VIT

(Autonomous College Affiliated to University of Mumbai)

Mid Semester Examination (CBSGS-C scheme) -(2022-23)

Date

Semester: V	Time: 1 Hr. & 15 Mints	Branch: CMPN
	Subject: DWM	Marks: 30

Q.		CO	BL
1)	Attempt any Five (2 Marks Each)		
a)	Define the terms Data Mining & KDD.	2	1
b)	Consider a dataset, D = {8, 11, 12, 12, 23, 29, 29, 35, 38, 45, 46, 66}. Give five number summary of this dataset.	2	3
c)	Consider two objects, A = {19, 22, 07, 34} and B = {23, 12, 17, 30}. Calculate Minkovaski distance (h=3) and supremum distance between these objects.	2	3
d)	Consider a dataset, D = {4, 8, 15, 21, 21, 24, 25, 28, 34}. Perform smoothing by bin boundaries. Let there be 3 bins and are equal frequency.	2	3
e)	Normalize the following group of data: 200,300,400,600,1000 using min-max normalization by setting min=0 and max=1	2	3
f)	What are the limitations of ID3 tree induction algorithm?	3	2
g)	Define the terms: Precision & Recall with respect to evaluation of classifiers	4	1
h)	Define the terms: Support & Confidence with respect to Association Analysis	3	1
Q. 2)	Attempt any two. (5 Marks Each)		
a)	Explain the architecture of a typical Data Mining system with neat diagram.	2	1
b)	Consider the following training set to calculate the Information gain for any one attribute using attribute relevance analysis.	3	3

Name	Hair	Height	Weight	Dublin	Result
Sarah	Blonde	Average	Light	No	Sunburned
Dana	Blonde	Tall	Average	Yes	None
Alex	Brown	Short	Average	Yes	None
Annie	Blonde	Short	Average	No	Sunburned
Emily	Red	Average	Heavy	No	Sunburned
Pete	Brown	Tall	Heavy	No	None
John	Brown	Average	Heavy	No	None
Katie	Brown	Short	Light	Yes	None

c)	Consider the training set in Q2 B, Classify a new sample X={Blonde, Short, Light, No} using the Naïve Bayes Classifier.	3	3
Q 3)	Attempt any One (10 Marks Each)		
a)	Explain different steps involved in data preprocessing	2	1
b)	Explain different data mining tasks in brief.	2	1

CO1	Understand data warehouse fundamentals, designdimensional model and apply OLAP operations.
CO2	Understand data mining principles, perform Data preprocessing and Visualization.
CO3	Identify & implement appropriate data mining algorithms to solve real world problems
CO4	Compare and evaluate different data mining techniques like classification, prediction, clustering and association rule mining
CO5	Describe complex information and social networks with respect to web mining.

VIT Vidyalankar Institute of Technology Accredited A+ by NAAC

Internal Assessment Test

Branch	Test Date	Semester	Div.	Roll No.	Student's Signature
CMPN	r	V	A		

~~IA Test No.~~	Subject
MSE	DWM

Junior Supervisor's full signature with date :

Question No.	1	2	3	Total 20	Examiners Signature	Student's Sign After receiving the assessed answer sheet
Marks obtained						

Q1(a) Define the term Data Mining and KDD

Data Mining is the process to find patterns by analysing large databases. The patterns should be

(a) valid: hold on new data with some certainty
(b) novel: non-obvious to the system
(c) useful: should be possible to act on the item
(d) understandable: humans should be able to interpret the pattern

KDD: Knowledge discovery in Databases also called as Datamining is the process of discovering useful Knowledge from a collection of data.

(b) Consider a dataset, D = {8, 11, 12, 12, 23, 29, 29, 35, 38, 45, 46, 66}
Give five number summary of this dataset

→ Dataset D = {8, 11, 12, 12, 23, 29, 29, 35, 38, 45, 46, 66}

The Median:- As the total number is even

$$\frac{29+29}{2} = 29$$

First Quartile =

$$Q_1 = \{8, 11, 12, 12, 23, 29\}$$
$$= (12+12)/2 = 12$$

Third Quartile

$$Q_3 = \{29, 35, 38, 45, 46, 61\}$$
$$(38+45)/2 = 42.$$

minimum = 8

maximum = 66.

