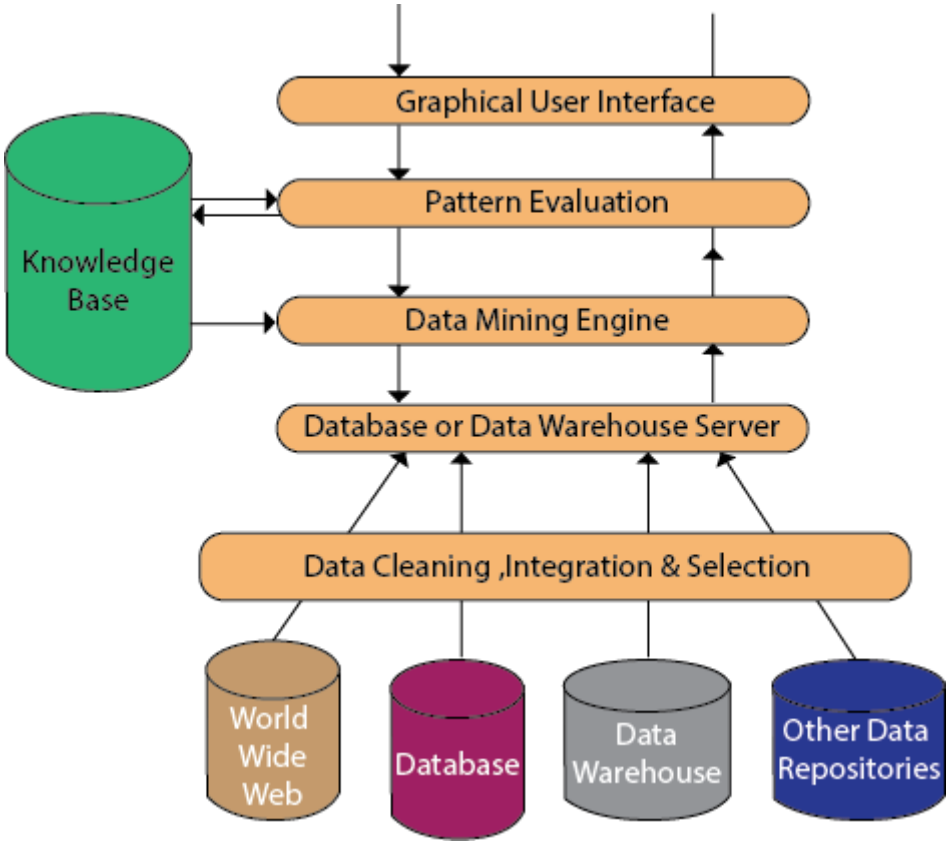
Name : Shruti Chaube
Roll No: 66

Assignment no: 01

Q1. With a neat labelled diagram explain the architecture of Typical Data Mining System.

Ans.



Data Mining Architecture

The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface, and knowledge base.

Data Source:

The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. You need a huge amount of historical data for data mining to be successful. Organizations typically store data in databases or data warehouses. Data warehouses may comprise one or more databases, text files spreadsheets, or other repositories of data. Sometimes, even plain text files or spreadsheets may contain information. Another primary source of data is the World Wide Web or the internet.

Different processes:

Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires to be cleaned and unified. More information than needed will be collected from various data sources, and only the data of interest will have to be selected and passed to the server. These procedures are not as easy as we think. Several methods may be performed on the data as part of selection, integration, and cleaning.

Database or Data Warehouse Server:

The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

Data Mining Engine:

The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.

In other words, we can say data mining is the root of our data mining architecture. It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

Pattern Evaluation Module:

The Pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value. It collaborates with the data mining engine to focus the search on exciting patterns.

This segment commonly employs stake measures that cooperate with the data mining modules to focus the search towards fascinating patterns. It might utilize a stake threshold to filter out discovered patterns. On the other hand, the pattern evaluation module might be coordinated with the mining module, depending on the implementation of the data mining techniques used. For efficient data mining, it is abnormally suggested to push the evaluation of pattern stake as much as possible into the mining procedure to confine the search to only fascinating patterns.

Graphical User Interface:

The graphical user interface (GUI) module communicates between the data mining system and the user. This module helps the user to easily and efficiently use the system without knowing the complexity of the process. This module cooperates with the data mining system when the user specifies a query or a task and displays the results.

Knowledge Base:

The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns. The knowledge base may even

contain user views and data from user experiences that might be helpful in the data mining process. The data mining engine may receive inputs from the knowledge base to make the result more accurate and reliable. The pattern assessment module regularly interacts with the knowledge base to get inputs, and also update it.

