**Technical Session - Exercise**

**Name: VIGNESH B**

**Emp.Id.: 2211443**

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

*Data normalization* is the process of reorganizing data within a database so that users can utilize it for further queries and analysis. Simply put, it is the process of developing clean data. This includes eliminating redundant and unstructured data and making the data appear similar across all records and fields.

Data normalization can be divided into different types of normal forms. The most popular ones are 1NF, 2NF, 3NF, and BCNF.

1NF - First Normal Form

The most basic form of data normalization is 1NF which ensures there are no two same entries in a group. For a table to be in the first normal form, it should satisfy the following rules:

* Each cell should contain a single value
* Each record should be unique

2NF - Second Normal Form

In a 2NF table, all the subsets of data that can be placed in multiple rows are placed in separate tables. For a table to be in the second normal form, it should satisfy the following rules:

* It should be in 1F
* The primary key should not be functionally dependant on any subset of candidate key

3NF - Third Normal Form

For a table to be in the third normal form, it should satisfy the following rules:

* It should be in 2F
* It should not have any transitive functional dependencies

A transitive functional dependency is when a change in a column (which is not a primary key) may cause any of the other columns to change.

BCNF - Boyce and Codd Normal Form

Boyce and Codd Normal Form is a higher version of 3NF and is also known as 3.5NF. A BCNF is a 3NF table that does not have multiple overlapping candidate keys. For a table to be in BCNF, it should satisfy the following rules:

* It should be in 3F
* For each functional dependency ( X → Y ), X should be a super key

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

As a statistical tool, a frequency distribution provides a visual representation of the distribution of observations within a particular test. Analysts often use a frequency distribution to visualize or illustrate the data collected in a sample.

The reasons for constructing a frequency distribution are as follows:

* To organize the data in a meaningful, intelligible way.
* To enable the reader to determine the nature or shape of the distribution.
* To facilitate computational procedures for measures of average and spread.

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

Statistical decisions are decisions made based on observations of a phenomenon that obeys probabilistic laws that are not completely known.

Inference is arriving at a decision or opinion by reasoning facts from known evidence or facts.

Statistical inference refers to the process of drawing conclusions from the model estimation.

Statistical inference consists in the use of statistics to draw conclusions about some unknown aspect of a population based on a random sample from that population. Some preliminary conclusions may be drawn by the use of EDA or by the computation of summary statistics as well, but formal statistical inference uses calculations based on probability theory to substantiate those conclusions. Statistical inference can be divided into two areas: estimation and hypothesis testing. In estimation, the goal is to describe an unknown aspect of a population. Estimation can be of two types, point estimation and interval estimation, depending on the goal of the application. The goal of hypothesis testing is to decide which of two complementary statements about a population is true. Point estimation, which addresses what particular value of a parameter is most consistent with the data. Interval estimation is concerned with quantifying the uncertainty or variability associated with the estimate. This approach supplements point estimation because it gives important information about the variability (or confidence) in the point estimate.

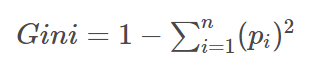
1. **Google- what is Gini in probability and explain in your own terms.**

Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen.

If all the elements belong to a single class, then it can be called pure. The degree of Gini index varies between 0 and 1,  
were,

* 0 denotes that all elements belong to a certain class or if there exists only one class, and
* 1 denotes that the elements are randomly distributed across various classes.

A Gini Index of 0.5 denotes equally distributed elements into some classes.



where *pi*is the probability of an object being classified to a particular class.

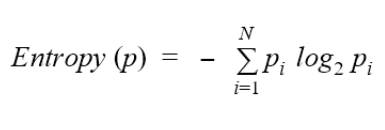
While building the decision tree, we would prefer choosing the attribute/feature with the least Gini index as the root node.

**5. What is entropy?**

The entropy measures the “amount of information” present in a variable. In information theory, the entropy of a random variable is the average level of “information“, “surprise”, or “uncertainty” inherent in the variable’s possible outcomes.

