EXP.NO:1	DATA EXPLORATION AND INTEGRATION WITH WEKA - IRIS DATASET

Introduction

The goal of this lab is to install and familiarize with Weka.

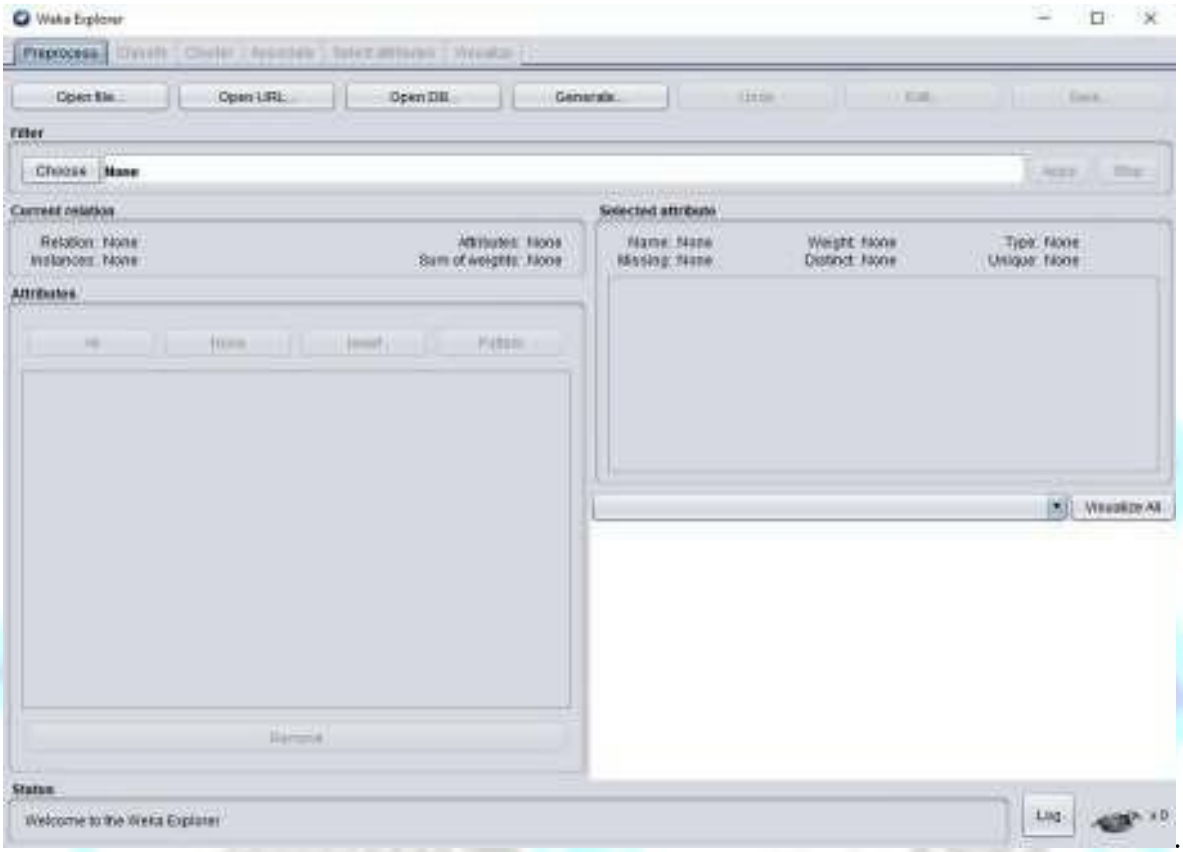
Steps:

1. Download and install Weka. You can find it here:
http://www.cs.waikato.ac.nz/ml/weka/downloading.html

2. Open Weka and have a look at the interface. It is an open-source project written in Java from the University of Waikato.



3. Click on the Explorer button on the right side:
lOMoAR cPSD|20149634



Check different tabsto familiarize with the tool.

5. Weka comes with a number of small datasets. Those files are located at C:\Program Files\Weka3-8 (If it is installed at this location. Or else, search for Weka-3-8 to find the installation location). In this folder, there is a subfolder named ‘data’. Open that folder to see all files that comes with Weka.
6. For easy access, copy the folder ‘data’ and paste it in your ‘Documents’ folder.
7. In this lab, we will work with the dataset Iris. To open Iris dataset, click on ‘Open file’ in the ‘Preprocess tab’. From your ‘data’ folder, select iris.arff and hit open.
8. To know more about the iris dataset, open iris.arff in notepad++ or in a similar tool and read the comments.
9. Click on visualize tab to see various 2D visualizations of the dataset.

    a. Click on some graphs to see more details about it.
    b. In any of the graph, click one’x’ to see details about that data record.
    lOMoAR cPSD|20149634

10. Fill this table:

Flower Type	Count
Iris Setosa	50
Iris Versicolour	50
Iris Virginica	50

11. Fill this table:

Attribute	Minimum	Maximum	Mean	StdDev
sepal length	4.3	7.9	5.84	0.83
sepal width	2.0	4.4	3.05	0.43
petal length	1.0	6.9	3.76	1.76
petal width:	0.1	2.5	1.20	0.76

Result:

Thus the data exploration and integration using weka is explored successfully.

EXP.NO: 2A	CONVERSION OF TEXT FILE INTO ARFF FILE

Aim:

To convert a text file to ARFF(Attribute-Relation File Format) using Weka3.8.2 tool.

Objectives:

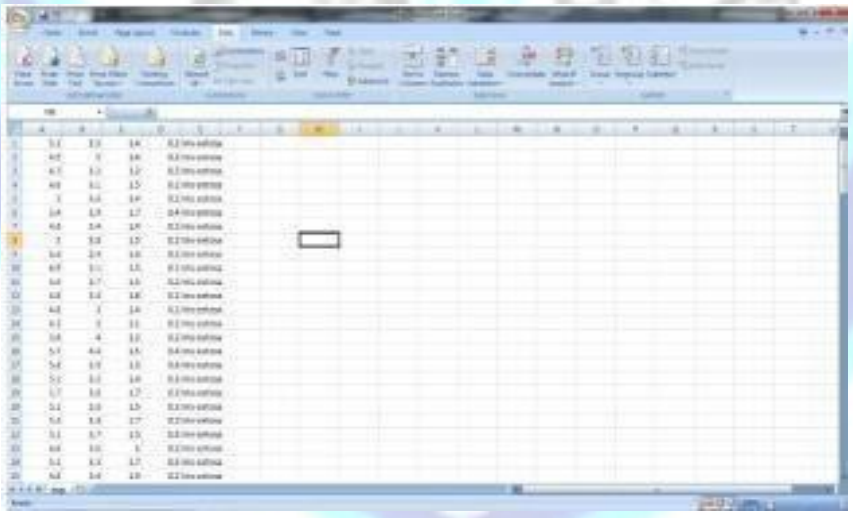
Most of the data that we have collected from public forum is in the text format that cannot be read byWeka tool. Since Weka (Data Mining tool) recognizes the data in ARFF format only we have to convert the text file into ARFF file.

Algorithm:

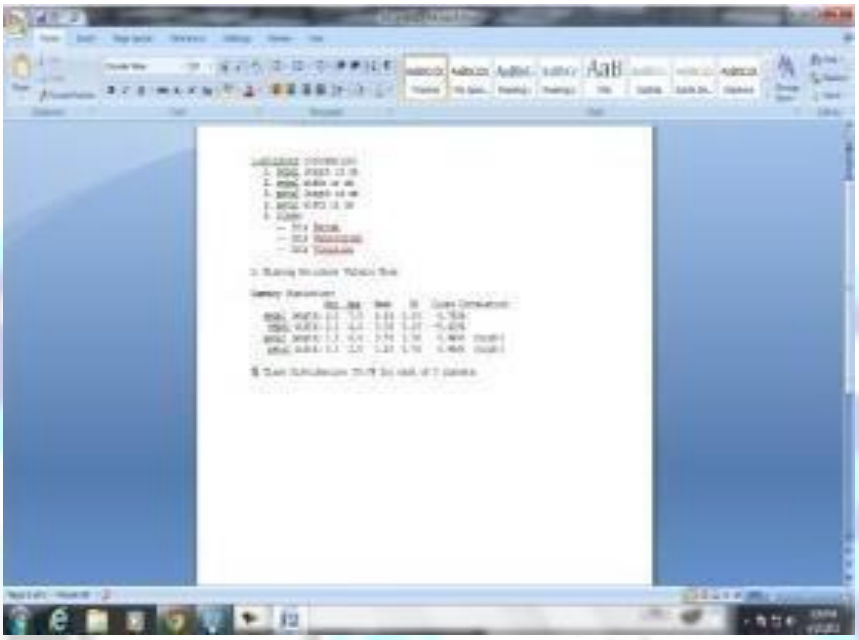
1. Download any data set from UCI data repository.
2. Open the same data file from excel. It will ask for delimiter (which produce column) in excel.
3. Add one row at the top of the data.
4. Enter header for each column.
5. Save file as .CSV (Comma Separated Values) format.
6. Open Weka tool and open the CSV file.
7. Save it as ARFF format.

Output:

Data Text File:



Data ARFF File:



Result:

Thus, conversion of a text file to ARFF(Attribute-Relation File Format) using Weka3.8.2 tool is implemented.

EXP.NO: 2B	CONVERSION OF ARFF TO TEXT FILE

Aim:

To convert ARFF (Attribute-Relation File Format) into text file.

Objectives:

Since the data in the Weka tool is in ARFF file format we have to convert the ARFF file to text format forfurther processing.

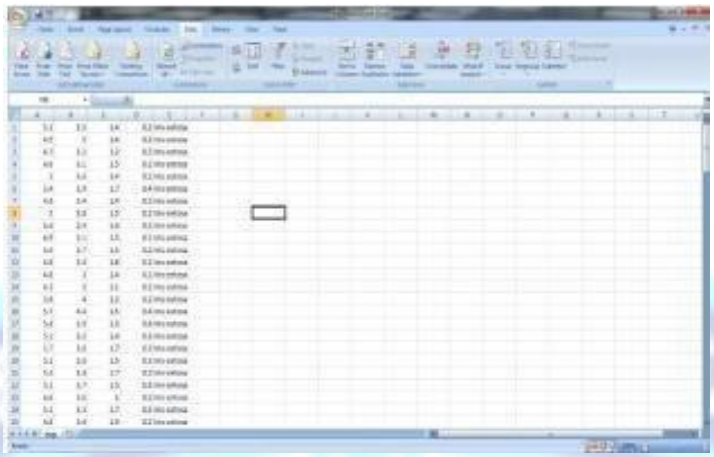
Algorithm:

1. Open any ARFF file in Weka tool.
2. Save the file as CSV format.
3. Open the CSV file in MS-EXCEL.
4. Remove some rows and add coreseponding header to the data.
5. Save it as text file with the desire delimiter.

Data ARFF File:



Data Text File:



Result:

Thus conversion of ARFF (Attribute-Relation File Format) into text file is implemented.

EXP.NO: 3A	TRAINING THE GIVEN DATASET FOR AN APPLICATION

Aim:

To apply the concept of Linear Regression for training the given dataset.

Algorithm:

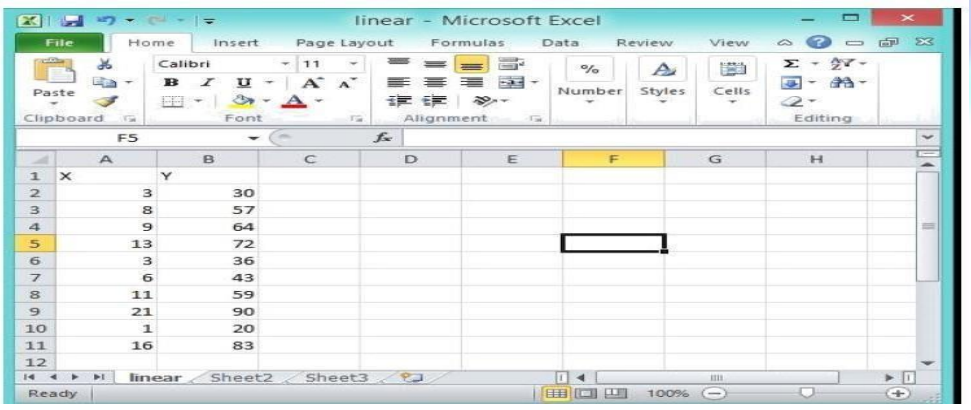
1. Open the weka tool.
2. Download a dataset by using UCI.
3. Apply replace missing values.
4. Apply normalize filter.
5. Click the Classify Tab.
6. Choose the Simple Linear Regression option.
7. Select the training set of data.
8. Start the validation process.
9. Note the output.

LINEAR REGRESSION:

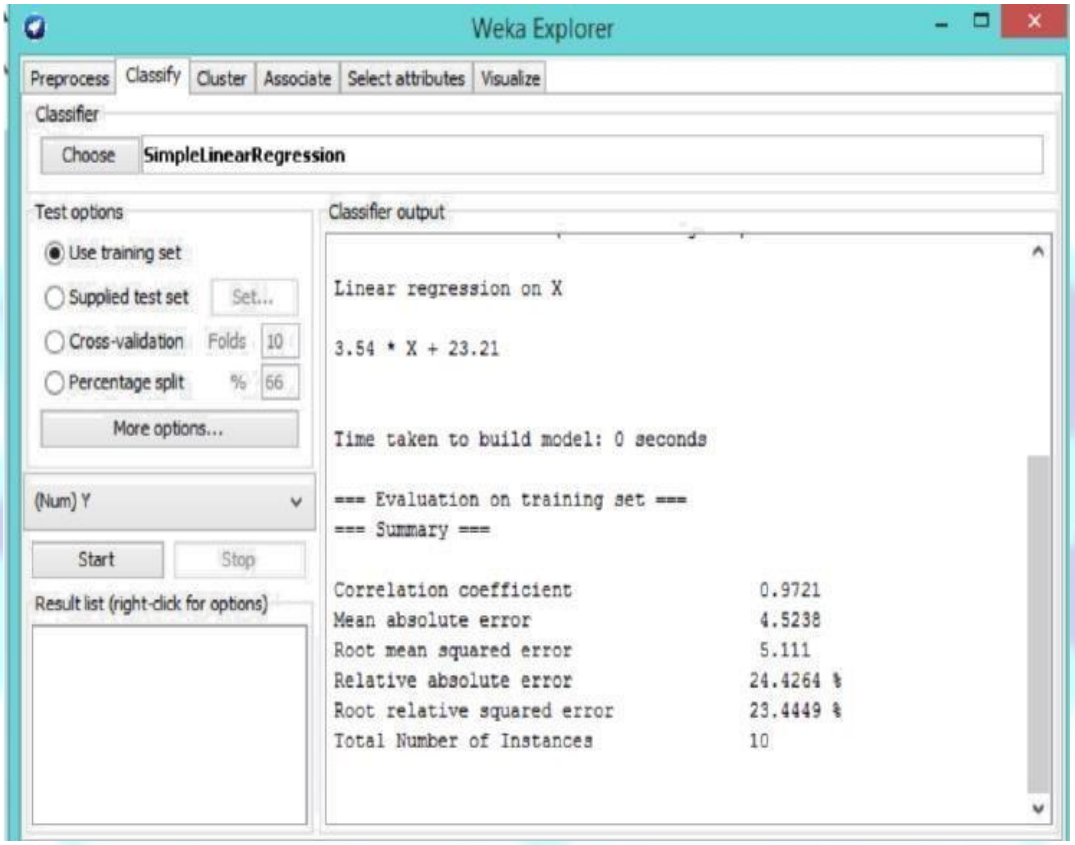
In statistics, Linear Regression is an approach for modeling a relationship between a scalar dependent variable Yand one or more explanatory variables denoted X.the case of explanatory variable is called Simple Linear Regression. Coefficient of Linear Regression is given by: **Y=ax+b**

PROBLEM:
Consider the dataset below where x is the number of working expeince of a college graduate and y is the corresponding salary of the graduate. Build a regression equation and predict the salary of college graduate whoseexperience is 10 years.