Q2. List and describe five primitives for specifying data mining task.

Ans. The data mining primitives specify the following:

The set of task-relevant data to be mined: This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest (referred to as the relevant attributes or dimensions).

The kind of knowledge to be mined: This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

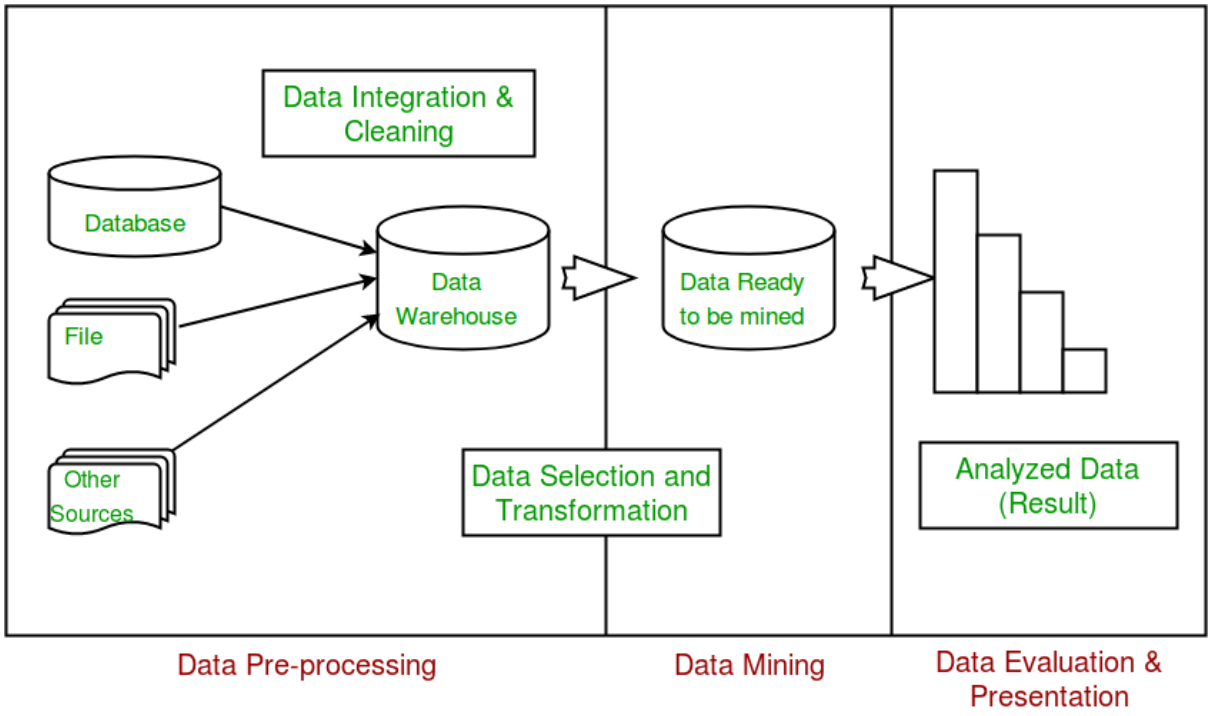
The background knowledge to be used in the discovery process: This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found. Concept hierarchies are a popular form of back- ground knowledge, which allow data to be mined at multiple levels of abstraction. An example of a concept hierarchy for the attribute (or dimension) age is shown in Figure 1.2. User beliefs regarding relationships in the data are another form of back- ground knowledge.

The interestingness measures and thresholds for pattern evaluation: They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures. For example, interestingness measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.

The expected representation for visualizing the discovered patterns: This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes.

Q3. Explain data mining. Give 2 examples where data mining can be used.

Ans. Data mining in the information system is like mining the earth. While mining finds out the hidden valuables in the earth, data mining not only finds out but also provides analysis of the hidden patterns of data in a data warehouse. Data mining aims at exploring knowledge from, data warehouses it organizes data in a manner so that it can derive the inherent meaning to contribute in the knowledge base. The data from a database or a data warehouse is first sorted to prepare the target data and then analysed to find out the structure, correlations and the meaning that it contains.



As shown in Figure, data mining contributes in exploring hidden patterns of data and enriches business KNOWLEDGE. Data mining has its application in almost every business area

Applications of Data Mining:

1.Data mining applications for Finance

2.Data mining applications for Healthcare

3.Data mining applications for Intelligence

4.Data mining applications for Telecommunication

5. Data mining applications for Energy

Q4. Explain different steps involved in Data Preprocessing.

