INDIAN SCHOOL AL WADI AL KABIR

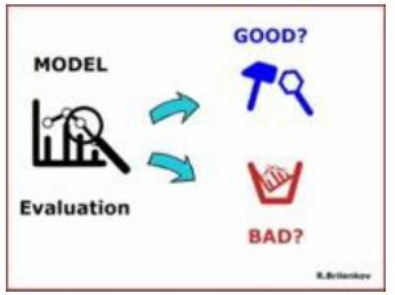
CLASS X -ARTIFICIAL INTELLIGENCE (417)

Part B Unit-3: Evaluating Models

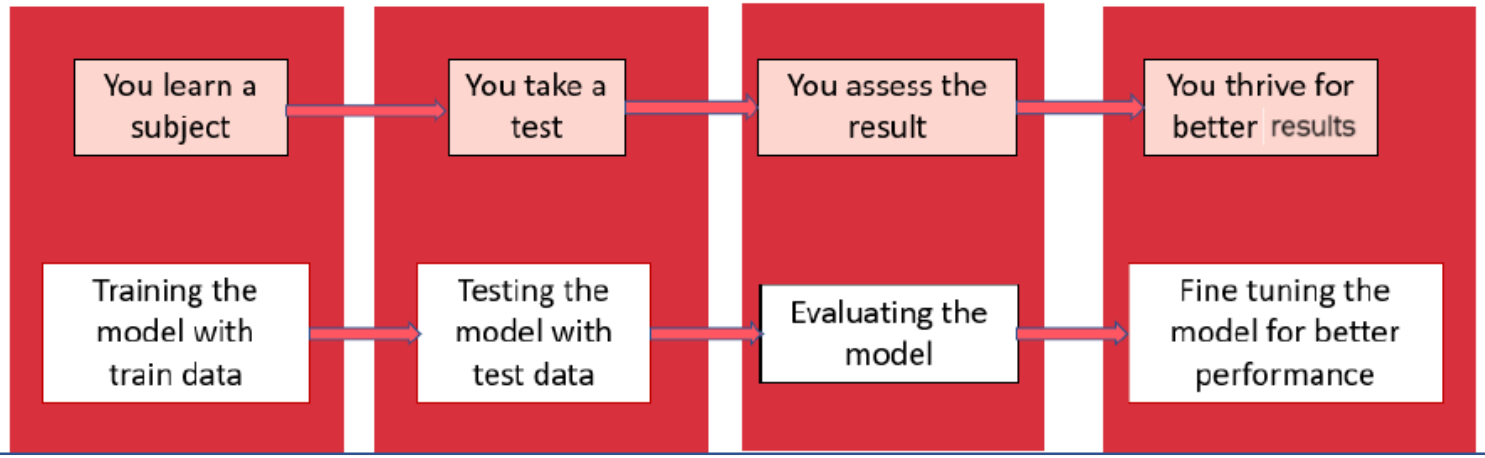
Introduction

- 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

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.

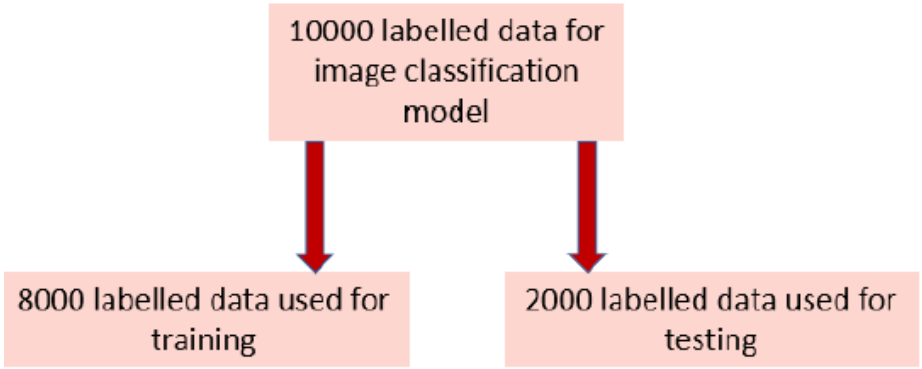


Need of model evaluation

- 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.

