Unit-3: Evaluating Models

Lesson Title: Evaluating Models	**Approach:** Session + Activity
Summary: In this module, students are introduced to the common metrics used to evaluate AI models. They will know how to derive and calculate the evaluation metrics and would also get an idea on how to improve the accuracy/efficiency of an AI Model. They will be introduced to the concept of Train/ Test Split, Common evaluation metrics such as Accuracy, Confusion Matrix, Precision, Recall, F1 Score) Learners will also be able to identify the use of this metrics in use cases encountered in everyday life.	
Learning Objectives: 1. To introduce students to the common metrics used to evaluate AI models 2. To familiarize students with deriving and calculating the evaluation metrics 3. To enable students to recognize the most suitable evaluation metric for a given application.	
Learning Outcomes: 1. Recognise common metrics used to evaluate AI models 2. Derive and calculate the evaluation metrics 3. Recognize the most suitable evaluation metric for a given application.	
Pre-requisites: Essential understanding of Artificial Intelligence	
Key-concepts: 1. Importance of model evaluation 2. Evaluation metrics for classification	

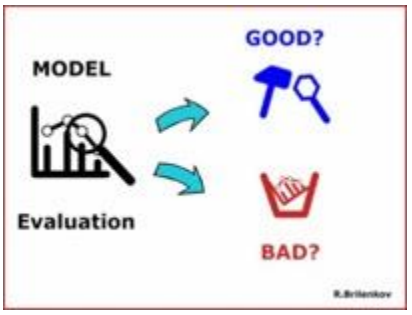
Introduction

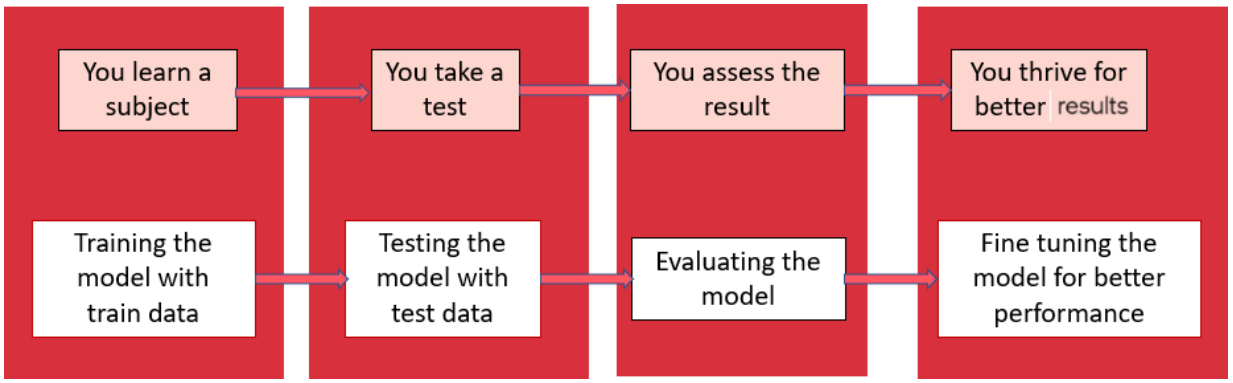
Till now we have learnt about the 4 stages of AI project cycle, viz. Problem scoping, Data acquisition, Data exploration and modelling. While in modelling we can make different types of models, how do we check if one's better than the other? That's where Evaluation comes into play. In the Evaluation stage, we will explore different methods of evaluating an AI model. Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future

3.1: Importance of Model Evaluation

What is evaluation?

- Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance
- An AI model gets better with constructive feedback
- You build a model, get feedback from metrics, make improvements and continue until you achieve a desirable accuracy





- It's like the report card of your school
- There are many parameters like grades, percentage, percentiles, ranks
- Your academic performance gets evaluated and you know where to work more to get better

* Images shown here are the property of individual organisations and are used here for reference purpose only.

Need of model evaluation

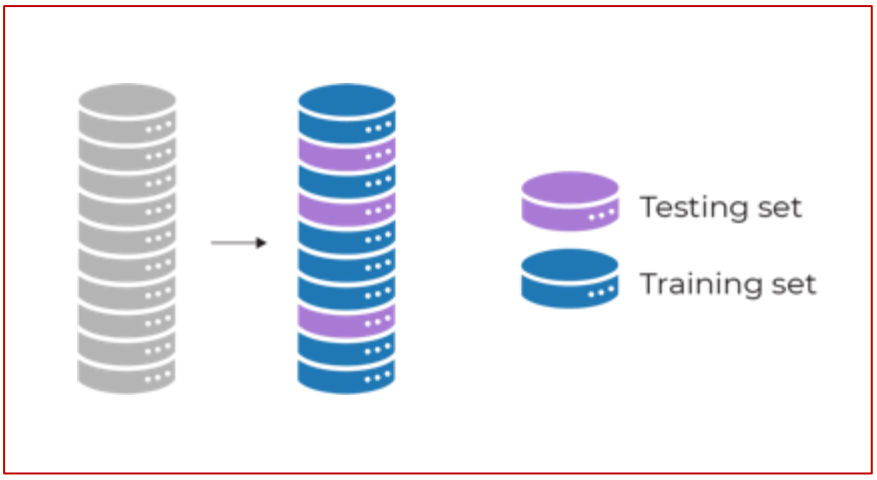
In essence, model evaluation is like giving your AI model a report card. It helps you understand its strengths, weaknesses, and suitability for the task at hand. This feedback loop is essential for building trustworthy and reliable AI systems.

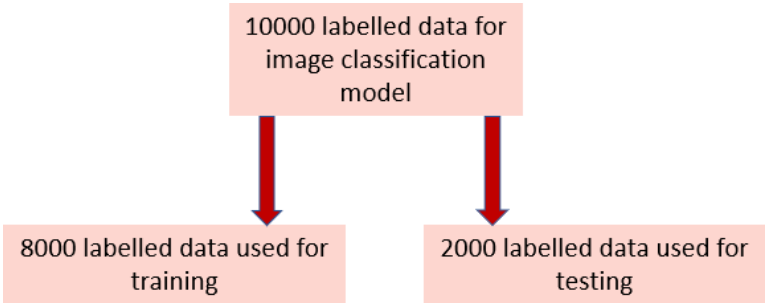
After understanding the need for Model Evaluation, let's know how to begin with the process. There can be different Evaluation techniques, depending of the type and purpose of the model.

