1) → Discrimination and Classification :-

- Difference :- Discrimination refers to the comparision of the general features of the target class object with the general features of objects from one or a set of differentiating classes.
  But, classification refers to the way of finding a set of functions that depict and recognize data classes for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
- Similarity: They both deal with the analysis of class data objects.

→ Characterization and clustering :-

- Difference :- Characterization refers to a summarization of the general characteristics or features of a target class of data while clustering deals with the analysis of data objects without consulting a known class label.
- Similarity :- They both deal with grouping together objects or data that are related or have high similarity in comparision to one another.

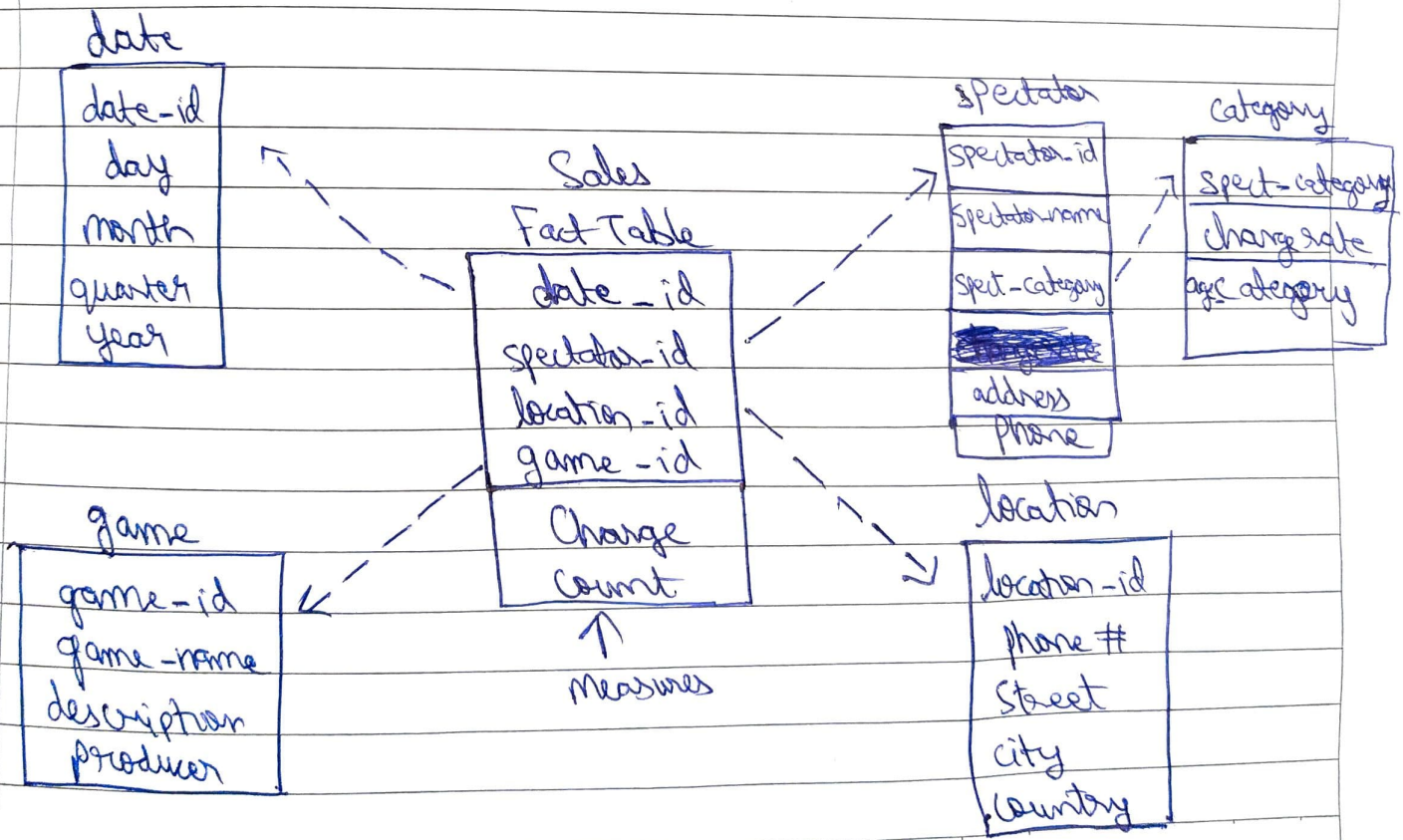
Expt. No. _____ Date _____

Page No. _____

→ Classification and regression:-

- Difference :- Classification predicts categorical labels which are discrete and unordered. Regression on the other hand predicts missing or unavailable and often numerical data values.
- Similarity :- This pair of tasks is similar in that they both are tools for prediction.