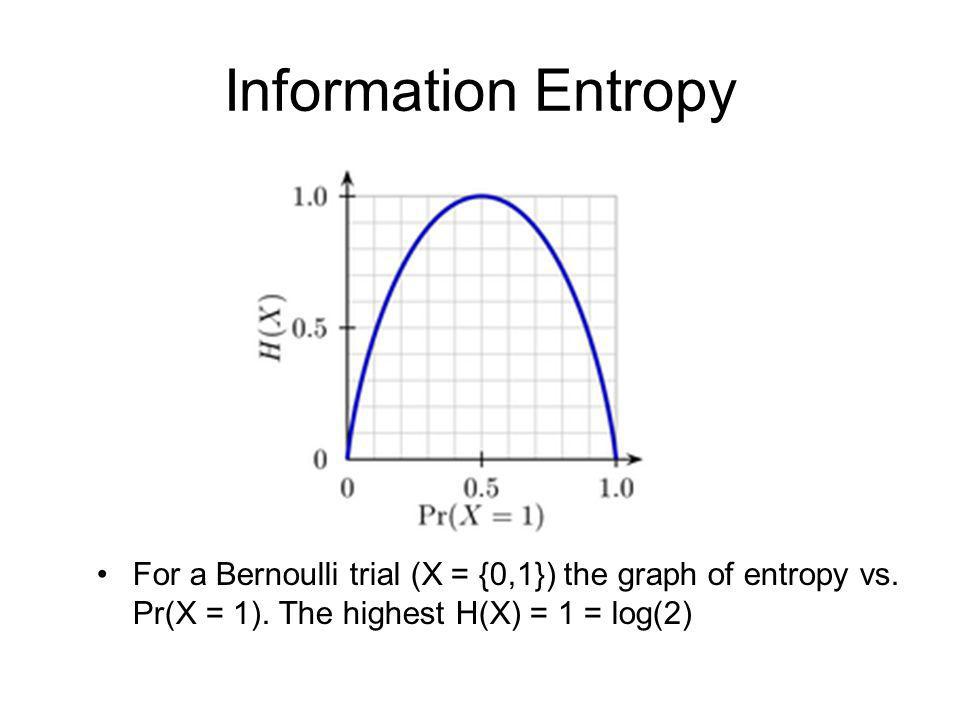
That is, the more certain or the more deterministic an event is, the less information it will contain. In a nutshell, the information is an increase in uncertainty or entropy.

Claude E. Shannon had expressed this relationship between the probability and the heterogeneity or impurity in the mathematical form with the help of the following equation:



The uncertainty or the impurity is represented as the log to base 2 of the probability of a category (pi). The index (i) refers to the number of possible categories. Here, i = 2 as our problem is a binary classification.

This equation is graphically depicted by a symmetric curve as shown below. On the x-axis is the probability of the event and the y-axis indicates the heterogeneity, or the impurity denoted by H(X).



**6 .** **What is Euclidean distance?**

In coordinate geometry, Euclidean distance is the distance between two points. To find the two points on a plane, the length of a segment connecting the two points is measured. We derive the Euclidean distance formula using the Pythagoras theorem.

The Euclidean distance formula gives the distance between two points (or) the straight-line distance. Assume that (x1, y1) (x1, y1) and (x2, y2) (x2, y2) are two points in a two-dimensional plane.

The Euclidean distance formula says:

d = √ [ (x2 – x1)2 + (y2– y1)2]

where,

* (x1, y1) are the coordinates of one point.
* (x2, y2) are the coordinates of the other point.
* d is the distance between (x1, y1) and (x2, y2).

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

In statistics, covariance and correlation are two mathematical notions. Both phrases are used to describe the relationship between two variables.

*Covariance* signifies the direction of the linear relationship between the two variables. By direction we mean if the *variables* are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).

The values of covariance can be any number between the two opposite infinities. Also, it’s important to mention that covariance only measures how two variables change together, not the dependency of one variable on another one.

The value of covariance between 2 variables is achieved by taking the summation of the product of the differences from the means of the variables as follows:



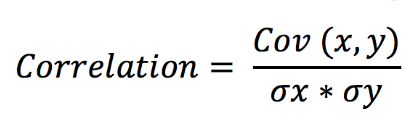
*Correlation* analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables.

It not only shows the kind of relation (in terms of direction) but also how strong the relationship is. Thus, we can say the correlation values have standardized notions, whereas the covariance values are not standardized and cannot be used to compare how strong or weak the relationship is because the magnitude has no direct significance. It can assume values from -1 to +1.

To determine whether the covariance of the two variables is large or small, we need to assess it relative to the standard deviations of the two variables.

To do so we must normalize the covariance by dividing it with the product of the standard deviations of the two variables, thus providing a correlation between the two variables.

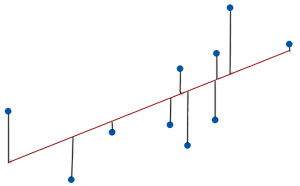
The main result of a correlation is called the correlation coefficient.



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).

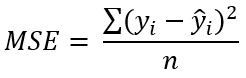
For example, in regression, the mean squared error represents the average squared residual.



This image depicts the relationship between the residuals and the mean squared error.

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:

* yi is the ith observed value.
* ŷi is the corresponding predicted value.
* n = 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 of those squared values and divide by the number of observations.