SQL_Worksheet_1

1. A (create), D(alter)

2. A (update), B(delete), C(select)

3. B (Structured query language)

4. B (Data definition language)

5. A (Data manipulation language)

6. C [Create Table A(B int, C float)]

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. C (Alter Table A Add Primary key B)

11. **Data-warehouse-** A central repository of information that can be analyzed to make more informed decisions.Basically it integrates data and information collected from various sources into one comprehensive database.
Four main components:- A central database, ELT tools (extract,transform,load), Metadata and Access tools.

12. **OLTP(** online transction processing): Are the systems that manage transction oriented applications which helps in operations. These systems are designed to support online transction and process query quickly on the internet. **eg.** POS (point of sale) system. While **OLAP**(online analytical processing): It refers to a set of software tools used for data analysis in businesses. It provides a platform for gaining insights from databases. **eg.** BI (business intelligence) applications rely on this technology.

13. Characteristic of Data-warehouse are: a. Subject oriented, b. Integrated. c. Time variant, d. Non-volatile.

14. **Star-schema:** It is a fundamental and the simplest schema among the data mart. It is widely used to develop a data warehouse and dimensional data marts. It is very efficient for handling basic queries. It is said to be star as its physical model resembles the star shape having a fact table in center and dimension table at its peripheral.

15. **SETL** (SET language): High level programming language developed by Jack Schwartz in 1970s. It is based on the mathematical theory of sets.

Machine Learning worksheet_1

1. B

2. D (1-Data points with outliers; 2-Data points with different densities; & 4-Data points with non-convex shapes)

3. D (Formulating the clustering problem)

4. A (Euclidean distance)

5. Divisive clustering

6. D (All answers are correct i.e., A- defined distance metric; B- Numbers of clusters; C-Initial guess as to cluster centroids)

7. A (Divide the data points into groups)

8. B (Unsupervised learning)

9. D (All of the above i.e., A- K-Means clustering; B- Hierarchical clustering; C- Diverse clustering)

10. A (K-Means clustering algorithm)

11. D (All of the above i.e., A, B,& C)

12. A (labeled data)

13.

14.

15. Cluster analysis is a multivariate data mining technique whose goal is to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data.

Types of cluster analysis:

1. Hierarchical cluster analysis
   a. Agglomerative method
   b. Divisive method
2. Centroid based clustering
   a. K-means
3. Distribution based clustering
4. Density based clustering

Statistics Worksheet_1

1. A(True)

2. A(Central limit theorem)

3. B(Modeling bound count data)

4. D (All of the mentioned)

5. C (Poisson)

6. B (False)

7. B (Hypothesis)

8. A (0)

9. C (Outliers can't conform to the regression relationship)

10. Normal distribution is a type of continuous probability distribution in which most data points cluster towards the middle of the range, while the rest shift symmetrically toward either extreme. It has symmetric bell shape curve.

11. We can handle missing data by two ways i.e.,

> **a.** Deletion: By deleting rown and column having null values. We can implement it in Python by simply using the pandas **.dropna** method like:- **df .dropna(axis=1, inplace=True)**
>
> **b.** Imputation: By filling the missing values with substitutes.To implement it in Python we can use **.fillna** method in pandas like:- **df .fillna(inplace=True)**

Recommended imputation techniques are:

> **a.** Regression imputation which includes, creating a model to predict the observed value of a variable based on another variable. Then we can use the model to fill the missing value of that variable. This technique is utilized for MAR(missing at random) and MCAR(missing completely at random).

b. Simple imputation involves utilizing a numerical summary of the variable where the missing value is occured. While using this method to fill the values, we need to evaluate the variables distribution to determine the central tendency summary. This method is utilized in MCAR category. And to implement it in Python we can use **SimpleImputer** transformer in the Scikit-learn library.

12. A/B testing is one of the most popular controlled experiments which is used to optimize web marketing strategies. It allows to choose best designs for a website by looking at the analytical results obtained with two possible alternatives.

13. No, because it ignores features correlation.

14. Linear regression is the most basic and commonly used predictive analysis. In statistics, it is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables.

15. Three branches of statistics are there i.e.,

a. Data collection

b. Descriptive statistics

c. Inferential statistics