D Y PATIL
RAMRAO ADIK INSTITUTE OF TECHNOLOGY
NAVI MUMBAI

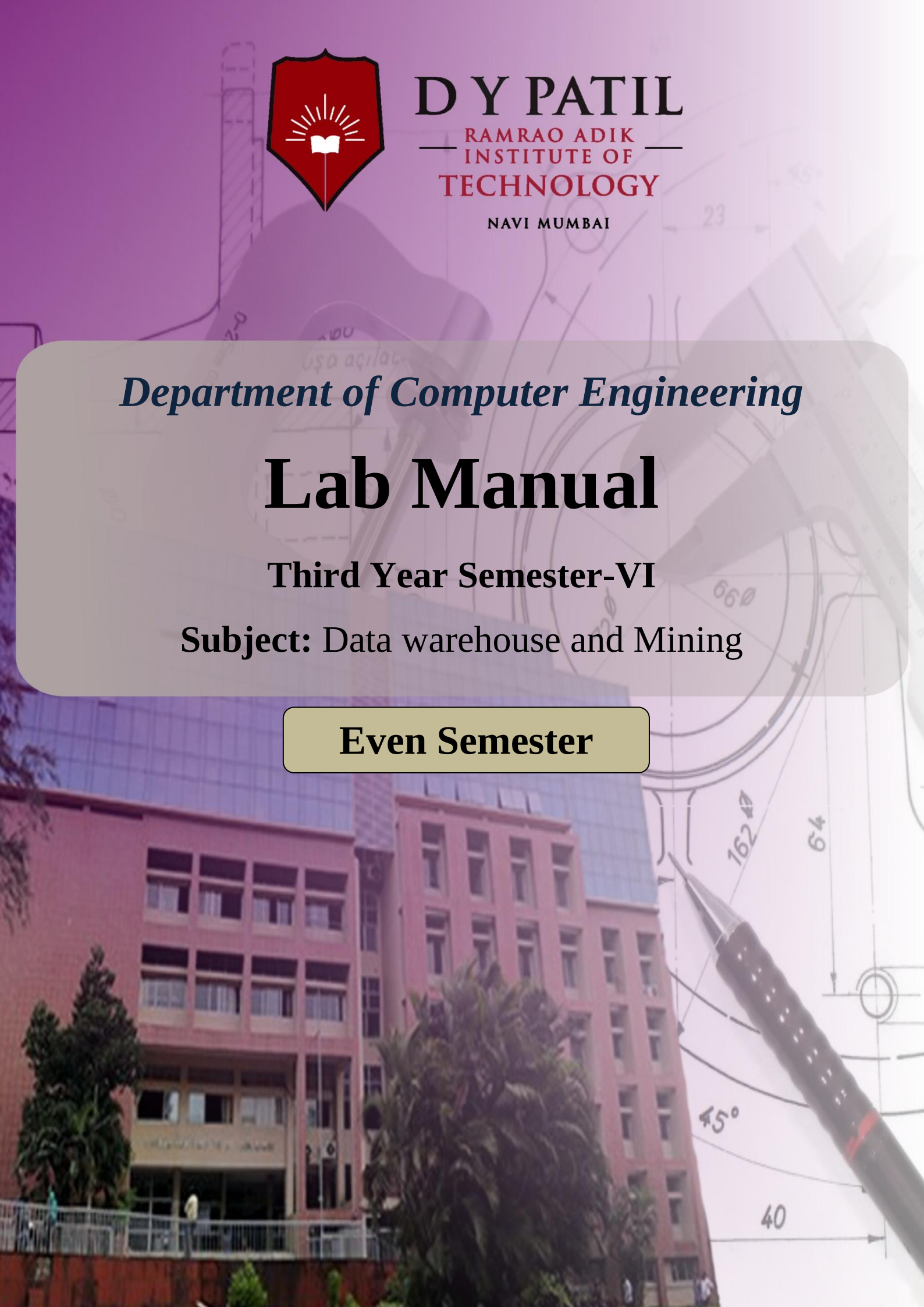
Department of Computer Engineering

Lab Manual

Third Year Semester-VI

Subject: Data warehouse and Mining

Even Semester



Institutional Vision and Mission

Our Vision

To foster and permeate higher and quality education with value added engineering, technology programs, providing all facilities in terms of technology and platforms for all round development with societal awareness and nurture the youth with international competencies and exemplary level of employability even under highly competitive environment so that they are innovative adaptable and capable of handling problems faced by our country and world at large.

Our Mission

The Institution is committed to mobilize the resources and equip itself with men and materials of excellence thereby ensuring that the Institution becomes pivotal center of service to Industry, academia, and society with the latest technology. RAIT engages different platforms such as technology enhancing Student Technical Societies, Cultural platforms, Sports excellence centers, Entrepreneurial Development Center and Societal Interaction Cell. To develop the college to become an autonomous Institution & deemed university at the earliest with facilities for advanced research and development programs on par with international standards. To invite international and reputed national Institutions and Universities to collaborate with our institution on the issues of common interest of teaching and learning sophistication.

Our Quality Policy

ज्ञानधीनं जगत् सर्वम।

Knowledge is supreme.

OurQuality Policy

Itisourearnestendeavourtoproducehighqualityengineeringprofessionalswhoare innovative and inspiring, thought and action leaders, competent to solve problems facedbysociety,nationandworldatlargebystrivingtowardsveryhighstandardsin learning,teaching andtrainingmethodologies.

OurMotto: If itis not of quality, itis NOTRAIT!

Departmental VisionandMission

Vision

To impart higher and quality education in computer science with value added engineering and technology programs to prepare technically sound, ethically strong engineers with social awareness. To extend the facilities, to meet the fast changing requirements and nurture the youths with international competencies and exemplary level of employability and research under highly competitive environments.

Mission

To mobilize the resources and equip the institution with men and materials of excellence to provide knowledge and develop technologies in the thrust areas of computer science and Engineering. To provide the diverse platforms of sports, technical, curricular and extracurricular activities for the overall development of student with ethical attitude. To prepare the students to sustain the impact of computer education for social needs encompassing industry, educational institutions and public service. To collaborate with IITs, reputed universities and industries for the technical and overall upliftment of students for continuing learning and entrepreneurship.

Departmental Program Educational Objectives (PEOs)

1. **Learn and Integrate**
   To provide Computer Engineering students with a strong foundation in the mathematical, scientific and engineering fundamentals necessary to formulate, solve and analyze engineering problems and to prepare them for graduate studies.

2. **Think and Create**
   To develop an ability to analyze the requirements of the software and hardware, understand the technical specifications, create a model, design, implement and verify a computing system to meet specified requirements while considering real-world constraints to solve real world problems.

3. **Broad Base**
   To provide broad education necessary to understand the science of computer engineering and the impact of it in a global and social context.

4. **Techno-leader**
   To provide exposure to emerging cutting edge technologies, adequate training & opportunities to work as teams on multidisciplinary projects with effective communication skills and leadership qualities.

5. **Practice citizenship**
   To provide knowledge of professional and ethical responsibility and to contribute to society through active engagement with professional societies, schools, civic organizations or other community activities.

6. **Clarify Purpose and Perspective**
   To provide strong in-depth education through electives and to promote student awareness on the life-long learning to adapt to innovation and change, and to be successful in their professional work or graduate studies.

Departmental Program Outcomes (POs)

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.

Program Specific Outcomes: PSO

PSO1: To build competencies towards problem solving with an ability to understand, identify, analyze and design the problem, implement and validate the solution including both hardware and software.

PSO2: To build appreciation and knowledge acquiring of current computer techniques with an ability to use skills and tools necessary for computing practice.

PSO3: To be able to match the industry requirements in the area of computer science and engineering. To equip skills to adopt and imbibe new technologies.

Index

List of Experiments

Expt. No.	Topic
1	Build Data Warehouse/Data Mart for a given problem statement i) Identifying the source tables and populating sample data ii) Design dimensional data model i.e. Star schema and Snowflake schema
2	To perform various OLAP operations such as slice, dice, drilldown, rollup, pivot
3	Implementation of Classification algorithm(Decision Tree/ Bayesian)
4	Implementation of Linear Regression.
5	Implementation of Clustering algorithm(K-means/ Agglomerative).
6	Implementation of Association Rule Mining algorithm(Apriori).
7	Perform data Pre-processing task and Demonstrate performing Classification, Clustering, Association algorithm on data sets using data mining using WEKA tool
8	Implementation of Spatial Clustering Algorithm- CLARANS Extensions
***9**	Case study on spatial data mining techniques from recent IEEE papers

Course Objectives & Course Outcome, Experiment Plan

Course Objectives:

1.	To study the methodology of engineering legacy databases for data warehousing.
2.	To study the design modeling of data warehouse.
3.	To study the preprocessing and online analytical processing of data.
4.	To study the methodology of engineering legacy of data mining to derive business rules for decision support systems.
5.	To analyze the data, identify the problems, and choose the relevant models and algorithms to apply

Course Outcomes:

CO1	Understand and Design models of data warehouse.
CO2	Apply steps of data exploration and implement data preprocessing using appropriate tools.
CO3	Identify and implement predication and association mining algorithms to solve real world problems
CO4	Design and analyze OLAP operations.
CO5	Identify and implement classification and clustering algorithms to solve real world problems.
CO6	Describe complex data type with respect to spatial and web mining and implement spatial mining algorithm.

Experiment Plan

Module No.	Week No.	Topic	CO Meet	Weightage
1	W1, W2	Build Data Warehouse/Data Mart for a given problem statement i) Identifying the source tables and populating sample data ii) Design dimensional data model i.e. Star schema and Snowflake schema	CO1	10
2	W3,W4	To perform various OLAP operations such as slice, dice, drilldown, rollup, pivot	CO4	10
3	W5	Implementation of Classification algorithm(Decision Tree/ Bayesian)	CO5	5
4	W6	Implementation of Linear Regression.	CO3	5
5	W7	Implementation of Clustering algorithm(K-means/ Agglomerative).	CO5	5
6	W8	Implementation of Association Rule Mining algorithm(Apriori).	CO3	5
7	W9,W10	Perform data Pre-processing task and Demonstrate performing Classification, Clustering, Association algorithm on data sets using data mining using WEKA tool	CO2	10
8	W11	Implementation of Spatial Clustering Algorithm-CLARANS Extensions	CO6	5
***9**	W12	Case study on spatial data mining techniques from recent IEEE papers	CO6	5

Mapping Course Outcomes (CO) - Program Outcomes (PO)

Subject Weight	Course Outcomes		Contribution to Program outcomes											
			PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
PRATICAL 100%	CO1	Understand and Design models of data warehouse.	1	2	3		1				2	1		
	CO2	Apply steps of data exploration and implement data preprocessing using appropriate tools.	1	3		2	2			1			1	
	CO3	Identify and implement predication and Association mining algorithms to solve real world problems.		2		3	1	1				2	1	
	CO4	Design and analyze OLAP operations.			3	2	2	2		1				
	CO5	Identify and implement classification and clustering algorithms to solve real world problems.			2	2	3	2						1
	CO6	Describe complex data type with respect to spatial and web mining and implement spatial mining algorithm.			2	2	3		1					2

Program Specific Outcomes:

PSO1: To build competencies towards problem solving with an ability to understand, identify, analyze and design the problem, implement and validate the solution including both hardware and software.