So the Five number Sumary is {8, 12, 29, 42, 66}

Q(c) Consider two objects

A = {19, 22, 07, 34} & B = {23, 12, 17, 30}

Calculate Minkowski distance (h=3) & supremum distance between these objects

Minkowski distance Formula

$$D(x_i, x_j) = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^{1/p} \right)^p$$

h=3 i.e p=3

= 12.862

The supremum distance = 24

d) Consider a dataset D = {4, 8, 15, 21, 21, 24, 24, 25, 28, 34}, perform Smoothing by bin boundaries. Let there be 3 bins and are equal frequency.

Bin 1 : {4, 8, 15}
Bin 2 : {21, 21, 24}
Bin 3 : {25, 28, 34}

Smoothing by bin

Bin 1 : {4, 8, 15}

Bin 2 : {21, 21, 24}

Bin 3 : {25, 25, 34}

e) Normalize the following group of data.
200, 300, 400, 600, 1000.

min-max normalization by setting min = 0 & max = 1

min = 200, max = 1000

$v' = \frac{(200 - 200)}{1000 - 200} \times (1 - 0) + 0 = 0$

v = 200

$v = 300 = \left(\frac{300 - 200}{1000 - 200}\right) \times (1 - 0) + 0 = 0.125$

$v = 400 = \left(\frac{400 - 200}{1000 - 200}\right) \times (1 - 0) + 0 = 0.25$

$v_1 = 600$

$$\frac{600-200}{(1000-200)} \times (1-0)+0 = 0.5$$

$v != 1000$

$$\frac{1000-200}{(1000-200)} \times (1-0)+0 = 1$$

After normalization.

(1) Limitation of ID3

The Limitation of ID3 is overfitting which can be overcomed by using pruning method.

Pruning method consists of two types.

1) pre-pruning
2) post-pruning

(2) Precision & Recall

Precision is a good measure to determine when the costs of false positive is high. For example spam detection a email is not spam but detected spam. The email we might lose important email if the precision is not high for the spam detection model

Precision is $\frac{TP}{TP+FP}$

$$\text{Recall} = \frac{TP}{TP+FN}$$

(m) Define the term: support & confidence with respect to Association Analysis

⇒ A support of 1% means that 1% of all the transaction under analysis shows that computer & s/w are purchased together.

$$\text{Support}(x \Rightarrow y) = \frac{\text{No of tuples that satisfies both } x \text{ \& } y}{\text{Total no of tuples}}$$

$$\text{Confidence}(x \Rightarrow y) = \frac{\text{no of tuples that satisfy both } x \text{ \& } y}{\text{No of tuples that statisfy } x.}$$

2(a) Explain the architecture of a typical DM system with neat diagram

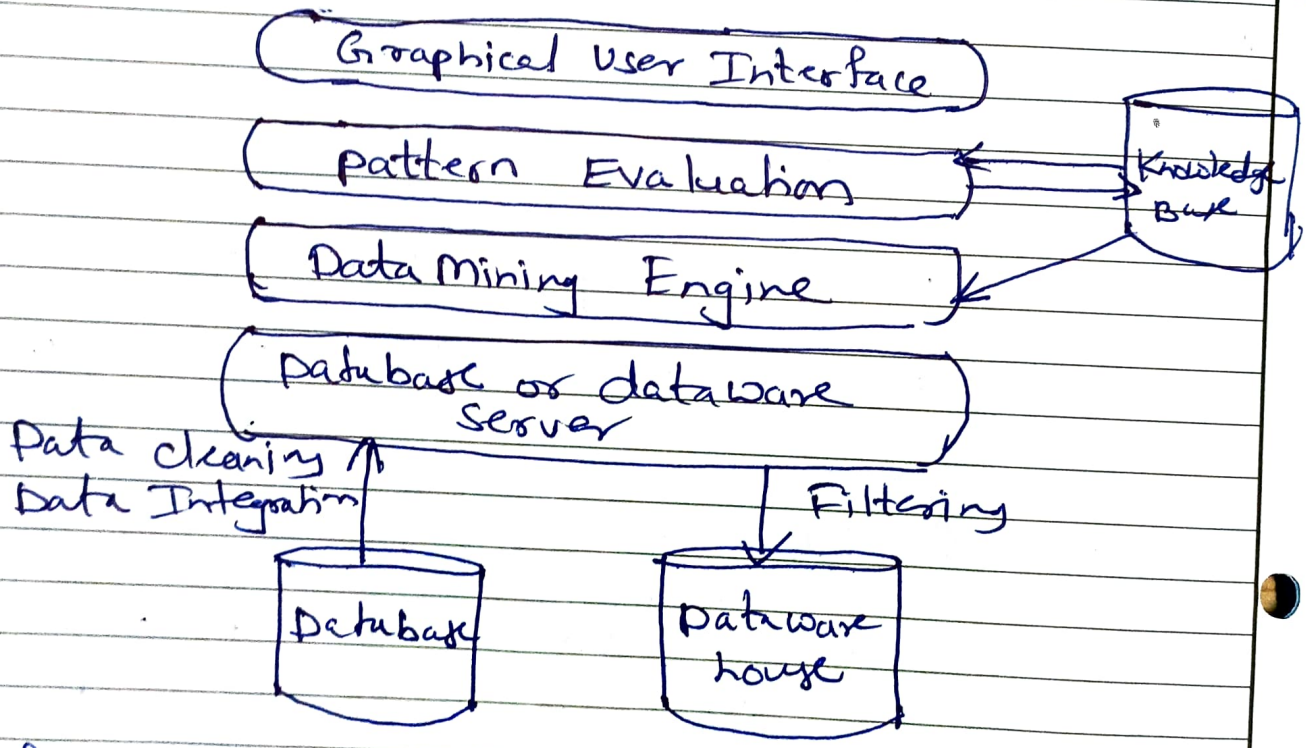
A DB is used to extract required data warehouse by executing SQL queries. If data is selected from a data ware house then no pre-processing is requirement because a data ware house contains cleared, transformed and integrated data.