Ans. 1. Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

2. Data integration: using multiple databases, data cubes, or files.

3.Data transformation: normalization and aggregation.

4.Data reduction: reducing the volume but producing the same or similar analytical results.

5.Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

Data cleaning

1. Fill in missing values (attribute or class value):

   • Ignore the tuple: usually done when class label is missing.

   • Use the attribute mean (or majority nominal value) to fill in the missing value

   • Use the attribute mean (or majority nominal value) for all samples belonging to the same class.

2. Predict the missing value by using a learning algorithm: consider the attribute with the missing value as a dependent (class) variable and run a learning algorithm (usually Bayes or decision tree) to predict the missing value. Identify outliers and smooth out noisy data

   • Binning

      1. Sort the attribute values and partition them into bins.
      2. Then smooth by bin means, bin median, or bin boundaries.

   • Clustering: group values in clusters and then detect and remove outliers (automatic or manual)

   • Regression: smooth by fitting the data into regression functions.

3. Correct inconsistent data: use domain knowledge or expert decision.

Data transformation

1. Normalization:

   • Scaling attribute values to fall within a specified range. Example: to transform V in [min, max] to V' in [0,1], apply V'=(V-Min)/(Max-Min)

   • Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers): V'=(V-Mean)/StDev

2. Aggregation: moving up in the concept hierarchy on numeric attributes.
3. Generalization: moving up in the concept hierarchy on nominal attributes.
4. Attribute construction: replacing or adding new attributes inferred by existing attributes.

Data reduction

1. Reducing the number of attributes • Data cube aggregation: applying roll-up, slice or dice operations. • Removing irrelevant attributes: attribute selection (filtering and wrapper methods), searching the attribute space

   • Principle component analysis (numeric attributes only): searching for a lower dimensional space that can best represent the data.

2. Reducing the number of attribute values

   • Binning (histograms): reducing the number of attributes by grouping them into intervals (bins).

- Clustering: grouping values in clusters.
- Aggregation or generalization

3. Reducing the number of tuples
   - Sampling

Q5. Describe the different types of attributes one may come across in data mining set with two examples for each type.

Ans.

1. **Nominal Attribute:**
   Nominal Attributes only provide enough attributes to differentiate between one object and another. Such as Student Roll No., Sex of the Person.
2. **Ordinal Attribute:**
   The ordinal attribute value provides sufficient information to order the objects. Such as Rankings, Grades, Height
3. **Binary Attribute:**
   These are 0 and 1. Where 0 is the absence of any features and 1 is the inclusion of any characteristics.
4. **Numeric attribute:** It is quantitative, such that quantity can be measured and represented in integer or real values, are of two types
   Interval Scaled attribute:
   It is measured on a scale of equal size units, these attributes allows us to compare such as temperature in C or F and thus values of attributes have order.
5. **Ratio Scaled attribute:**
   Both differences and ratios are significant for Ratio. For e.g. age, length, Weight.

Q6. Consider the datapoint :
13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70
Find mean, Median, mode, midrange, Q1,Q3, IQR, boxplot, outliers.

Ans.

Mean = 29.963

Median = 25

Mode = 35, 25 (each appeared 4 times)

Midrange = 57

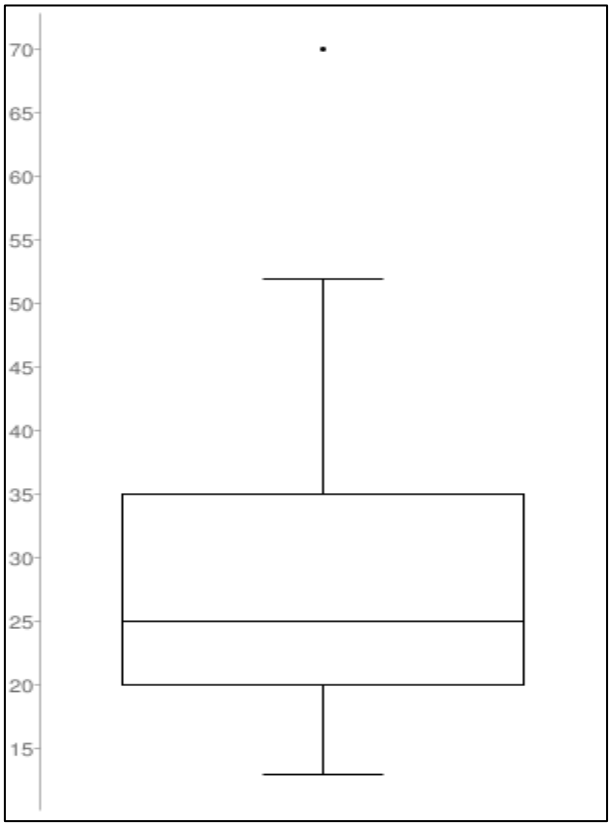
Q1 = 20

Q3 = 35

IQR = 15

Outlier = 70

Boxplot =



Q7. List different visualization techniques used in data mining.