PSO2: To build appreciation and knowledge acquiring of current computer techniques with an ability to use skills and tools necessary for computing practice.

PSO3: To be able to match the industry requirements in the area of computer science and engineering. To equip skills to adopt and imbibe new technologies.

Mapping of Course outcomes with Program Specific Outcomes:

Course Outcomes		Contribution to Program Specific outcomes		
		PSO1	PSO2	PSO3
CO1	Understand and Design models of data warehouse.	3		2
CO2	Apply steps of data exploration and implement data preprocessing using appropriate tools.	2	1	3
CO3	Identify and implement predication and Association mining algorithms to solve real world problems.		3	1
CO4	Design and analyze OLAP operations.	3	1	
CO5	Identify and implement classification and clustering algorithms to solve real world problems.	1	2	3
CO6	Describe complex data type with respect to spatial and web mining and implement spatial mining algorithm.	1	2	2

Study and Evaluation Scheme

Course Code	Course Name	Teaching Scheme			Credits Assigned			
CSL603	DataWarehouse and Mining Lab	Theory	Practical	Tutorial	Theory	Practical	Tutorial	Total
		-	02	--	-	01	--	01

Course Code	Course Name	Examination Scheme		
CSL603	Data Warehouse and Mining Lab	Term Work	Practical & Oral	Total
		25	25	50

Term Work:

Internal Assessment consists of two tests. Test 1, an Institution level central test, isfor 20 marks and is to be based on a minimum of 40% of the syllabus. Test 2 isalso for 20 marks and is to be based on the remaining syllabus. Test 2 may be either a class test or assignment on live problems or course project

Practical & Oral:

Oral examination is to be conducted by pair of internal and external examiners based on the syllabus.

Data Warehouse and Mining

Experiment No. : 1

Build Data Warehouse/Data Mart for a given problem statement
i) Identifying the source tables and populating sample data ii) Design dimensional data model i.e. Star schema, Snowflake schema and Fact Constellation schema (if applicable)

Experiment No. 1

1.Aim: Build Data Warehouse/Data Mart for a given problem statement i) Identifying the source tables and populating sample data ii) Design dimensional data model i.e. Star schema, Snowflake schema and Fact Constellation schema (if applicable)

1. **Objectives:** From this experiment, the student will be able to
    - Understand the basics of Data Warehouse
    - Understand the design model of Data Warehouse
    - Study methodology of engineering legacy databases for data warehousing

2. **Outcomes:** The learner will be able to
    - Apply knowledge of legacy databases in creating data warehouse
    - Understand, identify, analyse and design the warehouse
    - Use current techniques, skills and tools necessary for designing a data warehouse

3. **Software Required :**Oracle 11g
4. **Theory:**

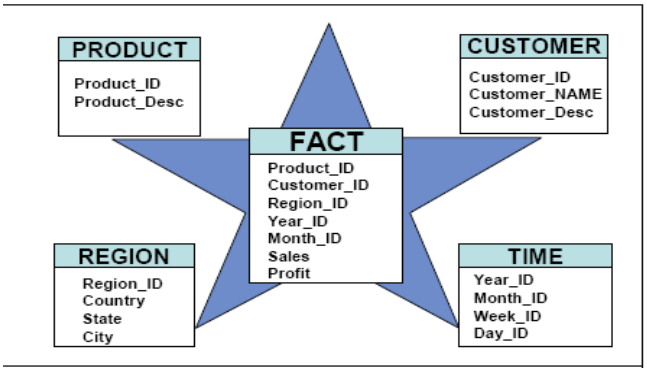
In computing, online analytical processing, or OLAP is an approach to answering multi-dimensional analytical (MDA) queries swiftly OLAP is part of the broader category of business intelligence which also encompasses relational database, report writing and data mining. Typical applicationsof OLAP include business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and similar areas, with new applications coming up, such as agriculture The term OLAP was created as a slight modification of the traditional database term online transaction processing.

Dimensional modeling-

Dimensional modeling (DM) names a set of techniques and concepts used in Dimensional modeling (DM) names a set of techniques and concepts used in datawarehouse design. It is considered to be different from Entity relationship (ER). Dimensional Modeling does not necessarily involve a relational database. The same modeling approach, at the logical level, can be used for any physical form, such as multidimensional database or even flat files. , DM is a design technique for databases intended to support end-user queries in a data warehouse. It is oriented around understandability and performance.

Star Schema

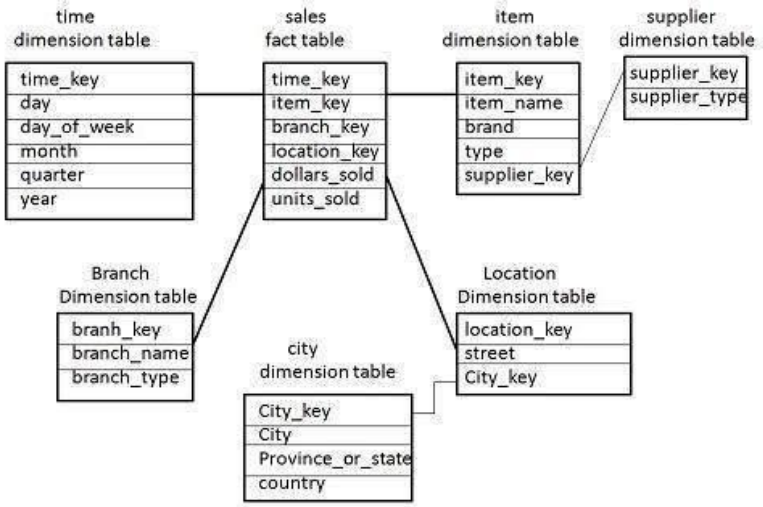
- Fact table is in middle and dimension tables are arranged around the fact table



Snowflake Schema

Normalization and expansion of the dimension tables in a star schema result in the implementation of a snowflake design.

Snowflaking in the dimensional model can impact understandability of the dimensional model and result in a decrease in performance because more tables will need to be joined to satisfy queries



5. Conclusion:

We have studied different schemas of data warehouse, and using the methodology of engineering legacy database, a new data warehouse was built. The normalization was applied wherever required on star schema and snowflake schema was designed.

7. Viva Questions:

- What is data warehouse?
- What is multi-dimensional data?
- What is difference between star and snowflake schema?

8. References:

- PaulrajPonniah, “Data Warehousing: Fundamentals for IT Professionals”, Wiley India
- ReemaTheraja “Data warehousing”, Oxford University Press

Data Warehouse and Mining

Experiment No. : 2

To perform various OLAP operations such as slice, dice, drilldown, rollup, pivot

Experiment No. 2

1. **Aim:** To perform various OLAP operations such as slice, dice, drilldown, rollup, pivot
2. **Objectives:** From this experiment, the student will be able to
    - Discover patterns from data warehouse
    - Online analytical processing of data
    - Obtain knowledge from data warehouse
3. **Outcomes:** Students will be able to discover patterns and knowledge from data warehouse.
4. **Software Required :** Oracle 11g

5. Theory:

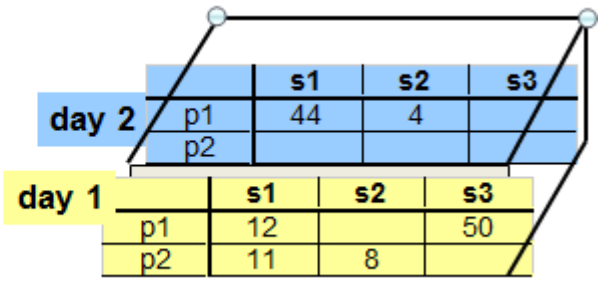
Following are the different OLAP operations

- Roll up (drill-up): summarize data
    - by climbing up hierarchy or by dimension reduction
- Drill down (roll down): reverse of roll-up
    - from higher level summary to lower level summary or detailed data, or introducing new dimensions
- Slice and dice:
    - project and select
- Pivot (rotate):
    - reorient the cube, visualization, 3D to series of 2D planes
    -

Fact table View

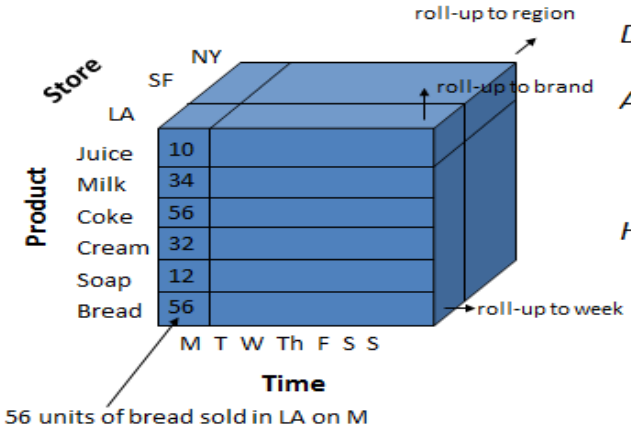
prodId	storeId	date	amt
p1	s1	1	12
p2	s1	1	11
p1	s3	1	50
p2	s2	1	8
p1	s1	2	44
p1	s2	2	4

Multi-dimensional cube



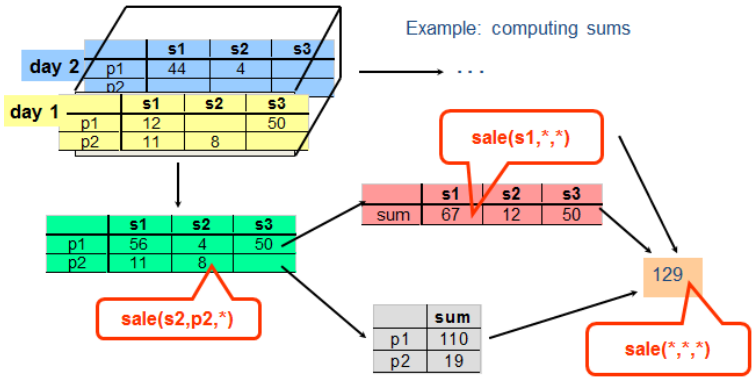
Dimension = 3

Example

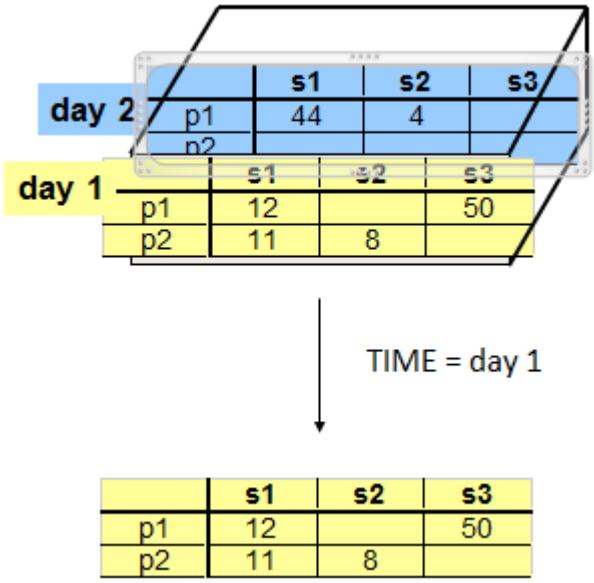


Dimensions:
Time, Product, Store
Attributes:
Product (upc, price, ...)
Store ...
...
Hierarchies:
Product → Brand → ...
Day → Week → Quarter
Store → Region → Country

Cube aggregation – roll up and drill down



Example – slicing



Example – slicing and pivoting

	Sales ($ millions)			
	Products	Time		
		d1	d2	
Store s1	Electronics	$5.2		
	Toys	$1.9		
	Clothing	$2.3		
	Cosmetics	$1.1		
Store s2	Electronics	$8.9		
	Toys	$0.75		
	Clothing	$4.6		
	Cosmetics	$1.5		

	Sales ($ millions)			
	Products	d1		
		Store s1	Store s2	
Store s1	Electronics	$5.2	$8.9	
	Toys	$1.9	$0.75	
	Clothing	$2.3	$4.6	
	Cosmetics	$1.1	$1.5	
Store s2	Electronics			
	Toys			
	Clothing			

2. Conclusion:

OLAP, which performs multidimensional analysis of business data and provides the capability for complex calculations, trend analysis, and sophisticated data modelingis studied.