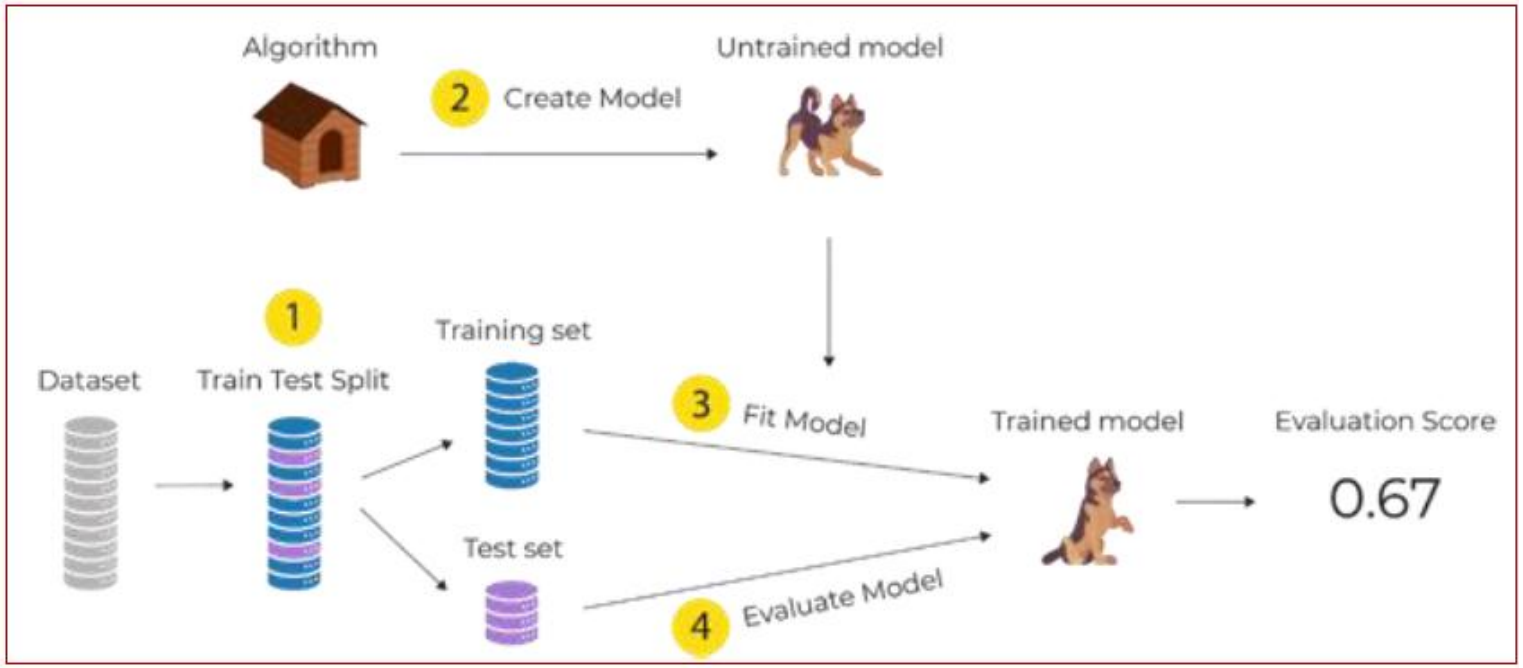
Splitting the training set data for Evaluation