3.2: Splitting the training set data for Evaluation

Train-test split

- The train-test split is a technique for evaluating the performance of a machine learning algorithm
- It can be used for any supervised learning algorithm
- The procedure involves taking a dataset and dividing it into two subsets: The training dataset and the testing dataset
- The train-test procedure is appropriate when there is a sufficiently large dataset available

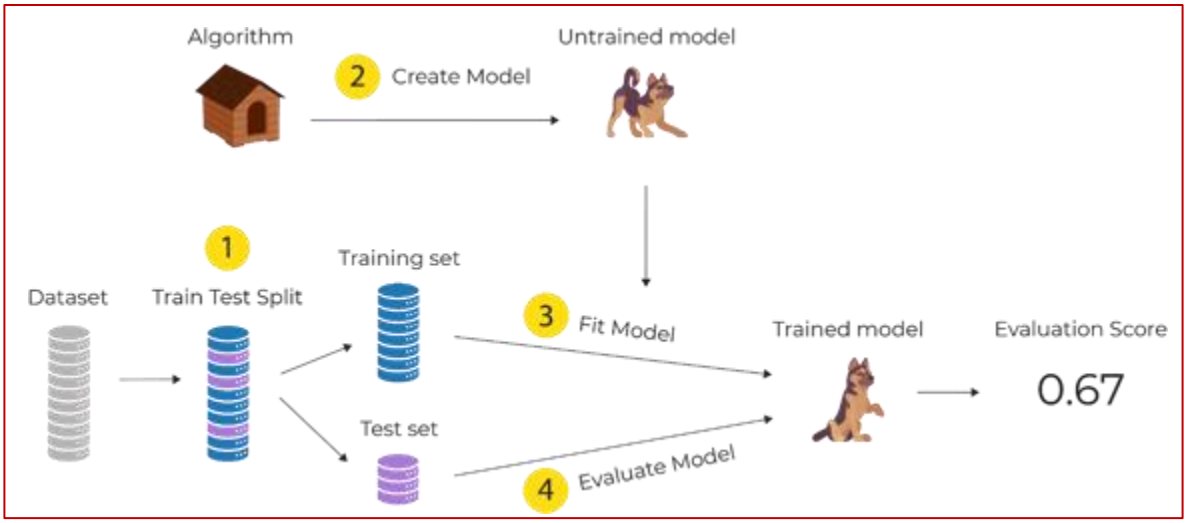




* Images shown here are the property of individual organisations and are used here for reference purpose only.

Need of Train-test split

- The train dataset is used to make the model learn
- The input elements of the test dataset are provided to the trained model. The model makes predictions, and the predicted values are compared to the expected values
- The objective is to estimate the performance of the machine learning model on new data: data not used to train the model



This is how we expect to use the model in practice. Namely, to fit it on available data with known inputs and outputs, then make predictions on new examples in the future where we do not have the expected output or target values.

Remember that It's not recommended to use the data we used to build the model to evaluate it. This is because our model will simply remember the whole training set, and will therefore always predict the correct label for any point in the training set. This is known as **overfitting**.

You will learn more about the concepts including train test split and cross validation in higher classes.

3.3: Accuracy and Error

- Bob and Billy went to a concert
- Bob brought Rs 300 and Billy brought Rs 550 as the entry fee for that
- The entry fee per person was Rs 500
- Can you tell:
  - Who is more accurate? Bob or Billy?
  - How much is the error for both Bob and Billy in estimating the concert entry fee?

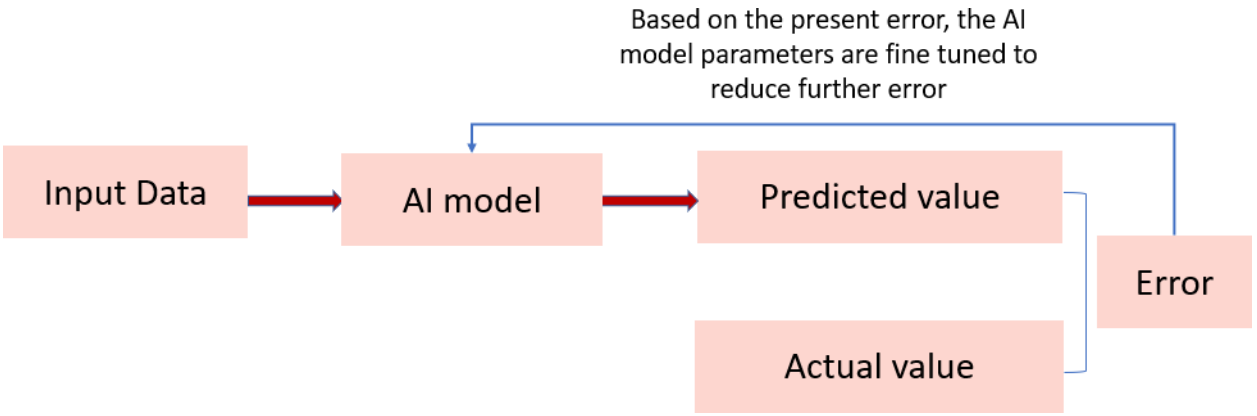
Accuracy

- Accuracy is an evaluation metric that allows you to measure the total number of predictions a model gets right.
- The accuracy of the model and performance of the model is directly proportional, and hence better the performance of the model, the more accurate are the predictions.

Error

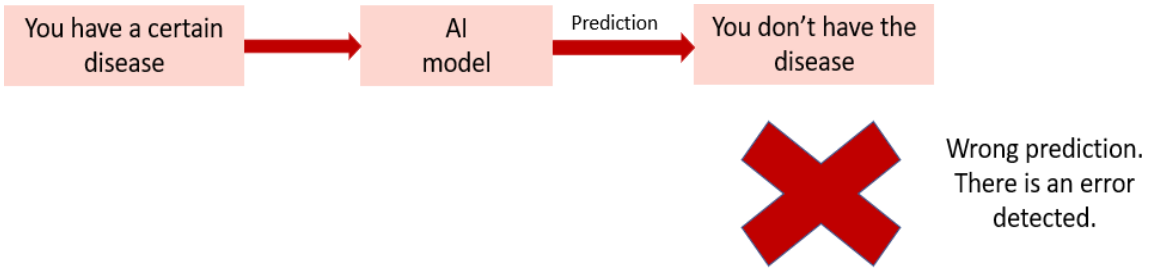
- Error can be described as an action that is inaccurate or wrong.
- In Machine Learning, the error is used to see how accurately our model can predict data it uses to learn new, unseen data.
- Based on our error, we choose the machine learning model which performs best for a particular dataset.

Error refers to the difference between a model's prediction and the actual outcome. It quantifies how often the model makes mistakes.



Imagine you're training a model to predict if you have a certain disease (classification task).

- **Error:** If the model predicts you don't have a disease but you actually have a disease, that's an error. The error quantifies how far off the prediction was from reality.



- **Accuracy:** If the model correctly predicts disease or no disease for a particular period, it has 100% accuracy for that period.

Key Points:

- Here the goal is to minimize error and maximize accuracy.
- Real-world data can be messy, and even the best models make mistakes.
- Sometimes, focusing solely on accuracy might not be ideal. For instance, in medical diagnosis, a model with slightly lower accuracy but a strong focus on avoiding incorrectly identifying a healthy person as sick might be preferable.
- Choosing the right error or accuracy metric depends on the specific task and its requirements.

Understanding both error and accuracy is crucial for effectively evaluating and improving AI models.

Activity 1: Find the accuracy of the AI model

Purpose: To understand how to calculate the error and the accuracy.

Say: "The youth will understand the concept of accuracy and error and practice it mathematically."

