Department of Computer Engineering

Jawahar Education Society's,
A. C. Patil College of Engineering
Kharghar, Navi Mumbai 410 210

Lab Manual

Programme (UG/PG): UG

Semester: **V**

Course Code: **CSL503**
Course Title: **Data Warehousing and Mining Lab**

Prepared By: Mr. A.R.Sonule

Approved By:

Institutional Vision, Mission

VISION

To create skilled professionals and engineers for catering the needs of industries and society.

MISSION

1. To provide qualified faculty and required infrastructure to impart quality education inculcating continuous learning attitude
2. To provide platform for the interaction between academia and industry.
3. To inculcate social values and responsible attitude amongst students through co-curricular and extracurricular activities.

Departmental Vision, Mission

Vision

To develop socially committed professionals in Computer engineering for fulfilling the needs of society and industries.

Mission

1. To provide theoretical foundation with laboratory exposure and essential infrastructure.
2. To provide platform for the interaction with industry personnel.
3. To inculcate social awareness through co-curricular and extracurricular activities.

Departmental Program Educational Objectives (PEOs)

- To engage in lifelong learning to adapt with rapidly changing technologies in the field of computer engineering.
- To work effectively in team and exhibit ethical responsibilities.
- To strengthen the knowledge in multidisciplinary areas of engineering.

PSO

- Demonstrate knowledge of programming, data science, operating system and computer network security.
- Apply professional computer engineering practices and strategies for the design, development, operation and maintenance of software.

Program Outcome Code	Program Outcome
PO1	**Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
PO2	**Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
PO3	**Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
PO4	**Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
PO5	**Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
PO6	**The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice
PO7	**Environment and sustainability**: Understand the impact of the professional

	engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
PO8	**Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
PO9	**Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
PO10	**Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
PO11	**Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
PO12	**Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change

Departmental Program Outcomes (POs)

Index

13.	Experiment No. 10	
14.	Mini project	

Lab Objectives

1. Learn and practice data modeling using the entity-relationship and developing database designs.
2. Understand the use of Structured Query Language (SQL) and learn SQL syntax.
3. Understand the needs of database processing and learn techniques for controlling the consequences of concurrent data access.

Lab Outcomes

Sr. No	CO Code	Course Outcomes
1.	CSL503.1	Design data warehouse and perform various OLAP operations.
2.	CSL503.2	Implement data mining algorithms like classification.
3.	CSL503.3	Implement clustering algorithms on a given set of data sample.
4	CSL503.4	Implement Association rule mining & web mining algorithm..

List of Experiments

Sr No	Name of Experiment	Mapped CO
1	Case study on building Data warehouse/Data Mart. Write detailed Problem statement and design dimensional modelling (creation of star and snowflake schema). Implementation of all dimension table and fact table based on case study	CSL503.1
2	Implementation of OLAP operations: Slice, Dice, Rollup, Drilldown and Pivot based on experiment 1 case study.	CSL503.1
3	Implementation of Data Discretization: Binning Methods.	CSL503.2
4	Implementation of Bayesian algorithm	CSL503.2
5	Implementation of K-means Clustering algorithm	CSL503.2
6	Perform data Pre-processing task and demonstrate Classification, Clustering, Association algorithm on data sets using data mining tool (WEKA/R tool)	CSL503.3
7	Implementation of Hierarchical Clustering method	CSL503.3
8	Implementation of Apriori Association Rule Mining algorithm	CSL503.4
9	Implementation of Page rank algorithm	CSL503.4
10	Implementation of HITS algorithm	CSL503.4

	Mini project- Apply Data mining with variation using Dataset after preprocessing.	CSL503.2, CSL503.3, CSL503.4

Mapping of Course Outcomes – Program Outcomes

Course Outcome	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	P012
CSL503.1	2	2	-	-	-	-	-	-	-	-	-	2
CSL503.2	2	2	-	-	-	-	-	-	-	-	-	2
CSL503.3	2	2	-	-	-	-	-	2	-	-	-	2
CSL503.4	2	2	-	-	-	-	-	-	-	-	-	2
CSL503.5	2	2	-	-	-	-	-	2	-	-	-	2
CSL503.6												
Average	**2**	**2**		**-**	**-**	**-**	**-**	**2**	**-**	**-**	**-**	**2**

Data Warehousing and Mining Lab

Experiment No.: 1

Case study on building Data warehouse/Data Mart. Write detailed Problem statement and design dimensional modelling (creation of star

and snowflake schema). Implementation of all dimension table and fact table based on case study

Experiment No. 1

1. **Aim:** Case study on building Data warehouse/Data Mart. Write detailed Problem statement and design dimensional modelling (creation of star and snowflake schema). Implementation of all dimension table and fact table based on case study. Implementation of all dimension tables and fact tables.
2. **Objectives:** Learn how to build a data warehouse and query it.
3. **Outcomes:** Design data warehouse and perform various OLAP operations.
4. **Hardware / Software Required: :** Pentium–V and above , UBUNTU LINUX, MySQL or Oracle or SQL Server.

5.Theory: A Data warehouse is a heterogeneous collection of different data sources organized under unified schema. Builders should take a broad view of the anticipated use of the warehouse while constructing a **data warehouse**. During the design phase, there is no way to anticipate all possible queries or analyses. Some characteristic of Data warehouse are:

- Subject oriented
- Integrated
- Time Variant
- Non-volatile

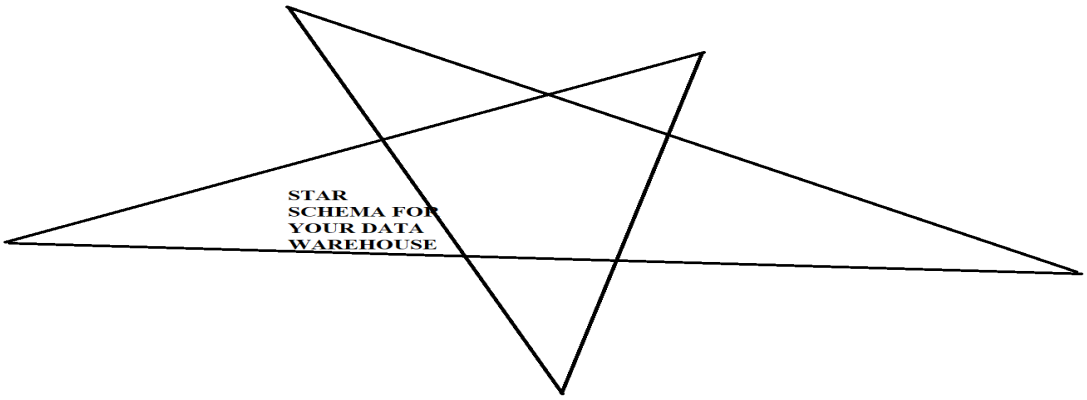
Building a Data Warehouse –
Some steps that are needed for building any data warehouse are as following below:

1. **To extract the data (transnational) from different data sources:**
   For building a data warehouse, a data is extracted from various data sources and that data is stored in central storage area.
2. **To transform the transnational data:**
   There are various DBMS where many of the companies stores their data. Some of them are: MS Access, MS SQL Server, Oracle, Sybase etc. Also these companies saves the data in spreadsheets, flat files, mail systems etc. Relating a data from all these sources is done while building a data warehouse.
3. **To load the data (transformed) into the dimensional database:**
   After building a dimensional model, the data is loaded in the dimensional database. This process combines the several columns together or it may split one field into the several columns. There are two stages at which transformation of the data can be performed and they are: while loading the data into the dimensional model or while data extraction from their origins.
4. **To purchase a front-end reporting tool:**
   There are top notch analytical tools are available in the market. These tools are provided by the several major vendors. A cost effective tool and Data Analyzer is released by the Microsoft on its own.