INPUT:



Output:



Result:

Thus the concept of Linear Regression for training the given dataset is applied and implemented.

EXP.NO: 3B	TESTING THE GIVEN DATASET FOR AN APPLICATION

Aim:

To apply the Navie Bayes Classification for testing the given dataset.

Algorithm:

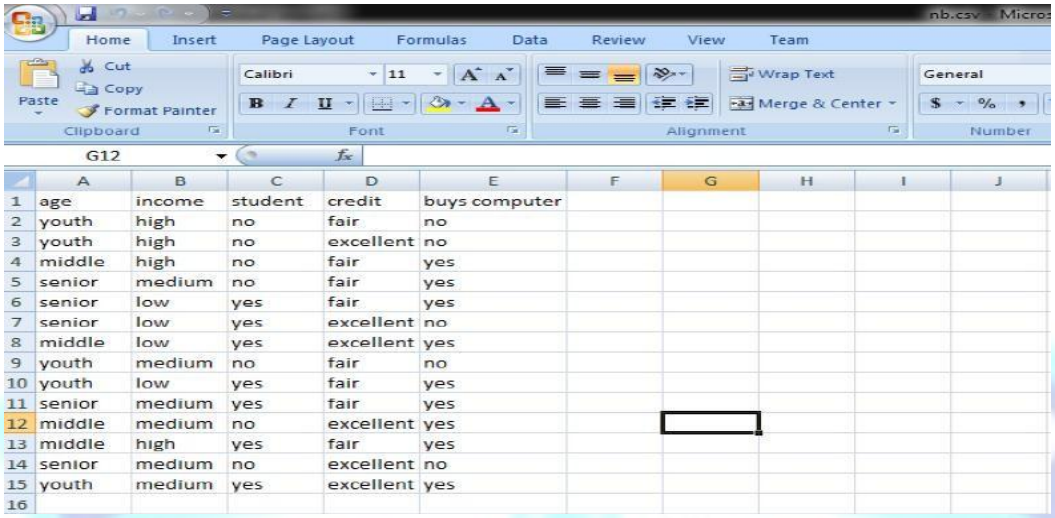
1. Open the weka tool.
2. Download a dataset by using UCI.
3. Apply replace missing values.
4. Apply normalize filter.
5. Click the Classification Tab.
6. Apply Navie Bayes Classification.
7. Find the Classified Value.
8. Note the output.

Bayes' Theorem In the Classification Context:

X is a data tuple. In Bayesian term it is considered "evidence".H is some hypothesis that X belongs to a specified class C .P(H|X) is the posterior probability of H conditioned on X .

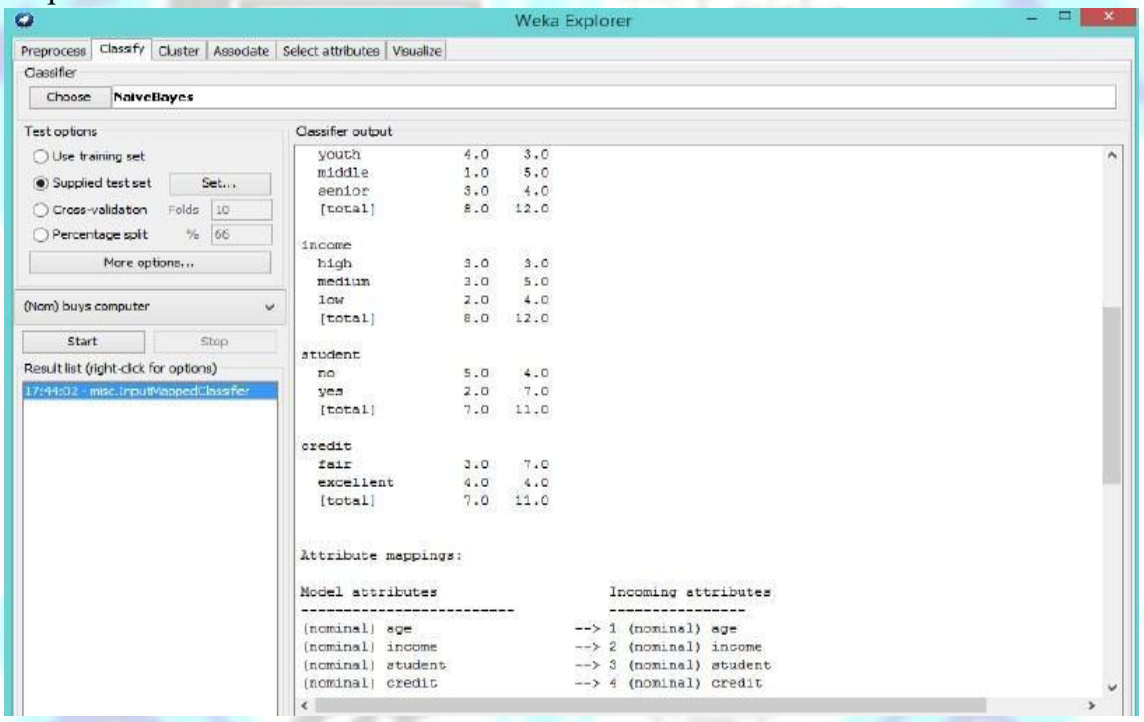
Example: predict whether a costumer will buy a computer or not " Costumers are described by two attributes: age and income " X is a 35 years-old costumer with an income of 40k " H is the hypothesis that the costumer will buy acomputer " P(H|X) reflects the probability that costumer X will buy a computer given that we know the costumers' age and income.

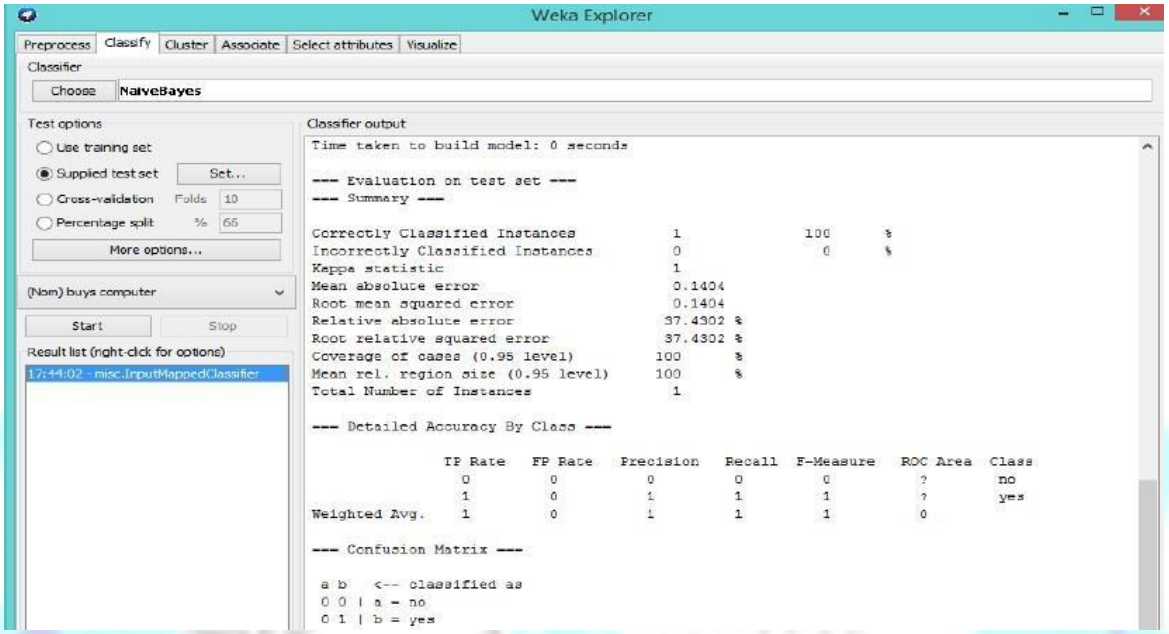
Input Data:



	A	B	C	D	E
1	age	income	student	credit	buys computer
2	youth	high	no	fair	no
3	youth	high	no	excellent	no
4	middle	high	no	fair	yes
5	senior	medium	no	fair	yes
6	senior	low	yes	fair	yes
7	senior	low	yes	excellent	no
8	middle	low	yes	excellent	yes
9	youth	medium	no	fair	no
10	youth	low	yes	fair	yes
11	senior	medium	yes	fair	yes
12	middle	medium	no	excellent	yes
13	middle	high	yes	fair	yes
14	senior	medium	no	excellent	no
15	youth	medium	yes	excellent	yes
16					

Output data:





Result:

Thus the Navie Bayes Classification for testing the given dataset is implemented.

EXP.NO: 4	Pre-process a given dataset based on Handling Missing Values

Aim:

To Pre-process a given dataset based on Handling Missing Values

Process: Replacing Missing Attribute Values by the Attribute Mean. This method is used fordata sets with numerical attributes. An example of such a data set is presented in fig no: 4.1

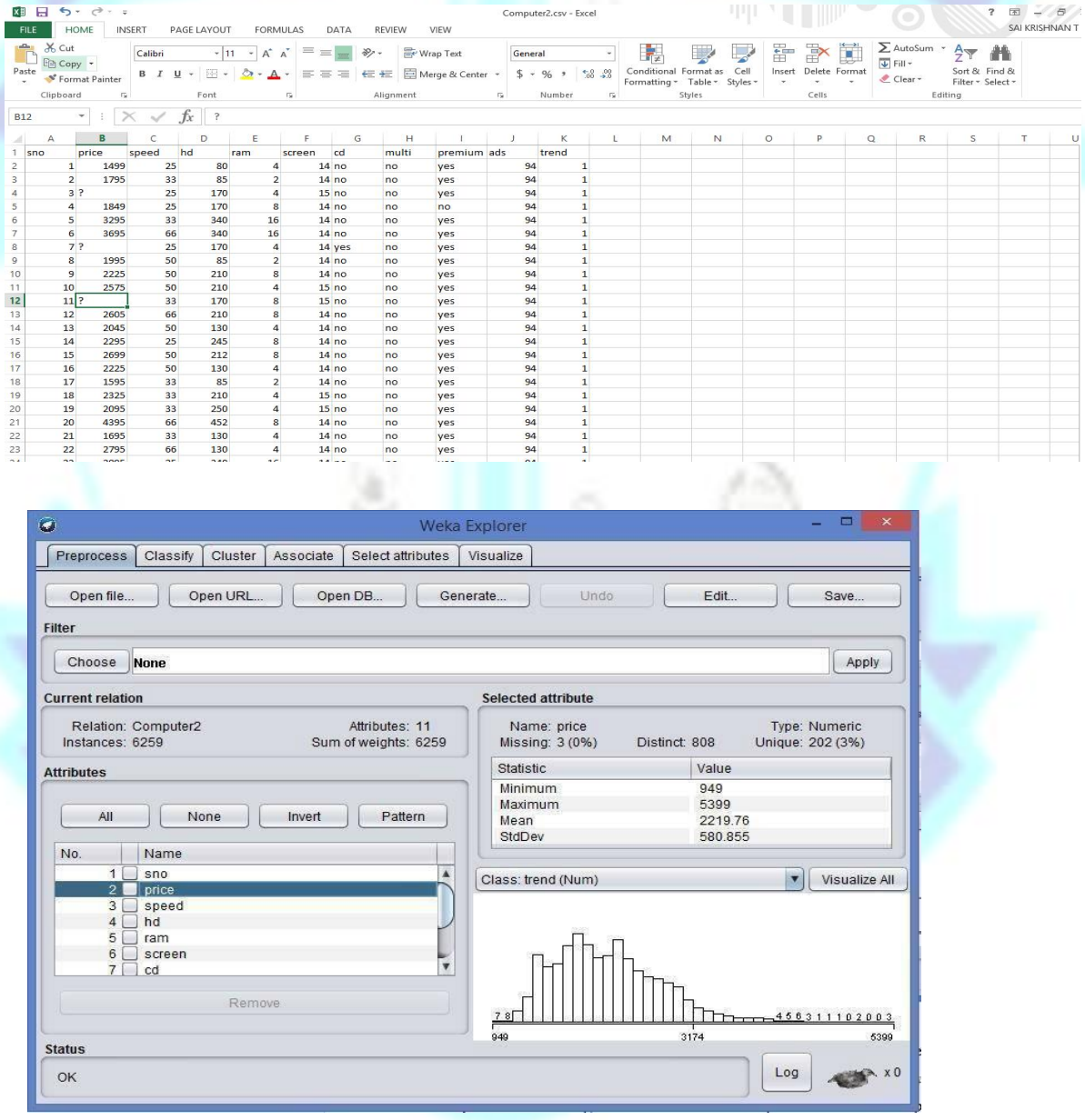
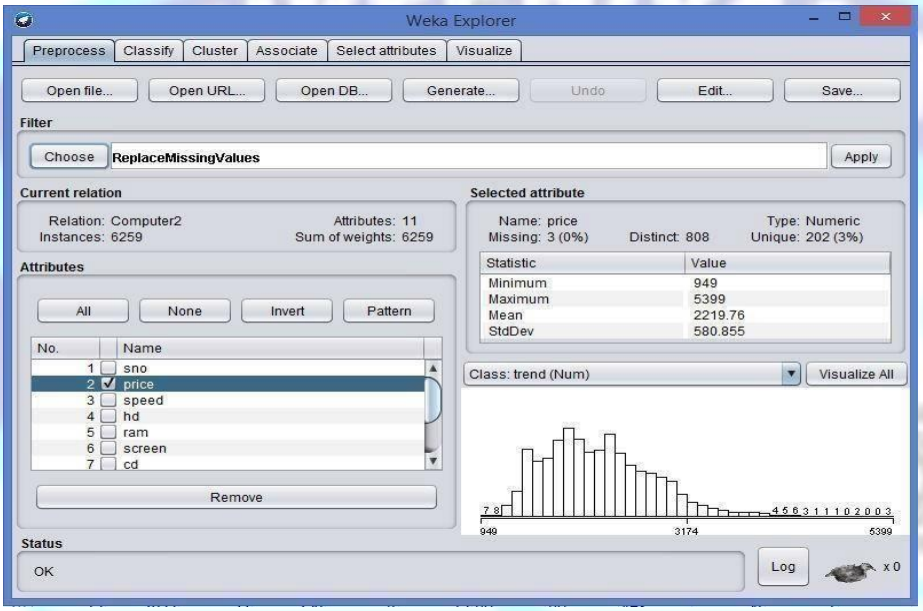
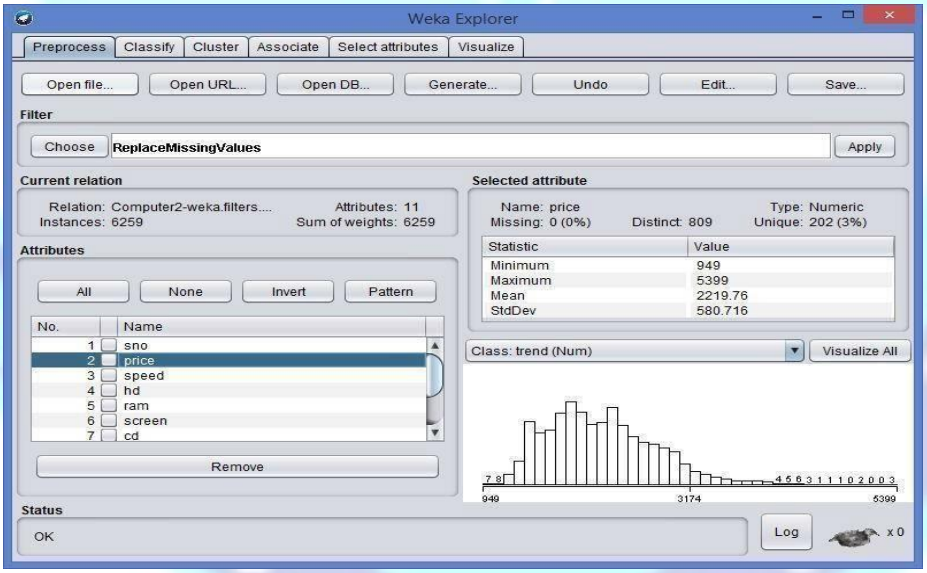
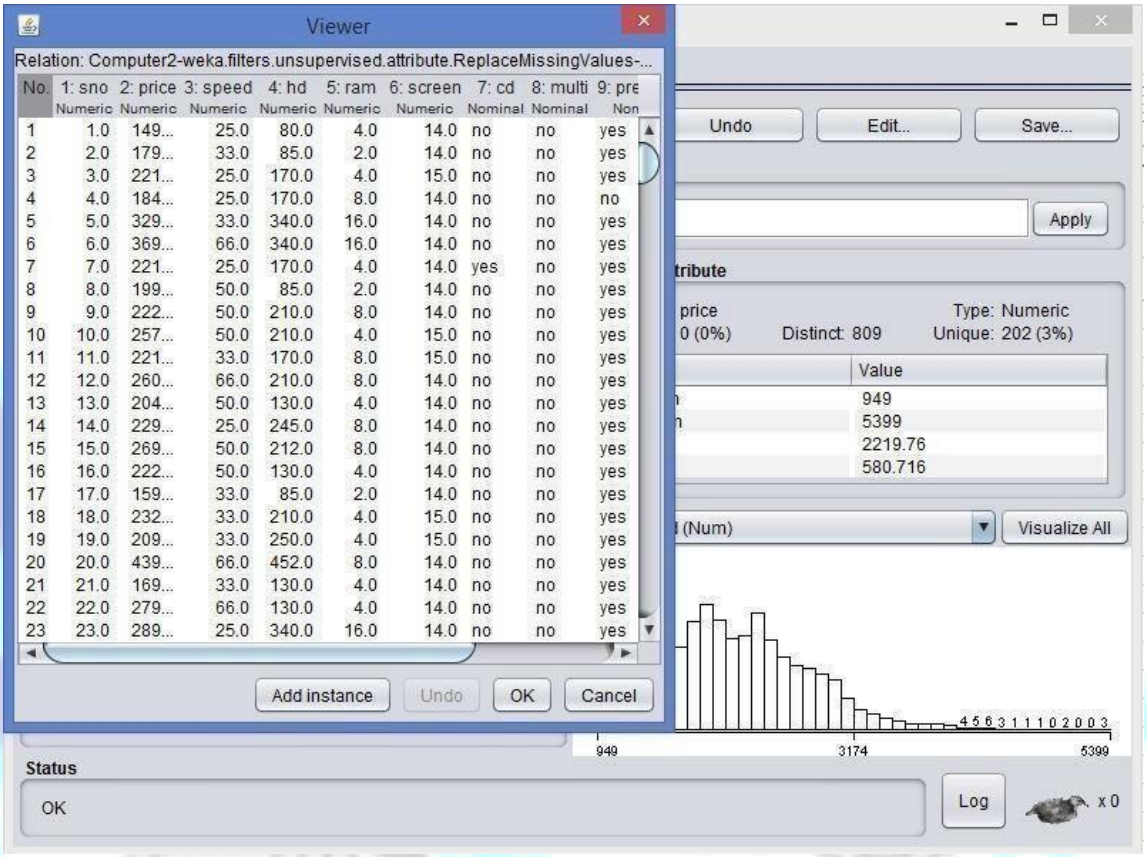