2. Viva Questions:

- What are OLAP operations?
- What is difference between OLTP and OLAP?
- What is difference between slicing and dicing?

3. References:

- Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd Edition
- M.H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education

Data Warehouse and Mining

Experiment No. : 3

Implementation of Classification algorithm(Decision Tree/ Bayesian)

Experiment No. 3

Aim: Implementation of Classification algorithm(Decision Tree/ Bayesian)

1.
2. **Objectives:** From this experiment, the student will be able to
    - Analyse the data, identify the problem and choose relevant algorithm to apply
    - Understand and implement classical algorithms in data mining
    - Identify the application of classification algorithm in data mining

3. **Outcomes:** Students will be able to understand and implement classical algorithms in data.

4. **Software Required :**JDK for JAVA

5. **Theory:**

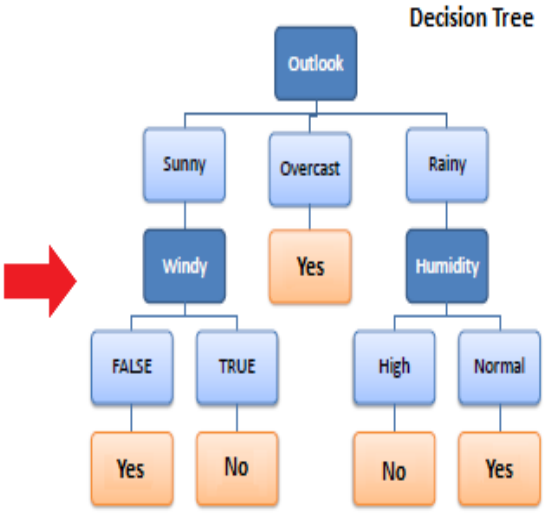
Decision Tree learning is one of the most widely used and practical methods for inductive inference over supervised data. A decision tree represents a procedure for classifying categorical data based on their attributes. It is also efficient for processing large amount of data, so is often use in data mining operations. The construction of decision tree does not require any domain knowledge or parameter setting, and therefore appropriate for exploratory knowledge discovery.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**

The core algorithm for building decision trees called **ID3** by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses *Entropy* and *Information Gain* to construct a decision tree.

Predictors | Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Entropy: A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one. To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

Information Gain: The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches).

6. Procedure/Program:

4. Calculate entropy of the target

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

5. The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy.

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

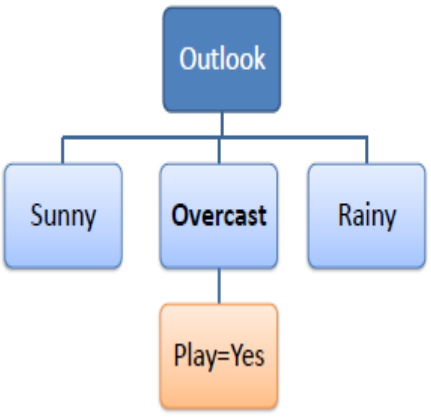
= 0.940 – 0.693 = 0.247

6. Choose attribute with the largest information gain as the decision node

★		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

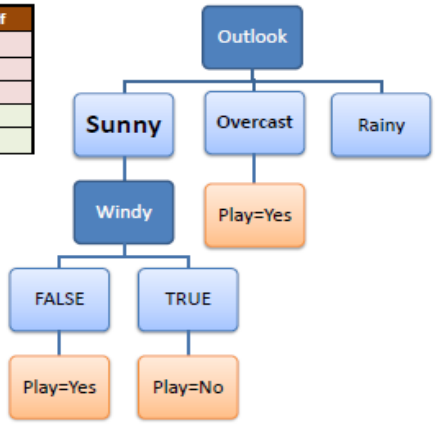
7. A. A branch with entropy of 0 is a leaf node

Temp	Humidity	Windy	Play Golf
Hot	High	FALSE	Yes
Cool	Normal	TRUE	Yes
Mild	High	TRUE	Yes
Hot	Normal	FALSE	Yes
Hot	High	FALSE	Yes



A. A branch with entropy more than 0 needs further splitting.

Temp	Humidity	Windy	Play Golf
Mild	High	FALSE	Yes
Cool	Normal	FALSE	Yes
Mild	Normal	FALSE	Yes
Cool	Normal	TRUE	No
Mild	High	TRUE	No



8. The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

7. Results:

Decision Tree to Decision Rules

A decision tree can easily be transformed to a set of rules by mapping from the root node to the leaf nodes one by one

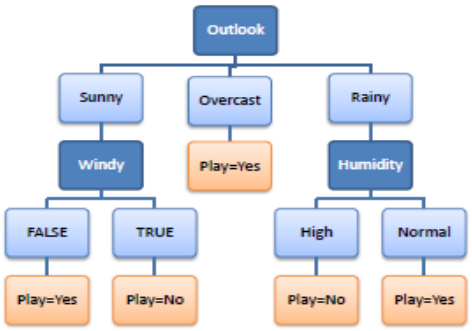
R_1: IF (Outlook=Sunny) AND (Windy=FALSE) THEN Play=Yes

R_2: IF (Outlook=Sunny) AND (Windy=TRUE) THEN Play=No

R_3: IF (Outlook=Overcast) THEN Play=Yes

R_4: IF (Outlook=Rainy) AND (Humidity=High) THEN Play=No

R_5: IF (Outlook=Rain) AND (Humidity=Normal) THEN Play=Yes



8. Conclusion:

The different classification algorithms of data mining were studied and one among them named decision tree (ID3) algorithm was implemented using JAVA. The need for classification algorithm was recognized and understood.

9. Viva Questions:

- What are various classification algorithms?
- What is entropy?
- How does u find information gain?

10. References:

- Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd Edition
- M.H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education

Data Warehouse and Mining

Experiment No. : 4

Implementation of Linear Regression.

Experiment No. 4

1. **Aim:** To implement Linear Regression.

2. **Objectives:** From this experiment, the student will be able to

   - To become familiar with regression methods.

3. **Outcomes:** The learner will be able to

   - How to estimate linear regression coefficients from data.
   - How to make predictions using linear regression for new data.

4. **Software Required:** Java

5. **Theory:**

Linear regression is a statistical approach for modeling relationship between a dependent variable with a given set of independent variables.

Steps for Simple Linear Regression

Step 1: We define

x as **feature vector**, i.e x =$[x_1, x_2, \ldots, x_n]$

y as **response vector**, i.e y = $[y_1, y_2, \ldots, y_n]$

Step 2: Need to find out best fit line for given scatter plot is called regression line.

$$h(x_i) = \beta_0 + \beta_1 x_i \qquad (1)$$

Here,

- $h(X_i)$ represents the **predicted response value** for i^{th} observation.
- β0 and β1are regression coefficients and represent **y-intercept** and **slope** of regression line respectively.

Step 3.Now, to estimate the values of regression coefficients β0 and β1. And once we've estimated these coefficients using Least Square Technique.

Now consider:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = h(x_i) + \epsilon_i \Rightarrow \epsilon_i = y_i - h(x_i) \qquad (2)$$

Here, ε_i is **residual error** in i^{th} observation. So, our aim is to minimize the total residual error.

Step 4. We define the squared error or cost function, J as:

$$J(\beta_0, \beta_1) = \frac{1}{2n}\sum_{i=1}^{n} \epsilon_i^2 \qquad (3)$$

To find the value of β_0 and $\beta_1$for which $J(\beta_0, \beta_1)$ is minimum

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}} \qquad (4)$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \qquad (5)$$

Where,

-SS_{xy} is the sum of cross-deviations of y and x:

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}$$ Equation no. 6

- and SS_{xx} is the sum of squared deviations of x:

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$ Equation no.7

Where, $\bar{x} - mean\ of\ x, \bar{y} -\ mean\ of\ y$

6. Conclusion :

We have studied how to estimate linear regression coefficients from data and how to make predictions using linear regression. Linear regression works well with continuous values.

Viva Questions:

- What is linear regression?
- What is the importance to conversion on text data into statistical manner?

7. References:

- PaulrajPonniah, "Data Warehousing: Fundamentals for IT Professionals", Wiley India
- ReemaTheraja "Data warehousing", Oxford University Press

Data Warehouse and Mining

Experiment No. : 5

Implementation of Clustering algorithm(K-means/ Agglomerative).

Experiment No. 5

Aim: Implementation of Clustering algorithm(K-means/ Agglomerative).

1. **Objectives:**From this experiment, the student will be able to
   - Analyse the data, identify the problem and choose relevant algorithm to apply
   - Understand and implement classical clustering algorithms in data mining
   - Identify the application of clustering algorithm in data mining

2. **Outcomes:** Students will be able to discover patterns and knowledge from data warehouse.

3. **Software Required :**JDK for JAVA

4. **Theory:**

Clustering is dividing data points into homogeneous classes or clusters:

- Points in the same group are as similar as possible
- Points in different group are as dissimilar as possible

When a collection of objects is given, we put objects into group based on similarity.
Clustering Algorithms:

A Clustering Algorithm tries to analyse natural groups of data on the basis of some similarity. It locates the centroid of the group of data points. To carry out effective clustering, the algorithm evaluates the distance between each point from the centroid of the cluster. The goal of clustering is to determine the intrinsic grouping in a set of unlabelled data theory:

K-means Clustering

K-means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining.

5. **Procedure:**

Input:

K: the number of clusters D: a data set containing n objects.

Output: A set of k clusters.

1. Arbitrarily choose K objects from D as the initial cluster centers
2. Partition of objects into k non-empty subsets
3. Identifying the cluster centroids (mean point) of the current partition.
4. Assigning each point to a specific cluster
5. Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
6. After re-allotting the points, find the centroid of the new cluster formed.

6. Conclusion:

The different clustering algorithms of data mining were studied and one among them named k-means clustering algorithm was implemented using JAVA. The need for clustering algorithm was recognized and understood.