- 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
- **Train-test split**



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



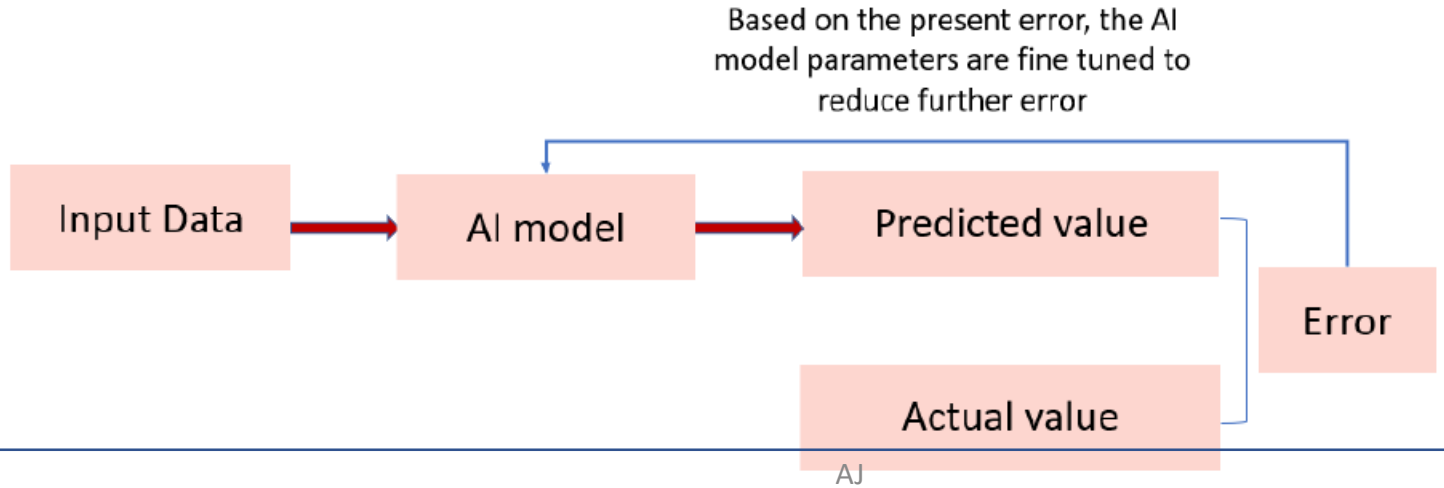
Remember that It's not recommended to use the data we used to build the model to evaluate it. This is because our model will simply remember the whole training set, and will therefore always predict the correct label for any point in the training set. This is known as **overfitting**.

Accuracy

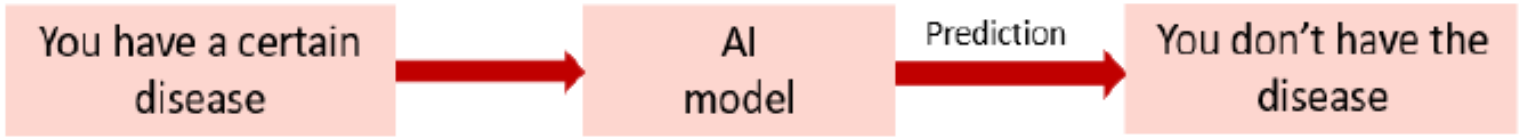
- Accuracy is an evaluation metric that allows you to measure the total number of predictions a model gets right.
- The accuracy of the model and performance of the model is directly proportional, and hence better the performance of the model, the more accurate are the predictions.

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



- **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.



Wrong prediction. There is an error detected.

- **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.

Activity 1: Find the accuracy of the AI model

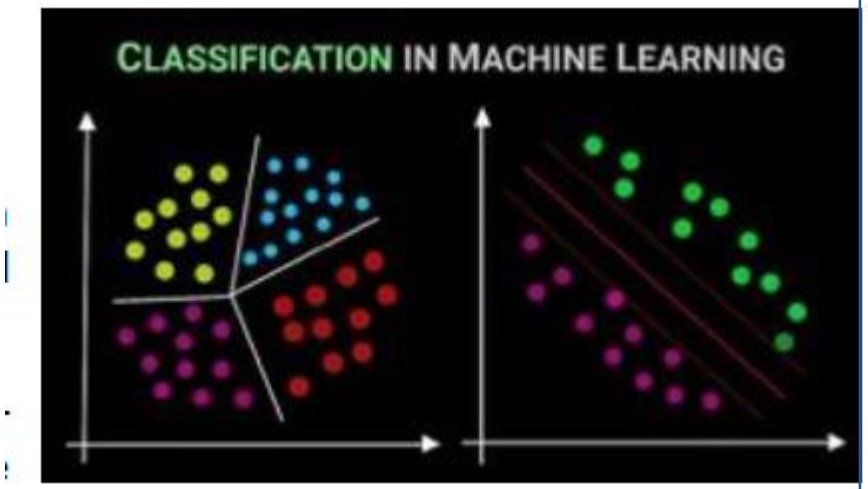
- **Calculate the accuracy of the House Price prediction AI model**
- Read the instructions and fill in the blank cells in the table.
- The formula for finding error and accuracy is shown in the table
- Accuracy of the AI model is the mean accuracy of all five samples
- Percentage accuracy can be seen 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)

Evaluation metrics for Classification

- **What is Classification?**
- 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,a vegetable- grocery-classifier model that predicts whether the item in the supermarket is a vegetable or a grocery item.