Fig: 4.1 Missing values

In this method, every missing attribute value for a numerical attribute is replaced by the arithmetic mean of known attribute values. In Fig, the mean of known attribute values for Temperature is 99.2, hence all missing attribute values for Temperature should be replaced by The table with missing attribute values replaced by the mean is presented in fig. For symbolic attributes Headache and Nausea, missing attribute values were replaced using the most common value of the Replace Missing Values.







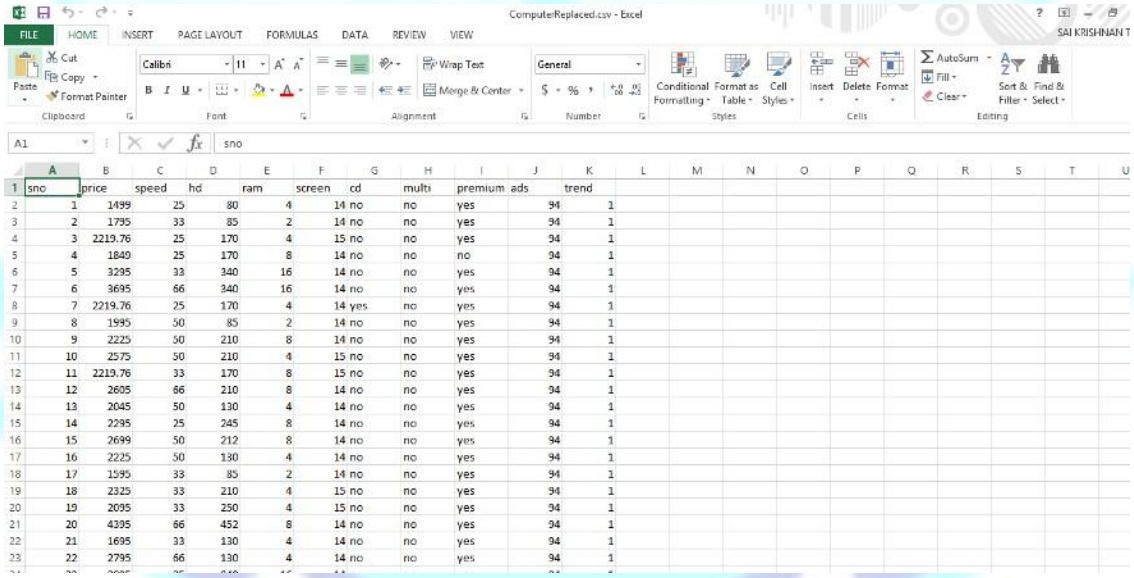


Fig: 4.2 Replaced values

Result:

Thus the preprocessing of handling missing values is filled successfully.

EXP.NO: 5	DATA PRE-PROCESSING – DATA FILTERS

Aim:

To perform the data pre-processing by applying filter.

Objectives:

The data collected from public fourms have plenty of noise or missing data. Weka provides filter to replace themissing values and to remove the noisy data. So that the result will be more accurate.

Algorithm:

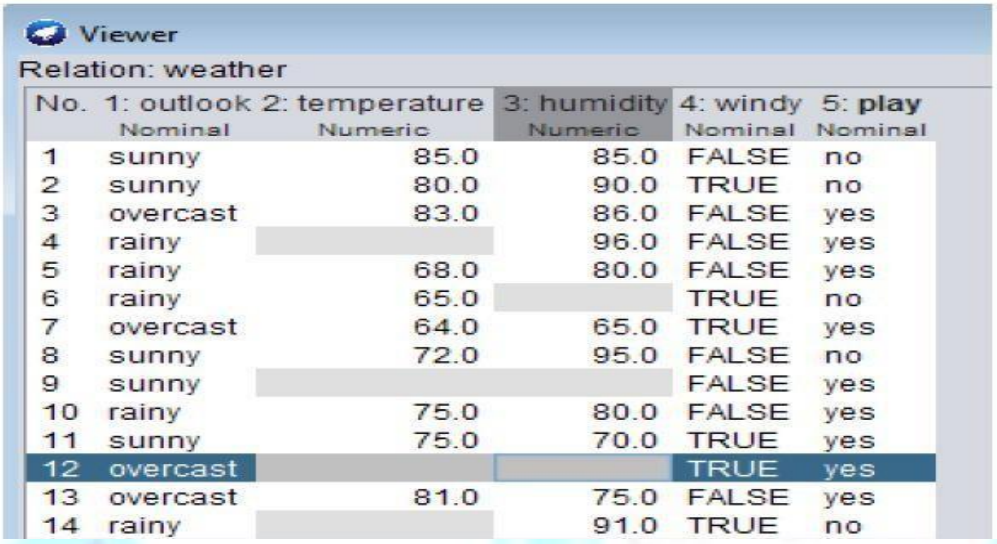
1. Download a complete data set (numeric) from UCI.
2. Open the data set in Weka tool.
3. Save the data set with missing values.
4. Apply replace missing value filter.
5. Calculate the accuracy using the formula

$$\text{Accuracy}=\sqrt{\sum (\text{old-new})^2}$$

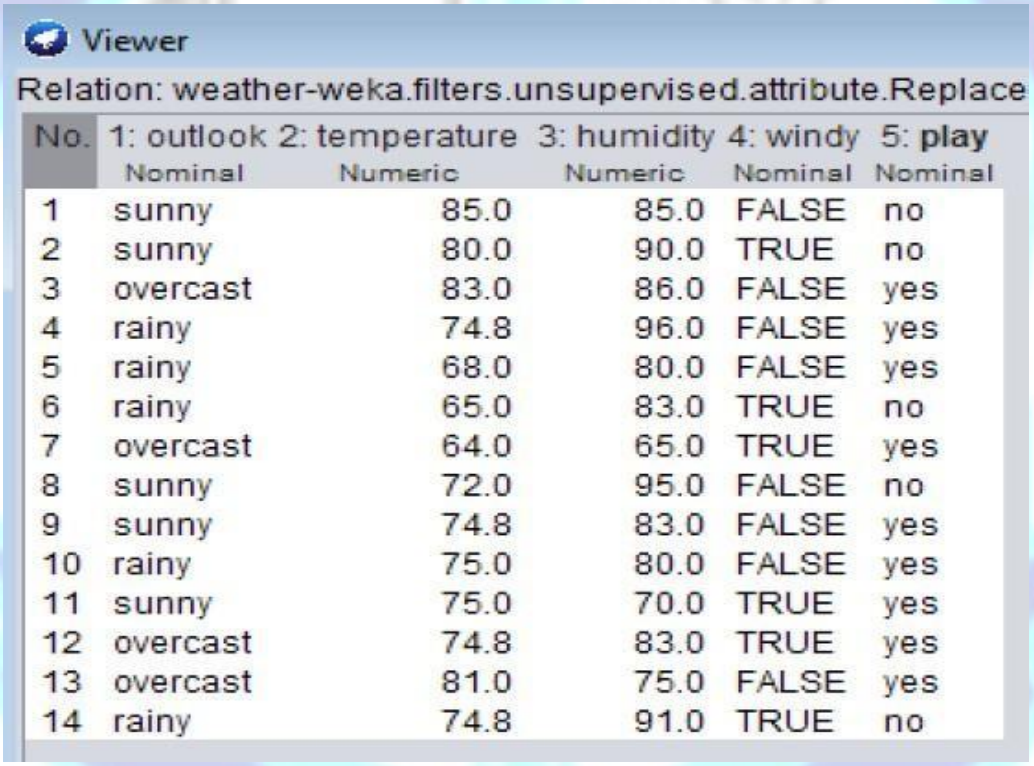
$$\text{Percentage of accuracy}=\frac{\text{Accuracy}}{\sum \text{old value}} \times 100$$

OUTPUT:

Student Details Table: Missing values



Student Details Table: Replace Missing values:



Result:

Thus the data pre-processing by applying filter is performed

EXP.NO: 6	PERFORM VARIOUS OLAP OPERATIONS

OLAP Operations:

Since OLAP servers are based on multidimensional view of data, we will discuss OLAPoperations in multidimensional data.

Here is the list of OLAP operations

- Roll-up (Drill-up)
- Drill-down
- Slice and dice
- Pivot (rotate)

Roll-up (Drill-up):

Roll-up performs aggregation on a data cube in any of the following ways

- By climbing up a concept hierarchy for a dimension
- By dimension reduction
- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level ofcity to the level of country.
- The data is grouped into cities rather than countries.
- When roll-up is performed, one or more dimensions from the data cube are removed.

Drill-down:

Drill-down is the reverse operation of roll-up. It is performed by either of the following ways

- By stepping down a concept hierarchy for a dimension
- By introducing a new dimension.
- Drill-down is performed by stepping down a concept hierarchy for the dimension time.
- Initially the concept hierarchy was "day < month < quarter < year."
- On drilling down, the time dimension is descended from the level of quarter to the levelof month.
- When drill-down is performed, one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

Slice:

The slice operation selects one particular dimension from a given cube and provides a new sub-cube.

Dice:

Dice selects two or more dimensions from a given cube and provides a new sub-cube.

Pivot (rotate):

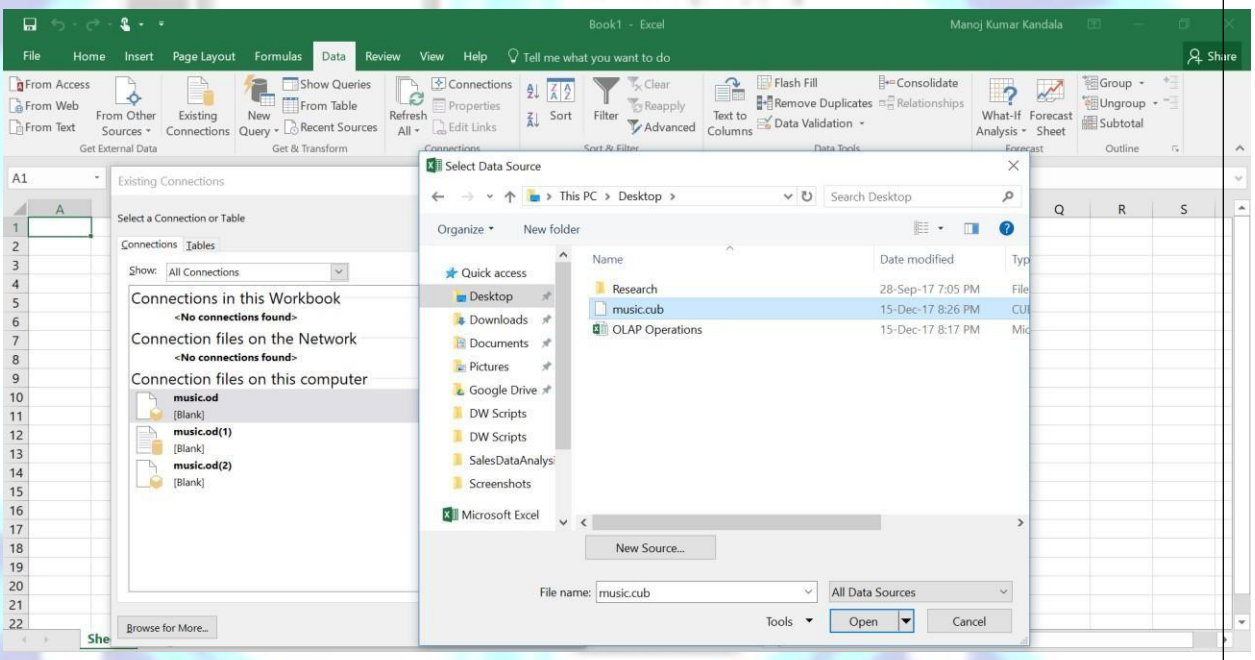
The pivot operation is also known as rotation. It rotates the data axes in view in order toprovide an alternative presentation of data.