A Data mining engine is used to extract all hidden pattern using intelligent algorithms

(3) Additional information required to form a particular Data mining task is called background knowledge and is available from knowledge Base.

(4) The pattern Evaluation Engine and the datamining engine are normally combined in a single component

(5) A GUI is used to convert extracted knowledge into required form. It allows users to interact with Datamining systems



The Architecture of a typical DM system

(b) For given data set calculate the IG for any one attribute using attribute relevance analysis

~~X = {Blonde, short, light, No}~~

The IG for the entire dataset using ID3 algorithm.

As there are 2 classes ↳ Sunburn = 3, None = 5

$$Info(D) = -\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8}$$

For attribute

$$Info(D)_{hair} = \left[\frac{3}{8} \times \left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)\right] +$$

Hair
- Blonde (3): Sunburn 2, None (1) 0
- Brown (4): None (4)
- Red (1)

$$\left[\frac{4}{8} \times \left(-\frac{4}{4}\log_2\frac{4}{4}\right) + 0\right] +$$

$$\left[\frac{1}{8} \times \left(-\frac{1}{1}\log_2\frac{1}{1}\right)\right]$$

$$IG_{hair} = Info(D) - Info(D)_{hair}$$

$$= 0.454$$

Q3) Classify a new Sample X = { Blonde, Short, Light, No } using Naive Bayes

⇒ Class =
C_1 = Sunburned = 3
C_2 = None = 5

P(Result = Sunburned) = 3/8
P(Result = None) = 5/8

For X = Blonde, Short, Light, no

P(Hair = Blonde | Result = Sunburned) = 2/3
P(Height = Short | Result = Sunburned) = 1/3
P(weight = Light | Result = Sunburned) = 1/3
P(Dublin = NO | Result = Sunburned) = 3/3

P(hair = Blonde | Result = None) = 1/5
P(Height = short | Result = None) = 2/5
P(weight = Light | Result = None) = 1/5
P(Dublin = NO | Result = None) = 2/5

$P(X \mid Result = Sunburned) = \frac{3}{8} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} \times 1 =$ 0.0277

$P(X \mid Result = None) = \frac{5}{8} \times \frac{1}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5} =$ 0.004

as probability of C_1 = sunburned) is greater than probability of C_2 = None)
∴ Unknown tuple X belongs to class C_1 = sunburned.

VIT Vidyalankar Institute of Technology
Accredited A+ by NAAC

Internal Assessment Test

Branch	~~Test Date~~ MSE	Semester	Div.	Roll No.	Student's Signature
CMPN		V	A		

~~IA Test No.~~ MSE	Subject
	DWM

Junior Supervisor's full signature with date :

Question No.	1	2	3	Total 20	Examiners Signature	Student's Sign After receiving the assessed answer sheet
Marks obtained						

Q3 (A) Explain different steps involved in preprocessing

(1) Data cleaning : Data in the real word is dirty as there is lots of potentially incorrect data.

eg occupation = '' '' missing data

noisy :- containing noise, errors or outliers

eg salary = '-10' (an error)

Age = 42, BD = 03/02/2010.