Calculate the accuracy of the House Price prediction AI model

- Read the instructions and fill in the blank cells in the table.
- The formula for finding error and accuracy is shown in the table
- Accuracy of the AI model is the mean accuracy of all five samples
- Percentage accuracy can be seen by multiplying the accuracy with 100

Predicted House Price (USD)	Actual House Price (USD)	Error Abs (Actual-Predicted)	Error Rate (Error/Actual)	Accuracy (1-Error rate)	Accuracy% (Accuracy*100)%
391k	402k	Abs (402k-391k)= 11k	11k/402k=0.027	1-0.027= 0.973	0.973*100%= 97.3%
453k	488k				
125k	97k				
871k	907k				
322k	425k				

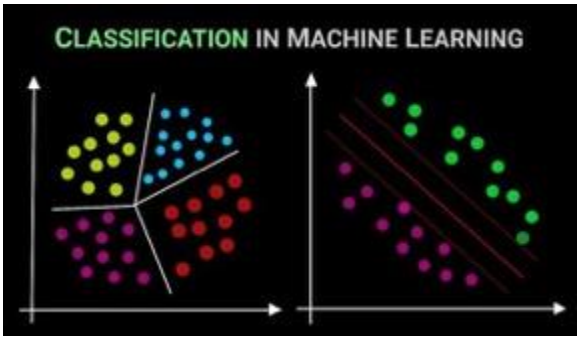
*Abs means the absolute value, which means only the magnitude of the difference without any negative sign (if any)

The Model Evaluation stands on the two pillars of accuracy and error. Let's understand some more metrics standing on these two pillars.

3.4: Evaluation metrics for Classification

What is Classification?

- You go to a supermarket and were given two trolleys
- In one, you have to place the fruits and vegetables; in the other, you must put the grocery items like bread, oil, egg, etc.
- So basically, you are classifying the items of the supermarket into two classes:
    - fruits and vegetables
    - grocery
- Classification usually refers to a problem where a specific type of class label is the result to be predicted from the given input field of data
- For example, here we are working on a vegetable-grocery-classifier model that predicts whether the item in the supermarket is a vegetable or a grocery item



Visualizing the concept of classification: Left 4 Classes; Right 2 classes

Try Yourself:

Which of this is a classification use case example?

House price prediction

Credit card fraud detection

Salary prediction

Classification Metrics

Popular metrics used for classification model

- Confusion matrix
- Classification accuracy
- Precision
- Recall

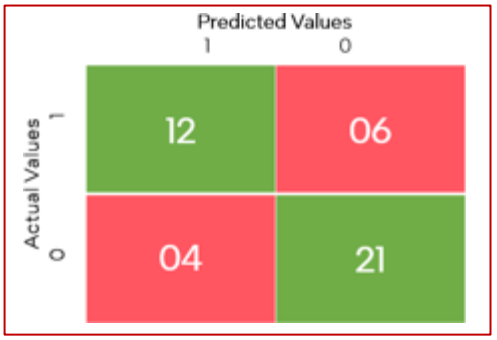
Let's understand these metrics in details:

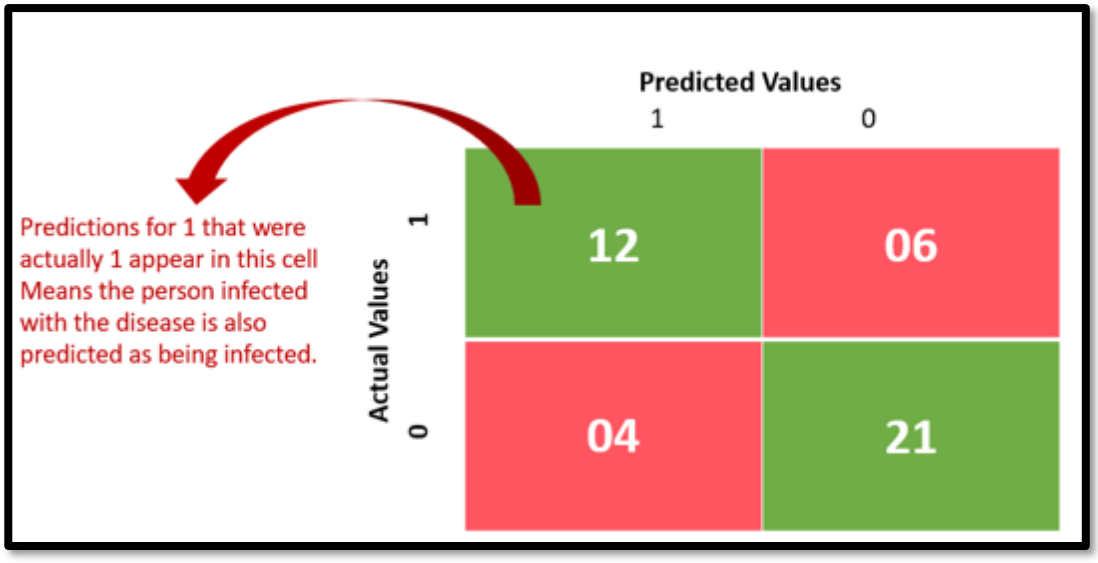
Confusion matrix

Let's say, based on some clinical parameters; you have designed a classifier that predicts whether a person is infected with a certain disease or not.

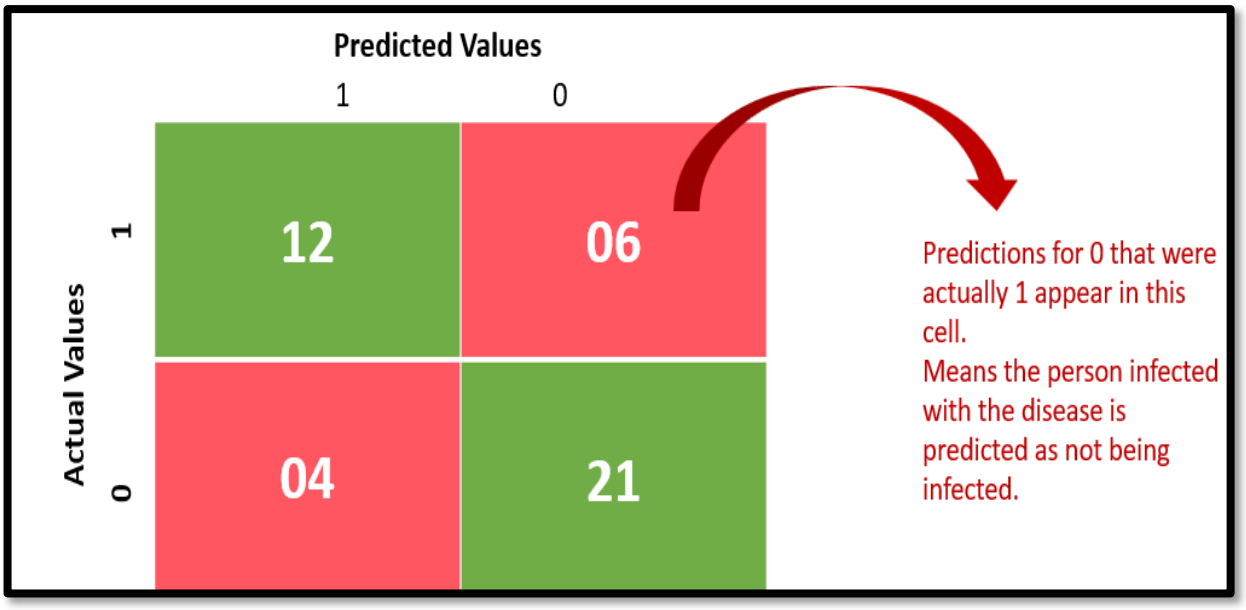
The output is 1 if the person is infected or 0 if the person is not infected. That is, 1 and 0 signify whether a person is infected or not.

