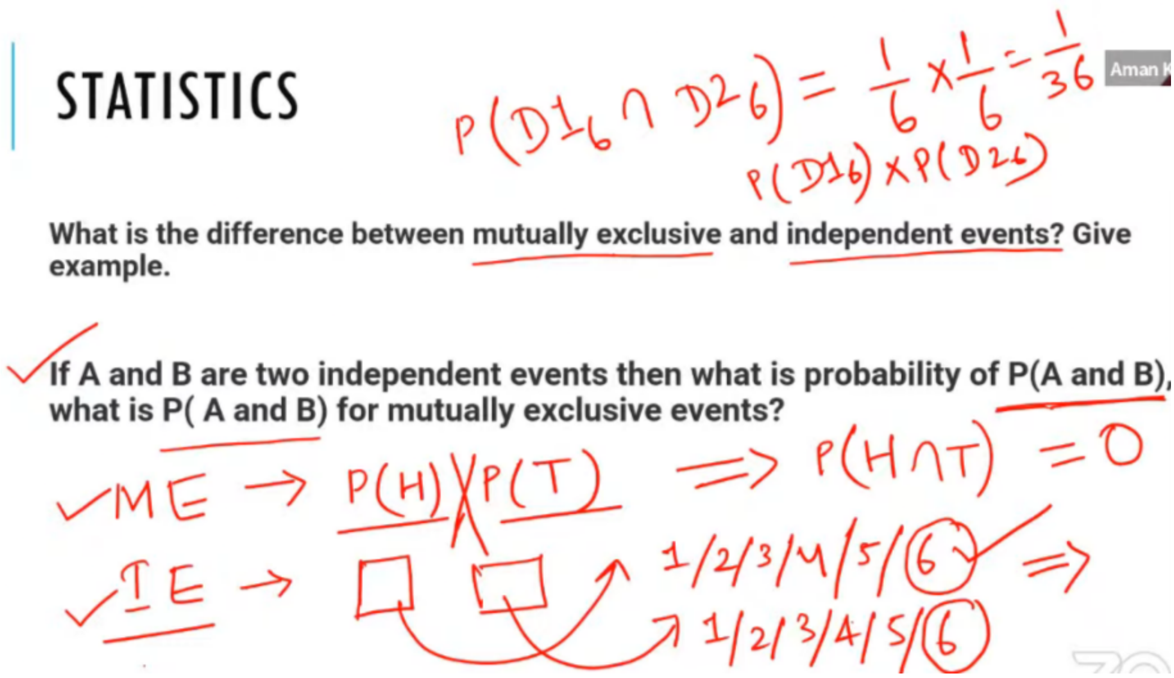
Session 15

Monday, March 27, 2023 2:01 PM



STATISTICS

$P(D1_6 \cap D2_6) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

$P(D1_6) \times P(D2_6)$

What is the difference between mutually exclusive and independent events? Give example.

✓ If A and B are two independent events then what is probability of P(A and B), what is P(A and B) for mutually exclusive events?