Ans. 1.Pixel oriented visualization techniques:

- A simple way to visualize the value of a dimension is to use a pixel where the color of the pixel reflects the dimension's value.
- For a data set of m dimensions pixel oriented techniques create m windows on the screen, one for each dimension.
- The m dimension values of a record are mapped to m pixels at the corresponding
- position in the windows.
- The color of the pixel reflects other corresponding values.
- Inside a window, the data values are arranged in some global order shared by all windows
- Eg: All Electronics maintains a customer information table, which consists of 4 dimensions: income, credit_limit, transaction_volume and age. We analyze the correlation between income and other attributes by visualization.
- We sort all customers in income in ascending order and use this order to layout the customer data in the 4 visualization windows as shown in fig.
- The pixel colors are chosen so that the smaller the value, the lighter the shading.

2.Geometric Projection visualization techniques

- A drawback of pixel-oriented visualization techniques is that they cannot help us much in understanding the distribution of data in a multidimensional space.
- Geometric projection techniques help users find interesting projections of multidimensional data sets.
- A scatter plot displays 2-D data point using Cartesian co-ordinates. A third dimension can be added using different colors of shapes to represent different data points.
- Eg. Where x and y are two spatial attributes and the third dimension is represented by different shapes
- Through this visualization, we can see that points of types “+” &”X” tend to be collocated.

Q8. Illustrate using Bayesian classification technique, show how we can classify a new tuple with (Homeowner =yes; status=employed; income=average)

Sr. no.	Homeowner	Status	income	Defaulted
1	yes	employed	high	no
2	no	business	average	no
3	no	employed	low	no
4	yes	unemployed	high	no
5	no	unemployed	average	yes
6	no	business	low	no
7	yes	unemployed	high	no
8	no	employed	average	yes
9	no	business	low	no
10	no	employed	average	yes

P(yes) = 3/10 = 0.3

P(no) = 7/10 = 0.7

Homeowner

P(yes| no) = 3/7

P(yes | yes) = 0/3

P(no | yes) = 3/3

P(no | no) = 4/7

Status

P(employed | yes) = 2/3

P(employed | no) = 2/7

P(unemployed | yes) = 1/3

P(unemployed | no) = 2/7

P(business | yes) = 0/3

P(business | no) = 3/7

Income

P(high | yes) = 0/3

P(high | no) = 3/7

P(low | yes) = 0/3

P(low | no) = 3/3

P(average | yes) = 3/3

P(average | yes) = 1/3

X=(Homeowner =yes; status=employed; income=average)

P(X | Yes) * P(yes)

= P(yes | yes) * P(employed | yes) * P(employed | yes) * P(average | yes) * P(yes)

= 0 * (2/3) * (3/3) * (3/3) * (3/10)

= 0

P(X | no) * P(no)

= P(yes | no) * P(employed | no) * P(employed | no) * P(average | no) * P(no)

= (3/7) * (2/7) * (1/7) * (7/10)

= 0.0122

Since, P(X | Yes) * P(yes) < P(X | no) * P(no)

i.e., 0 < 0.0122

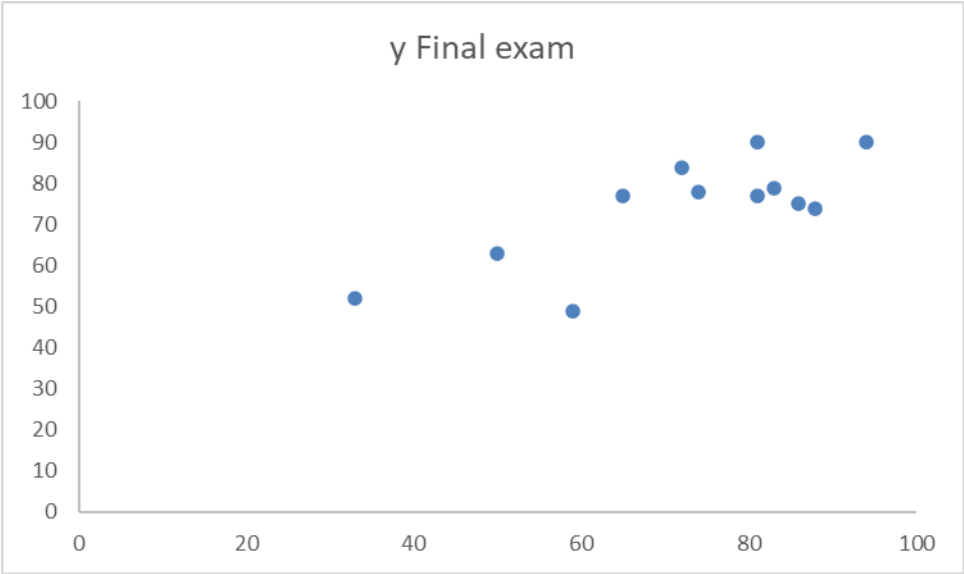
Therefore, the naive Bayesian classifier predicts Defaulted = “No” for sample X.