7. Viva Questions:

- What are different clustering techniques?
- What is difference between K-means and K-medoids?
- What is dendogram?

8. References:

- Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd Edition
- M.H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education

Data Warehouse and Mining

Experiment No. : 6

Implementation of Association Rule Mining algorithm(Apriori).

Experiment No. 6

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

2. **Objectives:** From this experiment, the student will be able to
   - Analyse the data, identify the problem and choose relevant algorithm to apply
   - Understand and implement classical association mining algorithms
   - Identify the application of association mining algorithms

3. **Outcomes:** Students will be able to understand and implement classical algorithms in data.

4. **Software Required :** JAVA OR Python

5. **Theory:**

   Apriori algorithm is well known association rule algorithm is used in most commercial product. It uses itemset property: Any subset of large item set must be large

6. **Procedure:**

Input:

I = // itemset
D = // db of transactions.
S= // support

Output:

L1

Apiriori Algorithm:

```
K=0;
L= #;
Ci= I;
repeat
  k=k+1;
Lk= #;
for each Ji belong to Ck do
Ci=0;
for each I,j belong to D do
for each Ii belong to tj then
Ci=Ci+1;
for each Ii belong to Ck do
ifCi>=(S*/D/)do
Lk=L U Ii;
   L=L U Lk;
```

Ck+1=Apriori-Gen(Lk) until Ck+1= # ;

7. Conclusion:

The different association mining algorithms of data mining were studied and one among them named Apriori association mining algorithm was implemented using JAVA. The need for association mining algorithm was recognized and understood.

8. Viva Questions:

- What is support and confidence?
- What are differenttypes association mining algorithms?
- What is the disadvantage of apriori algorithm?

9. References:

- Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd Edition
- M.H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education

Data Warehouse and Mining

Experiment No. : 7

Study of Weka Tool

Experiment No. 7

1. **Aim**: Perform data Pre-processing task and Demonstrate performing Classification, Clustering, Association algorithm on data sets using data mining using WEKA tool

1. **Objectives:** From this experiment, the student will be able to
    - Analyse the data, identify the problem and choose relevant algorithm to apply
    - Understand and implement classical algorithms in data mining
    - Understand and implement clustering algorithms in data mining
    - Understand and implement association algorithms in data mining

2. **Outcomes:** Students will be able to understand the analytical operations on data.
3. **Software Required :** WEKA tool.
4. **Theory:**

1.Implementation of ID3 algorithm using WEKA tool.-

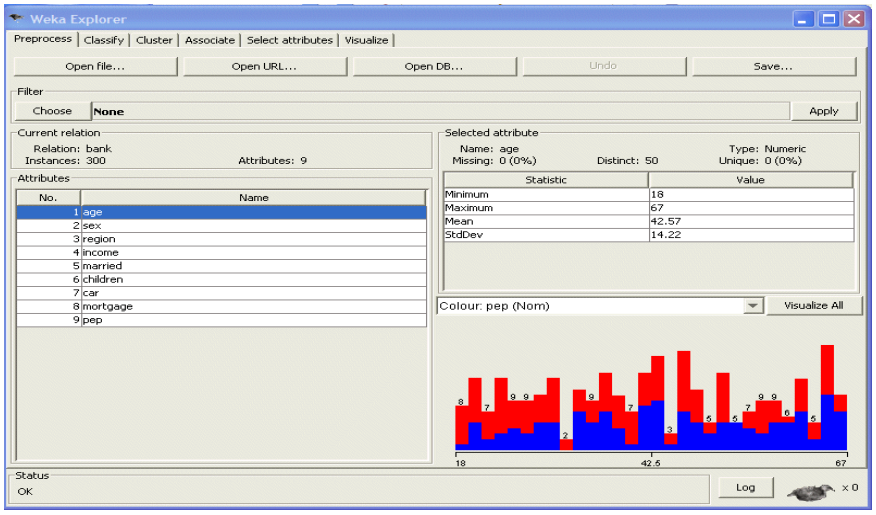
Decision tree learning is a method for assessing the most likely outcome value by taking into account the known values of the stored data instances. This learning method is among the most popular of inductive inference algorithms and has been successfully applied in broad range of tasks such as assessing the credit risk of applicants and improving loyality of regular customers

2. **Procedure:**

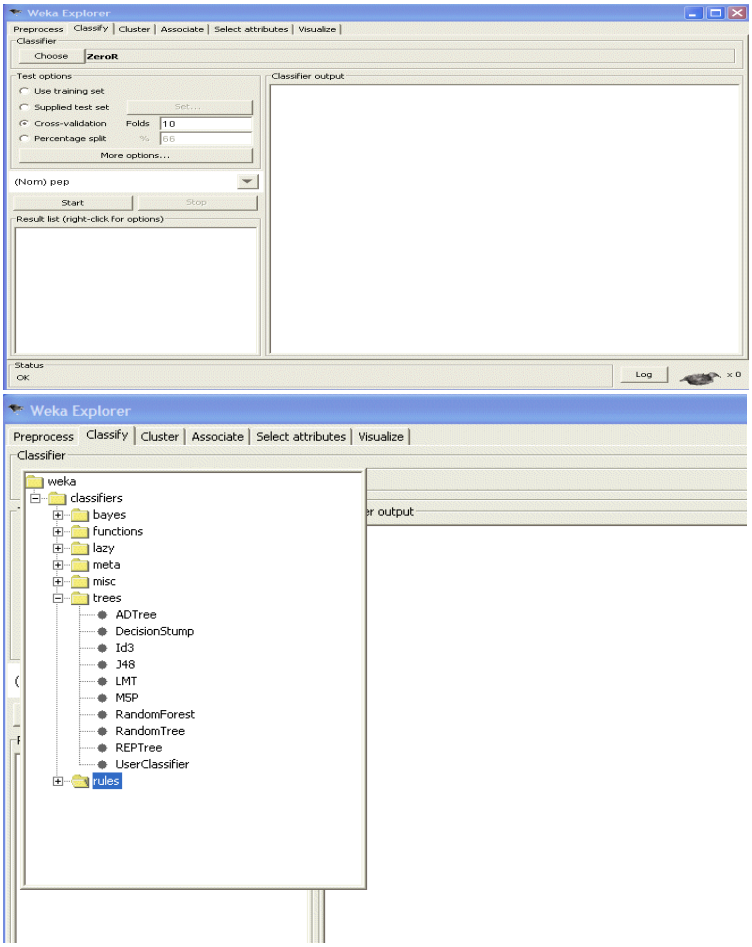
1. Download dataset for implementation of ID3 algorithm (.csv or .arff file). Here bank-data.csvdataset has taken fordecision tree analysis

id	age	sex	region	income	married	children	car	save_act	current_a	mortgage	pep
ID12101	48	FEMALE	INNER_CI	17546	NO	1	NO	NO	NO	NO	YES
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO	YES	YES	NO
ID12103	51	FEMALE	INNER_CI	16575.4	YES	0	YES	YES	YES	NO	NO
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO	YES	NO	NO
ID12105	57	FEMALE	RURAL	50576.3	YES	0	NO	YES	NO	NO	NO
ID12106	57	FEMALE	TOWN	37869.6	YES	2	NO	YES	YES	NO	YES
ID12107	22	MALE	RURAL	8877.07	NO	0	NO	NO	YES	NO	YES
ID12108	58	MALE	TOWN	24946.6	YES	0	YES	YES	YES	NO	NO
ID12109	37	FEMALE	SUBURBAI	25304.3	YES	2	YES	NO	NO	NO	NO
ID12110	54	MALE	TOWN	24212.1	YES	2	YES	YES	YES	NO	NO
ID12111	66	FEMALE	TOWN	59803.9	YES	0	NO	YES	YES	NO	NO
ID12112	52	FEMALE	INNER_CI	26658.8	NO	0	YES	YES	YES	YES	NO
ID12113	44	FEMALE	TOWN	15735.8	YES	1	NO	YES	YES	YES	YES
ID12114	66	FEMALE	TOWN	55204.7	YES	1	YES	YES	YES	YES	YES
ID12115	36	MALE	RURAL	19474.6	YES	0	NO	YES	YES	YES	NO
ID12116	38	FEMALE	INNER_CI	22342.1	YES	0	YES	YES	YES	YES	NO
ID12117	37	FEMALE	TOWN	17729.8	YES	2	NO	NO	NO	YES	NO
ID12118	46	FEMALE	SUBURBAI	41016	YES	0	NO	YES	NO	YES	NO
ID12119	62	FEMALE	INNER_CI	26909.2	YES	0	NO	YES	NO	NO	YES
ID12120	31	MALE	TOWN	22522.8	YES	0	YES	YES	YES	NO	NO
ID12121	61	MALE	INNER_CI	57880.7	YES	2	NO	YES	NO	NO	YES
ID12122	50	MALE	TOWN	16497.3	YES	2	NO	YES	YES	NO	NO
ID12123	54	MALE	INNER_CI	38446.6	YES	0	NO	YES	YES	NO	NO
ID12124	27	FEMALE	TOWN	15538.8	NO	0	YES	YES	YES	YES	NO
ID12125	22	MALE	INNER_CI	12640.3	NO	2	YES	YES	YES	NO	NO
ID12126	56	MALE	INNER_CI	41034	YES	0	YES	YES	YES	YES	NO
ID12127	45	MALE	INNER_CI	20809.7	YES	0	NO	YES	YES	YES	NO
ID12128	39	FEMALE	TOWN	20114	YES	1	NO	NO	YES	NO	YES
ID12129	39	FEMALE	INNER_CI	29359.1	NO	3	YES	NO	YES	YES	NO
ID12130	61	MALE	RURAL	24270.1	YES	1	NO	NO	YES	NO	YES
ID12131	61	FEMALE	RURAL	22942.9	YES	2	NO	YES	YES	NO	NO
ID12132	20	FEMALE	TOWN	16325.8	YES	2	NO	YES	NO	NO	NO