6. Procedure/ Program:

- **Choosing the Process:** Selecting the subjects from the information packages for the first set of logical structures to be designed.
- **Choosing the Grain:** Determining the level of detail for the data in the data structures.
- **Identifying and Conforming the Dimensions:** Choosing the business dimensions (such as product, market, time, etc.) to be included in the first set of structures and making sure that each particular data element in every business dimension is conformed to one another.
- **Choosing the Facts:** Selecting the metrics or units of measurements (such as product sale units, dollar sales, dollar revenue, etc.) to be included in the first set of structures.
- **Choosing the Duration of the Database**: Determining how far back in time we should go for historical data.

7.Results:



8.Conclusion: Thus, we have constructed Star Schema for XYZ enterprise.

Data Warehousing and Mining Lab

Experiment No.: 2

Implementation of OLAP operations: Slice, Dice, Rollup, Drilldown and Pivot

Experiment No. 2

1Aim: Implementation of OLAP operations: Slice, Dice, Rollup, Drilldown and Pivot

2. Objectives: Learn how to build a data warehouse and query it

3. Outcomes: Design data warehouse and perform various OLAP operations.

4. Hardware / Software Required: Pentium –V and above , UBUNTU LINUX, MySQL or Oracle or SQL Server.

5. Theory: **OLAP** is an acronym for Online Analytical Processing. **OLAP** performs multidimensional analysis of business **data** and provides the capability for complex calculations, trend analysis, and sophisticated **data** modelling
here are primary five types of analytical **operations** in **OLAP** 1) Roll-up 2) Drill-down 3) Slice 4) Dice and 5) Pivot.

Slicing : Slicing operation performs selection on one dimension of the given data cube resulting in a sub cube.

Dice: Dice operation defines a subcube by performing a selection on two or more dimension

Roll up

The roll-up operation (also called drill-up or aggregation operation) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by climbing down a concept hierarchy, i.e. dimension reduction. Let me explain roll up with an example:

Drill down **/Roll Down**

The roll down operation (also called drill down) is the reverse of roll up. It navigates from less detailed data to more detailed data. It can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions.

Pivot

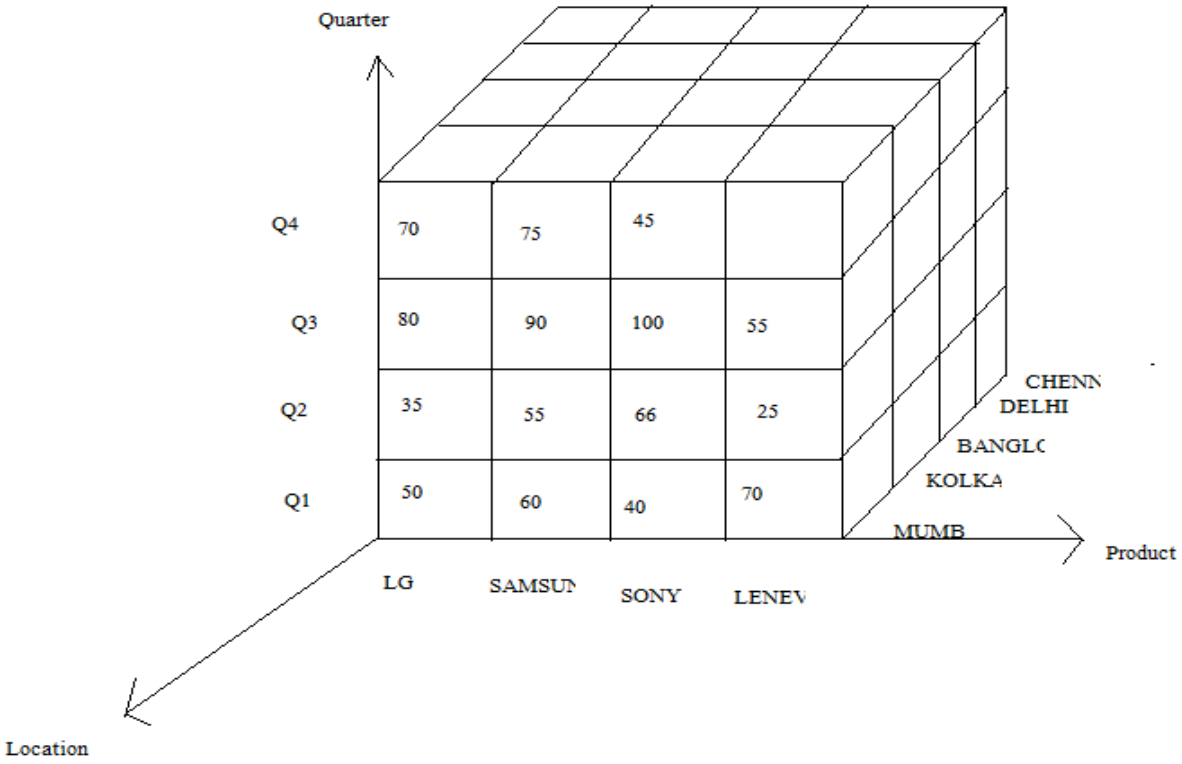
Pivot otheriwise known as Rotate changes the dimensional orientation of the cube, i.e. rotates the data axes to view the data from different perspectives. Pivot groups data with different dimensions. The below cubes shows 2D represntation of Pivot.

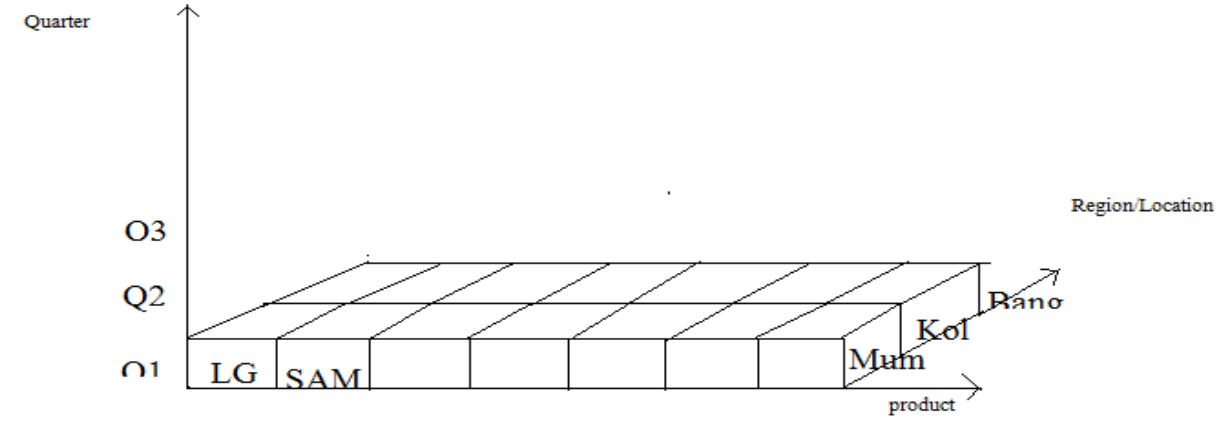
6. Procedure/ Program:

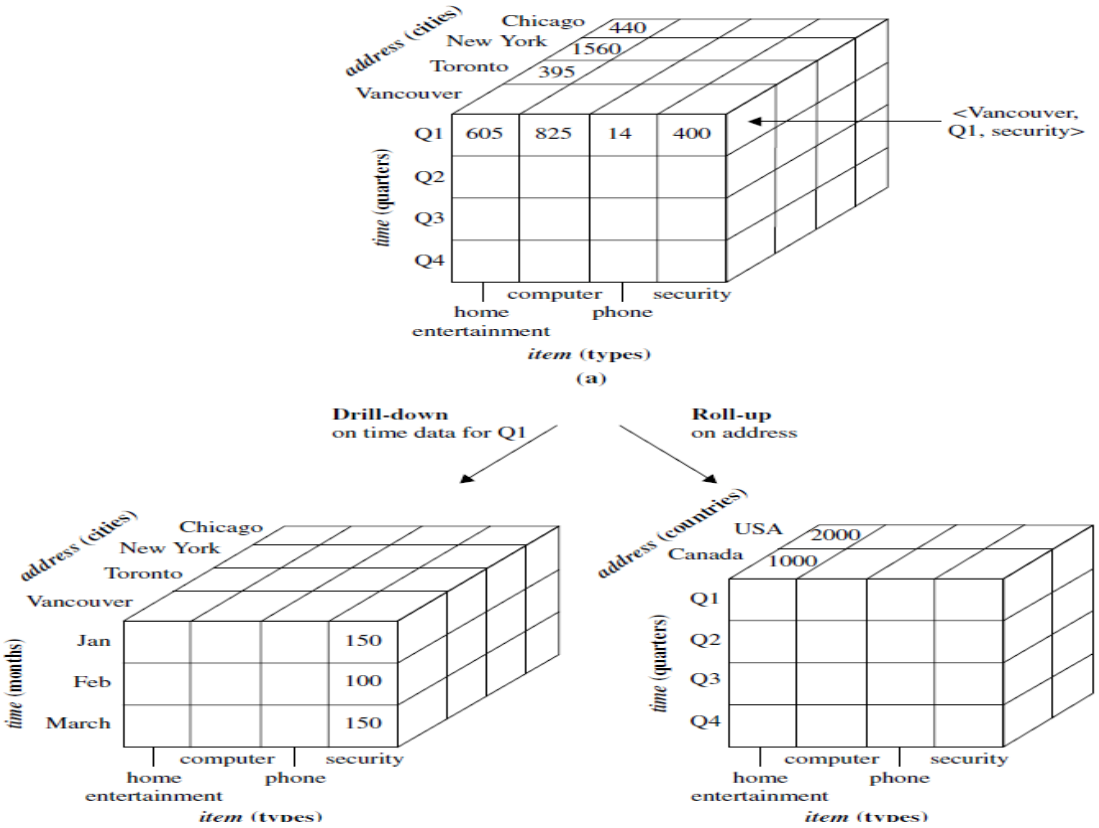
Using Aggregate Function for OLAP Operation

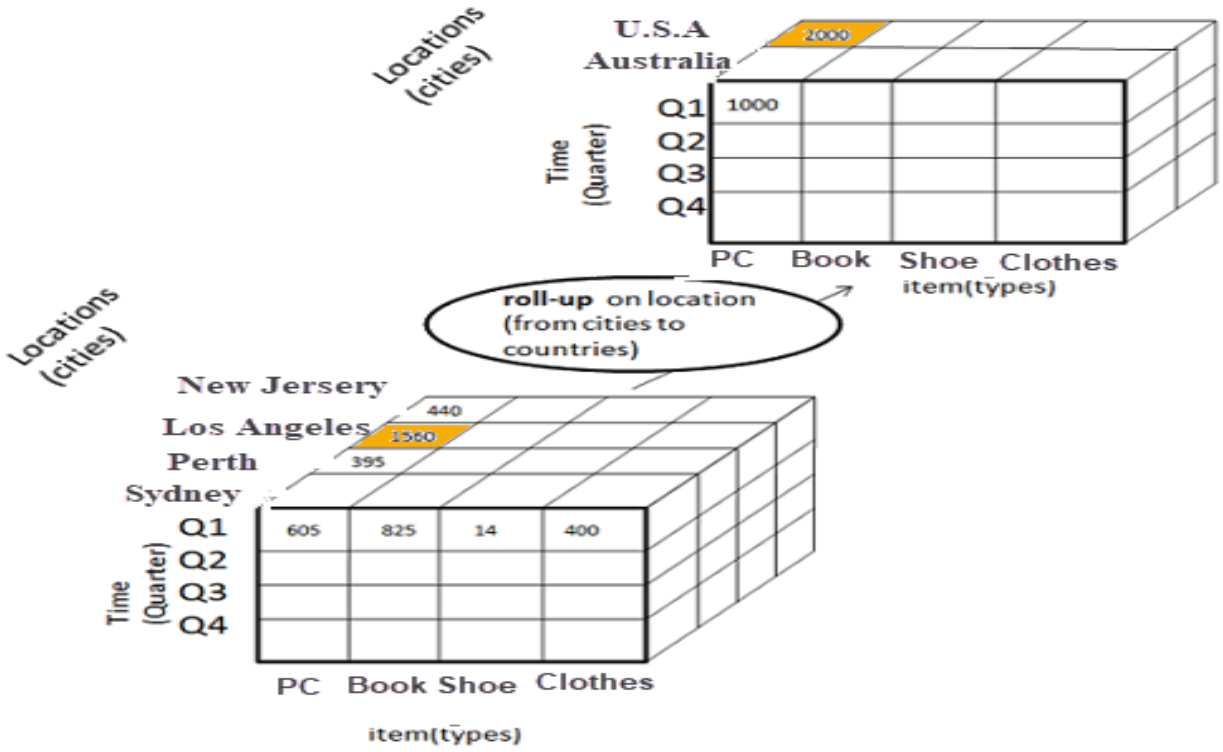
1. SUM
2. Group By
3. Count

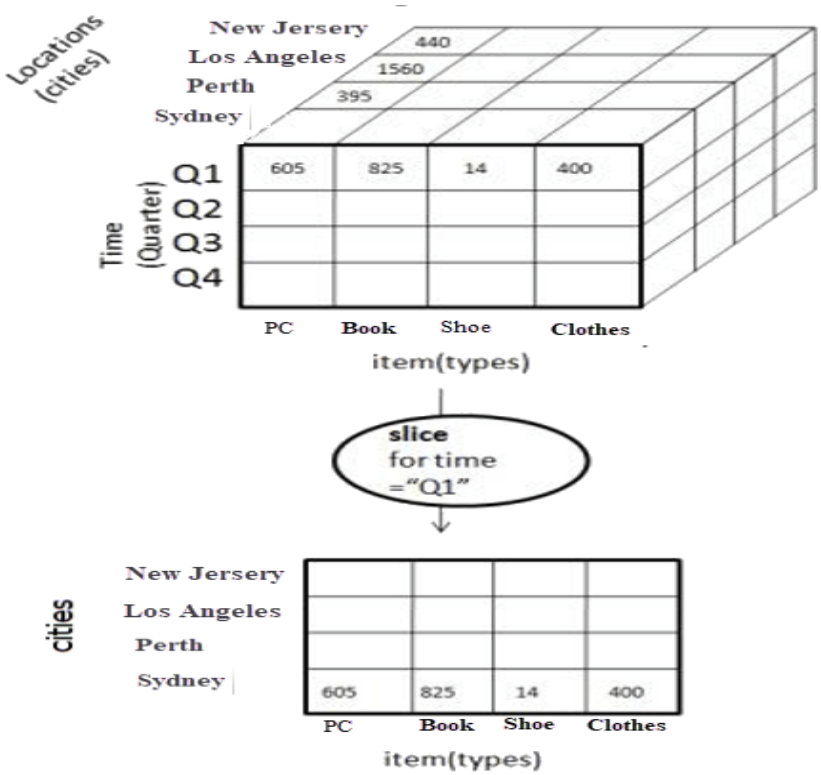
7.Results:

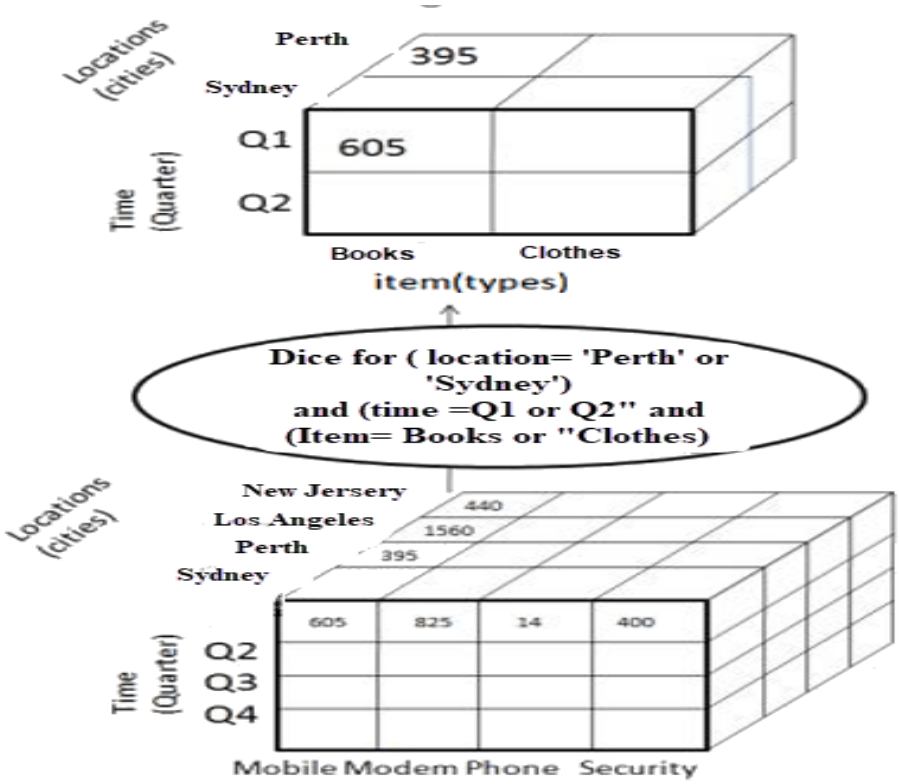

Quarter
Q4
Q3
Q2
Q1
70
75
45
80
90
100
55
35
55
66
25
50
60
40
70
CHENN
DELHI
BANGLC
KOLKA
MUMB
Product
LG
SAMSUN
SONY
LENEV
Location

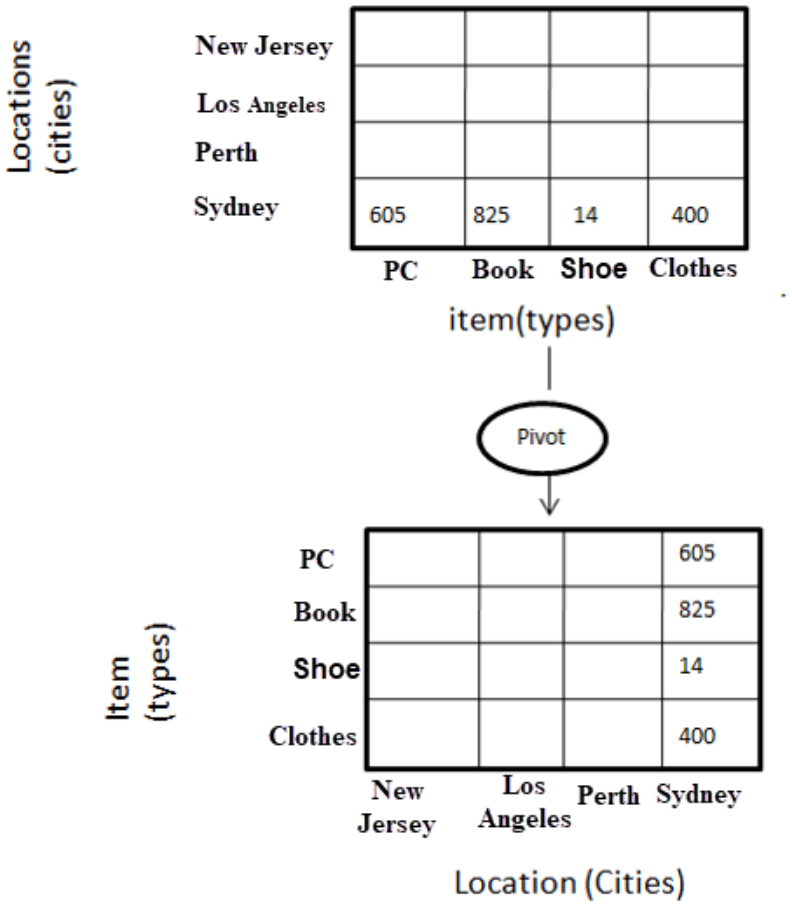

Quarter
Region/Location
O3
Q2
Q1
LG
SAM
Bang
Kol
Mum
product


address (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
Q2
Q3
Q4
605
825
14
400
<Vancouver, Q1, security>
time (quarters)
computer
security
home entertainment
phone
item (types)
(a)
Drill-down on time data for Q1
Roll-up on address
address (cities)
Chicago
New York
Toronto
Vancouver
Jan
Feb
March
150
100
150
time (months)
computer
security
home entertainment
phone
item (types)
address (countries)
USA
Canada
2000
1000
Q1
Q2
Q3
Q4
time (quarters)
computer
security
home entertainment
phone
item (types)


Locations (cities)
U.S.A
Australia
2000
Q1
Q2
Q3
Q4
1000
Time (Quarter)
PC
Book
Shoe
Clothes
item(types)
roll-up on location (from cities to countries)
Locations (cities)
New Jersery
Los Angeles
Perth
Sydney
440
1560
395
Q1
Q2
Q3
Q4
605
825
14
400
Time (Quarter)
PC
Book
Shoe
Clothes
item(types)


Locations (cities)
New Jersey
Los Angeles
Perth
Sydney
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
PC
Book
Shoe
Clothes
item(types)
slice for time ="Q1"
cities
New Jersey
Los Angeles
Perth
Sydney
605
825
14
400
PC
Book
Shoe
Clothes
item(types)


Locations (cities)
Perth
Sydney
395
Q1
605
Q2
Time (Quarter)
Books
Clothes
item(types)
Dice for (location= 'Perth' or 'Sydney') and (time =Q1 or Q2" and (Item= Books or "Clothes)
Locations (cities)
New Jersery
Los Angeles
Perth
Sydney
440
1560
395
605
825
14
400
Q2
Q3
Q4
Time (Quarter)
Mobile Modem Phone
Security



8.Conclusion: Thus, we have implemented OLAP operations.

Data Warehousing and Mining Lab

Experiment No.: 3

Implementation of Data Discretization: Binning Methods.

Experiment No. 3

1.Aim: Implementation of Data Discretization: Binning Methods.

2.Objectives: Learn about the data sets and data preprocessing.

3.Outcomes: Demonstrate the working of algorithms for data mining tasks.