✓ ME → $P(H) \times P(T) \Rightarrow P(H \cap T) = 0$

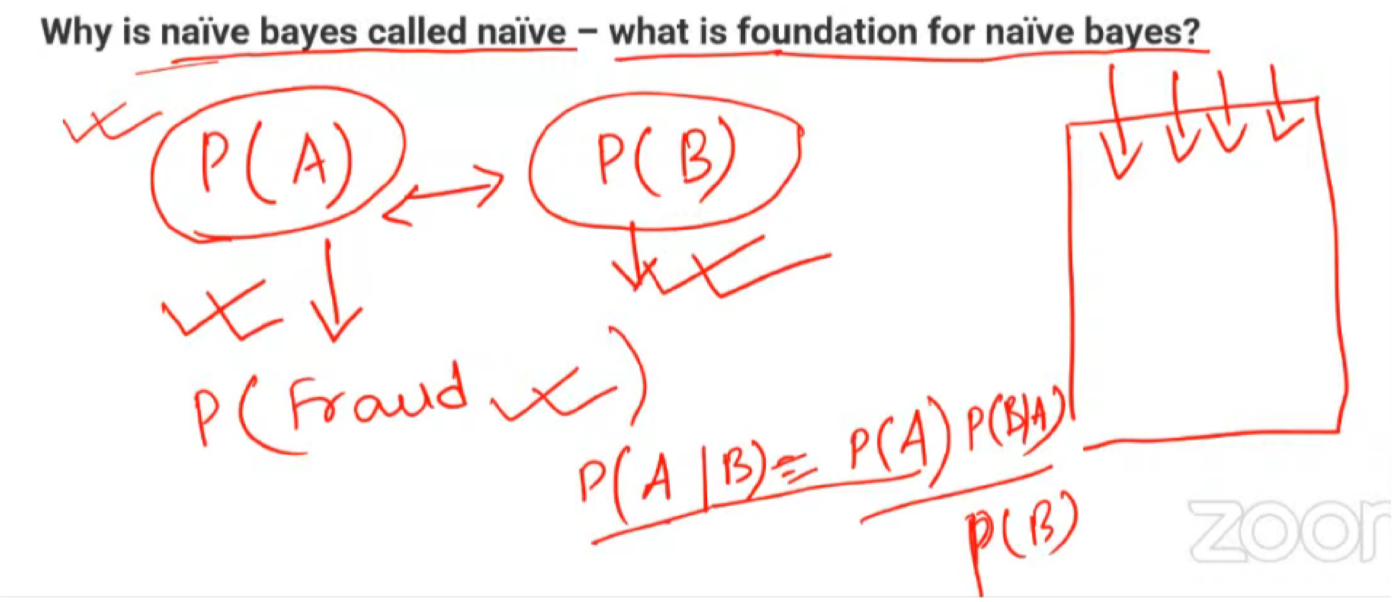
✓ IE → □ □ 1/2/3/4/5/6 ⇒ 1/2/3/4/5/6

Mutually exclusive events are events that cannot happen at the same time, meaning that the occurrence of one event makes the occu of the other event impossible. For example, when rolling a six-sided die, the events "rolling a 2" and "rolling a 4" are mutually exclusive because both events cannot occur simultaneously.

Independent events are events where the occurrence of one event does not affect the probability of the occurrence of the other event. example, if you flip a coin twice, the events "getting heads on the first flip" and "getting heads on the second flip" are independent even

If A and B are independent events, then the probability of both events happening (P(A and B)) is the product of their individual probabi (P(A) multiplied by P(B)). For example, if the probability of rolling a 2 on a die is 1/6 and the probability of rolling a 4 on a die is also 1/6 the probability of rolling both a 2 and a 4 on the same die roll is (1/6) x (1/6) = 1/36.

If A and B are mutually exclusive events, then the probability of both events happening (P(A and B)) is zero, because the events cann happen at the same time. For example, the probability of rolling both a 2 and a 4 on the same die roll is zero, because the events "rolli and "rolling a 4" are mutually exclusive.



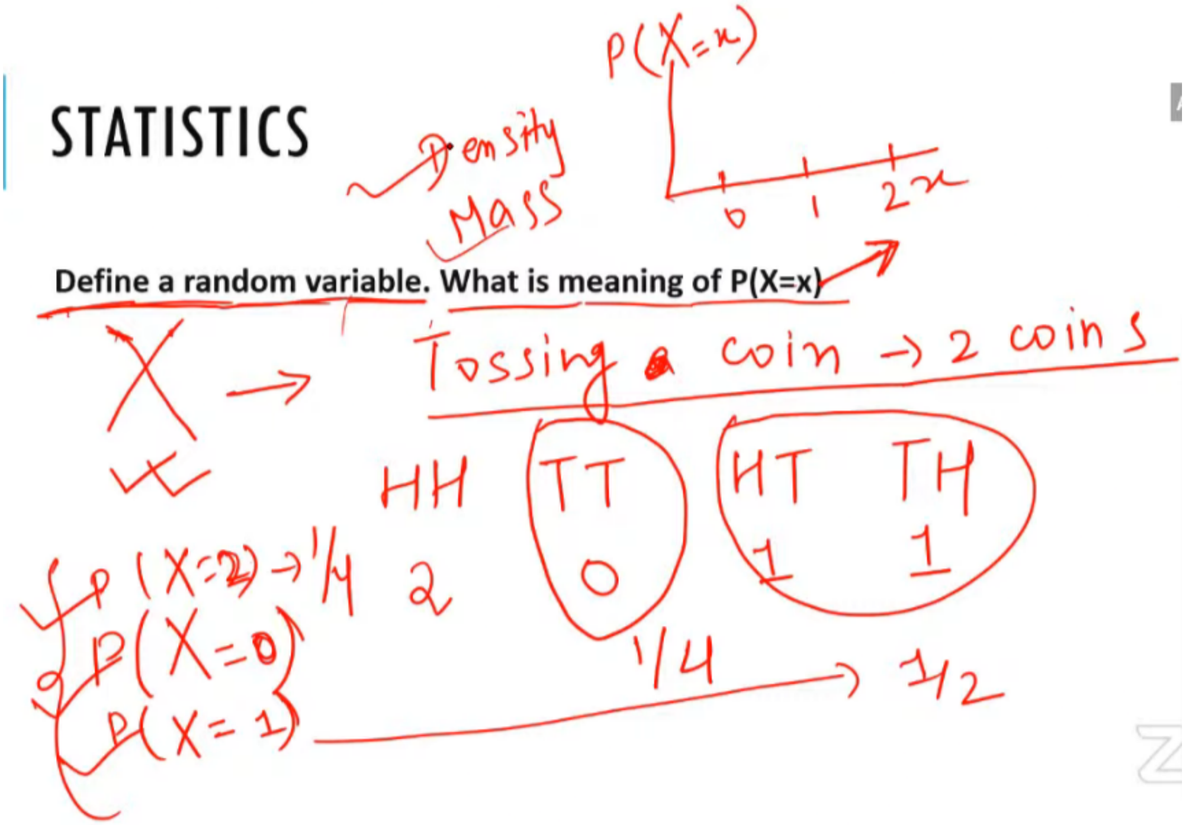
Why is naïve bayes called naïve – what is foundation for naïve bayes?

$P(A \mid B) = \frac{P(A)\, P(B|A)}{P(B)}$

Naive Bayes is called "naive" because it makes a simplifying assumption that the features (or attributes) being used to predict the class of an observation are conditionally independent of each other given the class label. This assumption is considered "naive" because it may not hold true in many real-world scenarios where the features are correlated.

The foundation for Naive Bayes is based on Bayes' theorem, which is a fundamental concept in probability theory. Bayes' theorem is used to calculate the probability of a hypothesis (such as a class label) given some observed evidence (such as feature values). In other words, it enables us to update our beliefs about the likelihood of a hypothesis given some new evidence.

Naive Bayes builds on this concept by assuming that the likelihood of each feature given the class label is independent of the other features, and then using Bayes' theorem to calculate the probability of the class label given the observed feature values. Despite the naive assumption, Naive Bayes has been shown to perform well in many real-world applications, especially when the number of features is large relative to the amount of training data available.



STATISTICS

Define a random variable. What is meaning of P(X=x)

A random variable is a variable whose possible values are determined by chance or probability. It is a mathematical function that assigns a numerical value to each possible outcome of a random process.

Consider the random experiment of flipping two coins simultaneously.

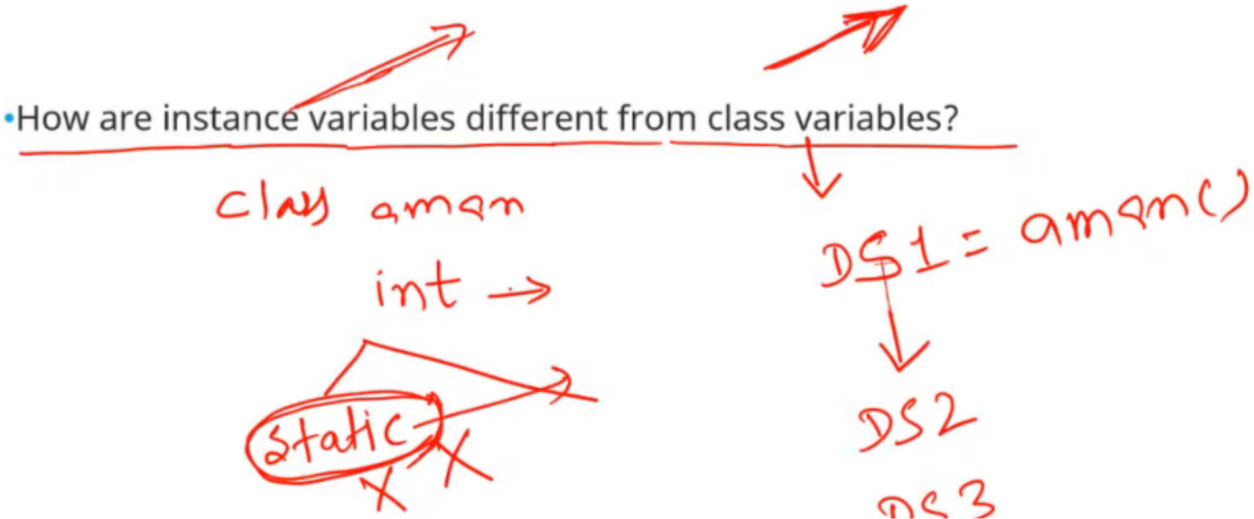
We can define the random variable X to be the number of heads that appear when the two coins are flipped. X can take on the values of 0, 1, or 2, depending on the number of heads that appear.

The probability distribution for X can be represented using a probability mass function (PMF) as follows:

P(X=0) = 1/4 (both coins show tails) P(X=1) = 1/2 (one coin shows heads and the other shows tails, or vice versa) P(X=2) = 1/4 (both coins show heads)

For example, if we want to find the probability that both coins show tails (i.e., X=0), we can use the PMF and calculate P(X=0) = 1/4.

Similarly, if we want to find the probability that at least one coin shows heads (i.e., X is either 1 or 2), we can use the PMF and calculate P(X=1) + P(X=2) = 1/2 + 1/4 = 3/4.



•How are instance variables different from class variables?

Instance variables and class variables are two types of variables in object-oriented programming that have different scopes and lifetimes.

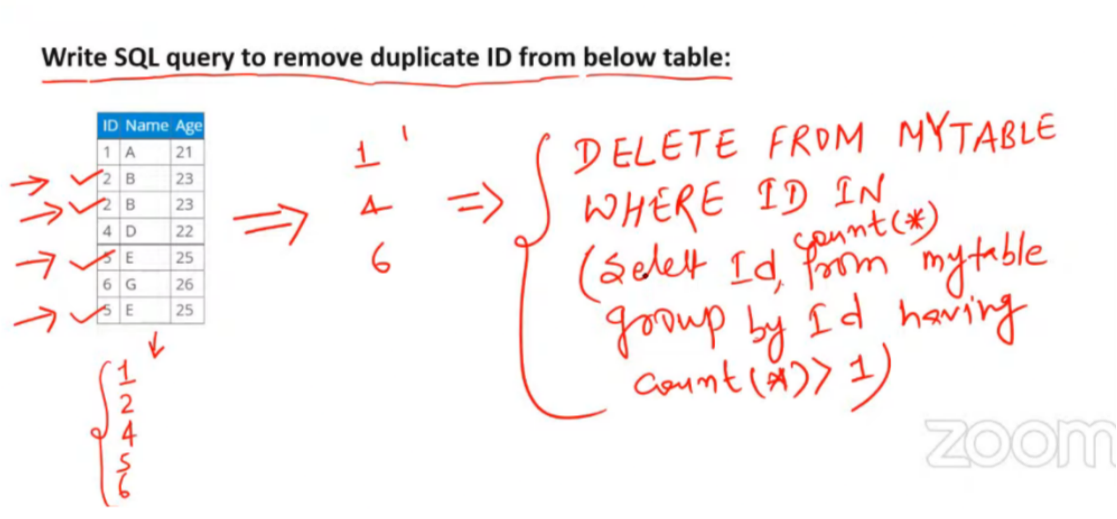
Instance variables are associated with an instance (or object) of a class and are defined within the class but outside of any methods. Each instance of the class has its own copy of the instance variables, which can have different values. Instance variables are used to store state information about an object and can be accessed and modified using the dot notation (object.variable).