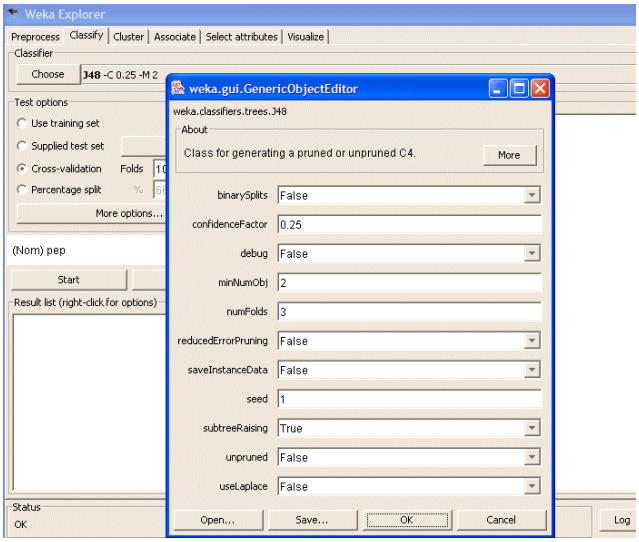
2. Load data in WEKA tool



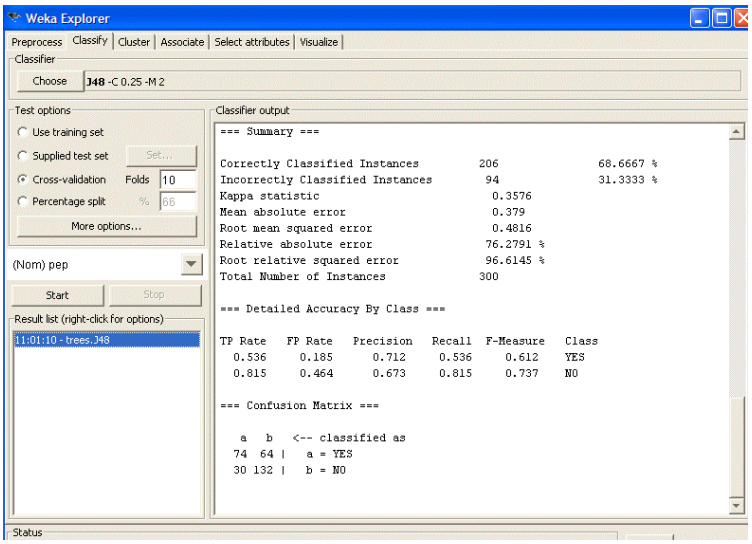
3. Select the "Classify" tab and click the "Choose" button to select the ID3 classifier



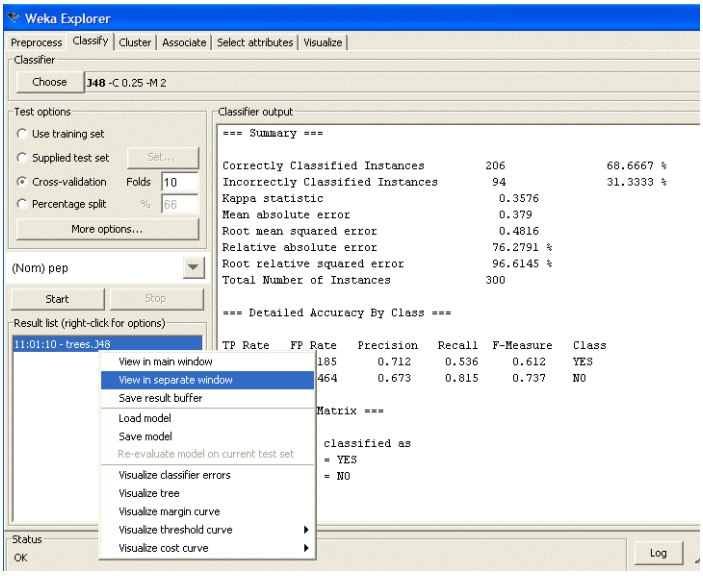
4. Specify the various parameters. These can be specified by clicking in the text box to the right of the "Choose" button. In this example we accept the default values. The default version does perform some pruning (using the sub tree raising approach), but does not perform error pruning



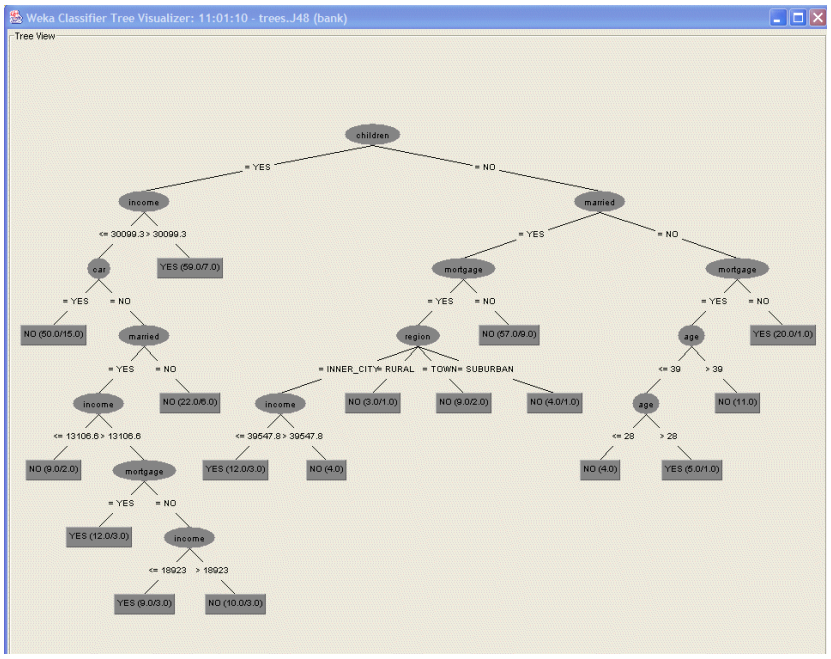
5. Under the "Test options" in the main panel we select 10-fold cross-validation as our evaluation approach. Since we do not have separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model. We now click "Start" to generate the model.



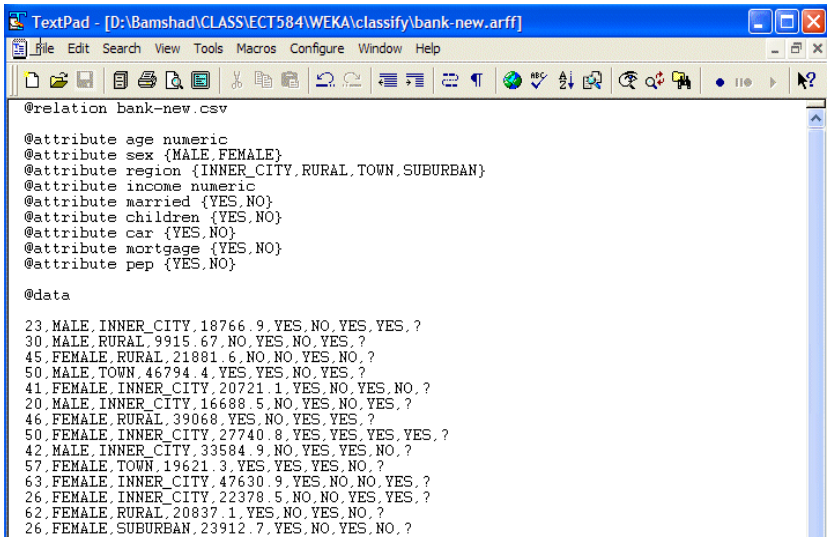
6. We can view this information in a separate window by right clicking the last result set (inside the "Result list" panel on the left) and selecting "View in separate window" from the pop-up menu.



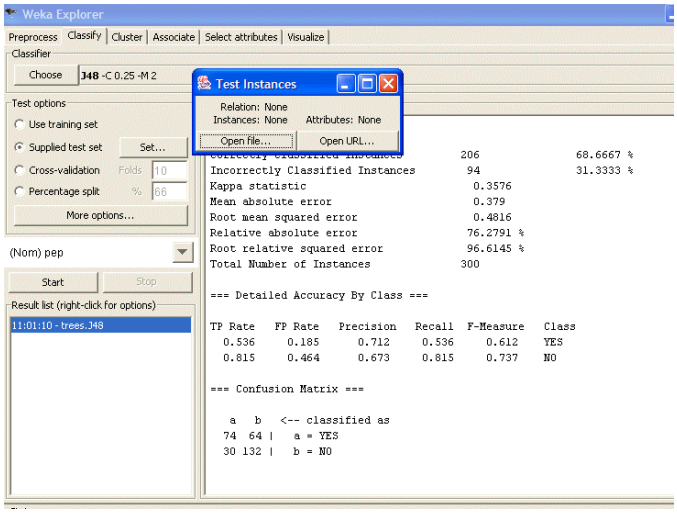
7. WEKA also provides view a graphical rendition of the classification tree. This can be done by right clicking the last result set (as before) and selecting "Visualize tree" from the pop-up menu.



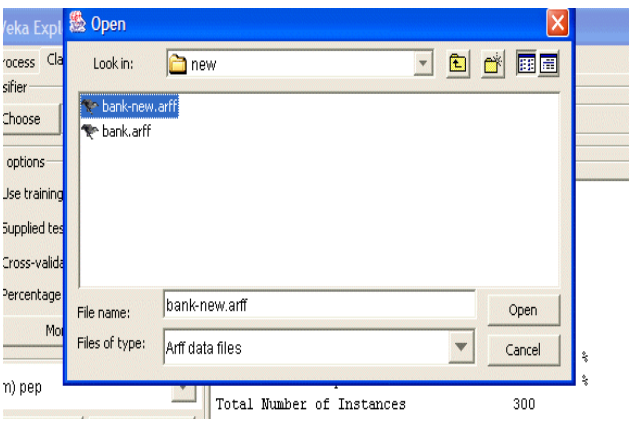
We will now use our model to classify the new instances. However, in data section the value of the "pep" attribute is "?" (orunknown).

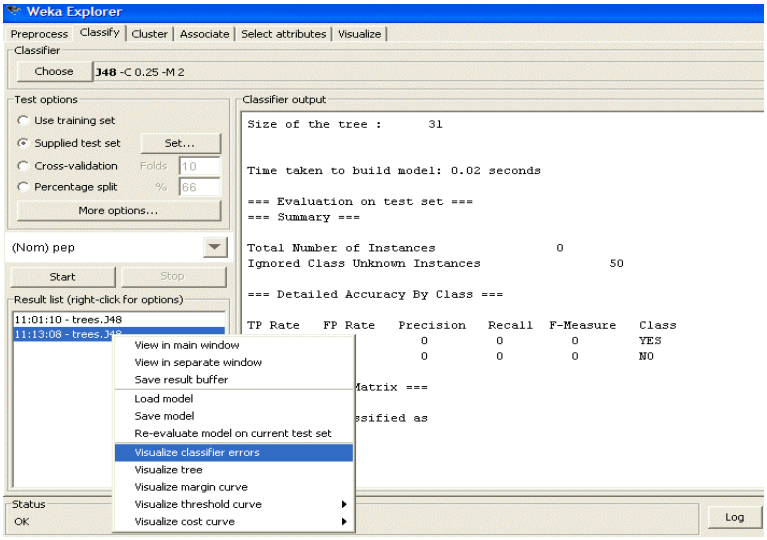


In the main panel, under "Test options" click the "Supplied test set" radio button, and then click the "Set..." button. This will pop up a window which allows you to open the file containing test instances.



In this case, we open the file "bank-new.arff" and upon returning to the main window, we click the "start" button. This, once again generates the models from our training data, but this time it applies the model to the new unclassified instances in the "bank-new.arff" file in order to predict the value of "pep" attribute.