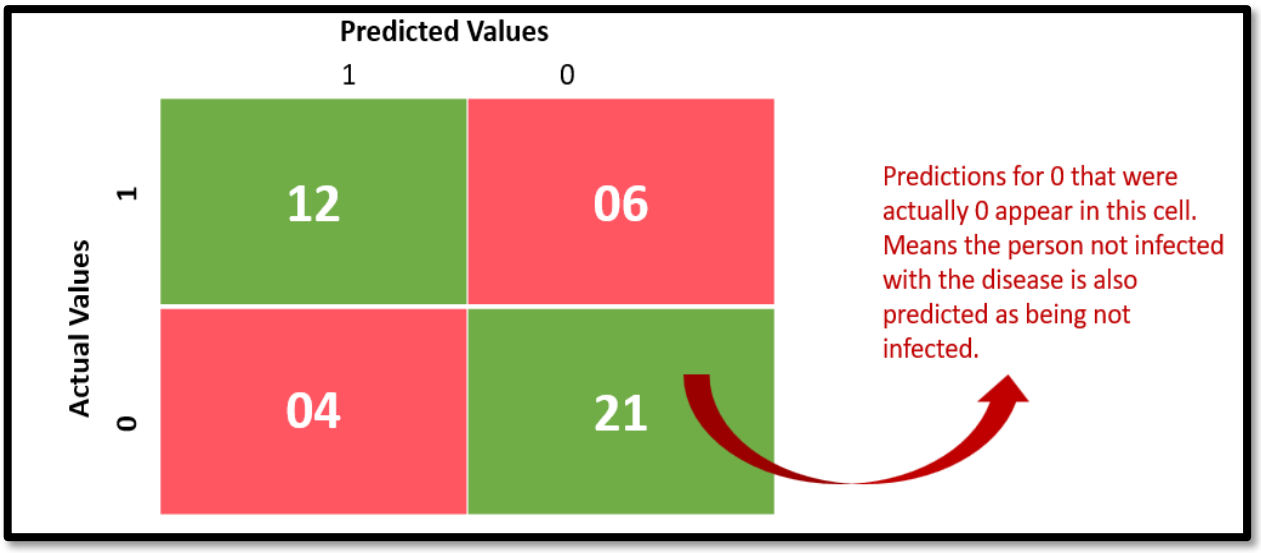
- The confusion matrix is a handy presentation of the accuracy of a model with two or more classes
- The table presents the actual values on the y-axis and predicted values on the x-axis
- The numbers in each cell represents the number of predictions made by a machine learning algorithm that falls into that particular category

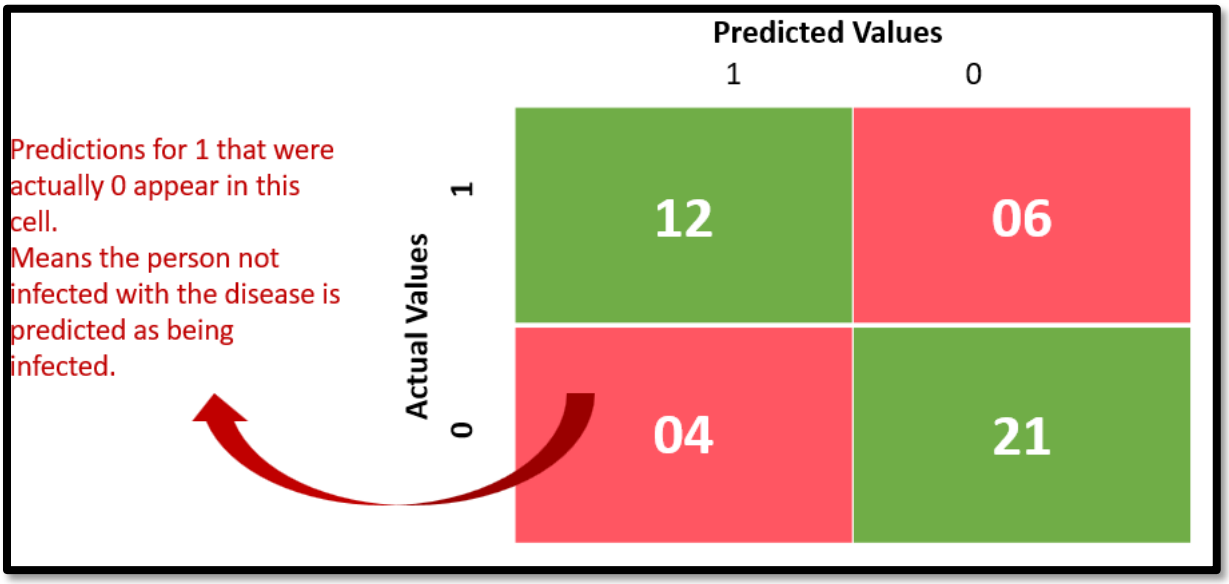




For example, a machine learning algorithm can predict 0 or 1 and each prediction may actually have been a 0 or 1.







Activity 2: Build the confusion matrix from scratch

Duration: 10 minutes

Purpose: Learn how to create confusion matrix from the scratch.

Say: “The youth need to analyze the situation and tabulate a non-numerical information a numerical one.”

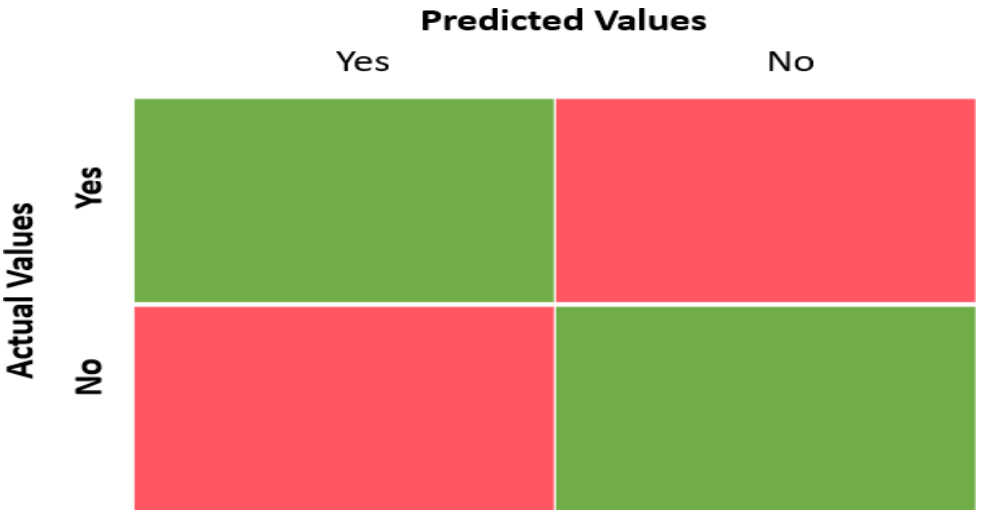
Activity Guidelines

- Let’s assume we were predicting the presence of a disease; for example, "yes" would mean they have the disease, and "no" would mean they don't have the disease
- So, the AI model will have output is Yes or No
- The following chart shows the actual values and the predicted values
    - Construct a confusion matrix.
    - Can you tell how many are correct predictions among all predictions?

