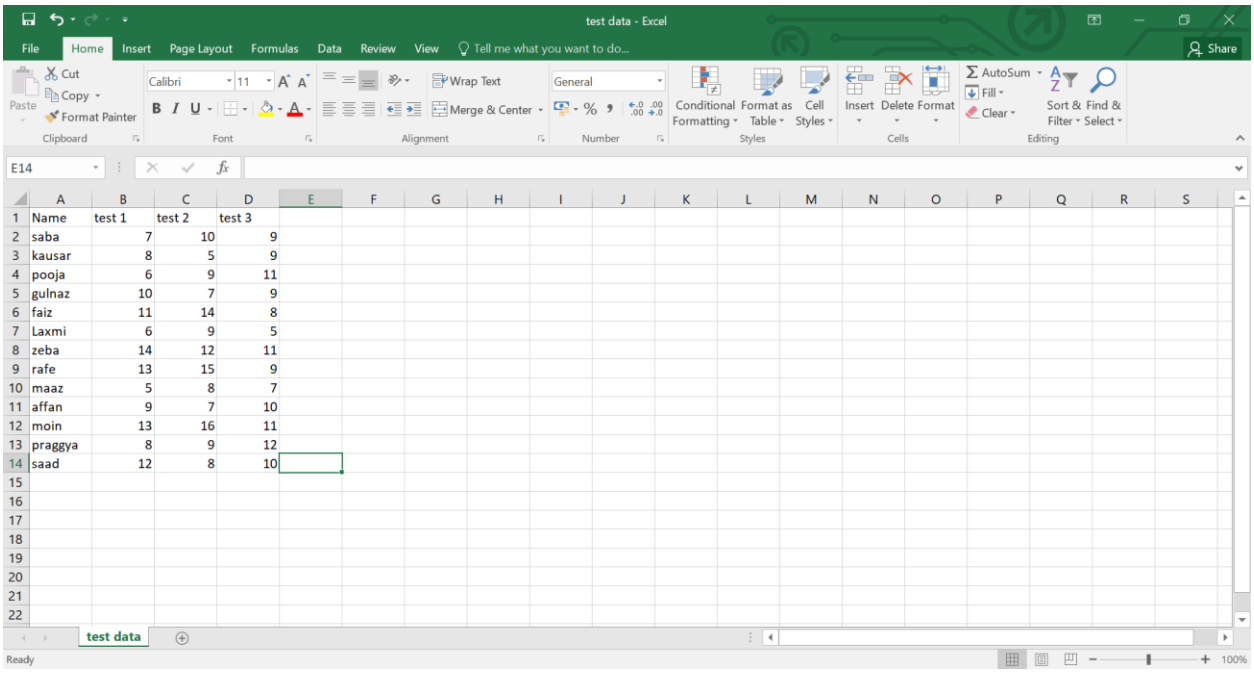
Data in a database or data warehouse can be viewed at different granularity or abstraction levels, or as different combinations of attributes or dimensions. Data can be presented in various visual forms, such as boxplots, 3-D cubes, data distribution charts, curves, surfaces, and link graphs, as shown 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 Exal Sheet