Which of this is a classification use case example?

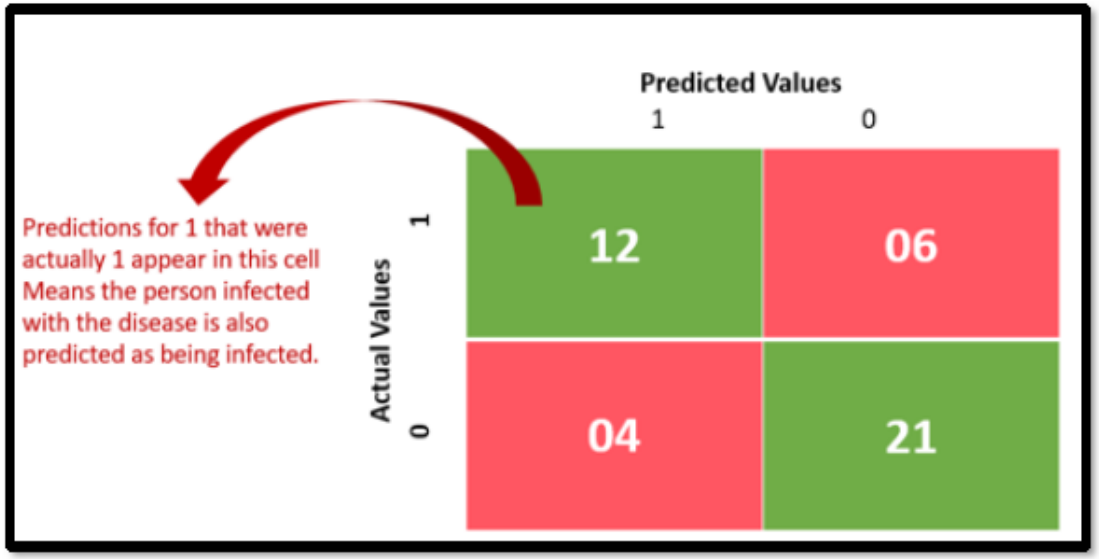
House price prediction

Credit card fraud detection

Salary prediction

Classification Metrics

- Popular metrics used for classification model
  - Confusion matrix
  - Classification accuracy
  - Precision
  - Recall



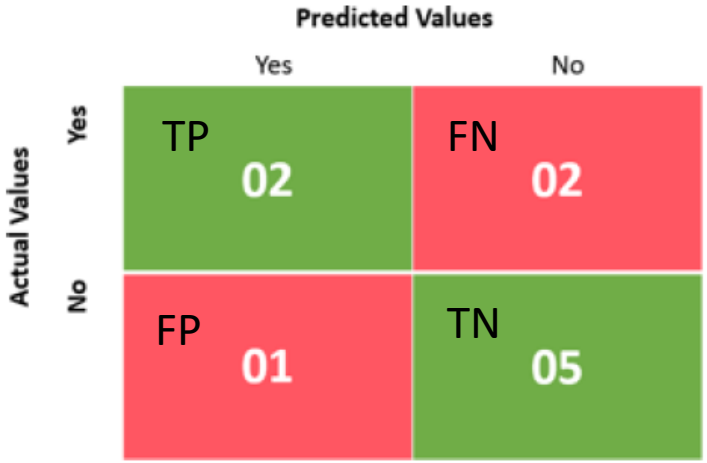
Confusion matrix

- 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:

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

True Positive (**TP**) is the outcome of the model correctly predicting the positive class.
Eg: the cases in which we predicted yes (they have the disease), and they do have the disease.



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

True Negative (**TN**) is the outcome of the model correctly predicting the negative class.
Eg: the cases in which we predicted No (they don’t have the disease), and they don’t have the disease

False Positive (FP) is the outcome of the model wrongly predicting the negative class as positive class.
Eg: the cases in which we predicted Yes (they have the disease), and they don’t have the disease.

False Negative (FN) is the outcome of the model wrongly predicting the positive class as the negative class.
Eg: the cases in which we predicted No (they don’t have the disease), and they have the disease.