Actual value	Predicted Value
Yes	Yes
No	No
No	Yes
Yes	No
No	No
Yes	Yes
Yes	No
No	No
No	No
No	No

* Images shown here are the property of individual organisations and are used here for reference purpose only.

Fill the matrix based on the table given here.



Count the number of rows having YES in both columns of the table and put the count in the first cell. Similarly, number of rows having YES in the first column and NO in the second column will be shown in the top right cell of confusion matrix. Number of rows having NO in the first column and YES in the second column will be shown in the down left cell of confusion matrix. Lastly, number of rows having NO in the first column and YES in the second column will be shown in the downright cell of confusion matrix.

Activity Guidelines – Solution

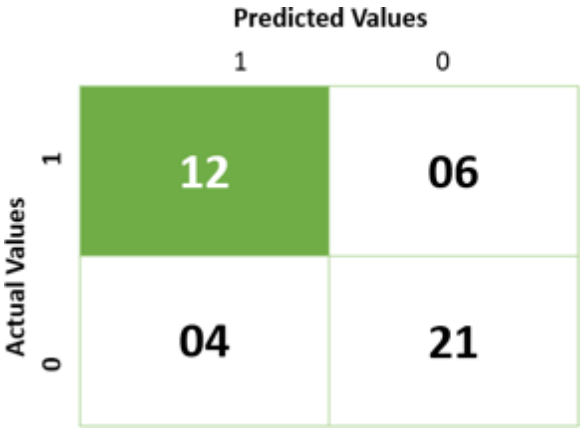
Activity Reflection

- So, there are 07 correct predictions out of 10 predictions.
- What do you think? How good is your model?

Now that you know how to construct a Confusion matrix, let's understand each cell of the matrix in details.

True Positive

- **True Positive** (**TP**) is the outcome of the model correctly predicting the positive class
- Any class can be assumed as a positive class, and the rest can be assumed as negative
- Let's say class 1 is assumed as the positive class
- Can you tell the TP value from this matrix?



Scenario 1:

Consider you are watching football world cup. Scenario 2:

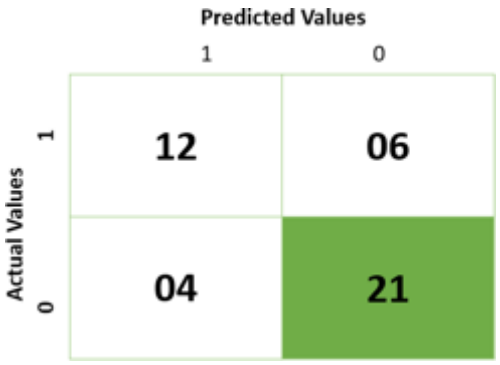
Consider the earlier example of medical diagnosis of an infected disease.

True Positive examples

- You had predicted that France would win the world cup, and it won.
- In the earlier activity, the cases in which we predicted yes (they have the disease), and they do have the disease.

True Negative

- **True Negative** (**TN**) is the outcome of the model correctly predicting the negative class.
- Since in the previous example, class 1 is assumed the positive class, class 0 should be assumed the negative class.
- Can you tell the TN value from this matrix?

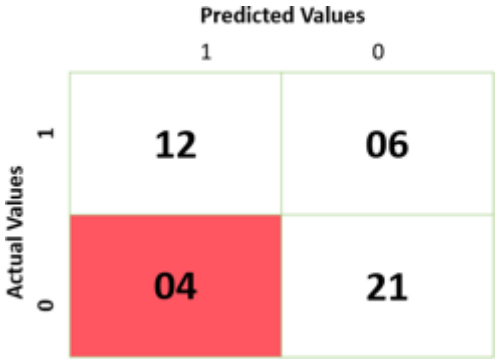


True Negative examples

- You had predicted that Germany would not win, and it lost
- In the earlier activity, the cases in which we predicted No (they don't have the disease), and they don't have the disease

False Positive

- **False Positive (FP)** is the outcome of the model wrongly predicting the negative class as positive class.
- Here, when a class 0 is predicted as class 1, it falls into the FP cell.
- Can you tell the FP value from this matrix?

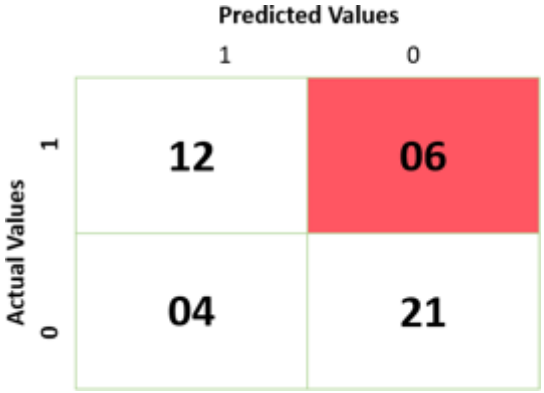


False Positive examples

- You had predicted that Germany would win, but it lost.
- In the earlier activity, the cases in which we predicted Yes (they have the disease), and they don't have the disease.

False Negative

- **False Negative (FN)** is the outcome of the model wrongly predicting the positive class as the negative class.
- Here, when class 1 is predicted as class 0, it falls into the FN cell.
- Can you tell the FN value from this matrix?



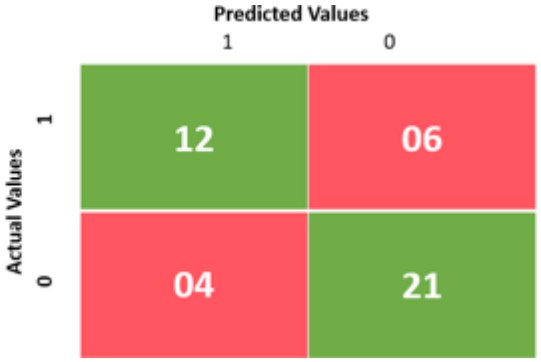
False Negative examples

- You had predicted that France would not win but it won
- In the earlier activity, the cases in which we predicted No (they don't have the disease), and they have the disease

Accuracy from Confusion matrix

Classification accuracy is the number of correct predictions made as a ratio of all predictions made.

Can you spot the correct predictions?



Calculate the Classification accuracy from this confusion matrix.

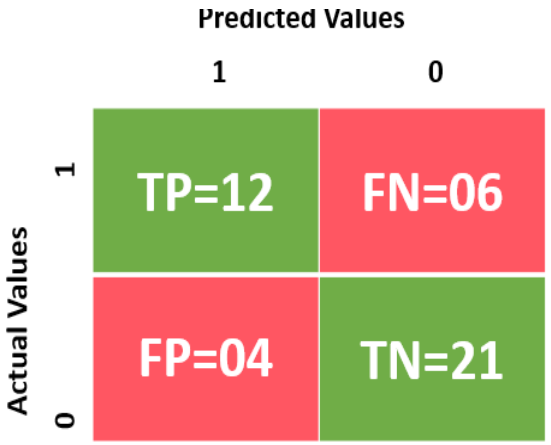
Correct predictions=TP+TN

Total predictions=TP+TN+FP+FN

$$\text{Classification accuracy} = \frac{\text{Correct predictions}}{\text{Total predictions}}$$

$$= \frac{TP+TN}{TP+TN+FP+FN}$$

$$= \frac{12+21}{12+21+04+06} = 0.767$$



Can we use Accuracy all the time?