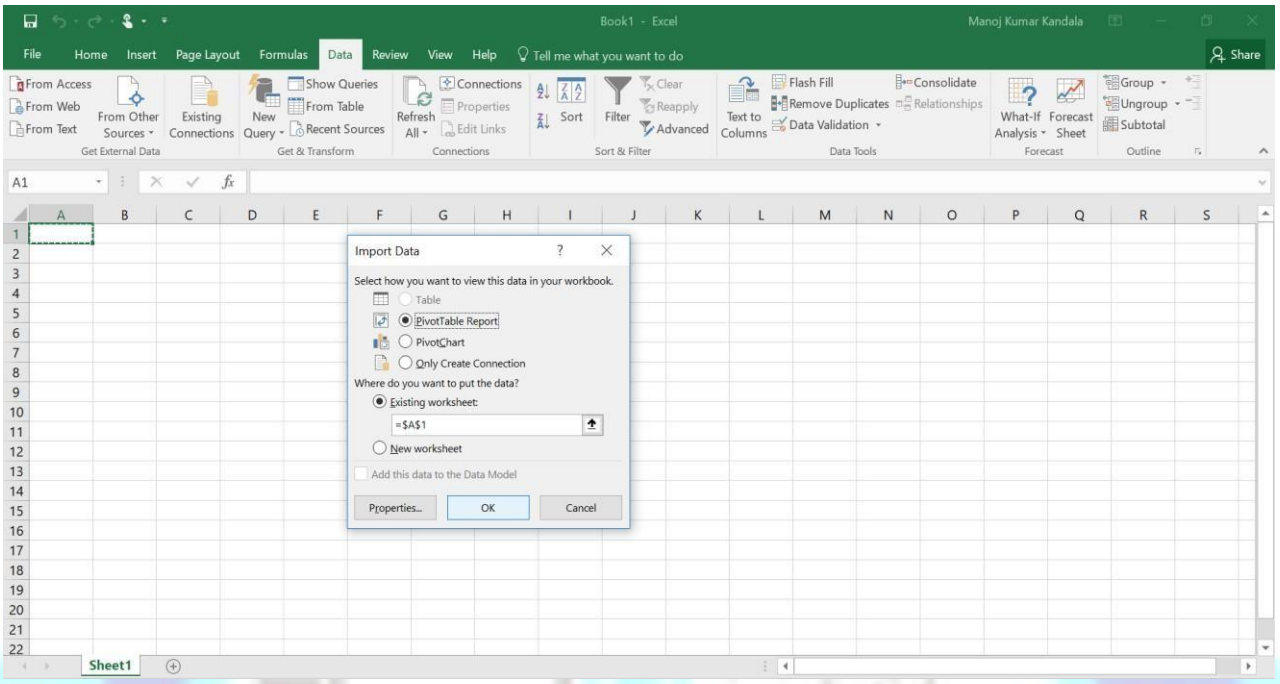
Now, we are practically implementing all these OLAP Operations using **Microsoft**

Excel.

Procedure for OLAP Operations:

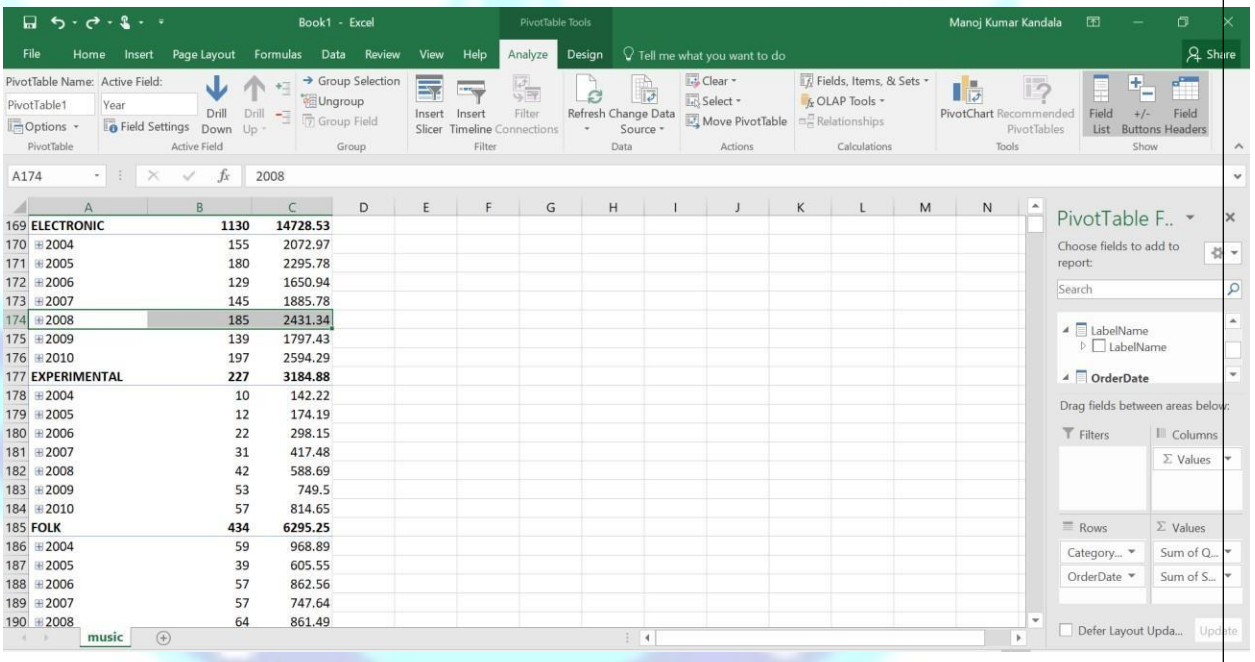
1. Open Microsoft Excel, go to**Data** tab in top & click on **—Existing Connections”**.
2. Existing Connections window will be opened, there **“Browse for more”**option should be clicked for importing **.cub extension file** for performing OLAP Operations. For sample, I took**music.cub** file.





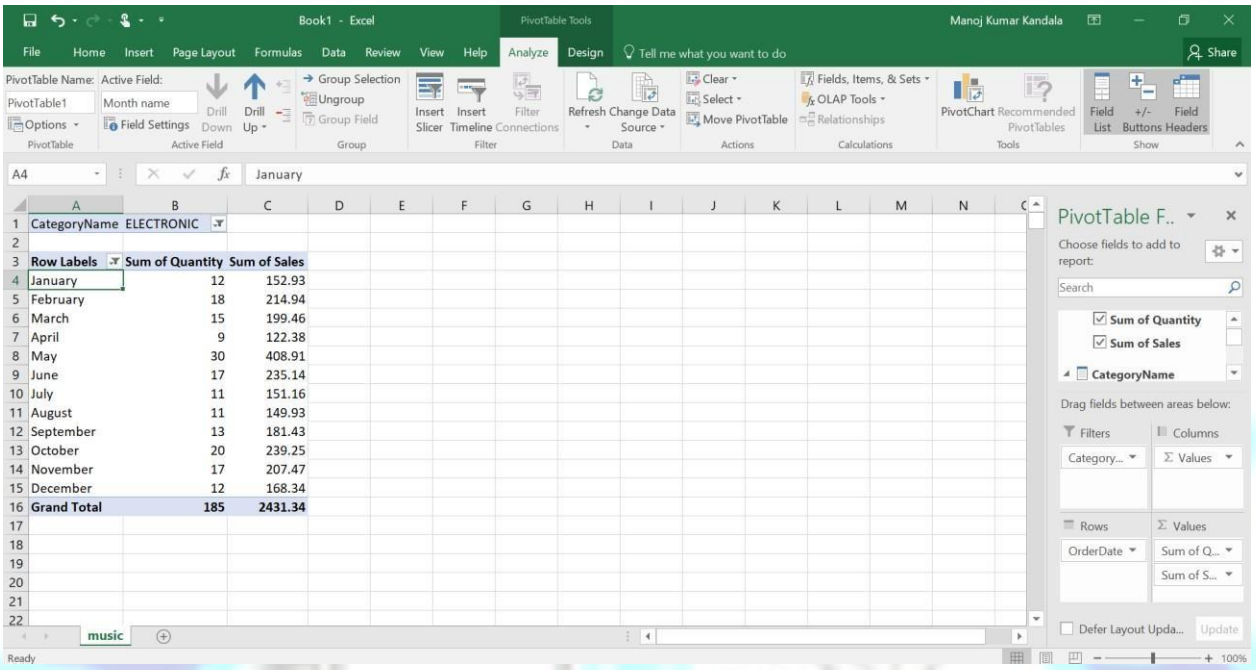
3. As shown in above window, select **—PivotTable Report"** and click **"OK".**

4. We got all the **music.cub** data for analyzing different OLAP Operations.Firstly, we performed **drill-down operation** as shown below.

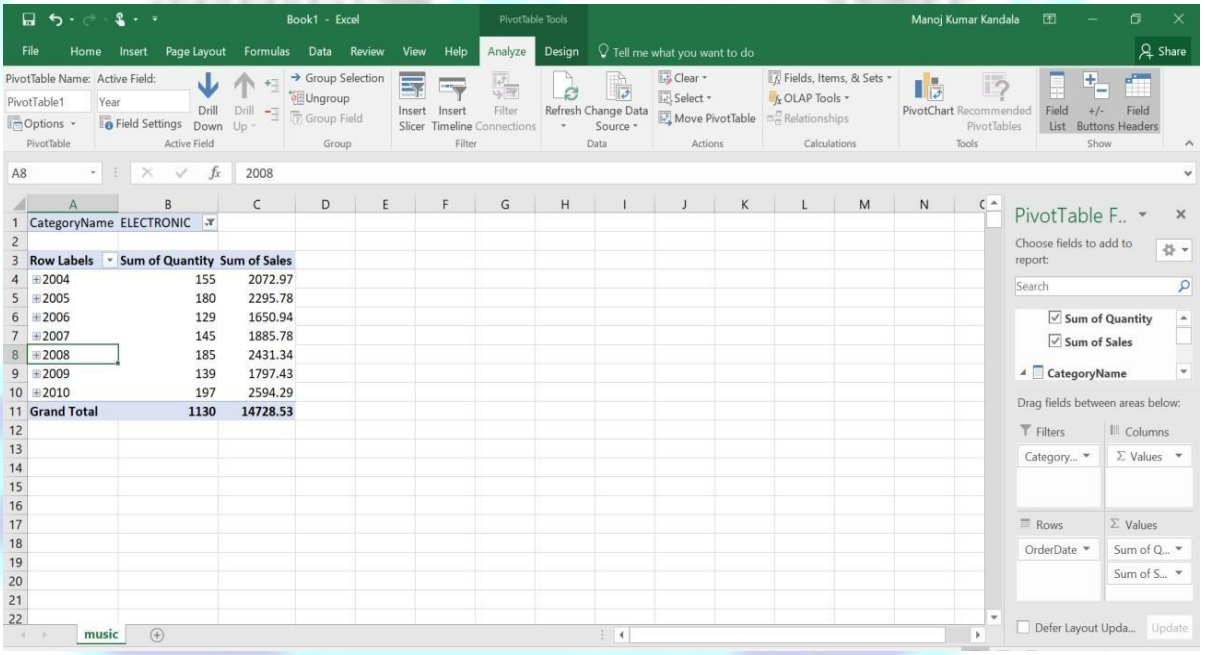


169	**ELECTRONIC**	**1130**	**14728.53**
170	2004	155	2072.97
171	2005	180	2295.78
172	2006	129	1650.94
173	2007	145	1885.78
174	2008	185	2431.34
175	2009	139	1797.43
176	2010	197	2594.29
177	**EXPERIMENTAL**	**227**	**3184.88**
178	2004	10	142.22
179	2005	12	174.19
180	2006	22	298.15
181	2007	31	417.48
182	2008	42	588.69
183	2009	53	749.5
184	2010	57	814.65
185	**FOLK**	**434**	**6295.25**
186	2004	59	968.89
187	2005	39	605.55
188	2006	57	862.56
189	2007	57	747.64
190	2008	64	861.49

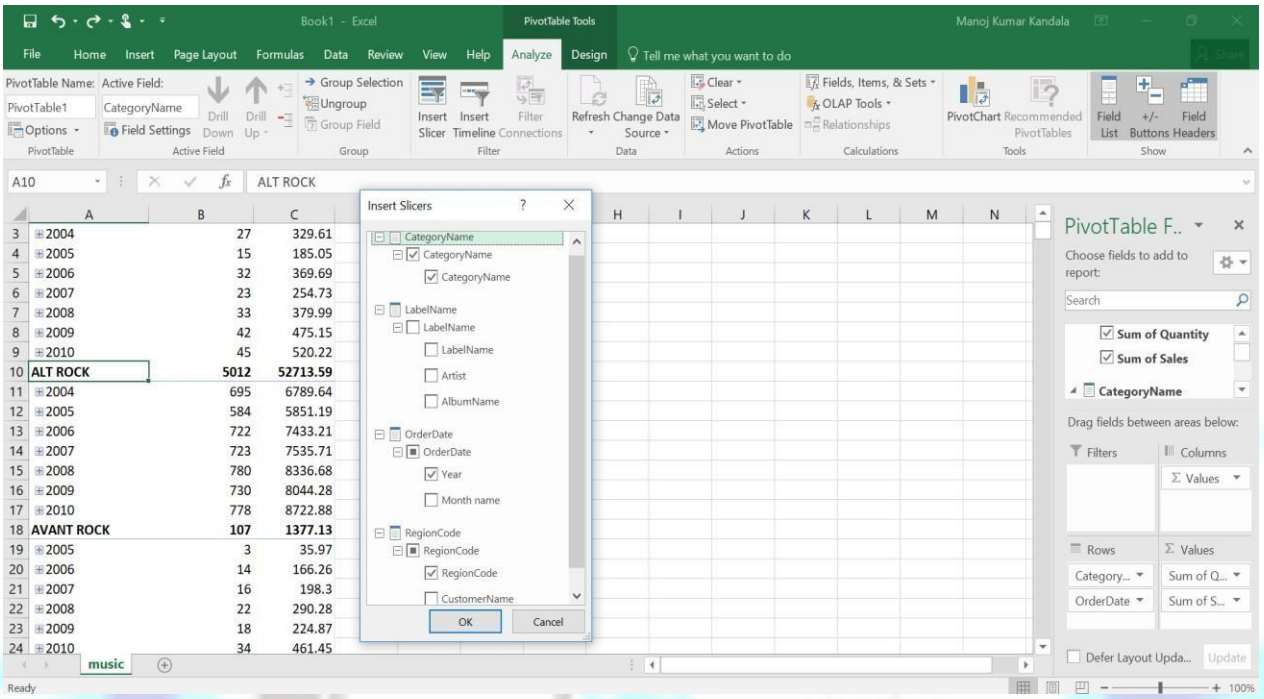
In the above window, we selected year **„2008"** in **„Electronic"** Category, then automatically the Drill-Down option is enabled on top navigation options. We will click on **„Drill-Down"** option, then the below window will be displayed.