2.) a) The ~~Star~~ snowflake schema diagram for the data warehouse is as follows:-



Teacher's Signature : _____

b) The DMQL for the corresponding schema would be :-

define cube sales - snowflake [date, game, spectator, location]:

charge = ~~sum~~ charge_rate,
count = count (*)

define dimension date as (date_id, day, month, quarter, year)

define dimension game as (game_id, game_name, description, producer)

define dimension spectator as (spectator_id, spectator_name, Category (spectator_category, charge_rate, age_category), address, phone)

define dimension location as (location_id, phone number, Street, city, country);

⇒

Date

Expt. No. Page No.

3.) Age: 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

Number of data points = 27

a) i) Smoothing by bin means :-

	Mean
bin 1: 13, 15, 16	14.7
bin 2: 16, 19, 20	18.3
bin 3: 20, 21, 22	21
bin 4: 22, 25, 25	24
bin 5: 25, 25, 30	26.7
bin 6: 33, 33, 35	33.7
bin 7: 35, 35, 35	35
bin 8: 36, 40, 45	40.3
bin 9: 46, 52, 70	56

∴ by smoothing, we get

bin 1: 14.7, 14.7, 14.7
bin 2: 18.3, 18.3, 18.3
bin 3: 21, 21, 21
bin 4: 24, 24, 24
bin 5: 26.7, 26.7, 26.7
bin 6: 33.7, 33.7, 33.7
bin 7: 35, 35, 35
bin 8: 40.3, 40.3, 40.3
bin 9: 56, 56, 56

Answer

Teacher's Signature : _____

i) Smoothing by bin boundaries :-

(take the value which is closer)

Bin 1: 13, 16, 16

Bin 2: 16, 20, 20

Bin 3: 20, 20, 22

Bin 4: 22, 25, 25

Bin 5: 25, 25, 30

Bin 6: 33, 33, 35

Bin 7: 35, 35, 35

Bin 8: 36, 36, 45

Bin 9: 46, 46, 70

b) Min-max :- Minimum = 13

Maximum = 70

Formula :- $v' = \frac{v - min_A}{max_A - min_A} * (new\_max_A - new\_min_A) + new\_min_A$

So, Old data :- 13, 15, 16, 16, 19, 20

So, normalization :-

$13 \rightarrow \quad 13 - 13 = 0$

$15 \rightarrow \quad \frac{15 - 13}{57} = 0.035087$

$16 \rightarrow \quad \frac{16 - 13}{57} = 0.052631$

$16 \rightarrow \quad \frac{16 - 13}{57} = 0.052631$

Expt. No. .....

Date .....

Page No. .....