- It is only suitable when there are an equal number of observations in each class, i.e., a balanced dataset (which is rarely the case), and that all predictions and prediction errors are equally important, which is often not the case.
- But why is that so? Let's understand it better from the next activity

Activity 3: Calculate the accuracy of the classifier model

Duration: 20 minutes

Purpose: To design an AI model that predicts whether a student will pass a test (Yes) or not pass a test (No).

Say: It classifies the input into two classes Yes and No. Also, calculate the accuracy of the classifier model and construct the confusion matrix for the model.

Activity Guidelines

- Let's assume you are testing your model on 1000 total test data.
- Out of which the actual values are 900 Yes and only 100 No (Unbalanced dataset).
- Let's assume that you have built a faulty model which, irrespective of any input, will give a prediction as Yes.
- Can you tell the classification accuracy of this model?

Step 1: Construct the Actual value vs Predicted value table

Actual value	Predicted Value

Consider '*Yes* 'as the positive class and '*No* 'as the negative class.

Step 2: Construct the confusion matrix.

Activity solution: Accuracy from Confusion matrix

So, the faulty model will predict all the 1000 input data as Yes.

Actual value	Predicted Value
Yes=900	Yes=1000
No=100	No=0

Consider '*Yes* ' as the positive class and '*No* ' as the negative class.

Construct the confusion matrix from the Actual vs Predicted table.

Activity solution: Accuracy from Confusion matrix

Step 3: Now calculate the accuracy from this matrix.

$$\text{Classification accuracy} = \frac{\text{Correct predictions}}{\text{Total predictions}}$$

$$= \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}}$$

Step 4: Converting the accuracy to percentage: = %

$$\text{Classification accuracy} = \frac{\text{Correct predictions}}{\text{Total predictions}}$$

$$= \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}}$$

$$= \frac{900}{\mathbf{900+0+100+0}} = 0.9$$

Converting the accuracy to percentage: 0.9x100 % = 90%

So, the faulty model you made is showing an accuracy of 90%. Does this make sense?

So, ***in cases of unbalanced data, we should use other metrics such as Precision, Recall or F1 score***. Let's understand them one by one…

Precision from Confusion matrix

- **Precision** is the ratio of the total number of correctly classified positive examples and the total number of predicted positive examples.
- Precision = 0.843 means that when our model predicts a patient has heart disease, it is correct around 84% of the time.

$$\textbf{Precision} = \frac{\textbf{Correct positive predictions}}{\textbf{Total positive predictions}}$$

$$= \frac{\textbf{TP}}{\textbf{TP+FP}}$$

Precision: where should we use it?

The metrics Precision is generally used for unbalanced datasets when dealing with the False Positives become important, and the model needs to reduce the FPs as much as possible.

Precision use case example

- For example, take the case of predicting a good day based on weather conditions to launch satellite.
- Let's assume a day with favorable weather condition is considered Positive class and a day with non-favorable weather condition is considered as Negative class.
- Missing out on predicting a good weather day is okay (low recall) but predicting the bad weather day (Negative class) as a good weather day (Positive class) to launch the satellite can be disastrous.
- So, in this case, the FPs need to be reduced as much as possible.

Recall from Confusion matrix

- The recall is the measure of our model correctly identifying True Positives
- Thus, for all the patients who actually have heart disease, recall tells us how many we correctly identified as having a heart disease. Recall = 0.86 tells us that out of the total patients who have heart disease 86% have been correctly identified.

$$\text{Recall} = \frac{\text{Correct positive predictions}}{\text{Total actual positive values}}$$

$$= \frac{TP}{TP+FN}$$

Do you know Recall is also called as Sensitivity or True Positive Rate?

Recall: Where we should we use it?

The metrics Recall is generally used for unbalanced dataset when dealing with the False Negatives become important and the model needs to reduce the FNs as much as possible.

Recall use case example

For example, for a covid-19 prediction classifier, let's consider detection of a covid-19 affected case as positive class and detection of covid-19 non-affected case as negative class.

- Imagine if a covid-19 affected person (Positive) is falsely predicted as non-affected of Covid-19 (Negative), the person if rely solely on the AI would not get any treatment and also may end up infecting many other persons.
- So, in this case, the FNs needs to be reduced as much as possible.
- Hence, Precision is a go-to metrics for this kind of use case.

* Images shown here are the property of individual organisations and are used here for reference purpose only.

F1 Score

- F1-Score provides a way to combine both precisions and recall into a single measure that captures both properties
- In those use cases, where the dataset is unbalanced, and we are unable to decide whether FP is more important or FN, we should use the F1 score as the suitable metric.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Activity 4: Decide the appropriate metric to evaluate the AI model

Duration: 30 minutes

Purpose: To work with the given scenario and choose the most appropriate evaluation metric to evaluate their model.

Say: "Different evaluation metrics are used for evaluation in different scenarios and it is important that we realize how to choose the correct one."

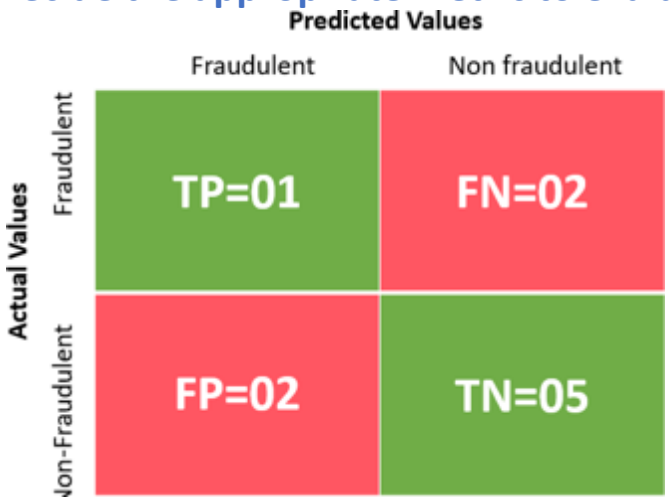
Scenario: Flagging fraudulent transactions

- You have designed a model to detect any fraudulent transactions with credit card.
- You are testing your model with highly unbalanced dataset.
- What is the metric to be considered in this case?
- It is okay to classify a legit transaction as fraudulent — it can always be re-verified by passing through additional checks.
- But it is definitely not okay to classify a fraudulent transaction as legit (false negative).
- So here false negatives should be reduced as much as possible.
- Hence in this case, Recall is more important.
- For the given data, construct the confusion matrix.
- Calculate the recall from the confusion matrix.

Transaction ID	Actual value	Predicted Value
1	Non-Fraudulent	Non-Fraudulent
2	Non-Fraudulent	Fraudulent
3	Non-Fraudulent	Non-Fraudulent
4	Fraudulent	Non-Fraudulent
5	Fraudulent	Fraudulent
6	Non-Fraudulent	Non-Fraudulent
7	Fraudulent	Non-Fraudulent
8	Non-Fraudulent	Fraudulent
9	Non-Fraudulent	Non-Fraudulent
10	Non-Fraudulent	Non-Fraudulent

Fill the matrix based on the table given above.

Activity solution: Decide the appropriate metric to evaluate the AI model



Calculate the recall from the confusion matrix based on.