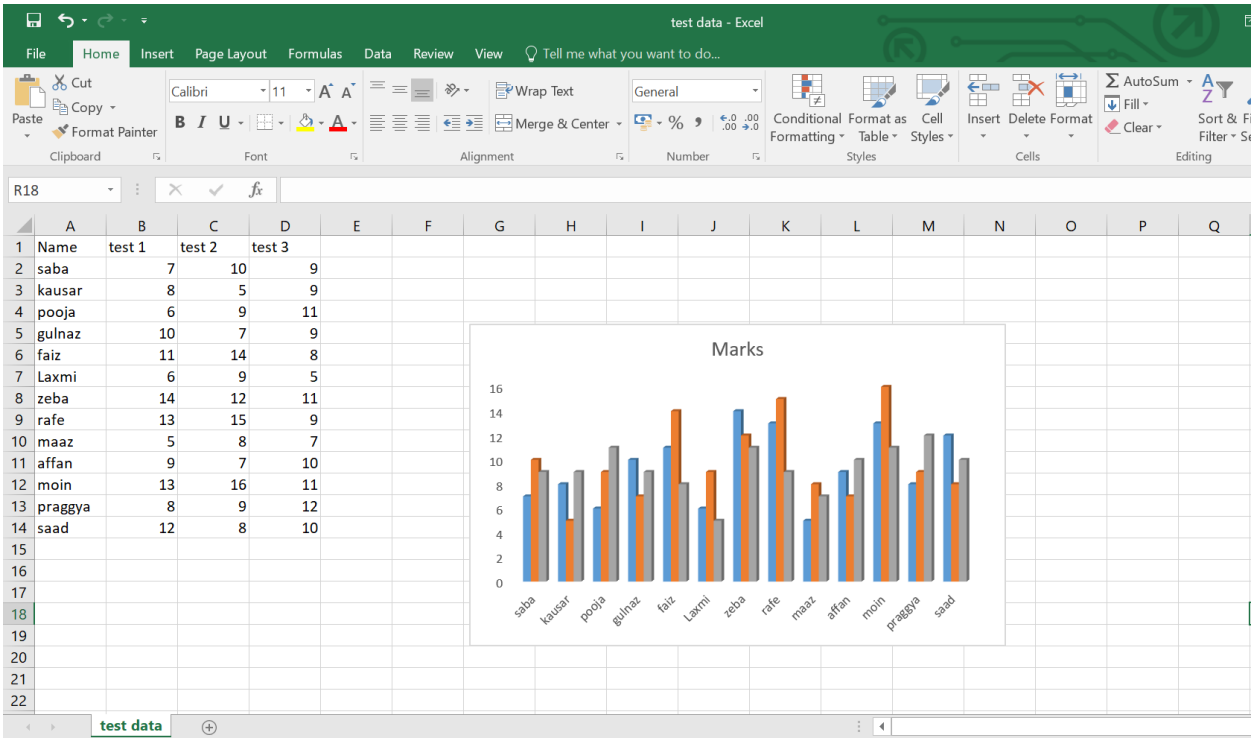
Step 2: Give some data

Name: Zeenat **Class: TYIT** **Roll no: 578**



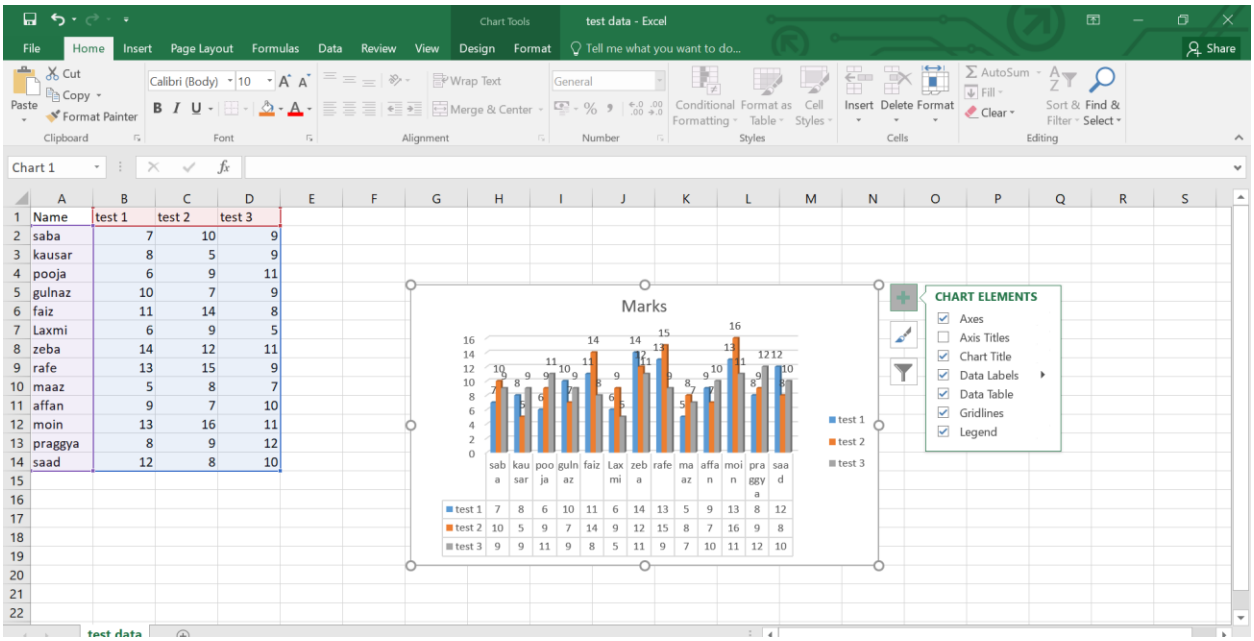
	A	B	C	D
1	Name	test 1	test 2	test 3
2	saba	7	10	9
3	kausar	8	5	9
4	pooja	6	9	11
5	gulnaz	10	7	9
6	faiz	11	14	8
7	Laxmi	6	9	5
8	zeba	14	12	11
9	rafe	13	15	9
10	maaz	5	8	7
11	affan	9	7	10
12	moin	13	16	11
13	praggya	8	9	12
14	saad	12	8	10

Step 3: Make a Graph for the test data.



Step 4: Add Chart Element

Name: Zeenat **Class: TYIT** **Roll no: 578**



Output: