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

**Ans:**

Data normalization is generally considered the development of clean data. Data normalization is the organization of data to appear similar across all records and fields. It increases the cohesion of entry types leading to cleansing, lead generation, segmentation, and higher quality data.

**1NF:**

The most basic form of data normalization is 1NFm which ensures there are no repeating entries in a group. To be considered 1NF, each entry must have only one single value for each cell and each record must be unique.

**2NF:**

To ensure no repeating entries, to be in the 2NF rule, the data must first apply to all the 1NF requirements. Following that, data must have only one primary key. To separate data to only have one primary key, all subsets of data that can be placed in multiple rows should be placed in separate tables. Then, relationships can be created through new foreign key labels.

**3NF:**

For data to be in this rule, it must first comply with all the 2NF requirements. Following that, data in a table must only be dependent on the primary key. If the primary key is changed, all data that is impacted must be put into a new table.

1. **What is a distribution? What are the uses for frequency and probability distribution?**

**ANS:**

A statistical distribution, or probability distribution, describes how values are distributed for a field. In other words, the statistical distribution shows which values are common and uncommon. There are many kinds of statistical distributions, including the bell-shaped normal distribution.

Frequency distribution is a curve that gives us the frequency of the occurrence of a particular data point in an experiment. This is usually the limit of a histogram of frequencies when the data points are very large and the results can be treated to be varying continuously instead of taking on discrete values.

Probability distribution yields the possible outcomes for any random event. It is also defined based on the underlying sample space as a set of possible outcomes of any random experiment. These settings could be a set of real numbers or a set of vectors or a set of any entities. It is a part of probability and statistics

1. **What is a decision? How's it different from inference?**

**Ans:**

In statistics, a set of quantitative methods for reaching optimal decisions. A solvable decision problem must be capable of being tightly formulated in terms of initial conditions and choices or courses of action, with their consequences.

The ladder of inference is a tool that explains how we make decisions. In computability theory and computational complexity theory, a decision problem is a problem that can be posed as a yes–no question of the input values. An example of a decision problem is deciding whether a given natural number is prime.

1. **what is Gini in probability and explain in your own terms?**

**Ans:**

The Gini coefficient is a statistic which quantifies the amount of inequality that exists in a population. The Gini coefficient is a number between 0 and 1, with 0 representing perfect equality and 1 perfect inequality. Sometimes these statistics are reported in terms of percentages, with numbers between 0 and 100.

**5. What is entropy?**

**Ans:**

Entropy is defined as the randomness or measuring the disorder of the information being processed in Machine Learning. Further, in other words, we can say that entropy is the machine learning metric that measures the unpredictability or impurity in the system.

**6. What is Euclidean distance?**

**Ans:**

The Euclidean distance formula is used to find the distance between two points on a plane . Euclidean distance is calculated as the square root of the sum of the squared differences between the two vectors.

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

**Ans:**

1. A measure used to indicate the extent to which two random variables change in tandem is known as covariance. A measure used to represent how strongly two random variables are related known as correlation.

2. Covariance is nothing but a measure of correlation. On the contrary, correlation refers to the scaled form of covariance.

3. The value of correlation takes place between -1 and +1. Conversely, the value of covariance lies between -∞ and +∞.

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

The mean squared error (MSE) tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs. It also gives more weight to larger differences. It’s called the mean squared error as you’re finding the average of a set of errors. The lower the MSE, the better the forecast.

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

variance is used in calculating standard deviation, which is a measure of how spread out a set of data is. Covariance is used in calculating correlation, which is a measure of how two variables relate to each other. whereas the MSE measures the average of the squares of the "errors", that is, the difference between the estimator and what is estimated.