TO handle missing value.

(1) Ignore the tuple

(2) Use the attribute mean to fill the missing value

(3) predict the missing value by using a learning algorithm.

(4) Identify the outliers & smooth out noisy data using following techniques

(1) Binning

(2) clustering

(3) Regression

(4) correct inconsistet data.

2) Data Integration:-
Combine data from multiple sources into a coherent store.
It merge the data from multiple heterogenous data source into a coherent data store. Data Integration may involve inconsistent data and therefore needs data cleaning.
Reduadant data occur when integration of multiple database that may be able to be detected by [redacted] co-relation analysis and covariance analysis.

3) Data Reduction:-
obtain a reduced representation of the data set that is much smaller in volume, but yet produce the same (or almost the same) analytical result
Different data reduction strategies are

(1) Dimensionality reduction: eg remove unimportant attributes. It can be done by using
(1) Wavelet transforms
(2) Principal Components Analysis (PCA)
(3) Feature subset selection or feature creation
(4) Parametric data Reduction: Using Linear regression and Log Linear model

Parametric methods assume the data fits only the parameters and discard the data dimentionality reduction to avoid the curse of dimentionality, help eliminate irrelevant features and reduce

noise. It also allow easier visualization.

(4) Data Transformation: A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values

Different methods are
(1) Smoothing, Remove noise from data

(2) Attribute/feature construction
New attribute constructed from the given ones

(3) Aggregation: Summarization, data cube construction

(4) Normalization: scaled to fall within a smaller specified range
(1) methods of Normalization are
(a) Min-max normalization
(b) Z-score normalization
normalization by decimal scaling

(5) Discretization: Concept hierarchy climbing
Different methods can be applied for recursively
(1) Binning: Top-down split, unsupervised
(2) Histogram analysis: Top down split or bottom up merge
(3) clustering analysis.

4) Decision tree Analysis
5) Correlation (eg X^2 analysis) :- Bottom up merged.

Q2) Explain different data mining task in brief

1) classification and Regression for predictive Analysis
→ classification is the proces of finding a model that describes and distinguishes data classes or concepts. The model are derived based on the analysis of a set of training data. (i.e data objects of which class labels are known). The model is used to predict the class label of objects for which the class label is unknown
Eg:- Classify countries based on climate

Methods used are Decision trees, naive Bayes, classification, svm.

Regression Analysis:- is a statistical methodology that is most often used for numerical prediction. Regression models continuous valued function i.e regression is used to predict missing or unavailable numerical data values.

(3) Clustering Analysis:

Clustering analyzes data objects without consulting class labels.

In many cases, class labeled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.

ie cluster of objects are similar within a cluster and dissimilar to objects in other cluster

Association Analysis:-

Association analysis is useful for discovering interesting relationships hidden in large datasets. The uncover relationship can be represented in the form of association rules or set of frequent items

Application of association Analysis
Market Basket Analysis

buys(X, 'computer') => buy('x', 'sl

To measure support and confidence are considered.

support (x=>y) = No of tuples that satisfy both x & y / Total No of tup

Confidence $(x \Rightarrow y)$ =

$$\frac{\text{No of tuples that satisfy both } x \text{ \& } y}{\text{No of tuples that satisfy } X}$$

(1) Single dimensional association rule:
Association rule that contain a single predicate are referred to single-dimensional association rule.

buys(X, 'computer') $\Rightarrow$ buys(X, 'S/w')

(2) Multidimensional association rule:
The association rule that contain multiple predicates/attribute are referred to multidimensional association rule.

eg: age(X, '20-29') $\wedge$ income(X '40K-49K') $\Rightarrow$ buys(X, 'laptop')

(5) Outlier Analysis:
A dataset may contain objects that do not comply with the general behaviour or model of the data. These data objects are outliers. Many applications discard outliers as noise or exception. However, in some applications (e.g. fraud detection) the rare event can be more interesting than the more regularly occuring ones. The

analysis may answer fraudulent usage of credit cards by detecting purchases of unusually large amount for a given account number.