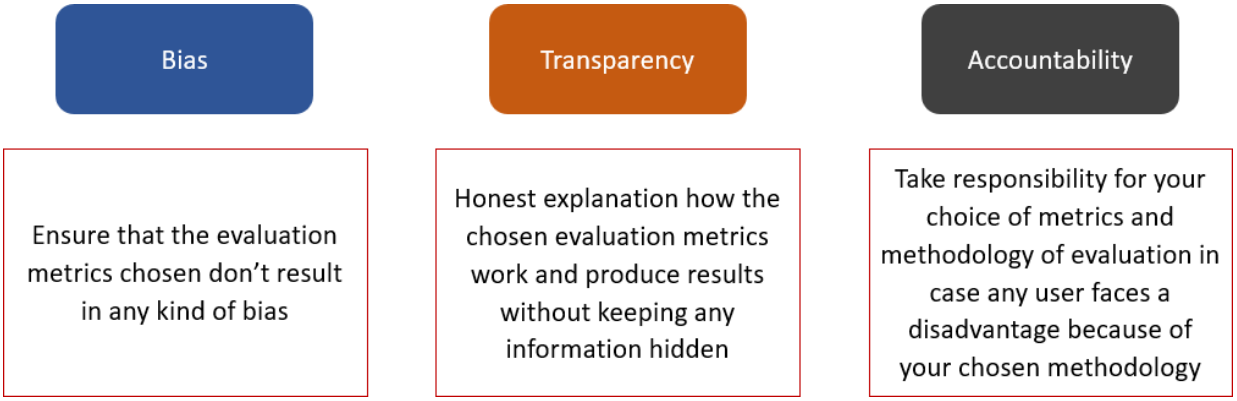
Write the formula for recall:

Calculate recall from the formula:

Are there any ethical concerns we need to keep in mind when performing model evaluation?

3.2 Ethical concerns around model evaluation

While evaluating an AI model, the following ethical concerns need to be kept in mind



Test Yourself

Choose the most appropriate answer for each question.

1. In a medical test for a rare disease, out of 1000 people tested, 50 actually have the disease while 950 do not. The test correctly identifies 40 out of the 50 people with the disease as positive, but it also wrongly identifies 30 of the healthy individuals as positive. What is the accuracy of the test?

   A) 97%

   B) 90%

   C) 85%

   D) 70%

2. A student solved 90 out of 100 questions correctly in a multiple-choice exam. What is the error rate of the student's answers?

   A) 10%

   B) 9%

   C) 8%

D) 11%

3. In a spam email detection system, out of 1000 emails received, 300 are spam. The system correctly identifies 240 spam emails as spam, but it also marks 60 legitimate emails as spam. What is the precision of the system?

   A) 80%

   B) 70%

   C) 75%

   D) 90%

4. In a binary classification problem, a model predicts 70 instances as positive out of which 50 are actually positive. What is the recall of the model?

   A) 50%

   B) 70%

   C) 80%

   D) 100%

5. In a sentiment analysis task, a model correctly predicts 120 positive sentiments out of 200 positive instances. However, it also incorrectly predicts 40 negative sentiments as positive. What is the F1 score of the model?

   A) 0.8

   B) 0.75

   C) 0.72

   D) 0.82

6. A medical diagnostic test is designed to detect a certain disease. Out of 1000 people tested, 100 have the disease, and the test identifies 90 of them correctly. However, it also wrongly identifies 50 healthy people as having the disease. What is the precision of the test?

   A) 90%

   B) 80%

   C) 70%

   D) 60%

7. A teacher's marks prediction system predicts the marks of a student as 75, but the actual marks obtained by the student are 80. What is the absolute error in the prediction?

   A) 5

   B) 10

   C) 15

   D) 20

8. The goal when evaluating an AI model is to:

A) Maximize error and minimize accuracy

B) Minimize error and maximize accuracy

C) Focus solely on the number of data points used

D) Prioritize the complexity of the model

9. A high F1 score generally suggests:

A) A significant imbalance between precision and recall

B) A good balance between precision and recall

C) A model that only performs well on specific data points

D) The need for more training data

10. How is the relationship between model performance and accuracy described?

A) Inversely proportional

B) Not related

C) Directly proportional

D) Randomly fluctuating

Reflection Time:

Q1. What will happen if you deploy an AI model without evaluating it with known test set data? Q2.

Do you think evaluating an AI model is that essential in an AI project cycle?

Q3. Explain train-test split with an example.

Q4. "Understanding both error and accuracy is crucial for effectively evaluating and improving AI models." Justify this statement.

Q5. What is classification accuracy? Can it be used all times for evaluating AI models?

Assertion and reasoning-based questions:

Q1. Assertion: Accuracy is an evaluation metric that allows you to measure the total number of predictions a model gets right.
Reasoning: The accuracy of the model and performance of the model is directly proportional, and hence better the performance of the model, the more accurate are the predictions.

Choose the correct option:
(a) Both A and R are true and R is the correct explanation for A
(b) Both A and R are true and R is not the correct explanation for A
(c) A is True but R is False
(d) A is false but R is True

Q2. Assertion: The sum of the values in a confusion matrix's row represents the total number of instances for a given actual class.
Reasoning: This enables the calculation of class-specific metrics such as precision and recall, which are essential for evaluating a model's performance across different classes.
Choose the correct option:
(a) Both A and R are true and R is the correct explanation for A
(b) Both A and R are true and R is not the correct explanation for A
(c) A is True but R is False
(d) A is false but R is True

Case study-based questions:

Q1. Identify which metric (Precision or Recall) is to be used in the following cases and why?

a) Email Spam Detection
b) Cancer Diagnosis
c) Legal Cases (Innocent until proven guilty)
d) Fraud Detection
e) Safe Content Filtering (like Kids YouTube)

Q2. Examine the following case studies. Draw the confusion matrix and calculate metrics such as accuracy, precision, recall, and F1-score for each one of them.

a. Case Study 1:

A spam email detection system is used to classify emails as either spam (1) or not spam (0). Out of 1000 emails:

- True Positives (TP): 150 emails were correctly classified as spam.

- False Positives (FP): 50 emails were incorrectly classified as spam.

- True Negatives (TN): 750 emails were correctly classified as not spam.

- False Negatives (FN): 50 emails were incorrectly classified as not spam.

-

b. Case Study 2:

A credit scoring model is used to predict whether an applicant is likely to default on a loan (1) or not (0). Out of 1000 loan applicants:

- True Positives (TP): 90 applicants were correctly predicted to default on the loan.

- False Positives (FP): 40 applicants were incorrectly predicted to default on the loan.

- True Negatives (TN): 820 applicants were correctly predicted not to default on the loan.

- False Negatives (FN): 50 applicants were incorrectly predicted not to default on the loan.