For example, consider a class called "Person" with instance variables "name" and "age". Each instance of the Person class will have its own values for these variables, such as "John" and 30 for one instance and "Jane" and 25 for another instance.

Class variables, on the other hand, are associated with the class itself and are shared among all instances of the class. Class variables are defined within the class but outside of any methods, and are typically used to store data that is common to all instances of the class. Class variables can be accessed using the class name (classname.variable) or an instance of the class (object.variable), but they are typically accessed using the class name.

For example, consider a class called "Rectangle" with a class variable "count" that keeps track of the number of Rectangle objects created. Each time a new Rectangle object is created, the count variable is incremented. All instances of the Rectangle class share the same count variable.

In summary, instance variables are specific to each instance of a class, while class variables are shared among all instances of the class. Instance variables are accessed using an object, while class variables are typically accessed using the class name.

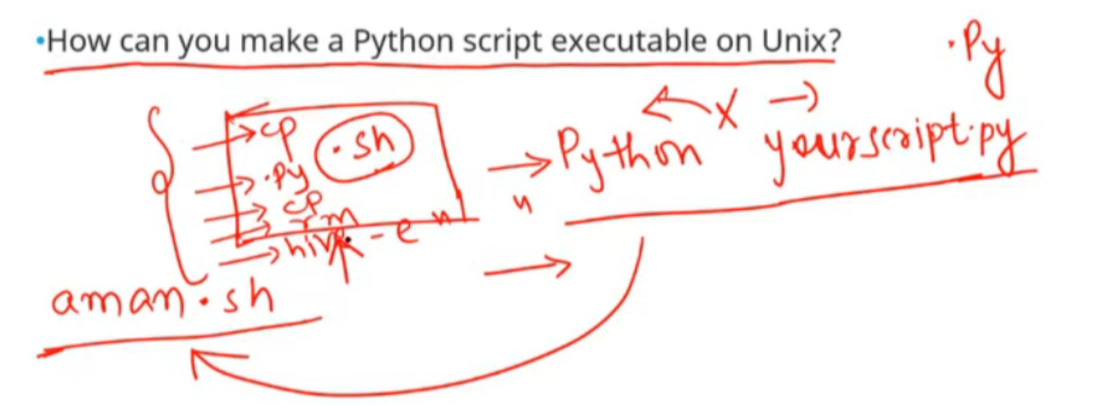


Write SQL query to remove duplicate ID from below table:

ID	Name	Age
1	A	21
2	B	23
2	B	23
4	D	22
5	E	25
6	G	26
5	E	25

DELETE FROM MYTABLE
WHERE ID IN
(Select Id, count(*) from mytable
group by Id having
Count(*)> 1)

DELETE FROM mytable Where id=id and not DELETE id from mytable WHERE id=id



•How can you make a Python script executable on Unix?