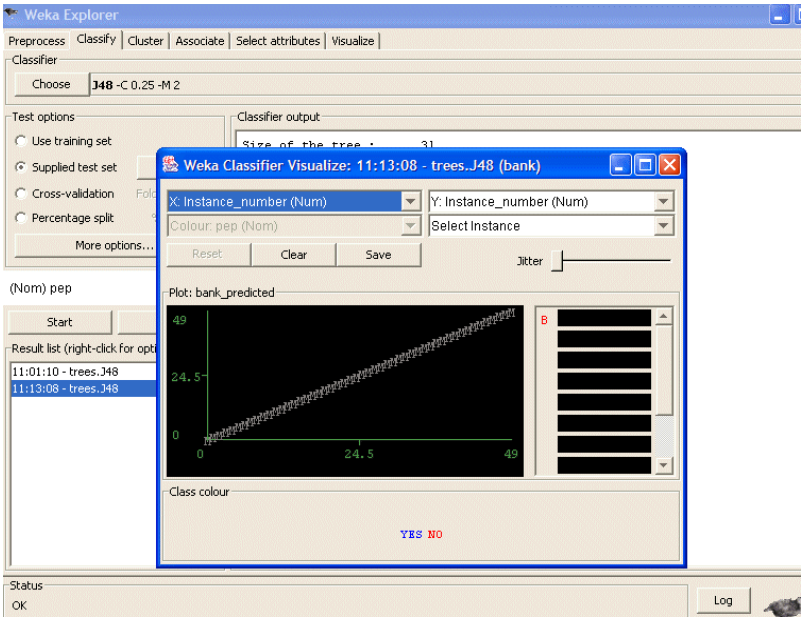




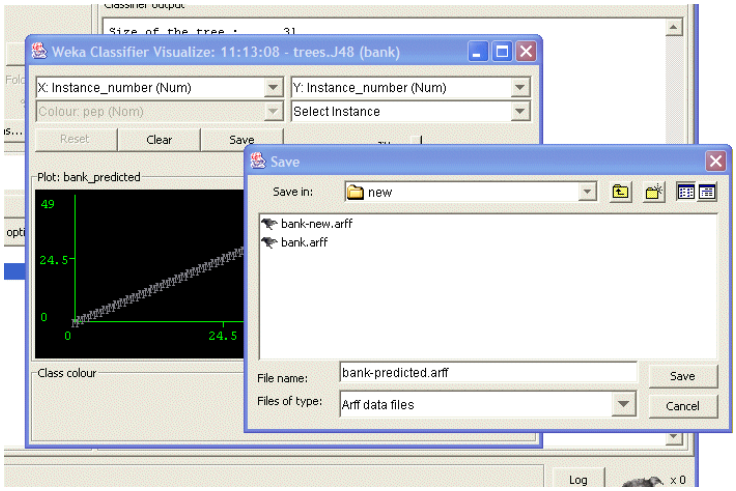
The summary of the results in the right panel does not show any statistics. This is because in our test instances the value of the class attribute ("pep") was left as "?", thus WEKA has no actual values to which it can compare the predicted values of new instances.

GUI version of WEKA is used to create a file containing all the new instances along with their predicted class value resulting from the application of the model.

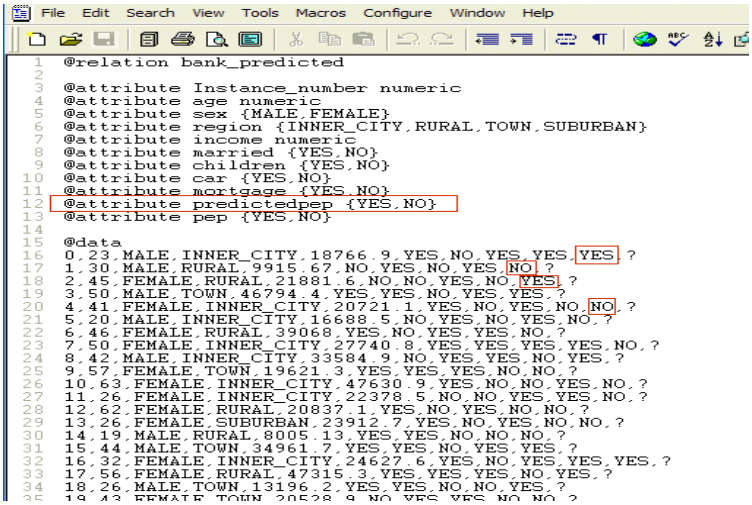
First, right-click the most recent result set in the left "Result list" panel. In the resulting pop-up window select the menu item "Visualize classifier errors". This brings up a separate window containing a two-dimensional graph.



8. To save the file: In the new window, we click on the "Save" button and save the result as the file: "bank-predicted.arff"



This file contains a copy of the new instances along with an additional column for the predicted value of "pep". The top portion of the file can be seen in below figure.



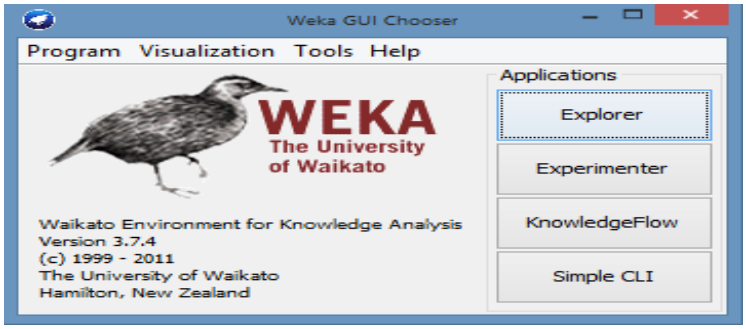
2.To implement the clustering algorithm, K-means using WEKA tool.-

Weka is a landmark system in the history of the data mining and machine learning research communities,because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time

The key features responsible for Weka's success are: –

- It provides many different algorithms for data mining and machine learning.
- Is is open source and freely available.
- It is platform-independent.
- It is easily useable by people who are not data mining specialists.
- It provides flexible facilities for scripting experiments – it has kept up-to-date, with new algorithms

WEKA INTERFACE



The GUI Chooser consists of four buttons—one for each of the four major Weka applications—and four menus.The buttons can be used to start the following applications:

• Explorer: An environment for exploring data with WEKA.

• Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes.

• Knowledge Flow: This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.

• Simple CLI: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

WEKA CLUSTERER

It contains "clusterers" for finding groups of similar instances in a dataset. Some implemented schemes are: *k*-Means, EM, Cobweb, *X*-means, FarthestFirst .Clusters can be visualized and compared to "true" clusters.

1. Procedure:

The basic step of k-means clustering is simple. In the beginning, we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids. Then the K means algorithm will do the three steps below until convergence. Iterate until *stable* (= no object move group):

1. Determine the centroid coordinate

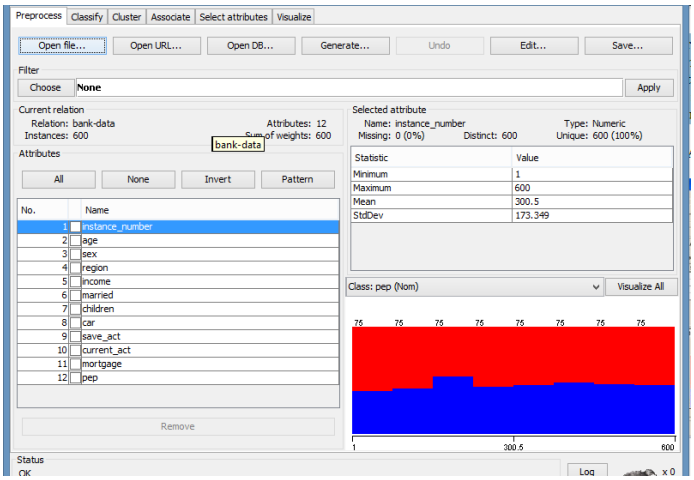
2. Determine the distance of each object to the centroids

3. Group the object based on minimum distance (find the closest centroid)

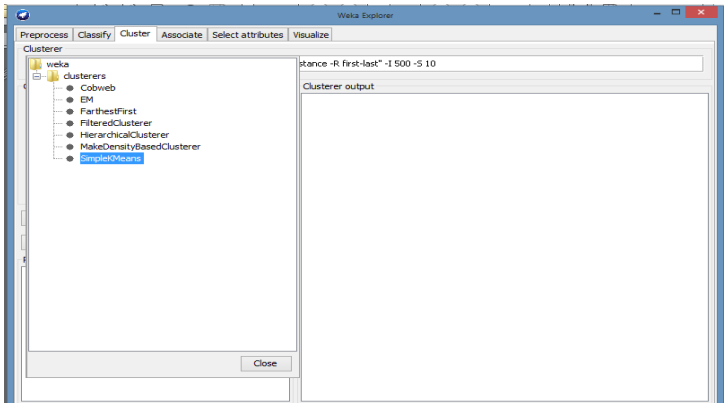
K-means in WEKA 3.7

The sample data set used is based on the "bank data" available in comma-separated format bank-data.csv. The resulting data file is "bank.arff" and includes 600 instances. As an illustration of performing clustering in WEKA, we will use its implementation of the K-

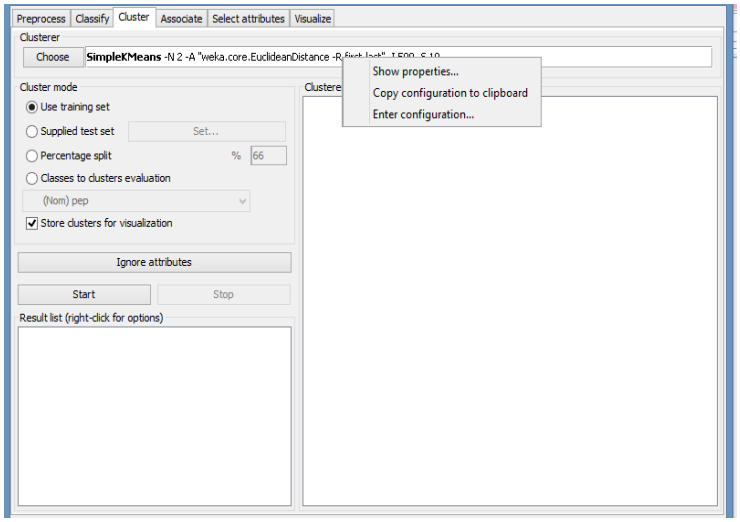
means algorithm to cluster the cutomers in this bank data set, and to characterize the resulting customer segments.



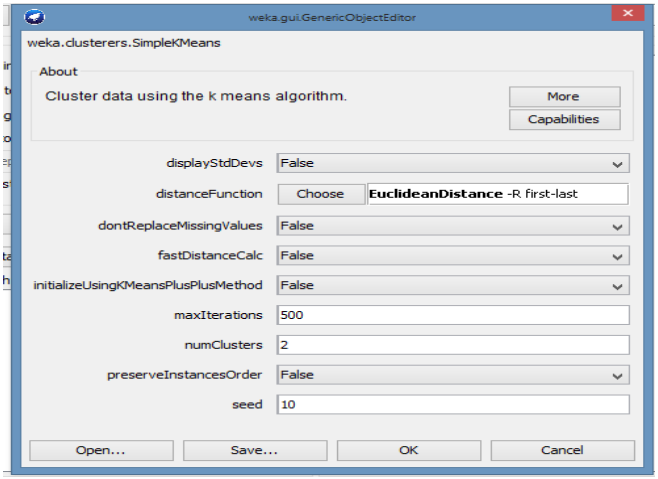
To perform clustering, select the "Cluster" tab in the Explorer and click on the "Choose" button. This results in a drop down list of available clustering algorithms. In this case we select "SimpleKMeans".