Accuracy from Confusion matrix

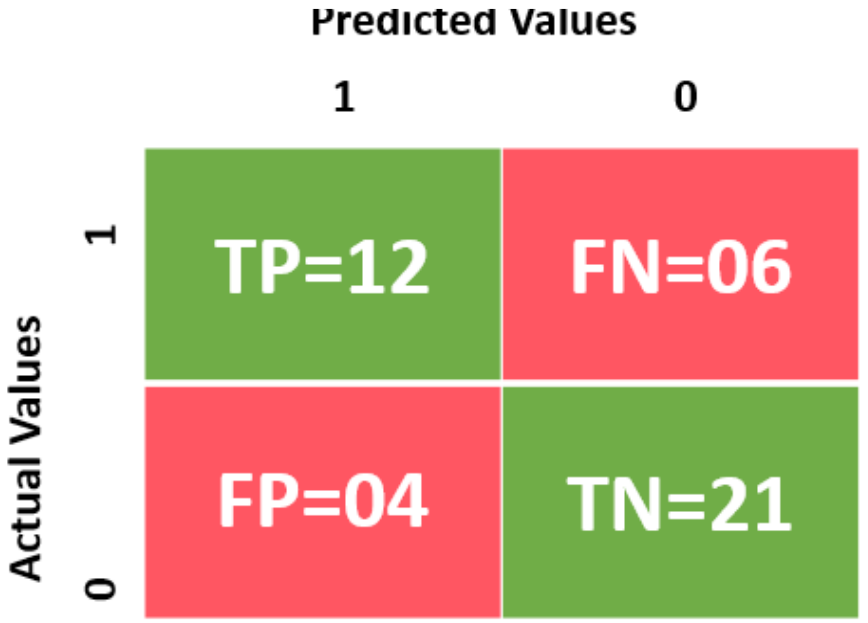
Correct predictions=TP+TN

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.

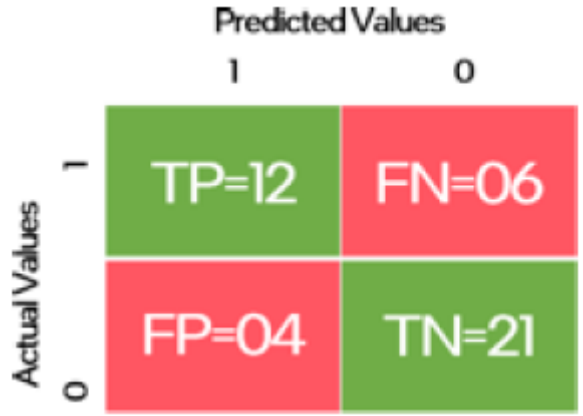
Precision from Confusion matrix

- **Precision** is the ratio of the total number of correctly classified positive examples a

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.



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

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

- **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

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

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

- 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.
- 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.

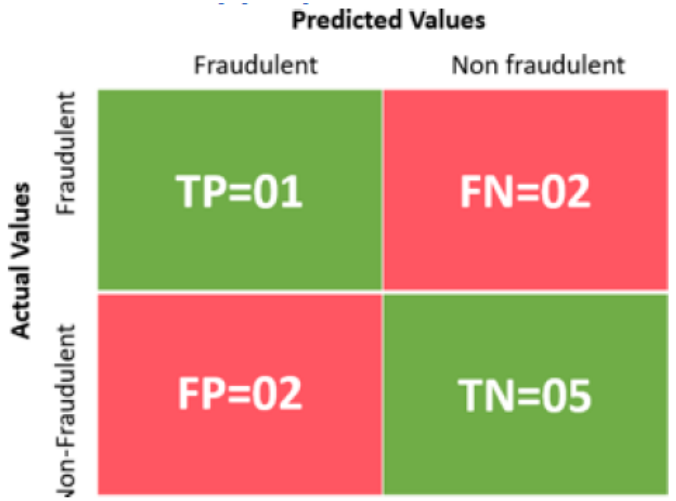
F1 Score

- 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.

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

- **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



Ethical concerns around model evaluation

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

Bias	Transparency	Accountability
Ensure that the evaluation metrics chosen don't result in any kind of bias	Honest explanation how the chosen evaluation metrics work and produce results without keeping any information hidden	Take responsibility for your choice of metrics and methodology of evaluation in case any user faces a disadvantage because of your chosen methodology

1.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.

2. 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) 96%

b) 90%

c) 85%

d) 70%

3. 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%

4. 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%

5. 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%