To make a Python script executable on Unix, you need to perform the following steps:

1. Add the shebang line at the beginning of the script. This line tells the Unix shell which interpreter to use to run the script. For Python scripts, the shebang line should be:

javascriptCopy code

```
#!/usr/bin/env python
```

This assumes that Python is installed in the default location (/usr/bin/) and can be invoked using the "python" command.

1. Set the file permissions to allow execution. This can be done using the "chmod" command in the terminal. For example, to make the script "myscript.py" executable, you can run:

bashCopy code

```bash
chmod +x myscript.py
```

This sets the executable flag for the file, allowing it to be executed as a program.

1. Optionally, you can move the script to a directory in your system's PATH environment variable. This allows you to run the script from anywhere in the terminal by simply typing its name, rather than specifying its full path. For example, you can move the script to /usr/local/bin/:

bashCopy code

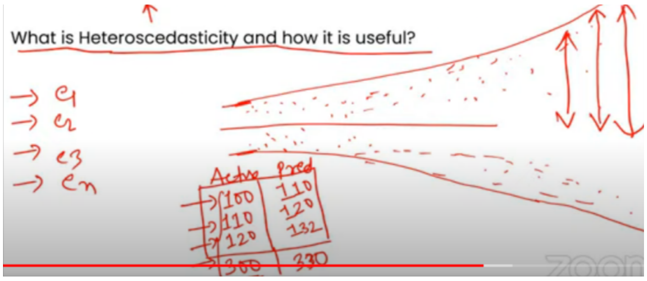
```bash
sudo mv myscript.py /usr/local/bin/
```

Now you can run the script by simply typing its name in the terminal:

Copy code

```
myscript.py
```

This will execute the script using the Python interpreter specified in the shebang line.



What is Heteroscedasticity and how it is useful?

Actual	Pred
100	110
110	120
120	132
300	330

Heteroscedasticity is a statistical term that refers to the phenomenon where the variance of the errors in a regression model is not constant across the range of predictor variables. In other words, the spread of the errors around the regression line varies across different levels of the independent variable(s).

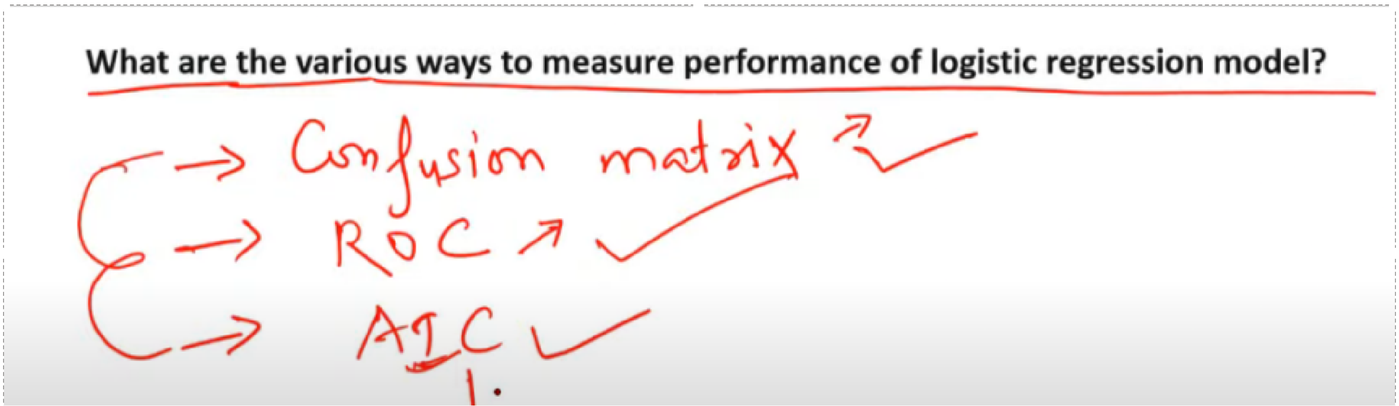
Heteroscedasticity can cause problems in regression analysis, such as biased and inconsistent parameter estimates, invalid hypothesis tests, and unreliable predictions. Therefore, it is important to detect and correct heteroscedasticity in regression models.