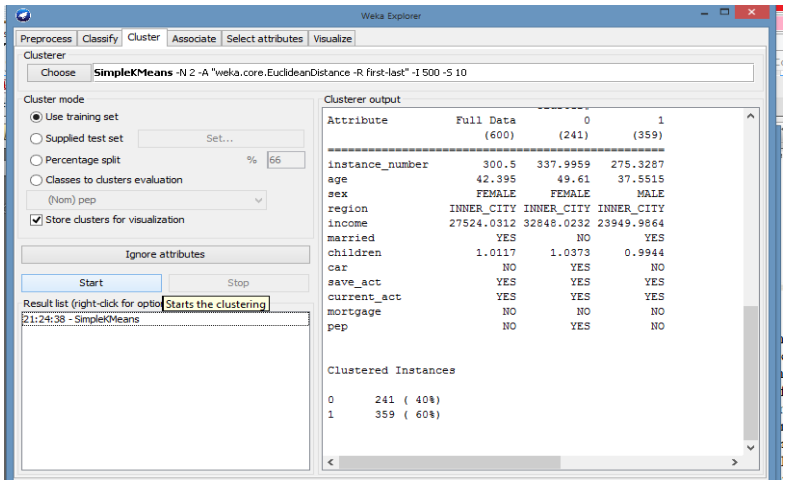
Next, click on the text box to the right of the "Choose" button to get the pop-up window shown below, for editing the clustering parameter.



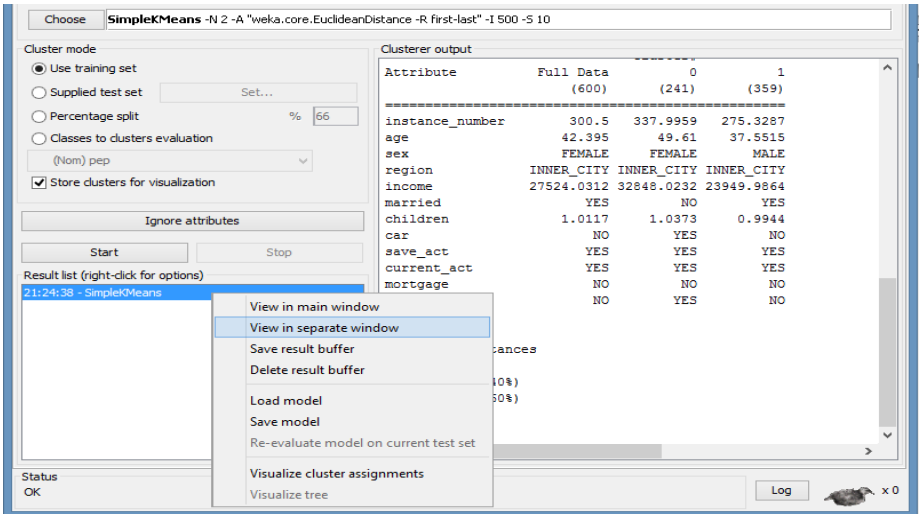
In the pop-up window we enter **2** as the number of clusters and we leave the value of "seed" as is.

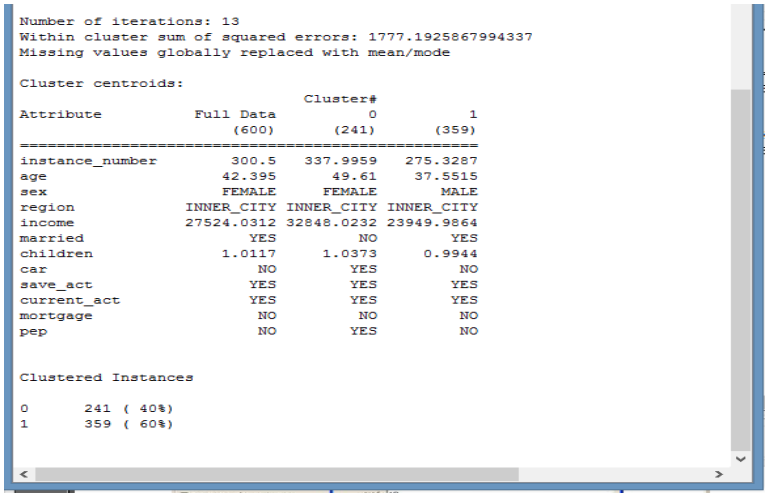


The seed value is used in generating a random number which is, in turn, used for making the initial assignment of instances to clusters.
Once the options have been specified, we can run the clustering algorithm. Here we make sure that in the "Cluster Mode" panel, the "Use training set" option is selected, and we click "Start".

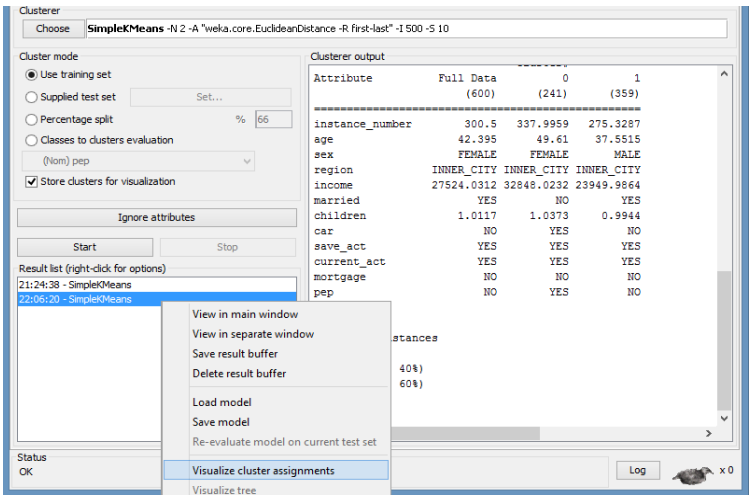


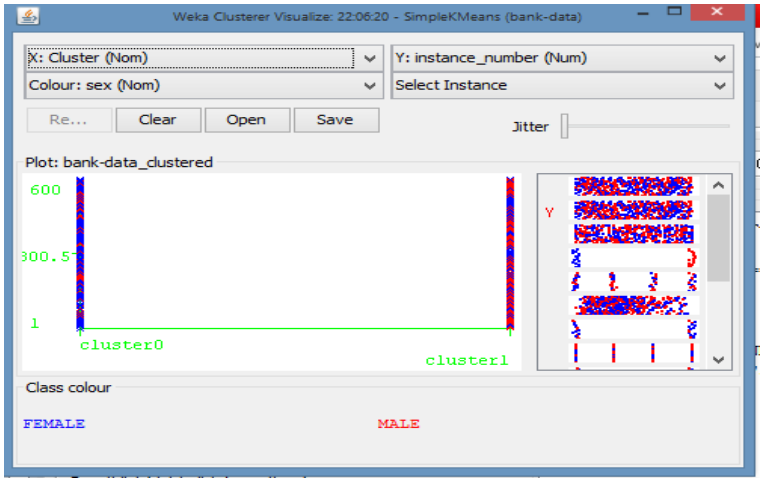
We can right click the result set in the "Result list" panel and view the results of clustering in a separate window.





We can even visualize the assigned cluster as below





You can choose the cluster number and any of the other attributes for each of the three different dimensions available (x-axis, y-axis, and color). Different combinations of choices will result in a visual rendering of different relationships within each cluster.

Note that in addition to the "instance_number" attribute, WEKA has also added "Cluster" attribute to the original data set. In the data portion, each instance now has its assigned cluster as the last attribute value (as shown below).

```
@relation bank-data_clustered

@attribute Instance_number numeric
@attribute instance_number numeric
@attribute age numeric
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income numeric
@attribute married {NO,YES}
@attribute children numeric
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}
@attribute Cluster {cluster0,cluster1}

@data
0,1,48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,NO,YES,cluster0
1,2,40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO,cluster1
2,3,51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO,cluster1
3,4,23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO,cluster1
4,5,57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO,cluster1
5,6,57,FEMALE,TOWN,37869.6,YES,2,NO,YES,YES,NO,YES,cluster0
6,7,22,MALE,RURAL,8877.07,NO,0,NO,NO,YES,NO,YES,cluster1
7,8,58,MALE,TOWN,24946.6,YES,0,YES,YES,YES,NO,NO,cluster1
8,9,37,FEMALE,SUBURBAN,25304.3,YES,2,YES,NO,NO,NO,NO,cluster1
9,10,54,MALE,TOWN,24212.1,YES,2,YES,YES,YES,NO,NO,cluster1
10,11,66,FEMALE,TOWN,59803.9,YES,0,NO,YES,YES,NO,NO,cluster1
11,12,52,FEMALE,INNER_CITY,26658.8,NO,0,YES,YES,YES,YES,NO,cluster0
12,13,44,FEMALE,TOWN,15735.8,YES,1,NO,YES,YES,YES,YES,cluster1
13,14,66,FEMALE,TOWN,55204.7,YES,1,YES,YES,YES,YES,YES,cluster0
14,15,36,MALE,RURAL,19474.6,YES,0,NO,YES,YES,YES,NO,cluster1
```

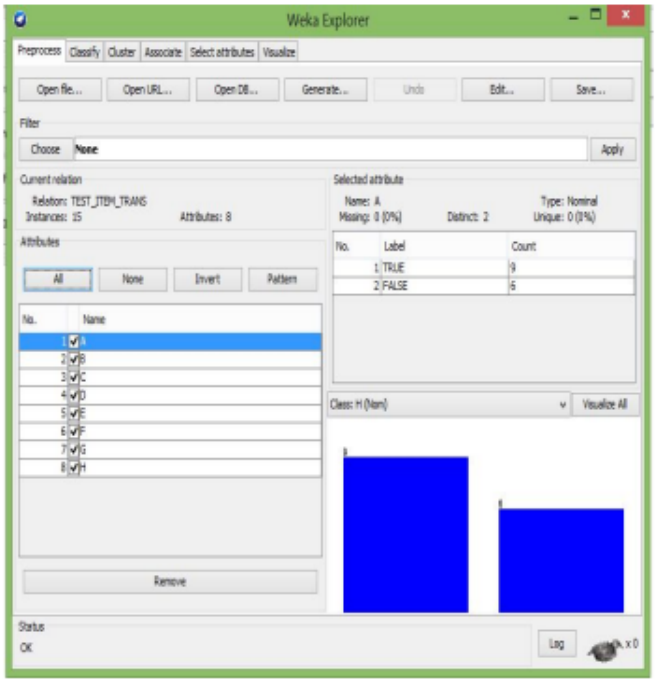
3. Implementation of Apriori algorithm in WEKA.-

WEKA implementation:

To learn the system, TEST_ITEM_TRANS.arff has been used.

Trans ID	Items
1	A,B,C,D,G,H
2	A,B,C,D,E,F,H
3	B,C,D,E,H
4	B,E,G,H
5	A,B,D,E,G,H
6	A,C,F,G,H
7	B,D,E,G,H
8	A,C,D,E,G,H
9	B,C,D,E,H
10	A,C,E,F,H
11	C,E,H
12	A,D,E,F,H
13	B,C,E,F,H
14	A,B,C,F,H
15	A,B,E,F,H

Using the Apriori Algorithm we want to find the association rules that have minSupport=50% and minimum confidence=50%. After we launch the WEKA application and open the TEST_ITEM_TRANS.arff file as shown in below figure.



