**Assignment-Technical Training**

**1)What is data normalization? How is it different from database normalization (1st/2nd/3rd)?**

Normalization is the process of organizing data in a proper manner, that is, the process of reducing data redundancy in a table and improving data integrity. It is a way of organizing data in a database.

Normalization involves organizing the columns and tables in the database to ensure that their dependencies are correctly implemented using database constraints. It is used to minimize the duplication of various relationships in the database. Normalization, also known as database normalization or data normalization, is an important part of relational database design because it helps to improve the speed, accuracy, and efficiency of the database.

First Normal Form:

A relation is in first normal form if every attribute in that relation is singled valued attribute.

A table is in 1 NF if and only if:

1. There are only Single Valued Attributes.
2. Attribute Domain does not change.
3. There is a unique name for every Attribute/Column.
4. The order in which data is stored does not matter.

Second Normal Form:

Second Normal Form (2NF) is based on the concept of full functional dependency. Second Normal Form applies to relations with composite keys, that is, relations with a primary key composed of two or more attributes.

A relation is in 2NF if it has No Partial Dependency, i.e., no non-prime attribute (attributes which are not part of any candidate key) is dependent on any proper subset of any candidate key of the table.

Third Normal Form:

A relation is in third normal form, if there is no transitive dependency for non-prime attributes as well as it is in second normal form.

**2) What is a distribution? What are the uses for frequency and probability distribution?**

A Statistical Distribution of data represents how often each value occurs in the dataset. Distributions of data may be frequency distributions or probability distributions. Distributions gives an idea of the central tendency(typical value), dispersion(deviations from the typical value) and most common and uncommon values present in the dataset.

A frequency distribution is a graph or dataset organized to show the frequency(number of times) of occurrence of each possible outcome of a repeated event observed many times. It is usually represented using a histogram.

A probability distribution is the mathematical function that gives the probabilities of occurrence of different possible **outcomes** for an experiment.

**3) What is a decision? How's it different from inference?**

In computational theory, a decision problem is a computational problem that can be posed as a yes–no question of the input values. For example, deciding by means of an algorithm whether a given natural number is even.

Statistical inference refers to the theory, methods, and practice of forming judgements about the parameters of a population and the reliability of statistical relationships, typically on the basis of random samples.

**4) Google- what is Gini in probability, and explain in your own terms**

In general, Gini coefficient (or index) is a synthetic indicator capturing the level of inequality for a given variable in a population. It lies between 0 and 1, with 1 implying highest degree of inequality. It is well used in measuring the income inequality along with the Lorenz Curve.

**5) What is entropy?**

Entropy is a scientific concept as well as a measurable physical property that is mostly associated with a state of disorder, randomness, or uncertainty.

Information Entropy or Shannon's entropy quantifies the amount of uncertainty (or surprise) involved in the value of a random variable or the outcome of a random process. Its significance in the decision tree is that it allows us to estimate the impurity or heterogeneity of the target variable

**6) What is euclidean distance?**

In mathematics, Euclidean distance between two points in the Euclidean space is the length of a line segment between the two points.

**7) What's the difference between correlation and covariance?**

Although both covariance and correlation are used in regression analysis to measure associativity of two variables, the basic difference is that covariance measures how the two variables differ, while correlation measures how the two variables are related.

**8) What is mean squared error?**

The mean squared error is the sum of the squared difference of each value of the dataset from the mean. It is given as:

It is used to see how well a model fits a dataset. The greater the MSE, the more unfit is the model for that data.

**9) What is the difference between covariance, standard deviation and mean squared error?**

Covariance measures the direction of the relationship between two variables. A positive covariance implies that both the variables have high or low values together. A negative covariance implies that if one variable has high value then the other has low values. This is given by:

Mean squared error measures the amount of error in a statistical model and is given as:

Standard Deviation is a measure of the dispersion of the actual values in the data from the central tendency. It conveys the spread of the data. It is given as the square root of variance.