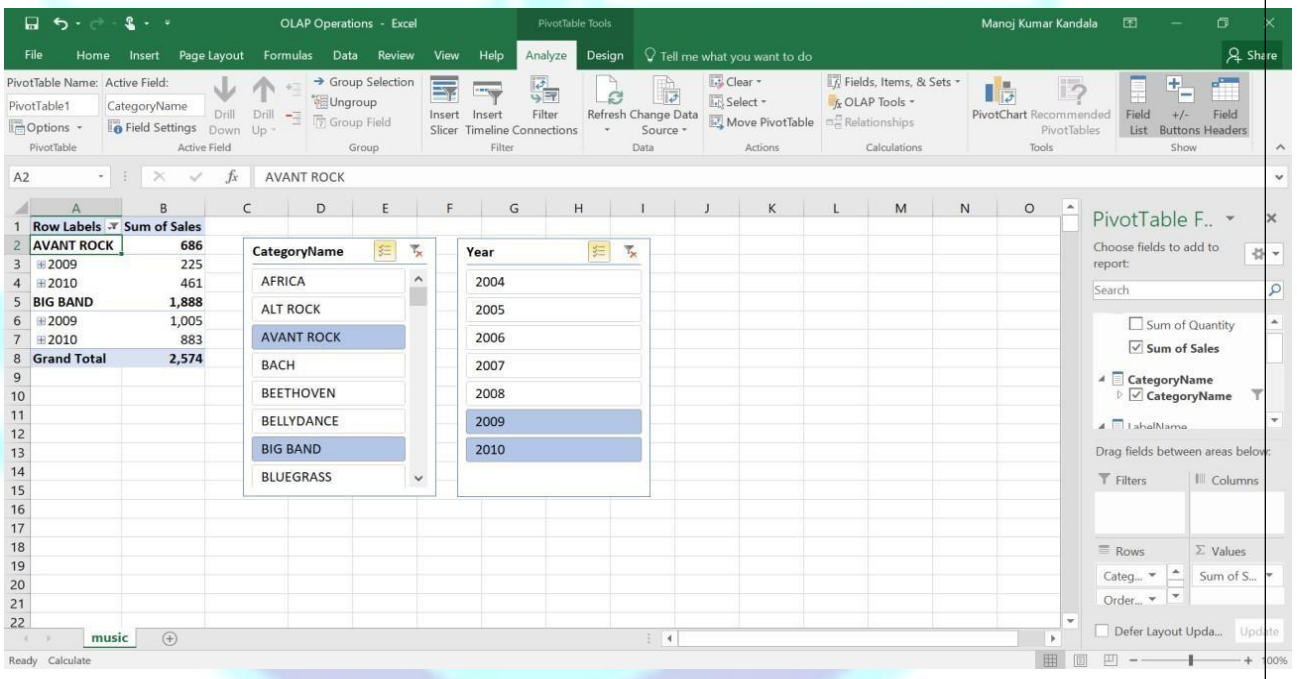
Now we are going to perform **roll-up (drill-up) operation**, in the above window I selected January month then automatically **Drill-up option** is enabled on top. We will click on **Drill-up** option, then the below window will be displayed.



5. Next OLAP operation **Slicing** is performed by inserting slicer as shown in top navigationoptions.

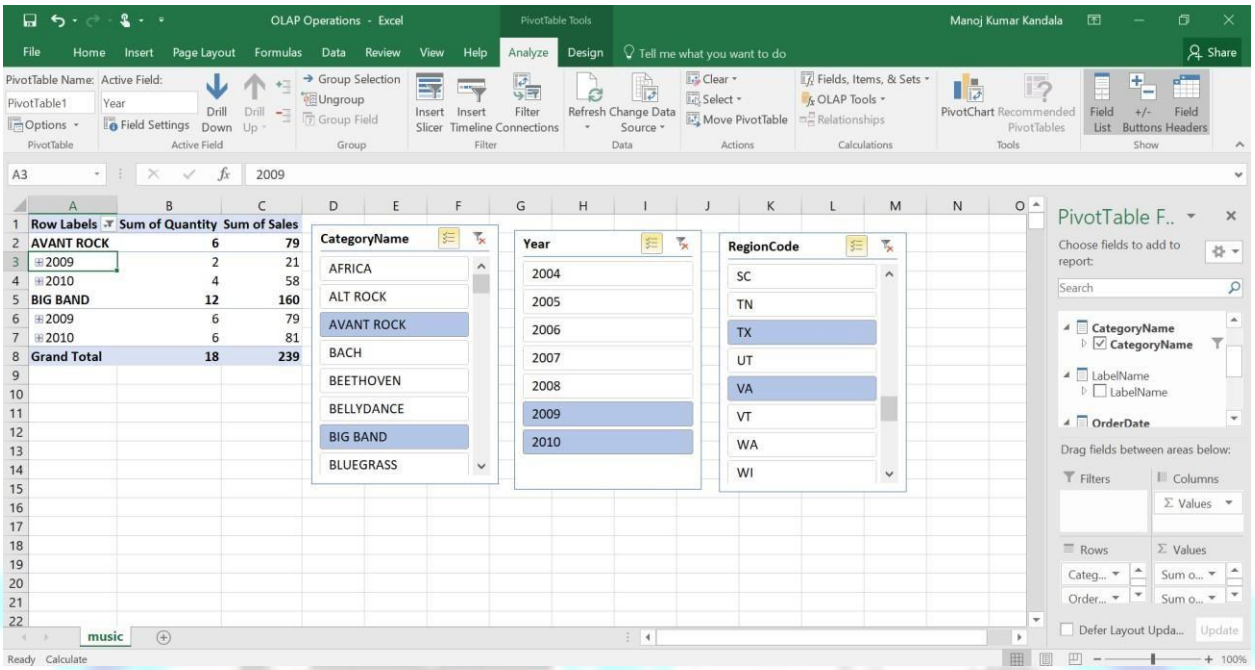


While inserting slicers for slicing operation, we select 2 Dimensions (for e.g. CategoryName & Year) only with one Measure (for e.g. Sum of sales).After inserting a slice& adding a filter (CategoryName: AVANT ROCK & BIG BAND; Year: 2009 & 2010), we will gettable as shown below.

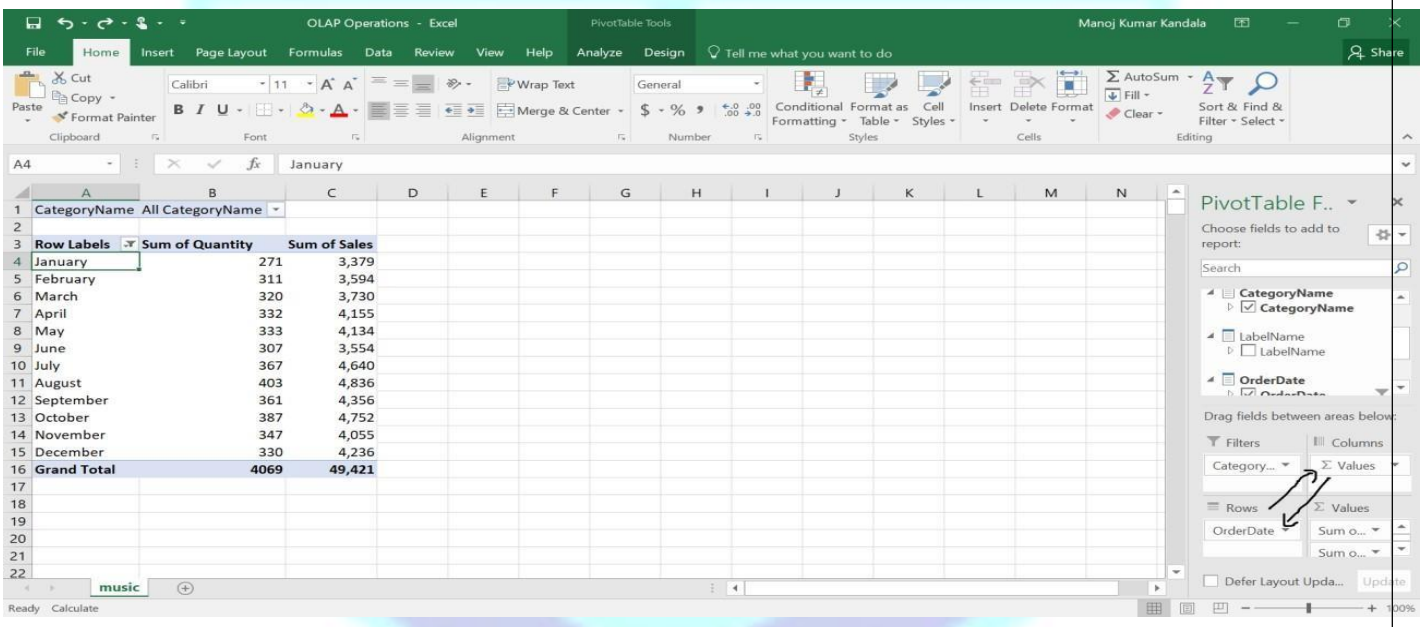


6. **Dicing operation** is similar to Slicing operation. Here we are selecting 3 dimensions (CategoryName, Year, RegionCode)& 2 Measures (Sum of Quantity, Sum of Sales) through **„insert slicer"** option. After that adding

a filter for CategoryName, Year & RegionCode asshown below.



7. Finally, the **Pivot (rotate)** OLAP operation is performed by swapping rows (Order Date-Year)& columns (Values-Sum of Quantity & Sum of Sales) through right side bottom navigation baras shown below.



Result:

Thus the olap operations such as roll up, drill down, slice , dice and pivot are implemented successfully.

EXP.NO: 7	WRITE ETL SCRIPTS AND IMPLEMENT USING DATA WAREHOUSE TOOLS

ETL (Extract-Transform-Load):

ETL comes from Data Warehousing and stands for Extract-Transform-Load. ETL coversa process of how the data are loaded from the source system to the data warehouse.
Currently, the ETL encompasses a cleaning step as a separate step. The sequence is thenExtract-CleanTransform-Load. Let us briefly describe each step of the ETL process.

PROCESS:

EXTRACT

The Extract step covers the data extraction from the source system and makes it accessible for further processing. The main objective of the extract step is to retrieve allthe required data from the source system with as little resources as possible. The extractstep should be designed in a way that it does not negatively affect the source system in terms or performance, response time or any kind of locking.

There are several ways to perform the extract:

- Update notification - if the source system is able to provide a notification that a record has been changed and describe the change, this is the easiest way to get the data.
- Incremental extract - some systems may not be able to provide notification that an update has occurred, but they are able to identify which records have been modified and provide an extract of such records. During further ETL steps, the system needs to identify changes and propagate it down. Note, that by using daily extract, we may not be able to handle deleted records properly.
- Full extract - some systems are not able to identify which data has been changed at all, so a full extract is the only way one can get the data out of the system. The full extract requires keeping a copy of the last extract in the same format in order to be able to identify changes. Full extract handles deletions as well.

When using Incremental or Full extracts, the extract frequency is extremely important.Particularly for full extracts; the data volumes can be in tens of gigabytes.

Clean:

The cleaning step is one of the most important as it ensures the quality of the

data in thedata warehouse. Cleaning should perform basic data unification rules, such as:

- Making identifiersunique(sexcategories Male/Female/Unknown, M/F/null, Man/Woman/Not Available are translated to standard Male/Female/Unknown)
- Convert null values into standardized Not Available/Not Provided value
- Convert phone numbers, ZIP codes to a standardized form
- Validate address fields, convert them into proper naming, e.g. Street/St/St./Str./Str
- Validate address fields against each other (State/Country, City/State, City/ZIP code,City/Street).

Transform:

The transform step applies a set of rules to transform the data from the source to the target. This includes converting any measured data to the same dimension (i.e. conformeddimension) using the same units so that they can later be joined. The transformation step also requires joining data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validationrules.

Load:

During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible. The target of the Load process is often a database. Inorder to make the load process efficient, it is helpful to disable any constraints and indexes before the load and enable them back only after the load completes. The referential integrity needs to be maintained by ETL tool to ensure consistency.

Managing ETL Process:

The ETL process seems quite straight forward. As with every application, there is a possibility that the ETL process fails. This can be caused by missing extracts from one ofthe systems, missing values in one of the reference tables, or simply a connection or power outage. Therefore, it is necessary to design the ETL process keeping fail-recoveryin mind.

Staging:

It should be possible to restart, at least, some of the phases independently from the others.For example, if the transformation step fails, it should not be necessary to restart the Extract step. We can ensure this by implementing proper staging. Staging

means that the data is simply dumped to the location (called the Staging Area) so that it can then be readby the next processing phase. The staging area is also used during ETL process to store intermediate results of processing. This is ok for the ETL process which uses for this purpose. However, tThe staging area should is be accessed by the load ETL process only.It should never be available to anyone else; particularly not to end users as it is not intended for data presentation to the end-user.may contain incomplete or in-the-middle- of-the-processing data.

ETL Tool Implementation:

When you are about to use an ETL tool, there is a fundamental decision to be made: willthe company build its own data transformation tool or will it use an existing tool?

Building your own data transformation tool (usually a set of shell scripts) is the preferredapproach for a small number of data sources which reside in storage of the same type.

The reason for that is the effort to implement the necessary transformation is little due tosimilar data structure and common system architecture. Also, this approach saves licensing cost and there is no need to train the staff in a new tool. This approach, however, is dangerous from the TOC point of view. If the transformations become more sophisticated during the time or there is a need to integrate other systems, the complexityof such an ETL system grows but the manageability drops significantly. Similarly, the implementation of your own tool often resembles re-inventing the wheel.

There are many ready-to-use ETL tools on the market. The main benefit of using off-the-shelf ETL tools is the fact that they are optimized for the ETL process by providing connectors to common data sources like databases, flat files, mainframe systems, xml, etc. They provide a means to implement data transformations easily and consistently across various data sources. This includes filtering, reformatting, sorting, joining, merging, aggregation and other operations ready to use. The tools also support transformation scheduling, version control, monitoring and unified metadata management. Some of the ETL tools are even integrated with BI tools.

Some of the Well Known ETL Tools:

The most well-known commercial tools are Ab Initio, IBM InfoSphereDataStage, Informatica, Oracle Data Integrator, and SAP Data Integrator. There are several open source ETL tools are
OpenRefine, Apatar, CloverETL, Pentaho and Talend.

In these above tools, we are going to use **OpenRefine 2.8 ETL tool**to different sampledatasets forextracting, data cleaning, transforming & loading.



Result:

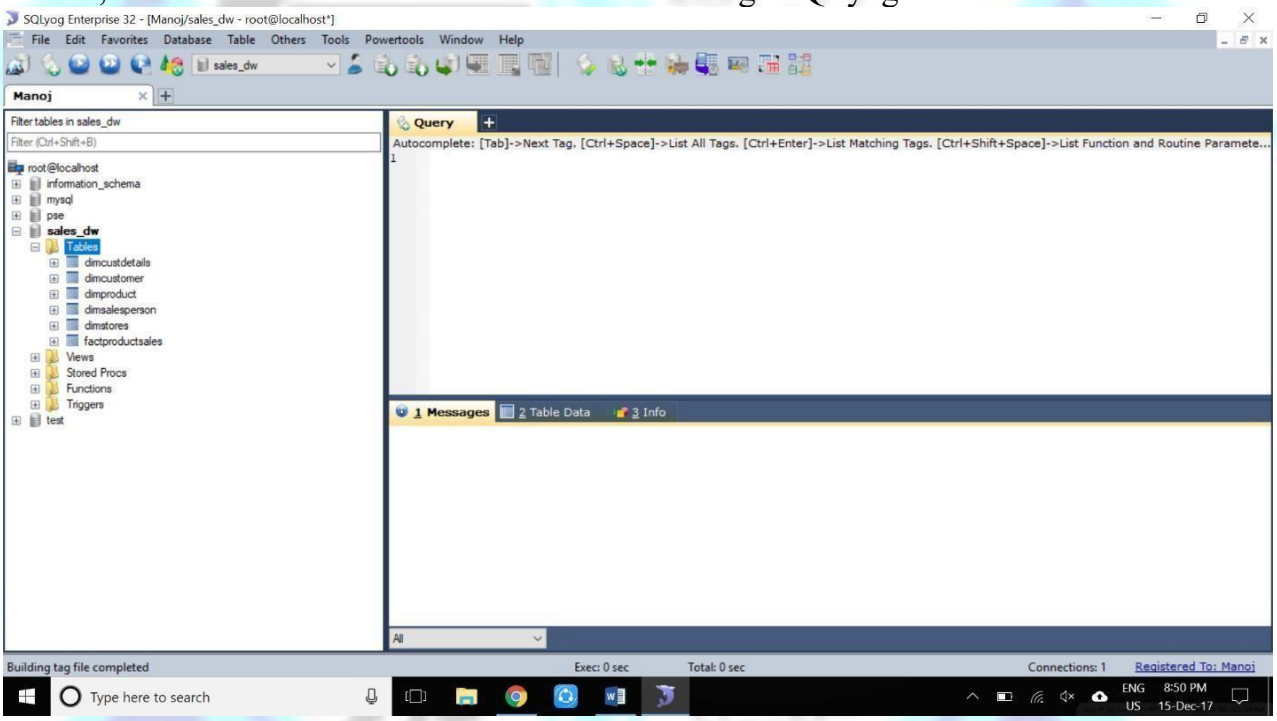
Thus the ELT script is implemented successfully.