Then we move to the Associate tab and we set up the configuration as shown below

After the algorithm is finished, we get the following results:
=== Run information ===
Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: TEST_ITEM_TRANS
Instances: 15
Attributes: 8
A B C D E F G H
=== Associator model (full training set) ===Apriori =======
Minimum support: 0.5 (7 instances)
Minimum metric: 0.5 Number of cycles performed: 10
Generated sets of large itemsets:
Size of set of large itemsetsL(1): 10
Size of set of large itemsetsL(2): 12
Size of set of large itemsetsL(3): 3
Best rules found
1. E=TRUE 11 ==> H=TRUE 11 conf:(1)
2. B=TRUE 10 ==> H=TRUE 10 conf:(1)
3. C=TRUE 10 ==> H=TRUE 10 conf:(1)
4. A=TRUE 9 ==> H=TRUE 9 conf:(1)
5. G=FALSE 9 ==> H=TRUE 9 conf:(1)
6. D=TRUE 8 ==> H=TRUE 8 conf:(1)
7. F=FALSE 8 ==> H=TRUE 8 conf:(1)
8. D=FALSE 7 ==> H=TRUE 7 conf:(1)
9. F=TRUE 7 ==> H=TRUE 7 conf:(1)
10. B=TRUE E=TRUE 7 ==> H=TRUE 7 conf:(1)
11. C=TRUE G=FALSE 7 ==> H=TRUE 7 conf:(1)
12. E=TRUE G=FALSE 7 ==> H=TRUE 7 conf:(1)
13. G=FALSE 9 ==> C=TRUE 7 conf:(0.78)
14. G=FALSE 9 ==> E=TRUE 7 conf:(0.78)

15. G=FALSE H=TRUE 9 ==> C=TRUE 7 conf:(0.78)
16. G=FALSE 9 ==> C=TRUE H=TRUE 7 conf:(0.78)
17. G=FALSE H=TRUE 9 ==> E=TRUE 7 conf:(0.78)
18. G=FALSE 9 ==> E=TRUE H=TRUE 7 conf:(0.78)
19. H=TRUE 15 ==> E=TRUE 11 conf:(0.73)
20. B=TRUE 10 ==> E=TRUE 7 conf:(0.7)

1. Conclusion:

The different mining algorithms of data mining were studied and The need for association mining algorithm was recognized and understood. Thus we perform data Pre-processing task and Demonstrate performing Classification, Clustering, Association algorithm on data sets using data mining using WEKA tool

2. References:

- Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd Edition
- M.H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education

Data Warehouse and Mining

Experiment No. : 8

Implementation of Spatial Clustering Algorithm- CLARANS Extensions

Experiment No. 8

1. **Aim:** Implementation of Spatial Clustering Algorithm- CLARANS Extensions
3. **Objectives:** From this experiment, the student will be able to
   - Analyse the data, identify the problem and choose CLARANS algorithm for Spatial Clustering
2. **Outcomes:** Students will be able to understand the clustering operations on spatial data.
3. **. Software Required :** Python /Java
4. **Theory:**

Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatialdatabases. This has three main contributions. First, we propose a new clustering method called CLARANS, whoseaim is to identify spatial structures that may be present in the data. Experimental results indicate that, when compared with existingclustering methods, CLARANS is very efficient and effective. Second, we investigate how CLARANS can handle not only pointsobjects, but also polygon objects efficiently. One of the methods considered, called the IR-approximation, is very efficient in clusteringconvex and nonconvex polygon objects. Third, building on top of CLARANS, we develop two spatial data mining algorithms that aim todiscover relationships between spatial and nonspatial attributes.

Algorithm CLARANS:

1.Input parameters numlocal and maxneighbor. Initi-alizeito 1, andmincostto a large number.

2.Setcurrentto an arbitrary node in$G_{n;k}$.

3.Setjto 1.

4.Consider a random neighborSofcurrent, and basedon 5, calculate the cost differential of the two nodes

5.IfShas a lower cost, setcurrenttoS, and go toStep 3.

6.Otherwise, incrementjby 1. Ifjmaxneighbor,goto Step 4.

7.Otherwise, whenj > maxneighbor, compare the costofcurrentwithmincost.

If the former is less thanmincost, setmincostto the cost ofcurrentand setbestnodetocurrent.

8.Incrementiby 1. Ifi > numlocal, outputbestnodeand halt. Otherwise, go to Step 2.

Steps 3 to 6 above search for nodes with progressively lowercosts. But, if the current node has already been comparedwith the maximum number of the neighbors of the node(specified bymaxneighbor) and is still of the lowest cost, thecurrent node is declared to be a "local" minimum. Then, inStep 7, the cost of this local minimum is compared with thelowest cost obtained so far. The lower of the two costs aboveis stored inmincost. Algorithm CLARANS then repeats tosearch for other local minima, untilnumlocalof them havebeen found.As shown

above, CLARANS has two parameters: themaximum number of neighbors examined (maxneighbor)and the number of local minima obtained (numlocal). Thehigher the value ofmaxneighbor, the closer is CLARANS toPAM, and the longer is each search of a local minima. But,the quality of such a local minima is higher and fewer localminima needs to be obtained. Like many applications ofrandomized search we rely on experiments todetermine the appropriate values of these parameters.

5. **Conclusion:** We have studied Spatial Clustering Algorithm- CLARANS Extensions

6. **Viva Questions:**
   - What is spatial data?
   - What are different types of clustering methods in spatial data mining?
   - What are advantages of CLARANS algorithm?

7. **References:**

   - Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd Edition
   - M.H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education

Data Warehouse and Mining

Experiment No. : 9

Case study on spatial data mining

Experiment No. 9

1. **Aim:** Case study on spatial data mining techniques from recent IEEE papers
2. **Objectives:** Obtain knowledge from spatial data mining techniques

3.Outcomes: Students will be able to understand concept of spatial data mining.

Theory:

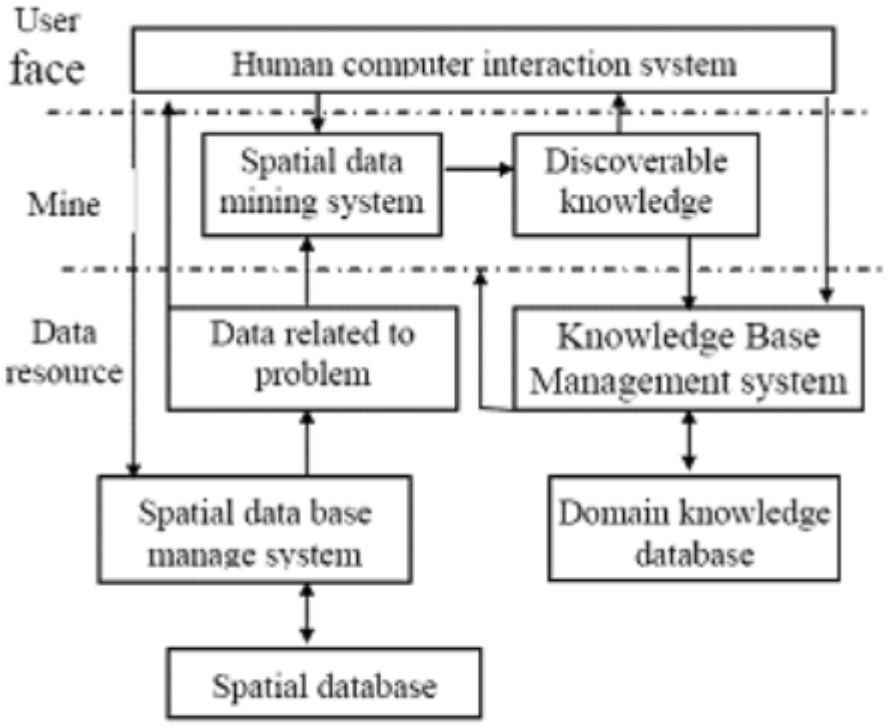
The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns.

Spatial data min ing is a special kind of data mining. The main difference between data mining and spatial data mining is that in spatial data mining tasks we use not only non-spatial attributes (as it is usual in data mining in non-spatial data), but also spatial attributes.Spatial data mining is the process of discovering interesting and previously un-known, but potentially useful patterns from large spatial datasets. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation.Specific features of geographical data that preclude the use of general purpose data mining algorithms are: •rich data types (e.g., extended spatial objects) •implicit spatial relat ionships among the variables•observations that are not independent, and spatial autocorrelation among the features.

Preprocessing spatial data: Spatial data mining techniques have been widely applied to the data in many application domains. However, research on the preprocessing of spatial data has lagged behind. Hence, there is a need for preprocessing techniques for spatial data to deal with problems such as treatment of missing location information and imprecise location specifications, cleaning of spatial data, feature selection, and data transformation.

The systematic structure of spatial data mining:

The spatial data mining can be used to understand spatial data, discover the relation between space and the non- space data, set up the spatial knowledge base, excel the query, reorganize spatial database and obtain concise total characteristic etc. The system structure of the spatial data mining can be divided into three layer structures mostly, such as the Figure 1 show .The customer interface layer is mainly used for input and output, the miner layer is mainly used to manage data, select algorithm and storage the mined knowledge, the data source layer, which mainly includes the spatial database and other related data and knowledge bases, is original data of the spatial data mining.



'igure 1: Systematic structure of spatial data mining

Spatial data mining tas ksBasic tasks of spatial data mining are:

•classification– finds a set of rules which determine the class of the classified object according to its attributes e. g. ”IF population of city = high AND economic power of city = high THEN unemployment of city = low” or classification of a pixel into one of classes, e. g. water, field, forest.

•association rules– find (spatially related) rules from the database. Association rules describe patterns, which are often in the database. The association rule has the following form: A →B(s%; c%), where s is the support of the rule (the probability, that A and B hold

together in all the possible cases) and c is the confidence (the conditional probability that B is true under the condition of A e. g. ”if the city is large, it is near the river (with probability 80%)” or ”if the neighboring pixels are classified as water, then central pixel is water (probability 80%).”

•characteristic rules– describe some part of database e. g. ”bridge is an object in the place where a road crosses a river.”

•discriminant rules– describe differences between two parts of database e. g. find differences between cities with high and low unemployment rate.

•clustering– groups the object from database into clusters in such a way that object in one cluster are similar and objects from different clusters are dissimilar e. g. we can find clusters of cities with similar level of unemployment or we can cluster pixels into similarity classes based on spectral characteristics.

•trend detection– finds trends in database. A trend is a temporal pattern in some time series data. A spatial trend is defined as a pattern of change of a non-spatial attribute in the neighborhood of a spatial object e. g. ”when moving away from Brno, the unemployment rate increases” or we can find changes of pixel classification of a given area in the last five years.

4. Procedure:

Students should decide their case study topics and attach one IEEE paper on their case study.

5. Conclusion: Different types of spatial data mining techniques are studied.

6. Viva Questions:

- What is spatial data mining ? How it is different than traditional data mining?
- What are different types of spatial data mining techniques?

7. References:

- Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd Edition

- M.H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education