Q9. The following table shows the midterm and final exam grades obtained for students in a database course.

x Midterm exam	y Final exam
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

(a) Plot the data. Do x and y seem to have a linear relationship?

Ans.



X and Y seem to have a moderately Positive relation with r = 0.78.

(b) Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course.

Ans.

X	Y	X^2	Y^2	XY
72	84	5184	7056	6048
50	63	2500	3969	3150
81	77	6561	5929	6237
74	78	5476	6084	5772
94	90	8836	8100	8460
86	75	7396	5625	6450
59	49	3481	2401	2891
83	79	6889	6241	6557
65	77	4225	5929	5005
33	52	1089	2704	1716
88	74	7744	5476	6512
81	90	6561	8100	7290
∑X = 866	∑Y = 888	∑X^2 = 65942	∑Y^2 = 67614	∑XY = 66088

The Least Squares Regression Line

The least squares regression line is given by $\hat{y} = a + bx$ where

$$\text{slope} = b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$

$$\text{y – intercept} = a = \bar{y} - b\bar{x}$$

Slope= b = [12 (66088) - (866) (888)] / [12 (65942) - (866) ^ 2]

=0.5816

a = (888/12) – (0.5812)*(866/12)

= 32.0367

Least squares regression line for regressing final exam scores, y , on midterm exam scores, x , is given by y = 32.0367 + 0.5816x.

(c) Predict the final exam grade of a student who received an 86 on the midterm exam.
Ans. Here value of x = 86

y = 32.0367 + 0.5816x

y = 32.0367 + 0.5816 (86)

y = 82.0543

Q10. Some nonlinear regression models can be converted to linear models by applying transformations to the predictor variables. Show how the nonlinear regression equation y = αX β can be converted to a linear regression equation solvable by the method of least squares.

Ans. To convert $Y = Ae^{bX}u$ to a linear equation, take the natural log of both sides:

$\ln(Y) = \ln(Ae^{bX}u)$ Apply the rules above and you get:

$\ln(Y) = \ln(A) + bX + \ln(u)$

To implement this, create a new variable y = ln(Y). (The Y in the original equation is "big Y." The newvariable is "little y.")

Also, define v as equaling ln(u) and a as equaling ln(A).

Substitute y for ln(Y), a for ln(A), and v for ln(u) in ln(Y) = ln(A) + bX + ln(u)

and you get: y= a + bX + v

This is a linear equation! It may seem like cheating to turn a curve equation into a straight-line equationthis way, but that is how it's done.

Doing the analysis this way is not exactly the same as fitting a curve to points using least squares. This method this method minimizes the sum of the squares of ln(Y) - (ln(A) + bX), not the sum of the squares of$Y - Ae^{bX}$. In our equation, the error term is something we multiply by rather than add. Our equation is $Y=Ae^{bX}u$, rather than $Y = Ae^{bX}+u$.

We make our equation $Y=Ae^{bX}u$, rather than $Y = Ae^{bX}+u$, so that when we take the log of both sides we getan error, which we call v, the logarithm of u in the original equation, that behaves like errors are supposed to behave for linear regression. For example, for linear regression we have to assume that the error's expected value is 0. To make this work, we assume that the expected value of u is 1. That makes the expected value of ln(u) equal zero, because ln(1) = 0. In $Y=Ae^{bX}u$, u is never 0 or negative, but v can takeon positive or negative values, because if u is less than 1, v=ln(u) is less than 0.

When you transform $Y=Ae^{bX}u$ to y = a + bX + v and do a least squares regression, here is how you interpret the coefficients: b̂is your estimate of the growth rate. â is your estimate of ln(A). To get an estimate of A, calculate e to the a-hat power.