EXP.NO: 8	DESIGN MULTI-DIMENSIONAL DATA MODELS

Multi-Dimensional model was developed for implementing data warehouses & it provides both a mechanism to store data and a way for business analysis. The primary components of dimensional model are dimensions & facts. There are different of types ofmulti-dimensional data models. They are:

- Star Schema Model
- SnowFlake Schema Model
- Fact Constellation Model.

Now, we are going to design these multi-dimensional models for the Marketing enterprise. First, we need to built the tables in a database through SQLyog as shown below.



In the above window, left side navigation bar consists of a database named as —sales_dw‖ in which there are six different tables (dimcustdetails, dimcustomer, dimproduct, dimsalesperson, dimstores, factproductsales) has been created.

After creating tables in database, here we are going to use a tool called as **"Microsoft Visual Studio 2012 for Business Intelligence"** for building multi-dimensional models.

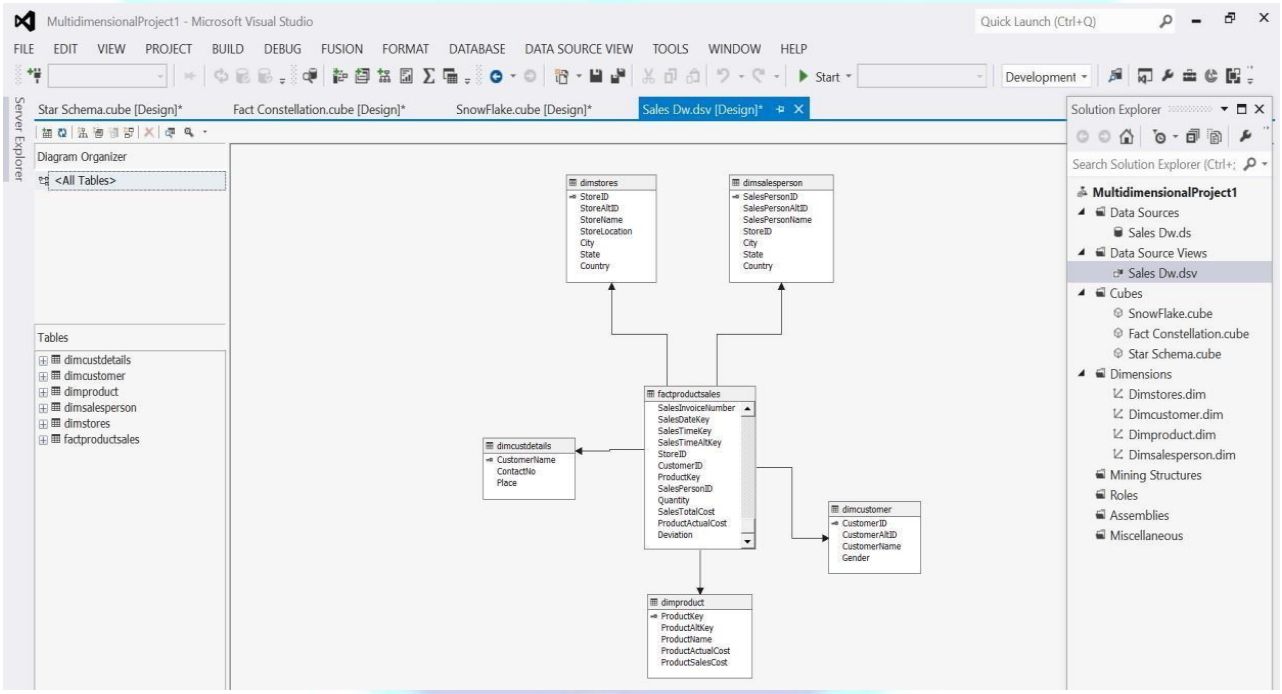
Through Data Sources, we can connect to our MySQL database named as **“sales_dw”.** Then, automatically all the tables in that database will be retrieved to this tool for creating multidimensional models.

By data source views & cubes, we can see our retrieved tables in multi-dimensional models. We need to add dimensions also through dimensions option. In general, Multidimensional models consists of dimension tables & fact tables.

Star Schema Model:

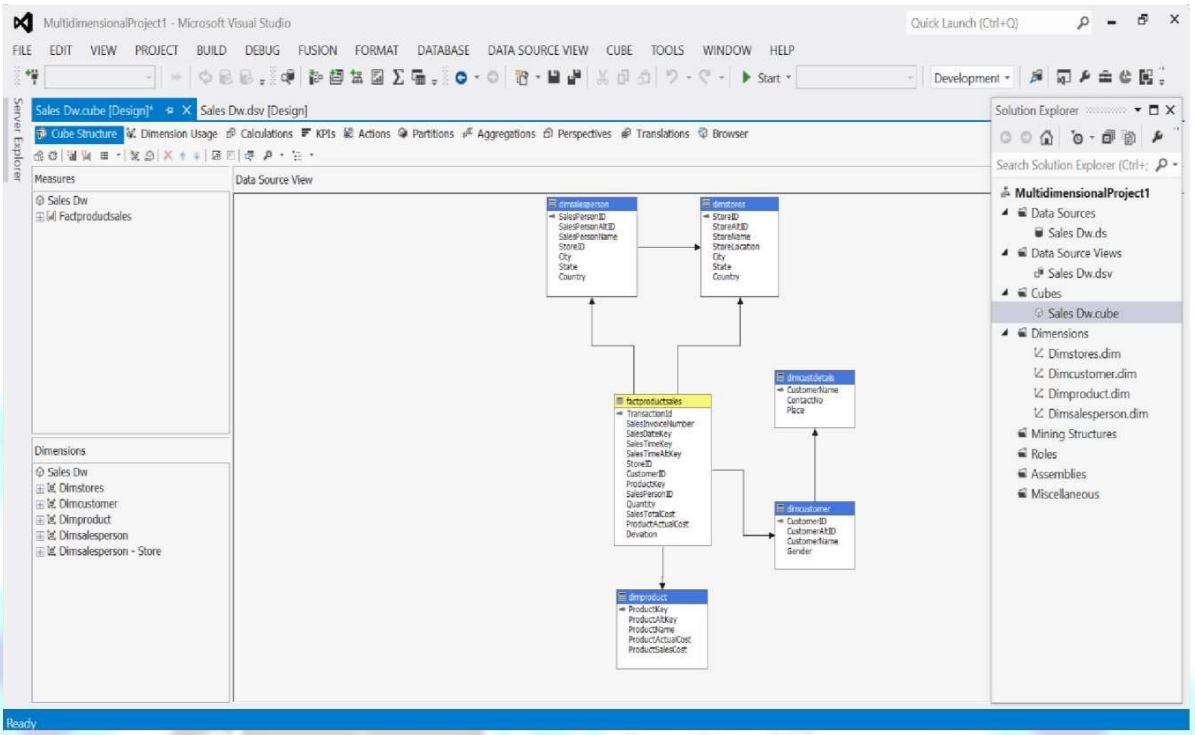
A Star schema model is a join between a fact table and a no. of dimension tables. Each dimensional table are joined to the fact table using primary key to foreign key join but dimensional tables are not joined to each other. It is the simplest style of dataware houseschema.

Star schema is a entity relationship diagram of this schema resembles a star with point radiating from central table as we seen in the below implemented window in visualstudio.



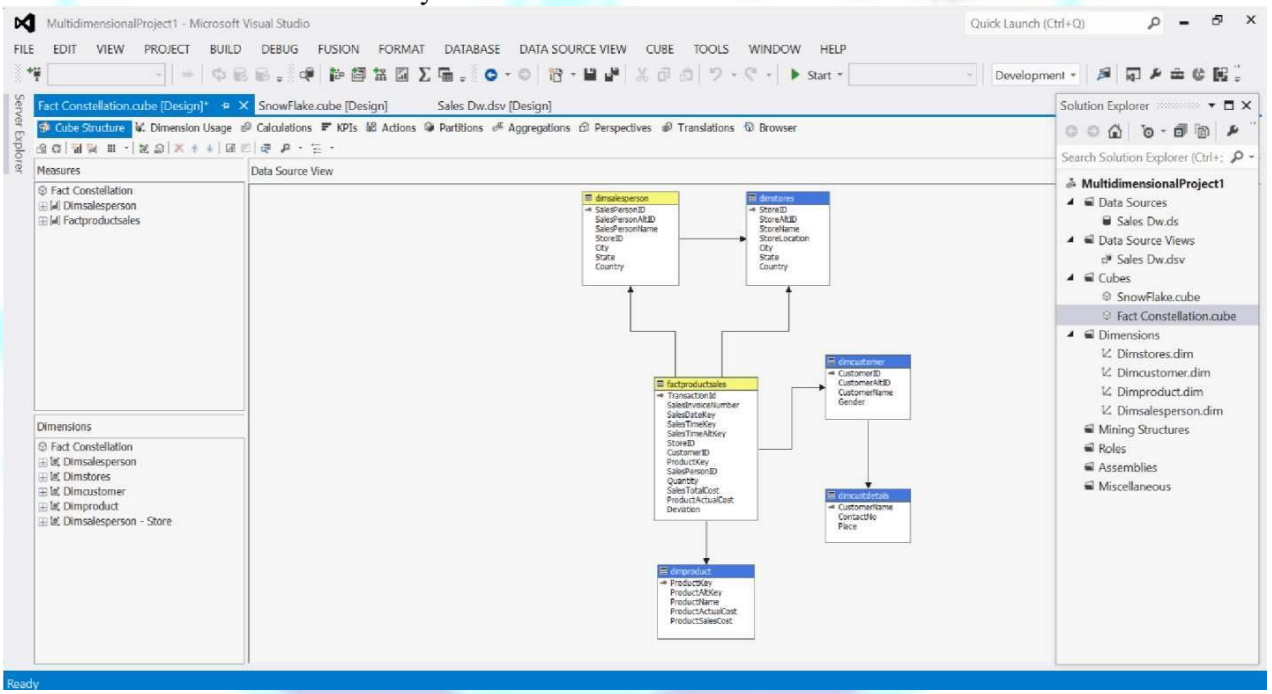
Snow Flake Schema:

It is slightly different from star schema in which dimensional tables from a star schema are organized into a hierarchy by normalizing them



Fact Constellation Schema:

Fact Constellation is a set of fact tables that share some dimension tables. In this schema there are two or more fact tables. We developed fact constellation in visual studio as shown below. Fact tables are labelled in yellow color.



Result:

Thus the multidimensional models are created successfully.

EXP.NO: 9	EXPLORE WEKA DATA MINING/MACHINE LEARNING TOOLKIT

(i). Downloading and/or installation of WEKA data mining toolkit

Procedure:

1. Go to the Weka website, http://www.cs.waikato.ac.nz/ml/weka/, and download the software. On the left-hand side, click on the link that says download.

2. Select the appropriate link corresponding to the version of the software based on youroperating system and whether or not you already have Java VM running on your machine (if you don‘t know what Java VM is, then you probably don‘t).

3. The link will forward you to a site where you can download the software from a mirrorsite. Save the self-extracting executable to disk and then double click on it to install Weka.Answer yes or next to the questions during the installation.

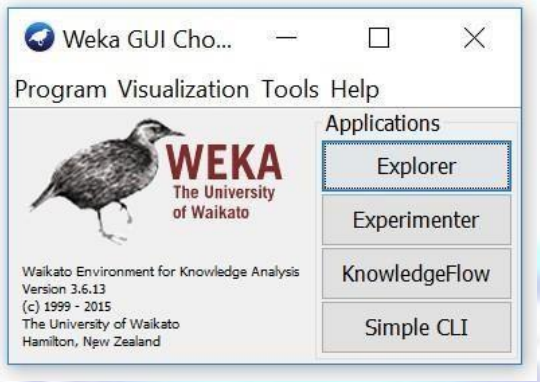
4. Click yes to accept the Java agreement if necessary. After you install the program Weka should appear on your start menu under Programs (if you are using Windows).

5. Running Weka from the start menu select Programs, then Weka.You will see the Weka GUI Chooser. Select Explorer. The Weka Explorer will then launch.

(ii). Understand the features of WEKA toolkit such as Explorer, Knowledge Flowinterface, Experimenter, command-line interface.

The Weka GUI Chooser (class weka.gui.GUIChooser) provides a starting pointfor launching Weka‘s main GUI applications and supporting tools. If one prefersa MDI (―multiple document interface‖) appearance, then this is provided by analternative launcher called ―Main‖ (class weka.gui.Main).

The GUI Chooser consists of four buttons—one for each of the four majorWeka applications— and four menus.



The buttons can be used to start the following applications:

Explorer- An environment for exploring data with WEKA

a) Click on —explorer‖ button to bring up the explorer window.
b) Make sure the —preprocess‖ tab is highlighted.
c) Open a new file by clicking on —Open New file‖ and choosing a file with —.arff‖extension from the —Data‖ directory.
d) Attributes appear in the window below.
e) Click on the attributes to see the visualization on the right.
f) Click —visualize all‖ to see them all

Experimenter- An environment for performing experiments and conducting statisticaltests between learning schemes.

a) Experimenter is for comparing results.
b) Under the —set up‖ tab click —New‖.
c) Click on —Add New‖ under —Data‖ frame. Choose a couple of arff format files from —Data‖ directory one at a time.
d) Click on —Add New‖ under —Algorithm‖ frame. Choose several algorithms, one ata time by clicking —OK‖ in the window and —Add New‖.
e) Under the —Run‖ tab click —Start‖.
f) Wait for WEKA to finish.
g) Under —Analyses‖ tab click on —Experiment‖ to see results.

Knowledge Flow- This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantageis that it supports incrementallearning.

SimpleCLI - Provides a simple command-line interface that allows directexecution

ofWEKA commands for operating systems that do not provide their own command line interface.