4.Hardware / Software Required: Pentium –V and above , UBUNTU LINUX, Python, Dataset with CSV or Oracle or SQL Server.

5.Theory: Discretization: Divide the range of a continuous attribute into intervals

- Interval labels can then be used to replace actual data values
- Reduce data size by discretization
- Supervised vs. unsupervised
- Split (top-down) vs. merge (bottom-up)
- Discretization can be performed recursively on an attribute
- Prepare for further analysis, e.g., classification

Simple Discretization: Binning

- Equal-width (distance) partitioning
  - Divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation

- Skewed data is not handled well

- Equal-depth (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling

Managing categorical attributes can be tricky

6 . Procedure/ Program and Output:

Sorted data

* Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1:
  - Bin 2:
  - Bin n:
* Smoothing by **bin means**:
  - Bin 1:
  - Bin 2:
  - Bin n:
* Smoothing by **bin boundaries**:
  - Bin 1:
  - Bin 2:
  - Bin n:

7.Results:

8.Conclusion : Thus we have implemented Binning Methods.

Data Warehousing and Mining Lab

Experiment No.: 4

Implementation of Bayesian algorithm

Experiment No. 4

1. Aim: Implementation of Bayesian algorithm

2. Objectives: a. Demonstrate the working of algorithms for data mining tasks such Classification.
b. Apply the data mining techniques with varied input values for different parameters.

3. Outcomes: Implement data mining algorithms like classification.

4. Hardware / Software Required: Pentium –V and above, UBUNTU LINUX, Python, Dataset with CSV or MySQL or Oracle or SQL Server.

5.Theory:

Bayes' Theorem

Bayes' theorem is named after Thomas Bayes, a nonconformist English clergyman who did early work in probability and decision theory during the 18th century.

Let X be a data tuple. In Bayesian terms, X is considered "evidence." As usual, it is described by measurements made on a set of n attributes.

Let H be some hypothesis such as that the data tuple X belongs to a specified class C.

For classification problems, we want to determine P(H/X), the probability that the hypothesis H holds given the "evidence" or observed data tuple X.

In other words, we are looking for the probability that tuple X belongs to class C, given that we know the attribute description of X.

Accuracy and Error measures

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Definition of the Terms:

• Positive (P) : Observation is positive (for example: is an apple).

• Negative (N) : Observation is not positive (for example: is not an apple).

• True Positive (TP) : Observation is positive, and is predicted to be positive.

• False Negative (FN) : Observation is positive, but is predicted negative.

• True Negative (TN) : Observation is negative, and is predicted to be negative.

• False Positive (FP) : Observation is negative, but is predicted positive.

- Accuracy =(TP+TN)/ (TP+FP+TN+FN)
- Precision =TP /(TP+FP)
- Recall =TP /(TP+FN)
- Sensitivity or True Positive Rate (TPR)=TP /(TP + FN)
- Specificity or True Negative Rate (TNR)=TN /(FP + TN)
- FPR =FP /(FP+TN)
- FNR = FN / (FN+TP)
- False Alarm Rate = FAR = (FPR + FNR) /2
- F1 Score = 2(Precision x Recall)/(Precision +Recall)

6.Procedure/ Program:

1. Use softwares Python/ Mysql/Oracle/SQL Server.

2. Use Jupyter/Pycharm/Spider.

3. Load dataset.

4. Write Python program without using libraries.

7.Results:

8.Conclusion : Thus we have applied Naïve Bayes on Dataset and got accuracy.. .

Data Warehousing and Mining Lab

Experiment No.: 5

Implementation of ID3 algorithm

Experiment No. 5

1. Aim: Implementation of ID3 algorithm

2. Objectives: a. Demonstrate the working of algorithms for data mining tasks such Classification.

b. Apply the data mining techniques with varied input values for different parameters.

3. Outcomes: Implement data mining algorithms like classification.

4. Hardware / Software Required: Pentium –V and above, UBUNTU LINUX, Python, Dataset with CSV or MySQL or Oracle or SQL Server.

5. Theory

Expected Information

Let S be a set consisting of s data samples.

Suppose the class label attribute has m distinct values defining m distinct classes, Ci (for i=.1..m)

Let si be the number of sample of S in class Ci

The expected information needed to classify a given sample is given by

$$I(s1,s2,..\ sm) = -\sum_{i=1}^{m} si/s.\log 2\ si/s$$

a log function to the base 2 is used since the information is encoded in bits.

Entropy:

- An attribute A with values $\{ a_1,a_2,.. a_r)$ can be used to partition S into the subsets $\{S_1,S_2 ..S_r \}$

where s_j contains those samples in S that value a_j of A.

- Let S_j contain s_{ij} samples of class Ci.
- The expected information based on this partitioning by A is known as the entropy of A. It is the weighted average.

$$I(A) = -\sum_{j=1}^{r} (s_{ij+\ldots}\ s_{mj})/s\ I(s_1,s_2,..\ s_m)$$

Information Gain

The information gain obtained by this partitioning on A is defined by

Gain(A)= I(s1,s2,.. sm) –E(A)

Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition *D*.

Input:

- Data partition, *D*, which is a set of training tuples and their associated class labels;
- *attribute_list*, the set of candidate attributes;
- *Attribute_selection_method*, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split point* or *splitting subset*.

Output: A decision tree.

Method:

```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C then
(3)      return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)      return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
         multiway splits allowed then // not restricted to binary trees
(9)      attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)     let Dj be the set of data tuples in D satisfying outcome j; // a partition
(12)     if Dj is empty then
(13)         attach a leaf labeled with the majority class in D to node N;
(14)     else attach the node returned by Generate_decision_tree(Dj, attribute_list) to node N;
     endfor
(15) return N;
```

Figure Basic algorithm for inducing a decision tree from training tuples.

6. Procedure/ Program:

1. Use software Python/ Mysql/Oracle/SQL Server.
2. Use Jupyter/Pycharm/Spider.
3. Load dataset.
4. Write Python program without using libraries.

7. Results:

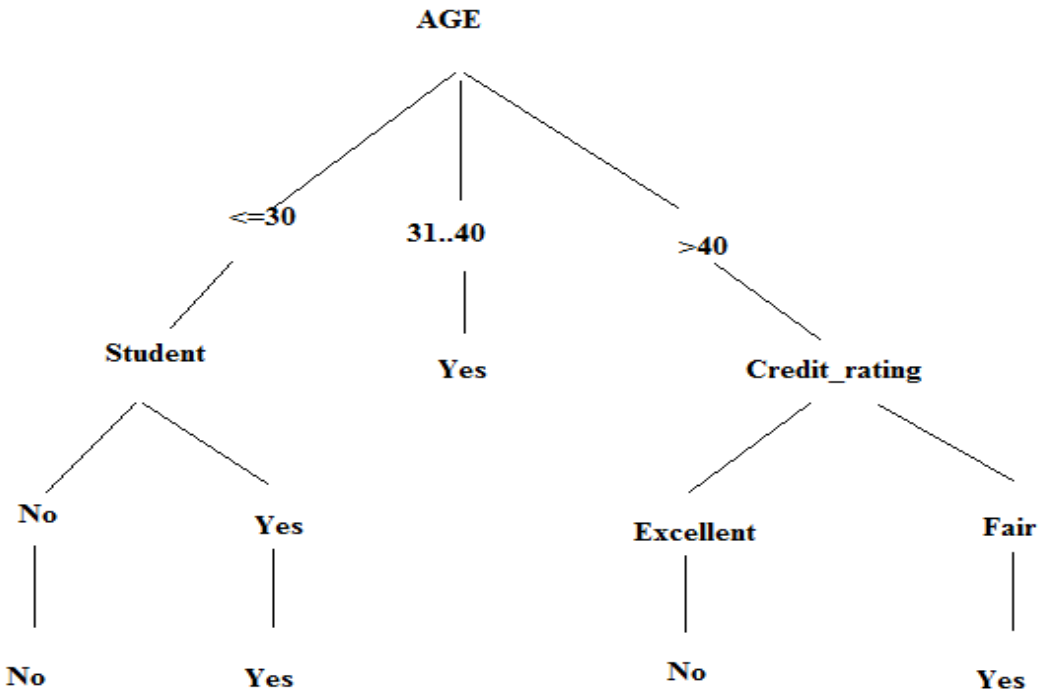


Fig: Decision tree for "buy_computer"

