**MACHINE LEARNING**

**1)** A : 2

**2)** D : 1,2 and 4

**3)** D : formulating the clustering problem

**4)** A : Euclidean distance

**5)** D : K-means clustering

**6)** A : Defined distance metric

**7)** D : All of the above

**8)** B : Unsupervised

**9)** A : K- Means clustering

**10)** A,B,C : K-mean, mode, medians clustering algorithm

**11)** D : all of the above

**12)** D : Categorical data

**Answers of WORKSHEET 1 SQL**

**1)** A,D : Create,Alter

**2)** A,B,C : Update,Delete,Select

**3)** B : Structured Query Language

**4)** B : Data Defination Lang.

**5)** A : Data Manipulation Languag

**6)** C : Create Table Tab\_name (Col datatype,col datatype)

**7)** B : Alter Table A ADD COLUMN D float

**8)** B : Alter Table A Drop Column D

**9)** B : Alter Table A Alter Column D int

**10)** A : Alter Table A Add Constraint Primary Key B

**11)**

Datawarehouse is the collection of multiple data marts of same Organisation.Datawarehouse can have historical data which we can use to analyse the business.

Example :- Consider Bank's DataWarehouse will have multiple Data-Marts like Saving Data-Mart,Loan Data-Mart,Insurance Data-Mart,

mutualFund DataMart etc.

**12)**

OLTP : Online Transcation Processing

Maintain Transcational data like ATM card swap. OLAP Normaly have day wise data into normalised form which will be use by business.

Basically OLTP used by bank is helping bank's business run.

OLAP : Online Analytical Processing

Basically it is nothing but historical data. Bank's business will never get affected by OALP

but it will help the management to decide best approach to grow bank's business.

**13)**

We can create datawarehouse using different methods like using ETL tool. So Using ETL tool we can perform the below tasks.

Data consolidation :- Daily,Monthly

Data Cleaning :- Can implement rules for IFSC code,Account no,country,email,mobile should be in correct format.

Data Integration :- For this we can use multip ETL tools

So features of Datawarehouse are

1) Time Varient :- Historical data storage. We can get the data from last 25 years or from the year when that organistion is started.

2) Subject Oriented :- Multiple DataWarehouses for different sections of banks. Like Loan (Home,Car,Morgage,Personal),Mutual Funds (SIP,ELSS),Saving (Deposits,FD etc)

3) Integrated :- Naming Standards,Columns Name this will help to understand the data. Because we can populate the data from multiple source like mainframe,SAP etc which we can not read. Once data is populated we can easily read the data.

4) Non-Volatile :- No deletion of data from datawahrehoue. SCD and upsert will take place. Even if data deletion is there then It will be like Soft delete(Set Flag as delete) but data will be there.

**14)**

STAR Schema

We can implement this schema on relational data base.

Basically it is start like structure of Fast tables and Dimension tables.

Dimension Table’s Primary Keys will be act as Foreign Key in Fact tables and data will be deformalized.

**15)**

No Idea of SETL in SQL.

**Answer STATISTICS WORKSHEET-1**

**1)** A : True

**2)** A : Central Limit Theorem

**3)** B : Modeling bounded count data

**4)** D : All options

**5)** C : Poisson

**6)** B : FALSE

**7)** A : Probability

**8)** A : 0

**9)** C : Outliers cannot conform to the regression relationship

**10)** Normal Distribution

Mean, Median and mode are same. Perfect symmetrical

**11)** Missing Values

Count is less then delete or replace it with mean or average value.

**12)** A/B testing

Same code with two versions A and B. Then analyze which one is more suitable or useful.

**13)** No, We should not use mean for missing values because it will change the dependent values.

**14)** Linear Regression

If our output continuous and depends on input values.

**13)** 2 types of Statistics

A) Descriptive ( 1. Central Tendency: Mean, Median, Mode

2. Dispersion : Range, Percentile, Std. Deviation)

B) Inferential ( Hypothesis Test, T test, F test)