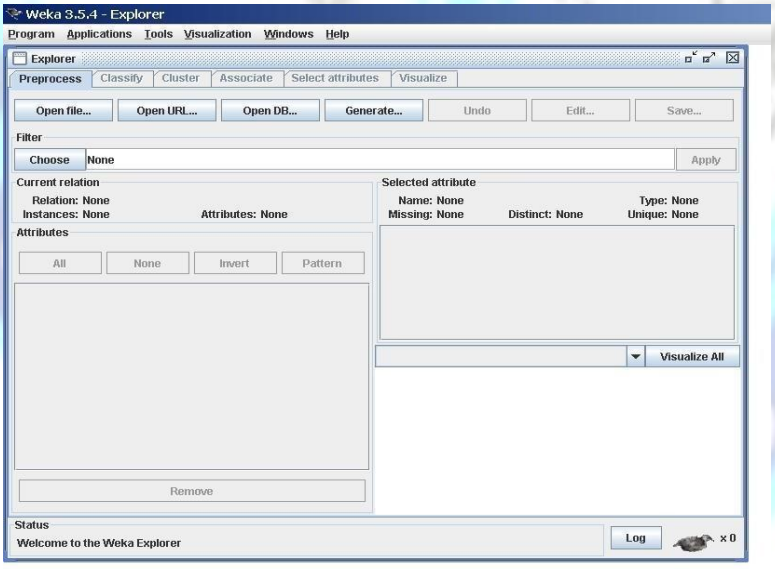
(iii). Navigate the options available in the WEKA (ex. Select attributes panel, Preprocess panel, classify panel, Cluster panel, Associate panel and Visualize panel)

When the Explorer is first started only the first tab is active; the others are greyed out. This is because it is necessary to open (and potentially pre-process) a data set before starting to explore the data.
The tabs are as follows:

1. Preprocess. Choose and modify the data being acted on.
2. Classify. Train and test learning schemes that classify or perform regression.
3. Cluster. Learn clusters for the data.
4. Associate. Learn association rules for the data.
5. Select attributes. Select the most relevant attributes in the data.
6. Visualize. View an interactive 2D plot of the data.

Once the tabs are active, clicking on them flicks between different screens, on which the respective actions can be performed. The bottom area of the window (including the statusbox, the log button, and the Weka bird) stays visible regardless of which section you are in.

1. Preprocessing



Loading Data:

The first four buttons at the top of the preprocess section enable you to load data into

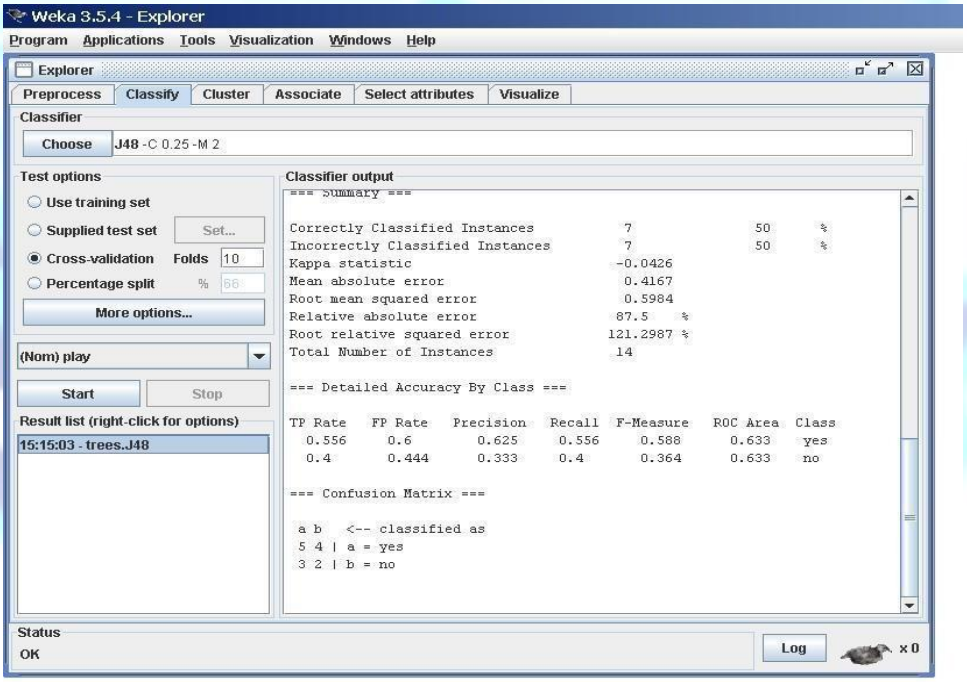
WEKA:

1. Open file..... Brings up a dialog box allowing you to browse for the datafile on the local
   file system.
2. Open URL..... Asks for a Uniform Resource Locator address for wherethe data is stored.
3. Open DB..... Reads data from a database. (Note that to make this workyou might have to
   edit the file in weka/experiment/DatabaseUtils.props.)
4. Generate. .....Enables you to generate artificial data from a variety of DataGenerators.

Using the Open file.....button you can read files in a variety of formats:

WEKA's ARFF format, CSV format, C4.5 format, or serialized Instances format. ARFF
files typically have a .arff extension, CSV files a .csv extension,C4.5 files a .data and.names extension, and serialized Instances objects a .bsiextension.

2. Classification:



Selecting a Classifier

At the top of the classify section is the Classifier box. This box has a text fieldthat gives the name of the currently selected classifier, and its options. Clickingon the text

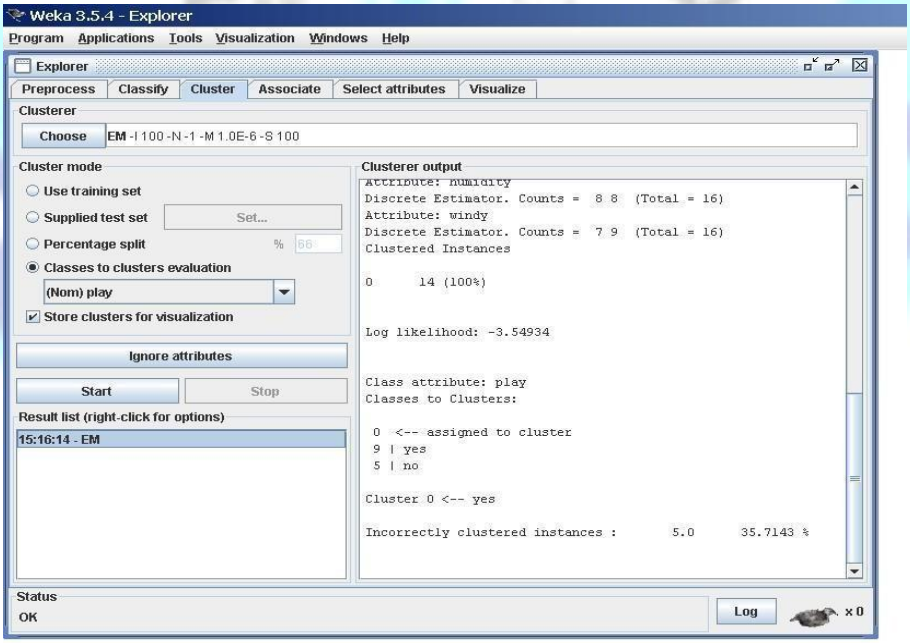
box withthe left mouse button brings up a GenericObjectEditordialog box, just the same as for filters, that you can use to configure the optionsof the current classifier. With a right click(or Alt+Shift+left click) you canonce again copy the setup string to the clipboard or display the properties in aGenericObjectEditor dialog box. The Choose button allows youto choose on4eof the classifiers that are available in WEKA.

Test Options

The result of applying the chosen classifier will be tested according to the optionsthat areset by clicking in the Test options box. There are four test modes:

1. **Use training set:** The classifier is evaluated on how well it predicts theclass of theinstances it was trained on.
2. **Supplied test set:** The classifier is evaluated on how well it predicts theclass of a setof instances loaded from a file. Clicking the Set... buttonbrings up a dialog allowing you to choose the file to test on.
3. **Cross-validation:** The classifier is evaluated by cross-validation, usingthe numberof folds that are entered in the Folds text field.
4. **Percentage split:** The classifier is evaluated on how well it predicts acertain percentage of the data which is held out for testing. The amountof data held out dependson the value entered in the % field.

3. Clustering:



Cluster Modes:

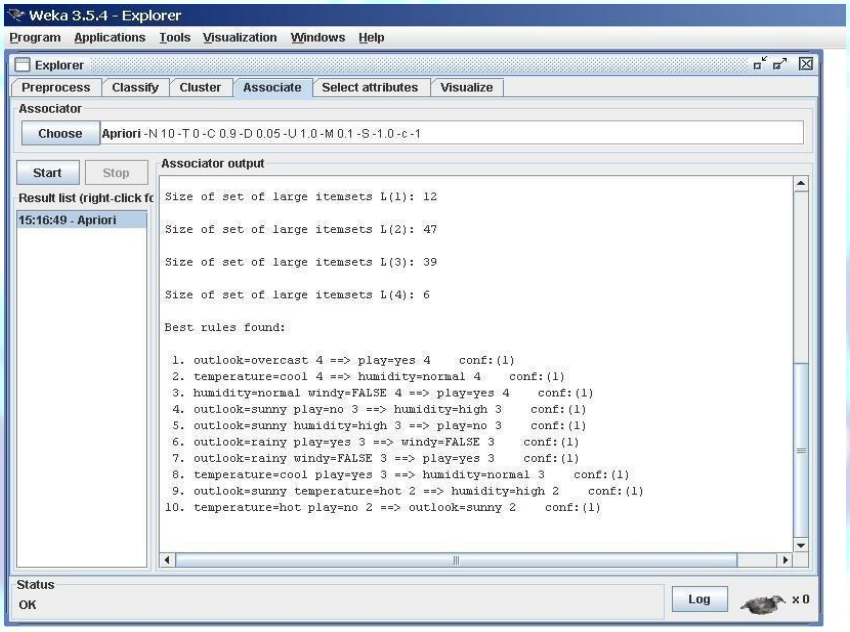
The Cluster mode box is used to choose what to cluster and how to evaluate the results. The first three options are the same as for classification: Use training set, Supplied test setand Percentage split.

Test Options

The result of applying the chosen classifier will be tested according to the optionsthat areset by clicking in the Test options box. There are four test modes:

1. **Use training set:** The classifier is evaluated on how well it predicts the class of theinstances it was trained on.
2. **Supplied test set:** The classifier is evaluated on how well it predicts theclass of a setof instances loaded from a file. Clicking the Set... buttonbrings up a dialog allowing you to choose the file to test on.
3. **Cross-validation:** The classifier is evaluated by cross-validation, using the numberof folds that are entered in the Folds text field.
4. **Percentage split:** The classifier is evaluated on how well it predicts acertain percentage of the data which is held out for testing. The amountof data held out dependson the value entered in the % field.

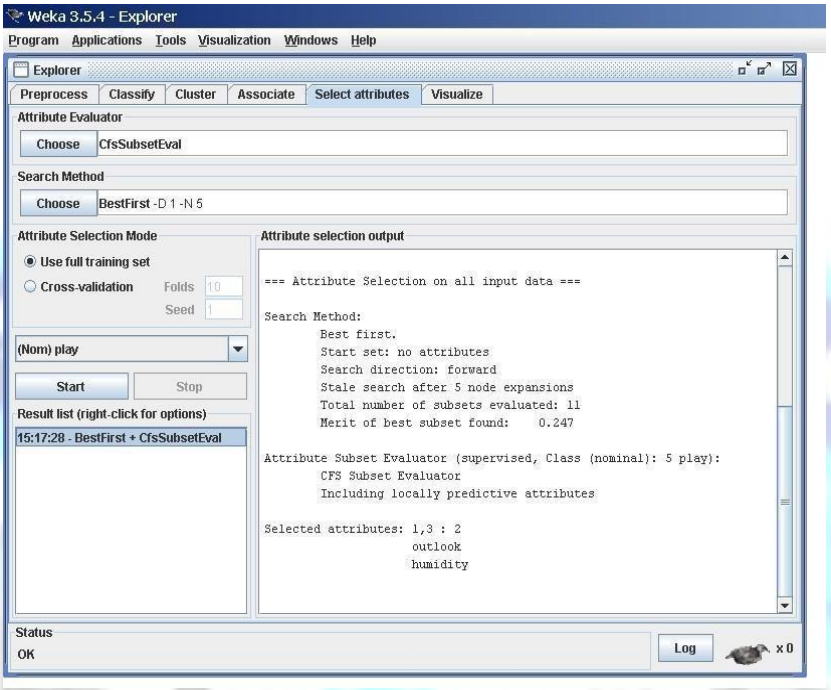
4. Associating:



Setting Up

This panel contains schemes for learning association rules, and the learners are chosenand configured in the same way as the clusterers, filters, and classifiers in the other panels.

5. Selecting Attributes:

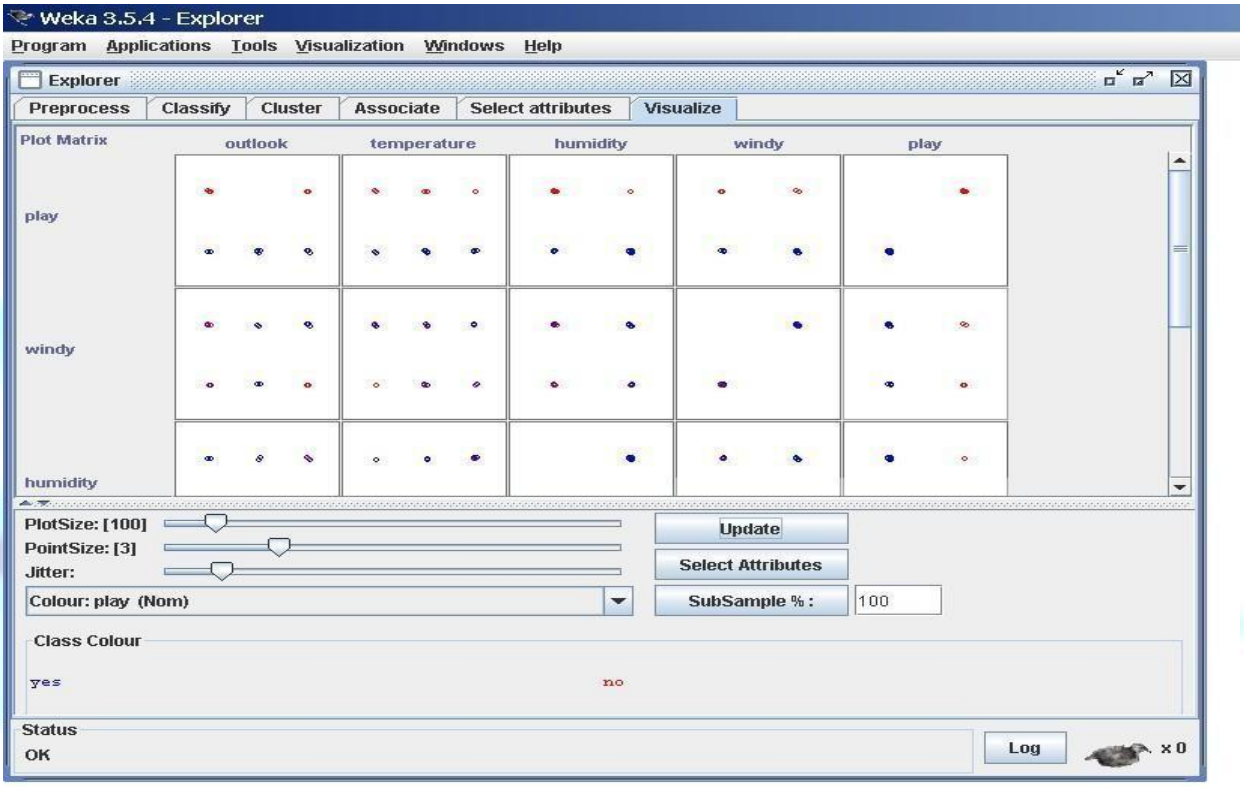


Searching and Evaluating

Attribute selection involves searching through all possible combinations of attributes inthe data to find which subset of attributes works best for prediction. To do this, two objects must be set up: an attribute evaluator and a searchmethod. The evaluator determines what method is used to assign a worth toeach subset of attributes. The searchmethod determines what style of searchis performed.