There are several methods to test for heteroscedasticity, such as visual inspection of residual plots, formal statistical tests (e.g., Breusch-Pagan test, White's test), and information criteria (e.g., AIC, BIC).

Once heteroscedasticity is detected, there are several techniques to correct it, such as:

1. Weighted least squares (WLS): This method assigns weights to the observations based on their variance, so that observations with high variance are given less weight and observations with low variance are given more weight.
2. Transformations: This method involves transforming the variables to reduce the heteroscedasticity. For example, taking the logarithm of a variable may reduce the heteroscedasticity.
3. Robust regression: This method uses alternative estimation techniques that are less sensitive to outliers and heteroscedasticity, such as M-estimation and iteratively reweighted least squares.

In summary, detecting and correcting heteroscedasticity is important to ensure the validity and reliability of regression models, and there are several techniques available to address this issue.



What are the various ways to measure performance of logistic regression model?

→ Confusion matrix
→ ROC
→ AIC

There are several ways to measure the performance of a logistic regression model, depending on the specific goals and requirements of the analysis. Here are some common measures of performance:

4. Accuracy: This is the proportion of correctly classified instances out of the total number of instances. Accuracy can be a useful measure when the classes are balanced and equally important.
5. Precision: This is the proportion of true positives out of the total number of predicted positives. Precision is a useful measure when the goal is to minimize false positives.
6. Recall: This is the proportion of true positives out of the total number of actual positives. Recall is a useful measure when the goal is to minimize false negatives.
7. F1 score: This is the harmonic mean of precision and recall. It is a useful measure when there is an unequal class distribution and you want to balance precision and recall.
8. Area under the receiver operating characteristic curve (AUC-ROC): This is a measure of the ability of the model to distinguish between the positive and negative classes across different probability thresholds. AUC-ROC ranges from 0 to 1, where 0.5 represents a random classifier and 1 represents a perfect classifier.
9. Confusion matrix: This is a table that shows the number of true positives, true negatives, false positives, and false negatives for a classifier. The confusion matrix can be used to calculate various performance metrics such as accuracy, precision, recall, and F1 score.
10. Log-loss: This is a measure of the difference between the predicted probabilities and the actual class labels. Log-loss penalizes incorrect and uncertain predictions more severely than correct and confident predictions.

In summary, there are several measures of performance for a logistic regression model, and the choice of measure depends on the specific objectives and trade-offs of the analysis. It is important to choose a measure that aligns with the business goals and interpret the results in context.

There are eight batteries, but only four of them work. You have to use them for a flashlight, which needs two working batteries. What is the minimum number of battery pairs you need to test to ensure that the flashlight is turned on?

Approach

Step 1: Divide the batteries into optimal segments, which in the worst case requires least number of tests to find a working pair.

Lets name the batteries A,B, C, D, E, F, G and H and we divide them into pairs of 3(ABC)–3(DEF)–2(GH)

Step 2: Let the testing begin

ABC can be **tested in 3 different ways**,

Possible testing combinations AB-AC-BC and possible type of batteries are as follows

1. all 3 does not work — all fails
2. 1 works and 2 does not work — all fails — *considering this for worst case*
3. 2 works and 1 does not work — one working pair found
4. all 3 works — one working pair found

DEF can also be **tested in 3 different way**,

Possible testing combinations DE-DF-EF and possible type of batteries are as follows

1. all 3 works — one working pair found
2. 2 works 1 does not — one working pair found
3. 2 does not work 1 works — all fails — *considering this for worst case*
4. all 3 does not work — cannot happen as we chose option 2. during ABC testing

GH can be **tested in 1 way** and both will be working as we chose option 2.

during ABC testing and option 3. during DEF testing

Solution

Minimum number of battery pair that will have to be tested is 7