8. Conclusion : Thus we have implemented ID3 and got accuracy.. .

Data Warehousing and Mining Lab

Experiment No.: 6

Implementation of Clustering algorithm (K-means/K-medoids)

Experiment No. 6

1.Aim: Implementation of Clustering algorithm (K-means/K-medoids

2.Objectives: a. Demonstrate the working of algorithms for data mining tasks such Classification.
b. Apply the data mining techniques with varied input values for different parameters.

3.Outcomes: Implement clustering algorithms on a given set of data sample.

4.Hardware / Software Required: : Pentium –V and above , UBUNTU LINUX, Python, Dataset with CSV or MySQL or Oracle or SQL Server.

5.Theory: The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing.

Algorithm: k-means. The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

k: the number of clusters,

D: a data set containing n objects.

Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects from D as the initial cluster centers;

(2) repeat

(3) (re)assign each object to the cluster to which the object is the most similar,based on the mean value of the objects in the cluster;

(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;

(5) until no change;

6.Procedure/ Program:

1. Use software Python/ Mysql/Oracle/SQL Server.
2. Use Jupyter/Pycharm/Spider.
3. Load dataset.
4. Write Python program without using libraries.

7. Result

8.Conclusion : Thus we have executed K mean algorithm.

Data Warehousing and Mining Lab

Experiment No.: 7

Perform data Pre-processing task and demonstrate Classification, Clustering, Association algorithm on data sets using data mining tool (WEKA/R tool)

Experiment No. 7

1.Aim: Perform data Pre-processing task and demonstrate Classification, Clustering, Association algorithm on data sets using data mining tool (WEKA/R tool)

2.Objectives: Explore open-source software (like WEKA) to perform data mining tasks.

3.Outcomes: Implement classificatiom/clustering algorithms on a given set of data sample.

4.Hardware / Software Required: : Pentium –V and above , UBUNTU LINUX, Python, Dataset with CSV or MySQL or Oracle or SQL Server.

5.Theory:

Data cleaning

Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

Integration of multiple databases, data cubes, or files

Data reduction

Dimensionality reduction

Numerosity reduction

Data compression

Data transformation and data discretization

Normalization

Concept hierarchy generation

Waikato Environment for Knowledge Analysis (Weka), developed at the University of Waikato, New Zealand, is free software licensed under the GNU General Public License, and the companion software to the book "Data Mining: Practical Machine Learning Tools and Techniques"

Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions.[1] The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains,[2][3] but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include:

Free availability under the GNU General Public License.

Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.

A comprehensive collection of data preprocessing and modeling techniques.

Ease of use due to its graphical user interfaces.

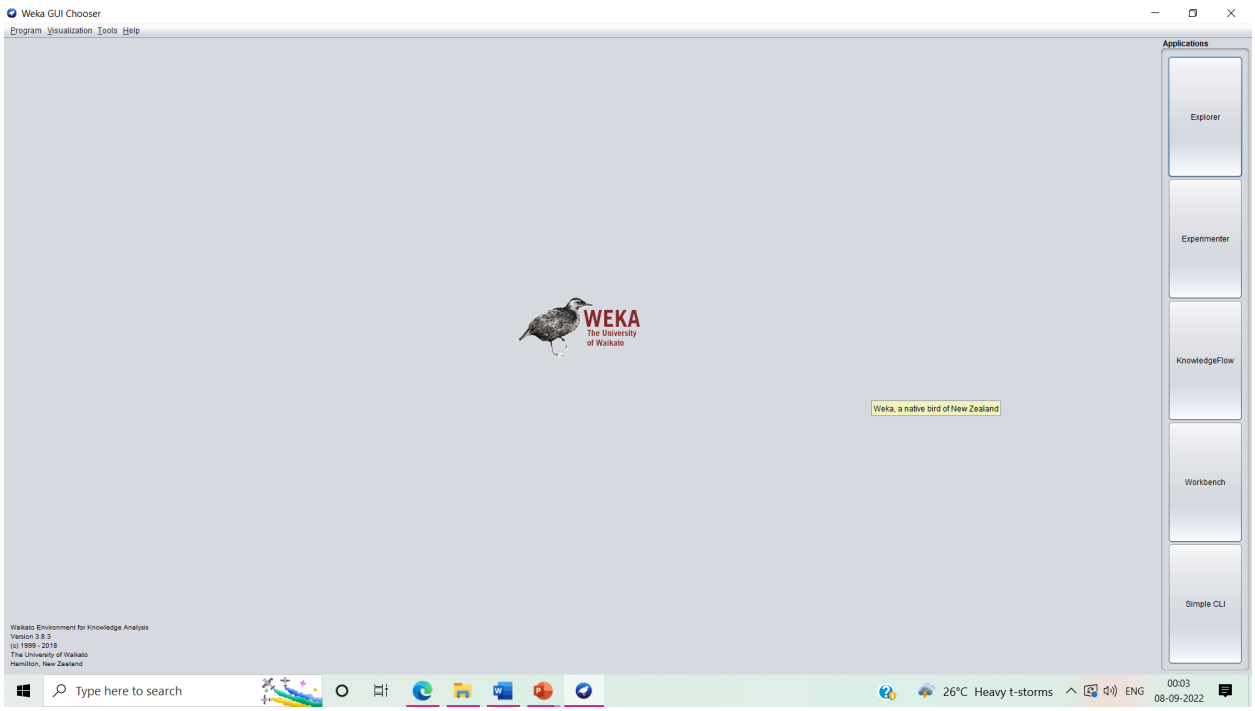
Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. Input to Weka is expected to be formatted according the Attribute-Relational File Format and with the filename bearing the .arff extension. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is