19 → $\frac{19-13}{57}$ = 0.105263

20 → $\frac{20-13}{57}$ = 0.122807

∴ Normalized data :-

0, 0.285714, 0.428571, 0.428571, 1

Normalized data :-

0, 0.035087, 0.0526316, 0.0526316, 0.105263, 0.122807

Answer

4.) Since we have 2-dimensional data, formula is

$$D = \sqrt{(i_{A_1} - j_{A_1})^2 + (i_{A_2} - j_{A_2})^2}$$

Given new data point $x = (1.4, 1.6)$

i) Euclidean distances are :-

$x_1 \rightarrow 0.1414 = \sqrt{(1.5-1.4)^2 + (1.7-1.6)^2}$

$x_2 \rightarrow 0.6708 = \sqrt{(2-1.4) + (1.9-1.6)^2}$

$x_3 \rightarrow 0.2828 = \sqrt{(1.6-1.4)^2 + (1.8-1.6)^2}$

$x_4 \rightarrow 0.2236 = \sqrt{(1.2-1.4)^2 + (1.5-1.6)^2}$

$x_5 \rightarrow 0.6083 = \sqrt{(1.5-1.4)^2 + (1.0-1.6)^2}$

Teacher's Signature : _____

ii) Formula for cosine similarity is

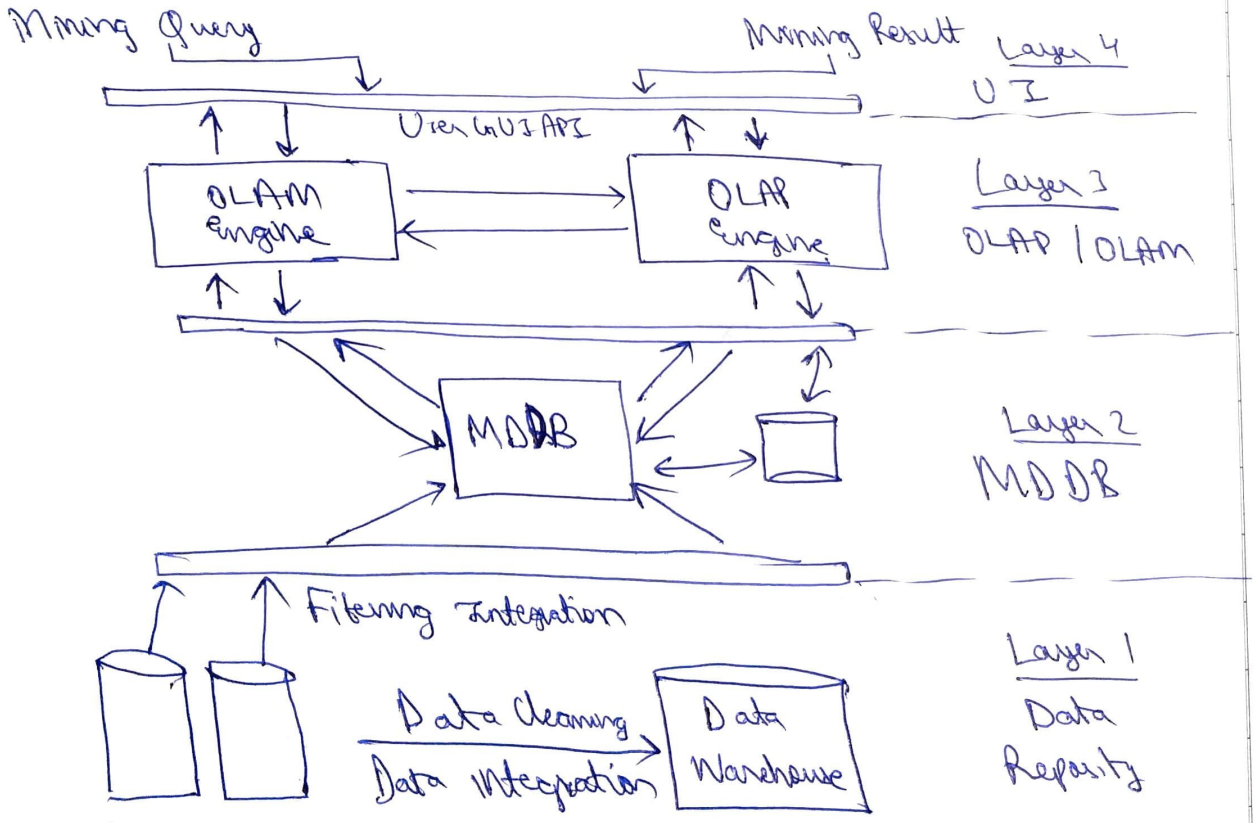
$$S(x, y) = \frac{x^t \cdot y}{\|x\| \|y\|}$$

5)

$X_1 \rightarrow 0.99999$
$X_2 \rightarrow 0.99575$
$X_3 \rightarrow 0.99997$
$X_4 \rightarrow 0.99903$
$X_5 \rightarrow 0.96536$

} Answer

5)



Date

Expt. No. Page No.

→ For banking and financial institutions, there is a requirement of huge data mining and data processing capabilities.

→ It would use a transactional database where each record represents a transaction.

→ OLAP & OLAM mechanisms accept users online requests through a graphical user interface (GUI) application programming interface.

→ We work with the data cube using cube API. data cube is created by integrating multiple databases.

→ OLAP is the technology behind many business intelligence (BI) applications. OLAP is a powerful technology for data discovery, including capabilities for limitless report viewing, complex analytical calculations and predictive scenarios (budget, forcast).

→ These technologies enable fast aggregations and calculations of underlying data sets, one can understand its usefulness in helping business leaders make better informed decisions.

Teacher's Signature : _____

→ OLTP is customer oriented and used for query & transaction processing. It also enables management of current data for decision making.

→ Enables short transactions with atomic concurrency control.

→ Hence, useful for banking and financial institutions.