MACHINE LEARNING ANSWERS

1 - a) 2

2 - d) 1, 2 and 4

3 - d) formulating the clustering problem

4 - a) Euclidean distance

5 - b) Divisive clustering

6 - d) All answers are correct

7 - a) Divide the data points into groups

8 - a) Supervised learning

9 - a) K- Means clustering

10 - a) K-means clustering algorithm

11 - d) All of the above

12 - a) Labeled data

13 - The hierarchical **cluster analysis** follows three basic steps: 1) **calculate** the distances, 2) link the **clusters**, and 3) choose a solution by selecting the right number of **clusters**

14- To **measure** a **cluster's** fitness within a **clustering**, we can compute the average silhouette coefficient value of all objects in the **cluster**. To **measure** the **quality** of a **clustering**, we can use the average silhouette coefficient value of all objects in the data set

15 - **Cluster analysis** is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. ... These **types** are Centroid **Clustering**, Density **Clustering** Distribution **Clustering**, and Connectivity **Clustering**

SQL ANSWERS

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 - C) Alter Table A D float int

10- C) Alter Table A Add Primary key B

11 - Data warehousing is the electronic storage of a large amount of information by a business or organization. A **data warehouse** is designed to run query and analysis on historical **data** derived from transactional sources for business intelligence and **data** mining purposes.

12 - Online transaction processing (**OLTP**) captures, stores, and processes data from transactions in real time. Online analytical processing (**OLAP**) uses complex queries to analyze aggregated historical data from **OLTP** systems

13 -

- Some **data** is denormalized for simplification and to improve performance.
- Large amounts of historical **data** are used.
- Queries often retrieve large amounts of **data**.
- Both planned and ad hoc queries are common.
- The **data** load is controlled.

14- A **star schema** is a data warehousing architecture model where one fact table references multiple dimension tables, which, when viewed as a diagram, looks like a **star** with the fact table in the center and the dimension tables radiating from it.

15- **SETL** (SET Language) is a very high-level programming language based on the mathematical theory of sets

STATISTICS ANSWERS

1 - a) True

2 - a) Central Limit Theorem

3 - b) Modeling bounded count data

4 - d) All of the mentioned

5 - c) Poisson

6 - b) False

7 - b) Hypothesis

8 - a) 0

9 - b) Outliers can be the result of spurious or real processes

10 **- Normal distribution**, also known as the Gaussian **distribution**, is a probability **distribution** that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, **normal distribution** will appear as a bell curve

11- You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value.

In Pandas, there are two very useful methods: isnull() and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value (for example, 0), you could use the fillna() method.

12 - **AB testing** is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal

13 - It is a non-standard, but a fairly flexible **imputation** algorithm. It uses RandomForest at its core to predict the **missing data**. It can be applied to both continuous and categorical variables which makes it advantageous over other **imputation** algorithms.

14 - In **statistics**, **linear regression** is a **linear** approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables)

15 - The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific **analysis** of data and both are equally important for the student of statistics.