Calculate metrics such as accuracy, precision, recall, and F1-score.

c. Case Study 3:

A fraud detection system is used to identify fraudulent transactions (1) from legitimate ones (0). Out of 1000 transactions:

- True Positives (TP): 80 transactions were correctly identified as fraudulent.
- False Positives (FP): 30 transactions were incorrectly identified as fraudulent.
- True Negatives (TN): 850 transactions were correctly identified as legitimate.
- False Negatives (FN): 40 transactions were incorrectly identified as legitimate.

d. Case Study 4:

A medical diagnosis system is used to classify patients as having a certain disease (1) or not having it (0). Out of 1000 patients:

- True Positives (TP): 120 patients were correctly diagnosed with the disease.
- False Positives (FP): 20 patients were incorrectly diagnosed with the disease.
- True Negatives (TN): 800 patients were correctly diagnosed as not having the disease.
- False Negatives (FN): 60 patients were incorrectly diagnosed as not having the disease.

e. Case Study 5:

An inventory management system is used to predict whether a product will be out of stock (1) or not (0) in the next month. Out of 1000 products:

- True Positives (TP): 100 products were correctly predicted to be out of stock.
- False Positives (FP): 50 products were incorrectly predicted to be out of stock.True Negatives (TN): 800 products were correctly predicted not to be out of stock.
- True Negatives (TN): 800 products were correctly predicted not to be out of stock.
-
- False Negatives (FN): 50 products were incorrectly predicted not to be out of stock.

Q1. In a medical test for a rare disease, out of 1000 people tested, 50 actually have the disease while 950 do not. The test correctly identifies 40 out of the 50 people with the disease as positive, but it also wrongly identifies 30 of the healthy individuals as positive. What is the accuracy of the test?

Given:

- Total people tested = **1000**
- People who actually have the disease = **50**
- People who don't have the disease = **950**
- Test correctly identifies **40 out of 50** = **True Positives (TP) = 40**
- Test wrongly identifies **30 healthy as positive** = **False Positives (FP) = 30**
- Therefore:
    - **False Negatives (FN)** = 50 - 40 = **10** (missed disease cases)
    - **True Negatives (TN)** = 950 - 30 = **920** (healthy people correctly identified)

- Correct Predictions = TP + TN = 40 + 920 = **960**
- Total Predictions = 1000

Accuracy = TP + TN / TP + TN + FP + FN = 960/1000 = .96 x 100 = 96 %

Q3. In a spam email detection system, out of 1000 emails received, 300 are spam. The system correctly identifies 240 spam emails as spam, but it also marks 60 legitimate emails as spam. What is the precision of the system?

Given:

- Total emails = **1000**
- Actual spam emails = **300**
- Spam emails correctly identified as spam = ✓ **True Positives (TP) = 240**
- Legitimate emails wrongly marked as spam = ✗ **False Positives (FP) = 60**

Precision = TP / TP +FP = 240 / 240 + 60 = 240 /300 = .80 x 100 = 80%

Q5. In a sentiment analysis task, a model correctly predicts 120 positive sentiments out of 200 positive instances. However, it also incorrectly predicts 40 negative sentiments as positive. What is the F1 score of the model?

Given:

- Total actual **positive** instances = **200**
- Model correctly predicts **120** of them → ✅ **True Positives (TP) = 120**
- Model wrongly predicts **40 negative** sentiments as positive → ❌ **False Positives (FP) = 40**
- So, the model **missed 80** actual positives → ❌ **False Negatives (FN) = 200 - 120 = 80**

Step 1: Calculate **Precision**

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{120}{120 + 40} = \frac{120}{160} = 0.75$$

Step 2: Calculate **Recall**

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{120}{120 + 80} = \frac{120}{200} = 0.6$$

Step 3: Calculate **F1 Score**

$$\text{F1} = 2 \times \frac{0.75 \times 0.6}{0.75 + 0.6} = 2 \times \frac{0.45}{1.35} = 2 \times 0.3333 = \boxed{0.6667 \text{ or } 66.67\%}$$

Q6. A medical diagnostic test is designed to detect a certain disease. Out of 1000 people tested, 100 have the disease, and the test identifies 90 of them correctly. However, it also wrongly identifies 50 healthy people as having the disease. What is the precision of the test?

Precision Formula:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- Total people tested = **1000**
- People who actually have the disease = **100**
- Test correctly identifies **90** of them → ✅ **True Positives (TP) = 90**
- Test wrongly says **50 healthy people** have the disease → ❌ **False Positives (FP) = 50**

Plug into the formula:

$$\text{Precision} = \frac{90}{90 + 50} = \frac{90}{140} \approx 0.6429$$

$$\boxed{\text{Precision} \approx 64.29\%}$$

Answer the following:

Q1. What will happen if you deploy an AI model without evaluating it with known test set data?

A1. If you **don't test your AI model** before using it in the real world, you won't know how **accurate, reliable, or fair** it is.

- **Poor performance:**
The model may make a lot of **mistakes**, like wrong predictions or classifications.

- **No idea of accuracy:**
You won't know if the model works well or not because it was never tested on data with known answers.

- **Risk in real-world use:**
In critical areas like **healthcare, finance, or safety**, a bad model can cause **harm or big losses**.

- **Bias and unfairness:**
The model might be **biased** (e.g., against certain groups) but you won't catch this unless you evaluate it.

- **Loss of trust:**
Users will **lose trust** if the AI gives wrong or unpredictable results.

Q2. Do you think evaluating an AI model is that essential in an AI project cycle?

- **To measure performance:** Evaluation tells us how well the model is doing — in terms of **accuracy, precision, recall, F1 score**, etc.

- **To identify mistakes:** Helps detect where the model is going wrong and how to improve it.

- **To ensure fairness and safety:** Prevents **biased or harmful decisions**, especially in fields like medicine, education, or hiring.

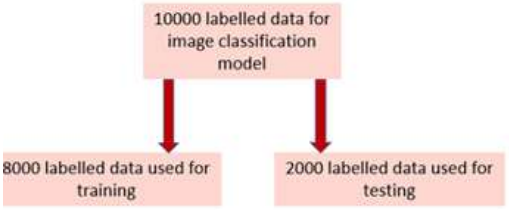
- **To compare different models:** Evaluation helps you choose the **best model** among several options.

Q3. Explain train-test split with an example.

Train-test split is a basic method used in machine learning to **evaluate how well a model will perform on new, unseen data**.

We split the available data into two parts:

- **Training set** – to teach the model
- **Testing set** – to check how well the model has learned



Q4. "Understanding both error and accuracy is crucial for effectively evaluating and improving AI models." Justify this statement.

To build a **good AI model**, you must know **how often it's right (accuracy)** and **how often it's wrong (error).**

- **Accuracy** tells you how many predictions were correct out of all predictions.
- A high accuracy means the model is doing well **overall**.

- **Error** shows where and how the model is going wrong.
    - It helps identify: **Wrong predictions, Biases in the data, Weak areas in the model**

Example: In a medical diagnosis AI:

- **Accuracy = 95%** → Sounds good!
- But if the model misses 5 out of 10 cancer cases (False Negatives), that's a **serious error**.

Understanding both accuracy and error gives a complete picture of how well your AI model works and what needs fixing. You can't improve what you don't measure.

Q5. What is classification accuracy? Can it be used all times for evaluating AI models?

Classification accuracy is the percentage of correct predictions made by a classification model.

It is a basic and useful metric, but it should **not be used alone** in every situation — especially when data is imbalanced or when different types of errors have different consequences.
Other metrics like **precision, recall, and F1 score** are often more helpful.

Formula:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Q1. Identify which metric (Precision or Recall) is to be used in the following cases and why?

a) Email Spam Detection — Precision

b) Cancer Diagnosis — Recall

c) Legal Cases (Innocent until proven guilty) — Precision

d) Fraud Detection — Recall

e) Safe Content Filtering (like Kids YouTube) — Precision