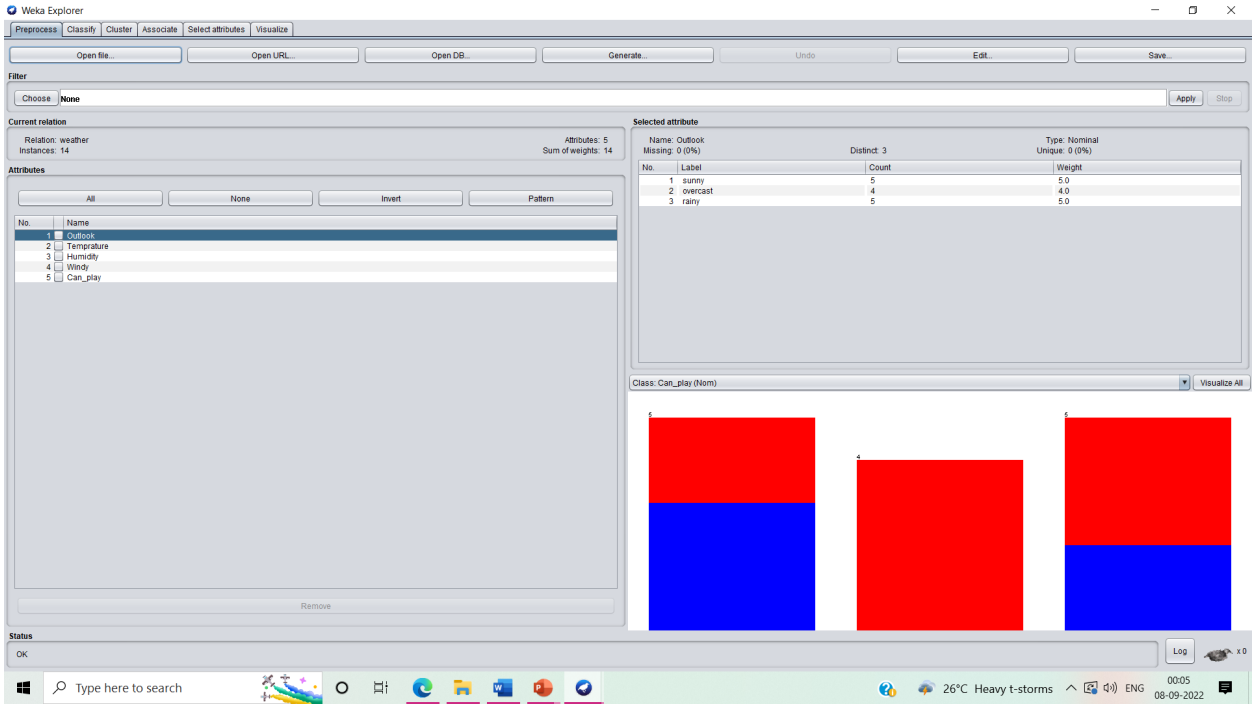
described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported).

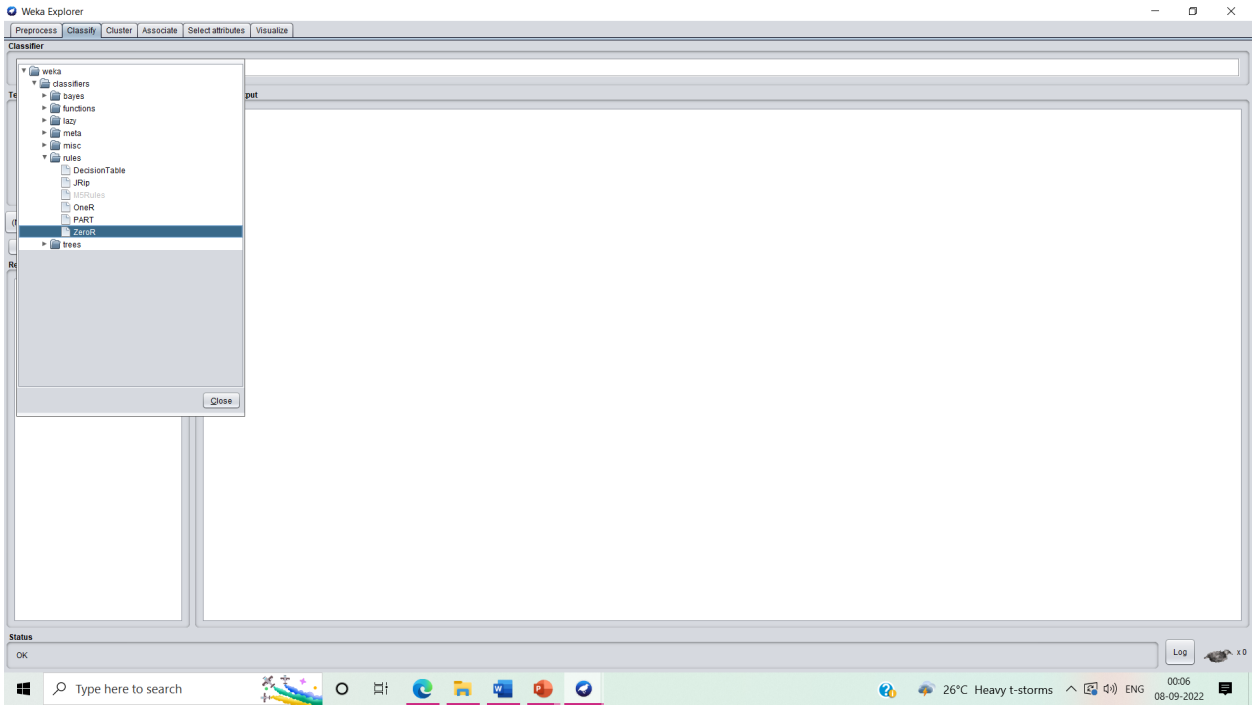
6.Procedure/ Program:

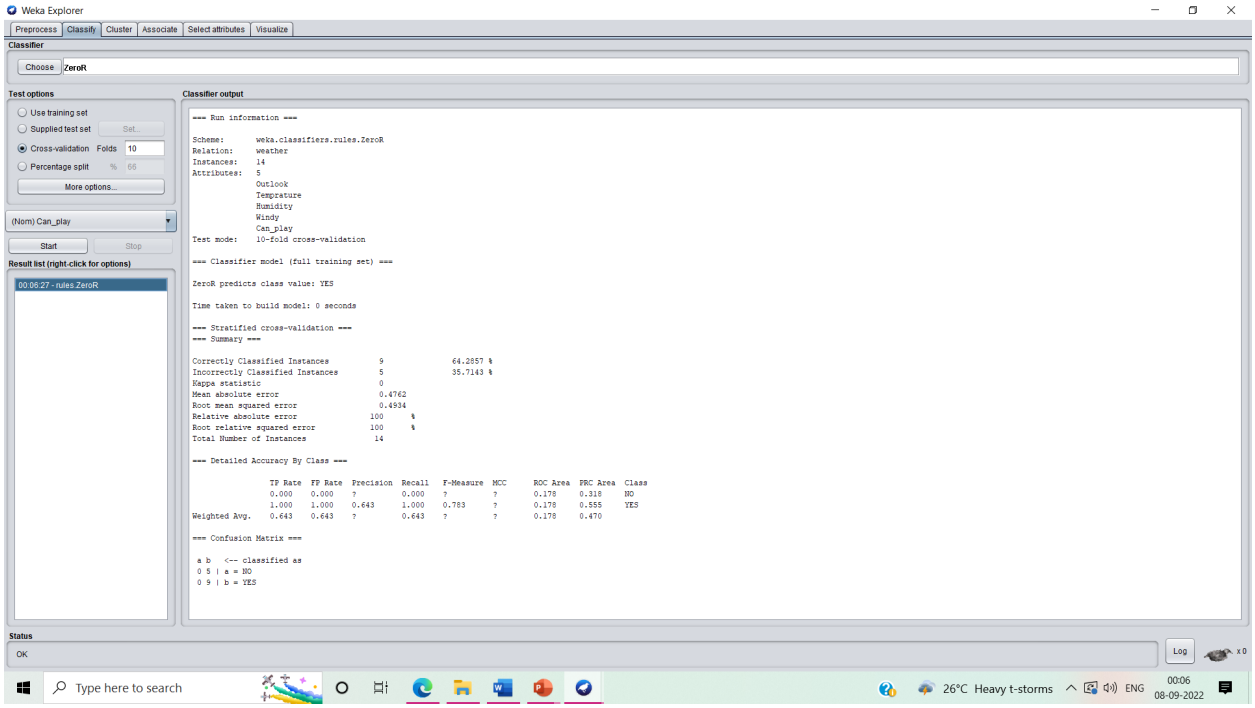
1. Use database software WEKA.
2. Open Dataset. Preprocess Dataset.
3. Select algorithm. View performance parameters.

