**Machine learning :-**

Que 1) B

Que 2) D

Que 3) D

Que 4) A

Que 5) B

Que 6) D

Que 7) A

Que 8) B

Que 9) A

Que 10) A

Que 11) D

Que 12) A

Que 13) How is cluster analysis calculated?

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.

Que 14). How is cluster quality measured?

Within the context of cluster analysis, [Purity is an external evaluation criterion of cluster quality.](http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html#fig:clustfg3) It is the percent of the total number of objects(data points) that were classified correctly, in the unit range [0..1]. 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.

We have a few methods to choose from for measuring the quality of a clustering. In general, these methods can be categorized into two groups according to whether ground truth is available. Here, ground truth is the ideal clustering that is often built using human experts.

If ground truth is available, it can be used by **extrinsic methods**, which compare the clustering against the group truth and measure. If the ground truth is unavailable, we can use **intrinsic methods**, which evaluate the goodness of a clustering by considering how well the clusters are separated. Ground truth can be considered as supervision in the form of “cluster labels.” Hence, extrinsic methods are also known as supervised methods, while intrinsic methods are [unsupervised methods](https://www.sciencedirect.com/topics/computer-science/unsupervised-method).

Que 15) What is cluster analysis and its types?

Clustering itself can be categorized into two types viz. Hard Clustering and Soft Clustering. In hard clustering, one data point can belong to one cluster only. But in soft clustering, the output provided is a probability likelihood of a data point belonging to each of the pre-defined numbers of clusters.

* **Hard Clustering:** In hard clustering, each data point either belongs to a cluster completely or not. For example, in the above example each customer is put into one group out of the 10 groups.
* **Soft Clustering**: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, from the above scenario each costumer is assigned a probability to be in either of 10 clusters of the retail store.

**SQL :-**

Que 1) A & D

Que 2) A, B & C

Que 3) B

Que 4) B

Que 5) A

Que 6) C

Que 7) B

Que 8) B

Que 9) B

Que 10) C

Que 11) What is data-warehouse?

A Data Warehousing 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.

Que 12) What is the difference between OLTP VS OLAP?

OLTP :- Online transaction processing provides transaction-oriented applications in a 3-tier architecture. OLTP administers day to day transaction of an organization.

Consists only operational current data. It provides a multi-dimensional view of different business tasks.

Uses of OLTP :- ATM center is an OLTP application. OLTP handles the ACID properties during data transaction via the application. It’s also used for Online banking, Online airline ticket booking, sending a text message, add a book to the shopping cart.

OLAP :- Online Analytical Processing consists of a type of software tools that are used for data analysis for business decisions. OLAP provides an environment to get insights from the database retrieved from multiple database systems at one time. Any type of Data warehouse system is an OLAP system.

Uses of OLAP :- Spotify analyzed songs by users to come up with the personalized homepage of their songs and playlist. Netflix movie recommendation system.

Que 13) What are the various characteristics of data-warehouse?

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.

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 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. 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: Data Loading, Data Access.

Que 14) What is Star-Schema?

A star schema is the elementary form of a dimensional model, in which data are organized into facts and dimensions. A fact is an event that is counted or measured, such as a sale or log in. A dimension includes reference data about the fact, such as date, item, or customer.

A star schema is a relational schema where a relational schema whose design represents a multidimensional data model. The star schema is the explicit data warehouse schema. It is known as star schema because the entity-relationship diagram of this schemas simulates a star, with points, diverge from a central table. The center of the schema consists of a large fact table, and the points of the star are the dimension tables.

The star schema is intensely suitable for data warehouse database design because of the following features:

It creates a DE-normalized database that can quickly provide query responses.

It provides a flexible design that can be changed easily or added to throughout the development cycle, and as the database grows.

It provides a parallel in design to how end-users typically think of and use the data.

It reduces the complexity of metadata for both developers and end-users.

Star Schemas are easy for end-users and application to understand and navigate. With a well-designed schema, the customer can instantly analyze large, multidimensional data sets.

The main advantage of star schemas in a decision-support environment are:

Query Performance

Load performance and administration

Built-in referential integrity

Easily Understood

Que 15) What do you mean by SETL?

