**Technical session – Assignment**

**Name: Iniya J**

**Emp Id: 2211477**

**Cohort code: CDB22DW022**

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

* Normalization 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.
* Normalization is the process of organizing data in a proper manner. It is used to minimize the duplication of various relationships in the database.
* It is also used to troubleshoot exceptions such as inserts, deletes, and updates in the table. It helps to split a large table into several small normalized tables.
* Relational links and links are used to reduce redundancy.
* 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.

**Normal form:**

* The process of refining the structure of a database to minimize redundancy and improve integrity of database is known as Normalization. When a database has been normalized, it is said to be in normal form.
* A relation is considered in 1NF if every domain attribute consists of one atomic or indiscreet value. it’s a really important property of a relationship that’s accessible within the RDBMS.
* A relation is in 2NF only if it is in 1NF also as all the non-key attributes in the tables are addicted to the table’s primary key. this type applies to those relations that are connected using composite keys. The relation of the tables is coupled with a key that is composed of over one attribute.
* A relation is in 3NF if and only if it’s in 2NF and there’s no transition dependency.
* The main use of normalization is to utilize in order to remove anomalies that are caused because of the transitive dependency. Normalization is to minimize the redundancy and remove Insert, Update and Delete Anomaly. It divides larger tables into smaller tables and links them using relationships.

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

* The distribution of a statistical dataset is the spread of the data which shows all possible values or intervals of the data and how they occur.
* A distribution is simply a collection of data or scores on a variable. Usually, these scores are arranged in order from ascending to descending and then they can be presented graphically.
* The distribution provides a parameterized mathematical function which will calculate the probability of any individual observation from the sample space.

**uses for frequency distribution**

* A frequency distribution in statistics is a representation that displays the number of observations within a given interval.
* The representation of a frequency distribution can be graphical or tabular so that it is easier to understand.
* Frequency distributions are particularly useful for normal distributions, which show the observations of probabilities divided among standard deviations.
* In finance, traders use frequency distributions to take note of price action and identify trends.

**uses for probability distribution**

* By using a discrete distribution, you can define the probability of occurrence of each value of a discrete random variable. A discrete random variable is a random variable that has only countable values, for example, a list of non-negative integers.
* By using a continuous distribution, you can define the probabilities of the possible values of a continuous random variable. A continuous random variable is a random variable that has a set of possible infinite and uncountable values

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

* Statistical decision theory is concerned with the making of decisions when in the presence of statistical knowledge (data) which sheds light on some of the uncertainties involved in the decision problem.
* The generality of these definitions is such that decision theory (dropping the qualifier ‘statistical’ for convenience) formally encompasses an enormous range of problems and disciplines. Any attempt at a general review of decision theory is thus doomed; all that can be done is to present a description of some of the underlying ideas.
* Statistical inference is the process of using data analysis to infer properties of an underlying distribution of probability.
* Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

1. **Google- what is Gini in probability, and explain in your own terms**

* The Gini index is a measure of the distribution of income across a population.
* A higher Gini index indicates greater inequality, with high-income individuals receiving much larger percentages of the population's total income.
* Global inequality, as measured by the Gini index, has steadily increased over the past few centuries and spiked during the COVID-19 pandemic.
* Because of data and other limitations, the Gini index may overstate income inequality and can obscure important information about income distribution.

1. **What is entropy?**

* Entropy is an information theory metric that measures the impurity or uncertainty in a group of observations. It determines how a decision tree chooses to split data. The image below gives a better description of the purity of a set.
* Consider a dataset with N classes. The entropy may be calculated using the formula below:

is the probability of randomly selecting an example in class . Let’s have an example to better our understanding of entropy and its calculation. Let’s have a dataset made up of three colors: red, purple, and yellow. If we have one red, three purple, and four yellow observations in our set, our equation becomes:

Where , and are the probabilities of choosing a red, purple and yellow example respectively.

1. **What is Euclidean distance?**

The Euclidean distance is defined as the distance between two points. In other words, the Euclidean distance between two points in the Euclidean space is defined as the length of the line segment between two points. As the Euclidean distance can be found by using the coordinate points and the Pythagoras theorem, it is occasionally called the Pythagorean distance.

The Euclidean distance formula helps to find the distance of a line segment. Let us assume two points, such as and in the two-dimensional coordinate plane.

Thus, the Euclidean distance formula is given by:

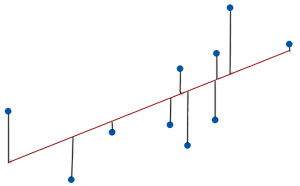
1. **What's the difference between correlation and covariance?**

|  |  |  |
| --- | --- | --- |
| **Basis for comparison** | **Covariance** | **Correlation** |
| Definition | Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. | Correlation is a statistical measure that indicates how strongly two variables are related. |
| Values | The value of covariance lies in the range of -∞ and +∞. | Correlation is limited to values between the range -1 and +1 |
| Change in scale | Affects covariance | Does not affect the correlation |
| Unit-free measure | No | Yes |

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

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).

For example, in regression, the mean squared error represents the average squared residual.



As the data points fall closer to the regression line, the model has less error, decreasing the MSE. A model with less error produces more precise predictions.

The formula for MSE is the following.

Where:

* is the observed value.
* is the corresponding predicted value.
* = the number of observations.

The calculations for the mean squared error are similar to the variance. To find the MSE, take the observed value, subtract the predicted value, and square that difference. Repeat that for all observations. Then, sum all those squared values and divide by the number of observations. Notice that the numerator is the sum of the squared errors (SSE), which linear regression minimizes. MSE simply divides the SSE by the sample size.