7.Results:




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: weather
Instances: 14
Attributes: 5
Sum of weights: 14
Attributes
All
None
Invert
Pattern
No.
Name
1 Outlook
2 Temprature
3 Humidity
4 Windy
5 Can_play
Remove
Selected attribute
Name: Outlook
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 sunny 5 5.0
2 overcast 4 4.0
3 rainy 5 5.0
Class: Can_play (Nom)
Visualize All
Status
OK
Log
x 0
Type here to search
26°C Heavy t-storms
ENG
00:05
08-09-2022


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
weka
classifiers
bayes
functions
lazy
meta
misc
rules
DecisionTable
JRip
MSRules
OneR
PART
ZeroR
trees
Close
Status
OK
Log
x 0
Type here to search
26°C Heavy t-storms
ENG
00:06
08-09-2022



8.Conclusion : Thus we have used WEKA Tool.

Data Warehousing and Mining Lab

Experiment No.: 8

Implementation of Hierarchical Clustering method

Experiment No. 8

1.Aim: Implementation of Hierarchical Clustering method

2.Objectives: a. Demonstrate the working of algorithms for data mining tasks such Classification.

b. Apply the data mining techniques with varied input values for different parameters.

3.Outcomes: Implement clustering algorithms on a given set of data sample**.**

4.Hardware / Software Required: Pentium –V and above, UBUNTU LINUX, Python, Dataset with CSV or MySQL or Oracle or SQL Server.

5.Theory:

Dissimilarity matrix (or object-by-object structure):

This stores a collection of proximities that are available for all pairs of n objects. It is often represented by an n-by-n table:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

where d(i, j) is the measured difference or dissimilarity between objects i and j. In general, d(i, j) is a nonnegative number that is close to 0 when objects i and j are highly similar or "near" each other, and becomes larger the more they differ.

- Single linkage: Minimum distance is considered
- Complete linkage: Maximum distance is considered
- Average linkage: Average distance is considered

Agglomerative Algorithm

Agglomerative Hierarchical Clustering (AHC) is **a clustering (or classification) method** which has the following advantages: It works from the dissimilarities between the objects to be grouped together. A type of dissimilarity can be suited to the subject studied and the nature of the data.

The agglomerative clustering is the most common type of hierarchical clustering used **to group objects in clusters based on their similarity**. It's also known as AGNES (Agglomerative Nesting). The algorithm starts by treating each object as a singleton cluster.

The step that Agglomerative Clustering take are: **Each data point is assigned as a single cluster**. Determine the distance measurement and calculate the distance matrix. Determine the linkage criteria to merge the clusters.

What is agglomerative clustering in machine learning?

Agglomerative Clustering is **a bottom-up strategy in which each data point is originally a cluster of its own, and as one travels up the hierarchy, more pairs of clusters are combined**. In it, two nearest clusters are taken and joined to form one single cluste

Divisive Clustering

- **DIANA** (Devise Analysis)
- **AGNES** (Agglomerative Nesting)
- The divisive clustering algorithm is **a top-down clustering approach**, initially, all the points in the dataset belong to one cluster and split is performed recursively as one moves down the hierarchy.

6.Procedure/ Program:

7.Results:

8.Conclusion: Thus, we have implemented Hierarchical clustering.

Data Warehousing and Mining Lab

Experiment No.: 9

Implementation of Association Rule Mining algorithm (Apriori)

Experiment No. 9

1.Aim: Implementation of Association Rule Mining algorithm (Apriori).

2.Objectives: a. Demonstrate the working of algorithms for data mining tasks such association rule mining.
b. Apply the data mining techniques with varied input values for different parameters.

3.Outcomes: Implement Association rule mining & web mining algorithm.

4.Hardware / Software Required: Pentium –V and above, UBUNTU LINUX, Python, Dataset with CSV or MySQL or Oracle or SQL Server.

5.Theory:

SUPPORT

The support of an association pattern refers to the percentages of task relevant data tuples for which the pattern is true.

For association rules of the form 'A ☾ B' where A & B are sets of items, the support is defined as

Support (A ☾ B) = #tuples containing both A & B/ total number of tuples

= P(A U B)

Confidence

A certainty measure for association rules of the form

'A ☾ B' where A & B are sets of items, is confidence.

Given a set of task relevant data tuples the confidence

of A ☾ B is defined as

Confidence (A ☾ B) = #tuples containing both A & B / #tuples containing A

= P(B|A)

= P(A U B) / P (A)

6.Procedure/ Program:

Join Step: Ck is generated by joining Lk-1with itself

Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

Pseudo-code:

Ck: Candidate itemset of size k

Lk : frequent itemset of size k

L1 = {frequent items};
for (k = 1; Lk !=∅; k++) do begin
Ck+1 = candidates generated from Lk;
for each transaction t in database do
increment the count of all candidates in Ck+1 that are contained in t
Lk+1 = candidates in Ck+1 with min_support
end
return ∪k Lk;

7.Results:

8.Conclusion: Thus, we have implemented Apriori algorithm.

Data Warehousing and Mining Lab

Experiment No.: 10

Implementation of Page rank algorithm / HITS algorithm

Experiment No. 10

1.Aim: Implementation of Page rank algorithm / HITS algorithm

2.Objectives: a. Demonstrate the working of algorithms for data mining tasks such association rule mining.

b. Apply the data mining techniques with varied input values for different parameters.

3.Outcomes: Implement Association rule mining & web mining algorithm.

4.Hardware / Software Required: Pentium –V and above, UBUNTU LINUX, Python, Dataset with CSV or MySQL or Oracle or SQL Server.

5.Theory:

PageRank:

This algorithm is used by Google to rank search results. The name of this algorithm is given by Google-founder Larry Page. The rank of a page is decided by the number of links pointing to the target node.

HITS Algorithm

Input:
W //WWW viewed as a directed graph.
q //Query.
s // Support.
Output:
A // Set of authority pages.
H // Set of hub pages.
HITS Algorithm
$R = SE(W, q)$;
$B = R \cup \{pages\ linked\ to\ from\ R\} \cup \{pages\ which\ link\ to\ pages\ in\ R\}$;
$G(B, L) = Subgraph\ of\ W\ induced\ by\ B$;
$G(B, L^1) = Delete\ links\ in\ G\ within\ same\ site$;
$x_p = \sum_{q\ where\ <q,p>\in L^1} y_q$; // Find authority weights.
$y_p = \sum_{q\ where\ <p,q>\in L^1} x_q$; // Find hub weights.
$A = \{p \mid p\ has\ one\ of\ the\ highest\ x_p\}$;
$H = \{p \mid p\ has\ one\ of\ the\ highest\ y_p\}$;

6.Procedure/ Program:

7.Results:

8.Conclusion : Thus we have implemented web mining algorithm.

Data Warehousing and Mining Lab

Mini Project

Apply Data mining with variation using Dataset after preprocessing.

Mini Project

1.Aim: Apply Data mining with variation using Dataset after preprocessing.

2.Objectives: Learn about the data sets and data preprocessing.

a

3.Outcomes:

a.

4.Hardware / Software Required: : Pentium –V and above , UBUNTU LINUX, Python, MySQL, Oracle or SQL Server, Front Software

5.Theory:

6.Procedure/ Program:

7.Results:

8.Conclusion : Thus we have done mini project.