**Confusion Matrix:-**

A Confusion Matrix helps us understand how well a classification model is performing by comparing its predictionswith the actual results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Actual \ Predicted** | **Positive (Yes)** | |  | | --- | |  |  |  | | --- | | **Negative (No)** | |
| **Positive (Yes)** | True Positive (TP) | False Negative (FN) |
| **Negative (No)** | False Positive (FP) | True Negative (TN) |

**Explanation of Terms :-**

1. **True Positive (TP) →** The model correctly predicted positive (e.g., predicted "pass" and the student actually passed).
2. **False Positive (FP) [Type I Error] →** The model incorrectly predicted positive (e.g., predicted "pass" but the student actually failed).
3. **False Negative (FN) [Type II Error] →** The model incorrectly predicted negative (e.g., predicted "fail" but the student actually passed).
4. **True Negative (TN) →** The model correctly predicted negative (e.g., predicted "fail" and the student actually failed).

**Example: Predicting Whether a Student Will Pass an Exam**

|  |  |  |
| --- | --- | --- |
| **Actual/predicted** | **Pass (Positive)** | **Fail (Negative)** |
| **Pass (Actual)** | TruePositive=50 | FlaseNegative=10 |
| **Fail (Actual)** | FalsePositive=5 | TrueNegative=35 |

**🡪 True Positive (TP = 50):** 50 students correctly predicted as passing**.**

**🡪False Positive (FP = 5):** 5 students wrongly predicted as passing but actually

Failed.

**🡪False Negative (FN = 10):** 10 students wrongly predicted as failing but

actually passed.

**🡪True Negative (TN = 35):** 35 students correctly predicted as failing.

**Accuracy :-**

Accuracy=TP+TN/TP+TN+FP+FN

Accuracy=50+35/50+35+5+10

=85/100

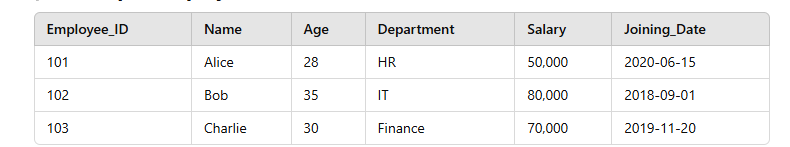
=85%

**Precision:-**

Precision=TP/FN+TP​ Precision

=50/50+10=50/60=90.9%

**Columns:**

Columns are the **structural components** of a dataset (typically in tabular format). Each column represents a **specific type of information** and stores **values of the same attribute** for different rows (records).

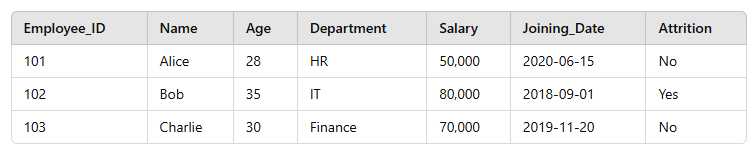
* **Columns** = Employee\_ID, Name, Age, Department, Salary, Joining\_Date
* **Rows** = Individual records (employees)

**Key characteristics of columns:**

* **Hold raw data** and maintain a structured format.
* Each column represents a single attribute (e.g., Age, Department).
* Different data types:
  + Age → Numeric
  + Name → Text
  + Joining\_Date → Date

**Features :-**

Features are **input variables** used to train a machine learning model. They represent **meaningful** or **relevant** information extracted from the dataset, often preprocessed for better accuracy.

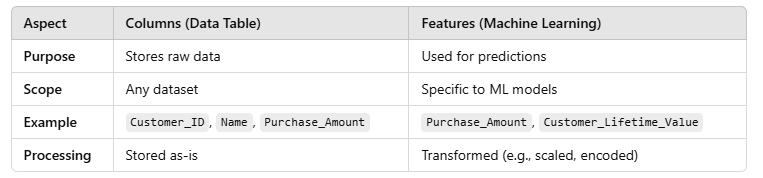


**Relevant Features**:

* Age
* Salary
* Department (converted to numbers via encoding)
* **New Feature**: Years\_at\_Company (calculated from Joining\_Date)

**Non-relevant Features**:

* Employee\_ID (Just an identifier, does not impact attrition)
* Name (Personal data, not useful for prediction)

**Difference between Columns and features:-**