**ML ASSIGNMENT-1**

1. Option (b) 4 clusters
2. Option(d) Data point with outliers, Data point with outliers & Data point with non-convex shapes
3. Option(d) Formulating the clustering problem
4. Option(a) Euclidean Distance
5. Option(b) Divisive Clustering
6. Option(d) All the answers are correct
7. Option(a) Dividing data points into groups
8. Option(b) Unsupervised Learning
9. Option(a) K-Means Clustering
10. Option(a) K-Means Clustering
11. Option(d) All of the above
12. Option(a) Labeled Data
13. Cluster Analysis is calculated by:
14. Calculating distance between two points
15. Linking the clusters
16. Choosing a perfect solution by selecting the right number of clusters
17. There are majorly two types of measures to assess the clustering performance.
18. **Extrinsic Measures** which require ground truth labels. Examples are Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure.
19. **Intrinsic Measures** that does not require ground truth labels. Some of the clustering performance measures are Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.
20. 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 techniques create clusters that allow us to understand how our data is related. The most common applications of cluster analysis in a business setting is to segment customers or activities.

There are four basic types of cluster analysis used in data science. These types are Centroid Clustering, Density Clustering Distribution Clustering, and Connectivity Clustering.

**SQL WORKSHEET-1**

1. Options A and D : Create and Alter
2. Options A, B & C: Update, Delete & Select
3. Option B: Structured Query Language
4. Option B: Data Definition Language
5. Option A: Data Manipulation Language
6. Option C: Create Table A(B int, C float)
7. Option B: Alter Table A Add column D float
8. Option B: Alter Table A Drop Column D
9. Option B: Alter Table A Alter Column D int
10. Option A: Alter Table A Add Constraint Primary Key B
11. A **Data Warehousing** (DW) is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting.
12. Difference between OLAP and OLTP

|  |  |  |
| --- | --- | --- |
| **S. no.** | **OLAP** | **OLTP** |
| 1 | It is an online transactional system and manages database modification. | It is an online data retrieving and data analysis system. |
| 2 | Insert, Update, Delete information from the database. | Extract data for analyzing that helps in decision making. |
| 3 | OLTP and its transactions are the original source of data. | Different OLTPs database becomes the source of data for OLAP. |
| 4 | The processing time of a transaction is comparatively less in OLTP. | The processing time of a transaction is comparatively more in OLAP. |

1. Some characteristics of data warehousing are:

**Subject-oriented –**  
A data warehouse is always a subject oriented as it delivers information about a theme instead of organization’s current operations. It can be achieved on specific theme. That means the data warehousing process is proposed to handle with a specific theme which is more defined. These themes can be sales, distributions, marketing etc.  
  
A data warehouse never put emphasis only current operations. Instead, it focuses on demonstrating and analysis of data to make various decision. It also delivers an easy and precise demonstration around particular theme by eliminating data which is not required to make the decisions.

**Integrated –**  
It is somewhere same as subject orientation which is made in a reliable format. Integration means founding a shared entity to scale the all similar data from the different databases. The data also required to be resided into various data warehouse in shared and generally granted manner.  
  
A data warehouse is built by integrating data from various sources of data such that a mainframe and a relational database. In addition, it must have reliable naming conventions, format and codes. Integration of data warehouse benefits in effective analysis of data. Reliability in naming conventions, column scaling, encoding structure etc. should be confirmed. Integration of data warehouse handles various subject related warehouse.

**Time-Variant –**  
In this data is maintained via different intervals of time such as weekly, monthly, or annually etc. It founds various time limit which are structured between the large datasets and are held in online transaction process (OLTP). The time limits for data warehouse is wide-ranged than that of operational systems. The data resided in data warehouse is predictable with a specific interval of time and delivers information from the historical perspective. It comprises elements of time explicitly or implicitly. Another feature of time-variance is that once data is stored in the data warehouse then it cannot be modified, alter, or updated.

**Non-Volatile –**  
As the name defines the data resided in data warehouse is permanent. It also means that data is not erased or deleted when new data is inserted. It includes the mammoth quantity of data that is inserted into modification between the selected quantity on logical business. It evaluates the analysis within the technologies of warehouse.  
  
In this, data is read-only and refreshed at particular intervals. This is beneficial in analysing historical data and in comprehension the functionality. It does not need transaction process, recapture and concurrency control mechanism. Functionalities such as delete, update, and insert that are done in an operational application are lost in data warehouse environment. Two types of data operations done in the data warehouse are:

* 1. Data Loading
  2. Data Access

1. **Star Schema** in data warehouse, in which the center of the star can have one fact table and a number of associated dimension tables. It is known as star schema as its structure resembles a star. The Star Schema data model is the simplest type of Data Warehouse schema. It is also known as Star Join Schema and is optimized for querying large data sets.
2. SETL or SET Language is a high-level programming language that’s based on the mathematical theory of sets.

**STATISTICS WORKSHEET-1**

1. Option A: True
2. Option A: Central Limit Theorem
3. Option B: Modelling bounded count data
4. Option D: All of the above
5. Option C: Poisson Distribution
6. Option B: False
7. Option B: Hypothesis Testing
8. Option A: 0
9. Option C: Outliers cannot conform to the regression relationship
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. Some ways by which we can handle the missing data are:

Listwise or case deletion, Pair-wise deletion, Mean substitution, Regression imputation, Last observation carried forward, Maximum likelihood, Expectation-Maximization, and Multiple imputation.

I will be recommending mean/median imputation technique for simple data whereas it solely depends upon data that which technique should be used.

1. A/B testing is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing". A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.
2. It is not acceptable every time as the choosing an imputation technique depends upon the nature of data.
3. **Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

1. The two main branches of statistics are [descriptive statistics](https://explorable.com/descriptive-statistics) and [inferential statistics](https://explorable.com/inferential-statistics).

[**Descriptive statistics**](https://explorable.com/descriptive-statistics) deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid [biases](https://explorable.com/research-bias) that are so easy to creep into the [experiment](https://explorable.com/conducting-an-experiment).

[**Inferential statistics**](https://explorable.com/inferential-statistics), as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.