6. Visualizing:

WEKA‘s visualization section allows you to visualize 2D plots of the currentrelation.



Result :

Thus the tools are explored and analyzed successfully.

EXP.NO: 10	DESIGN OF FACT AND DIMENSION TABLES

Aim:

To design fact and dimension tables.

Fact Table :

A fact table is used in the dimensional model in data warehouse design. A fact table is found at the center of a star schema or snowflake schema surrounded by dimension tables.A fact table consists of facts of a particular business process e.g., sales revenue by month by product. Facts are also known as measurements or metrics. A fact table record captures a measurement or a metric.

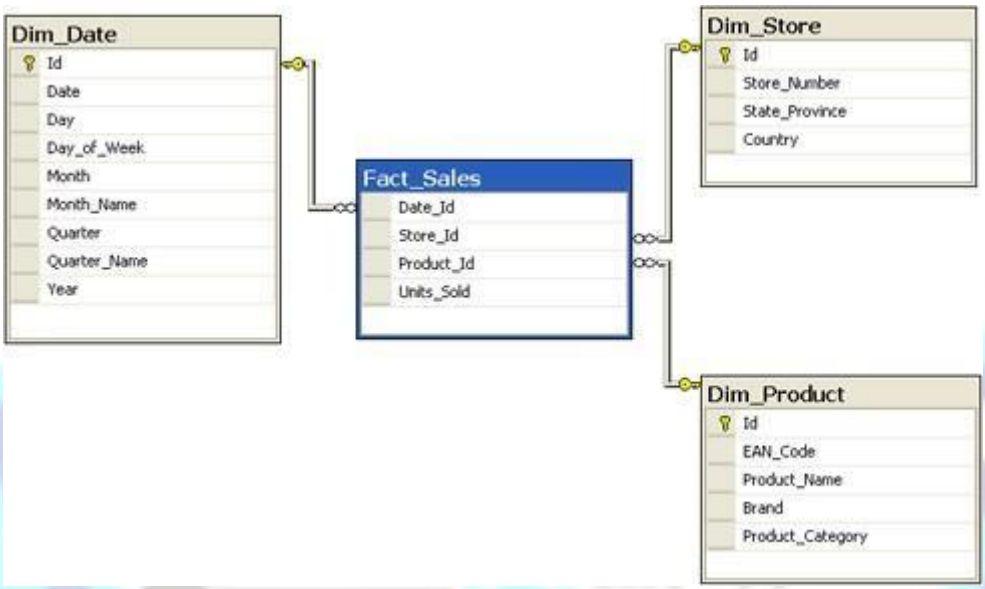
Designing fact table steps

Here is overview of four steps to designing a fact table:

1. **Choosing business process to model** – The first step is to decide what business process to model by gathering and understanding business needs and available data
2. **Declare the grain** – by declaring a grain means describing exactly what a fact table record represents
3. **Choose the dimensions** – once grain of fact table is stated clearly, it is time to determine dimensions forthe fact table.
4. **Identify facts** – identify carefully which facts will appear in the fact table.

Fact table FACT_SALES that has a grain which gives us a number of units sold by date, by store and by product.

All other tables such as DIM_DATE, DIM_STORE and DIM_PRODUCT are dimensions tables. This schema isknown as the star schema.



Result: Thus design fact and dimension tables are created.

EXP.NO: 11	NORMALIZE EMPLOYEE TABLE DATA USING KNOWLEDGE FLOW

Aim:

Normalize Employee Table data using Knowledge Flow.

Description:

The knowledge flow provides an alternative way to the explorer as a graphical front end to WEKA'salgorithm. Knowledge flow is a working progress. So, some of the functionality from explorer is not yet available. So, on the other hand there are the things that can be done in knowledge flow, but not in explorer. Knowledge flow presents a dataflow interface to WEKA. The user can select WEKA components from a toolbar placed them on a layout campus and connect them together in order to form a knowledge flow for processing and analyzing the data.

Creation of Employee Table:

Procedure:

1) Open Start Programs Accessories Notepad
2) Type the following training data set with the help of Notepad for Employee Table.

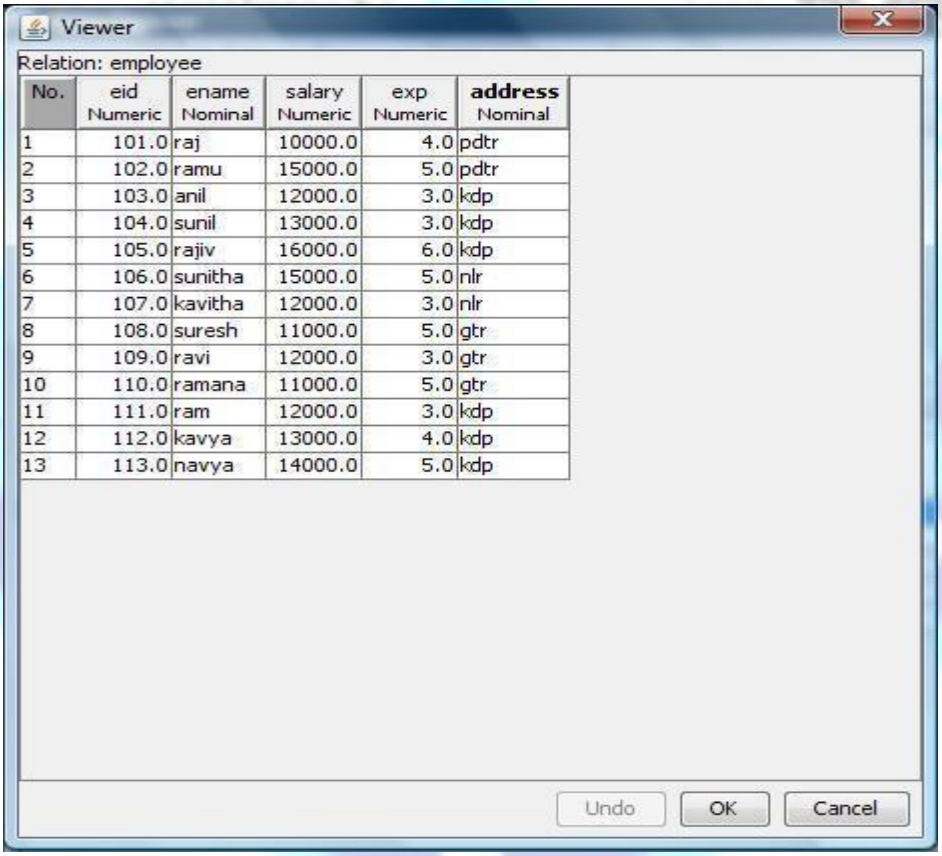
```
@relation employee
@attribute eid numeric
@attribute ename
{raj,ramu,anil,sunil,rajiv,sunitha,kavitha,suresh,ravi,ramana,ram,kavya,navya}@attribute salary numeric
@attribute exp numeric
@attribute address
{pdtr,kdp,nlr,gtr} @data
101,raj,10000,4,pdtr
102,ramu,15000,5,pdtr
103,anil,12000,3,kdp
104,sunil,13000,3,kdp
105,rajiv,16000,6,kdp
106,sunitha,15000,5,nlr
```

107,kavitha,12000,3,nlr
108,suresh,11000,5,gtr 109,ravi,12000,3,gtr
110,ramana,11000,5,gtr
111,ram,12000,3,kdp
112,kavya,13000,4,kdp
113,navya,14000,5,kdp

3) After that the file is saved with **.arff** file format.
4) Minimize the arff file and then open Start Programs weka-3-4.
5) Click on **weka-3-4**, then Weka dialog box is displayed on the screen.
6) In that dialog box there are four modes, click on **explorer**.
7) Explorer shows many options. In that click on **'open file'** and select the arff file
8) Click on **edit button** which shows employee table on weka.

Output:

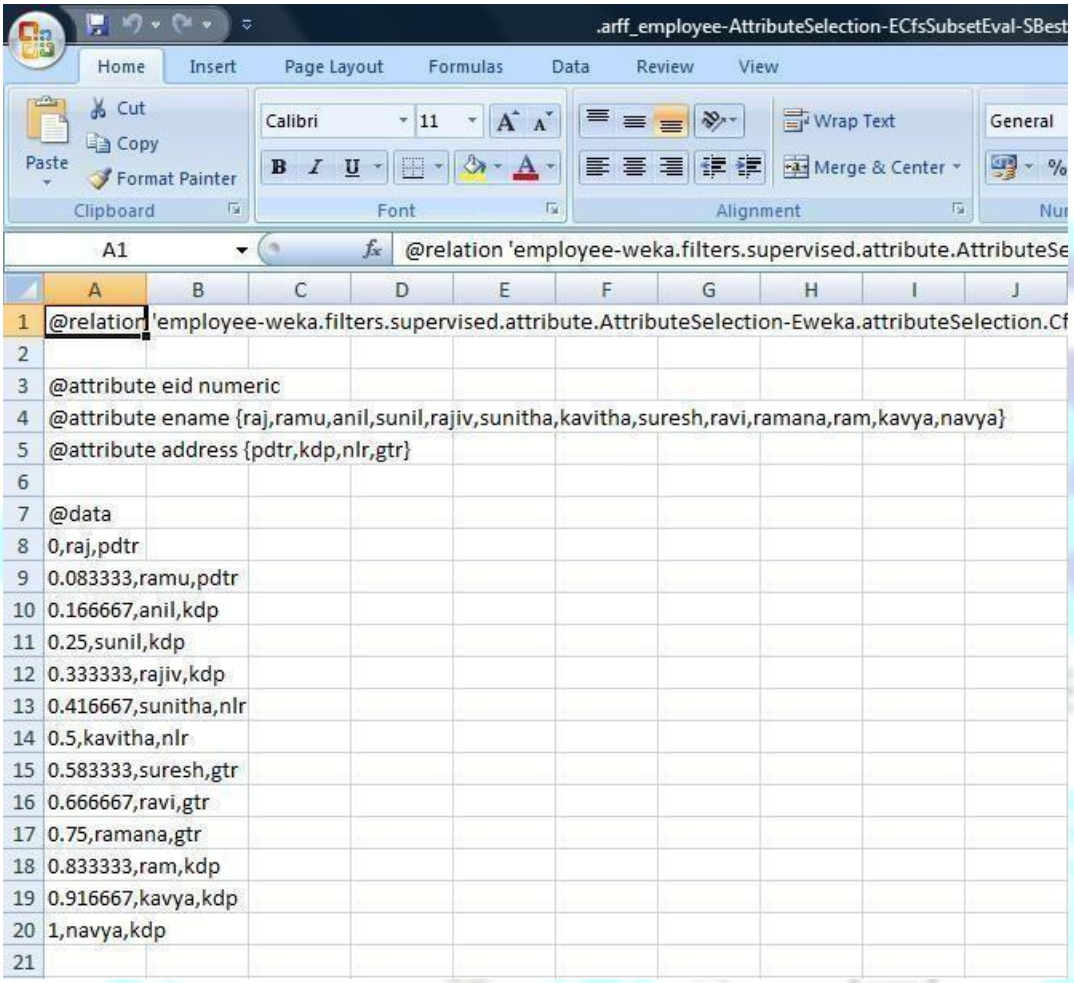
Training Data Set Employee Table

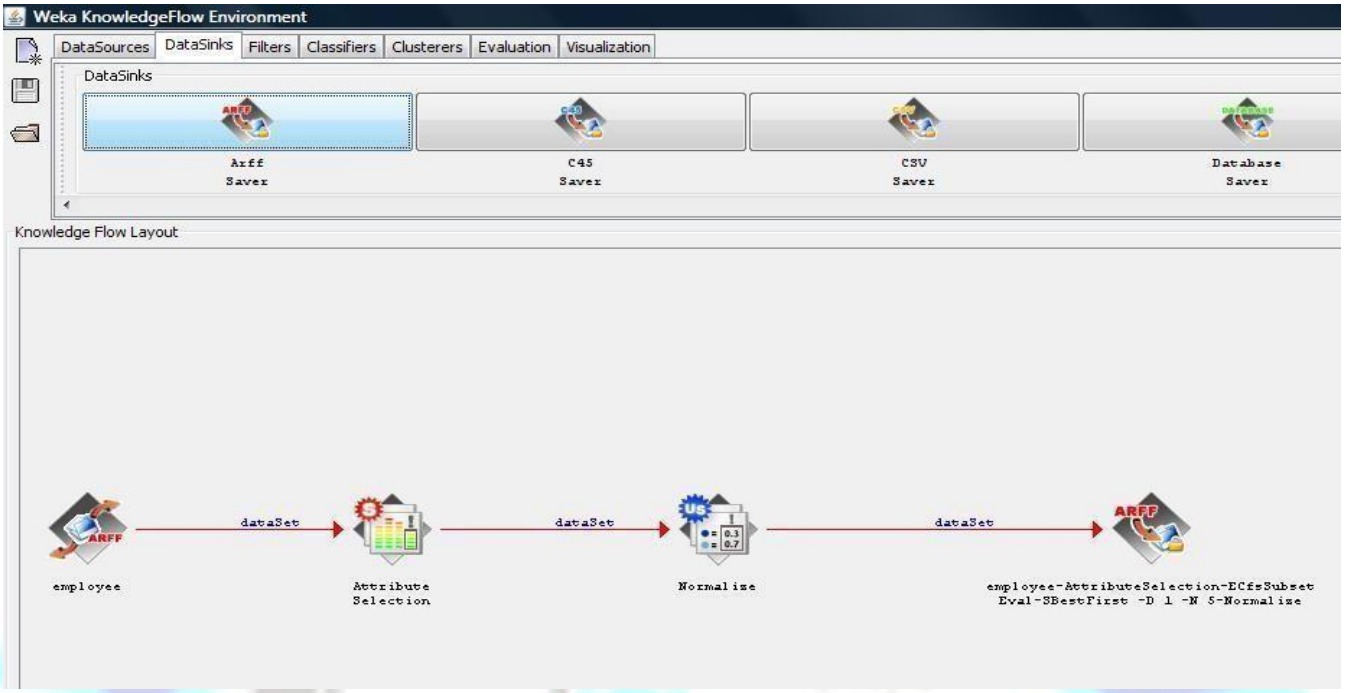


Procedure for Knowledge Flow:

1) Open Start □Programs □Weka-3-4 □Weka-3-4
2) Open the **Knowledge Flow**.
3) Select the **Data Source component** and **add Arff Loader** into the **knowledge layout canvas**.
4) Select the **Filters component** and **add Attribute Selection** and N**ormalize** into the knowledge layout canvas.
5) Select the **Data Sinks** component and **add Arff Saver** into the knowledge layout canvas.
6) Right click on **Arff Loader** and select **Configure option** then the new window will be opened and select **Employee.arff**
7) Right click on **Arff Loader** and select **Dataset option** then establish a link between **Arff Loader** and**Attribute Selection**.
8) Right click on **Attribute Selection** and select **Dataset option** then establish a link between **Attribute Selection** and **Normalize**.
9) Right click on **Attribute Selection** and select **Configure option** and choose the best attribute for Employeedata.
10) Right click on **Normalize** and select **Dataset option** then establish a link between **Normalize** and **Arff Saver**.
11) Right click on **Arff Saver** and select **Configure option** then new window will be opened and set the path,enter **.arff** in look in dialog box to save normalize data.
12) Right click on **Arff Loader** and click on **Start Loading option** then everything will be executed one by one.
13) Check whether output is created or not by selecting the preferred path.
14) Rename the data name as **a.arff**

15) Double click on **a.arff** then automatically the output will be opened in **MS-Excel**.





Result:

This program has been successfully executed.