SETL is a high-level programming language that’s based on the mathematical theory of sets. It was developed in the early 1970’s by mathematician Professor J. Schwartz. SETL is an interpreted language with a syntax that is resembles C and in many cases similar to Perl. In SETL every statement is terminated by a semicolon. Variable names are case-insensitive and are automatically determined by their last assignment.

SETL provides quantified boolean expressions constructed using the universal and existential quantifiers of first-order predicate logic.

SETL provides several iterators to produce a variety of loops over aggregate data structures.

**Statistics :-**

Que 1) A

Que 2) A

Que 3) B

Que 4) C

Que 5) C

Que 6) B

Que 7) B

Que 8) A

Que 9) C

Que 10) What do you understand by the term Normal Distribution?

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.

A normal distribution is the proper term for a probability bell curve.

In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.

Normal distributions are symmetrical, but not all symmetrical distributions are normal.

In reality, most pricing distributions are not perfectly normal.

Que 11) How do you handle missing data? What imputation techniques do you recommend?

When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be Missing completely at random

The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is Missing At Random.

When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

Imputation Techniques:-

1) Mean or Median Imputation:-

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

There may not be enough observations with non-missing data to produce a reliable analysis

In predictive analytics, missing data can prevent the predictions for those observations which have missing data

External factors may require specific observations to be part of the analysis.

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.

2) Multivariate Imputation by Chained Equations (MICE):-

MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

3) Random Forest:-

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

Random forest works best with large datasets and using random forest on small datasets runs the risk of overfitting.

.

Que 12) What is A/B testing?

**A/B testing (also known as split testing) is a process of showing two variants of the same web page to different segments of website visitors at the same time and comparing which variant drives more conversions.**

If B2B businesses today are unhappy with all the unqualified leads they get per month, eCommerce stores, on the other hand, are struggling with a high cart abandonment rate. Meanwhile, media and publishing houses are also dealing with low viewer engagement. These core conversion metrics are affected by some common problems like leaks in the conversion funnel, drop-offs on the payment page, etc.

A/B testing offers a very systematic way of finding out what works and what doesn’t work in any given marketing campaign. Most marketing efforts are geared toward driving more traffic. As traffic acquisition becomes more difficult and expensive, it becomes paramount to offer your users the best experience who comes to your website. This will help them achieve their goals and convert in the fastest and most efficient manner possible. A/B testing in marketing allows you to make the most out of your existing traffic.

Que 13) Is mean imputation of missing data acceptable practice?

* Bad practice in general
* If just estimating means: mean imputation preserves the mean of the observed data
* Leads to an underestimate of the standard deviation
* Distorts relationships between variables by “pulling” estimates of the correlation toward zero

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.

Que 14) What is linear regression in statistics?

Linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?  (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?  These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.  The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

**Simple linear regression**  
1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

[**Multiple linear regression**](https://www.statisticssolutions.com/data-analysis-plan-multiple-linear-regression/)  
1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

[**Logistic regression**](https://www.statisticssolutions.com/data-analysis-plan-logistic-regression/)  
1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

[**Ordinal regression**](https://www.statisticssolutions.com/data-analysis-plan-ordinal-regression/)  
1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

[**Multinomial regression**](https://www.statisticssolutions.com/data-analysis-plan-multinominal-logistic-regression/)  
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

[**Discriminant analysis**](https://www.statisticssolutions.com/discriminant-analysis-independent-variables/)  
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

Que 15) What are the various branches of statistics?

Statistics may be divided into two main branches:  
**(1)** Descriptive Statistics **(2)** Inferential Statistics  
 **(1) Descriptive Statistics**

Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data.

**For example:** Industrial statistics, population statistics, trade statistics, etc. Businessmen make use of descriptive statistics in presenting their annual reports, final accounts, and bank statements.

**(2) Inferential Statistics**

Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates.

**For example:** Suppose we want to have an idea about the percentage of the illiterate population of our country. We take a sample from the population and find the proportion of illiterate individuals in the sample. With the help of probability, this sample proportion enables us to make some inferences about the population proportion. This study belongs to inferential statistics.