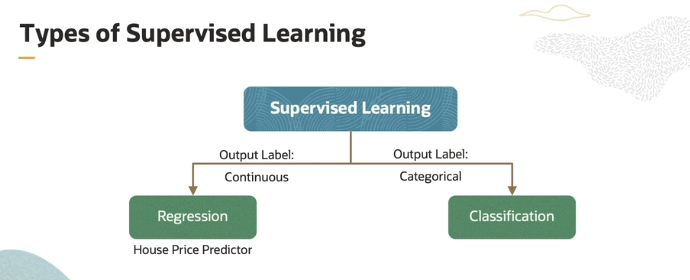
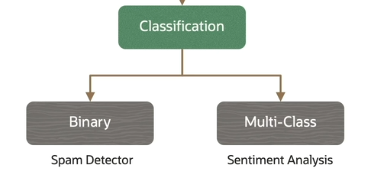
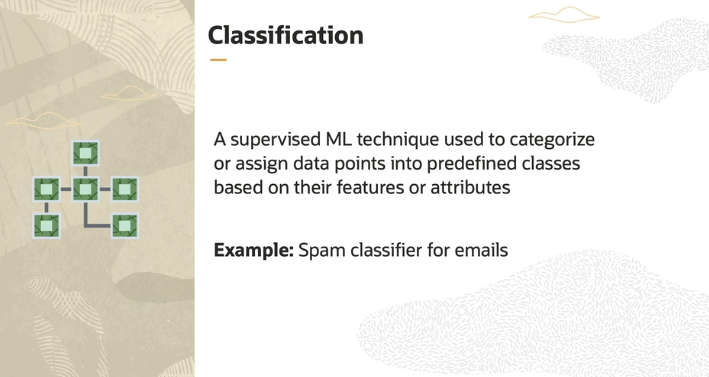
**Supervise ML-Classification**



Now as we already know when the output is continuous (numeric form) so we use **Regression** which we have covered in previous document, and when the output is categorical so then we use **Classification** that we are going to learn in this document.



Now there are two types of classifications one is **Binary** and other is **Multiclass.** Now Binary means asi classification jisme srf two category hoti hai it is like **true or false,** now if we take example of **spam Detector** toh usme jo output hai wo 2 category may lie krta hai that are **Spam** or **not-spam** and isi may ager koi Gender Detector hai toh wobi Binary classification may lie krega kay **Male or Female.** Now what is **multiclass-classification** ? It means that when we have output in more than 2 category then it is known as **multiclass-classification** , which means like if we are creating a **Customer review Detector** toh usme humaray pass jo output hai it lies in three category **Positive,Negative or neutral** which means kay jo Detector jo review detect krega wo Positive hoga ya Negative ya neutral So this type of classification is known as **Multiclass classification**.



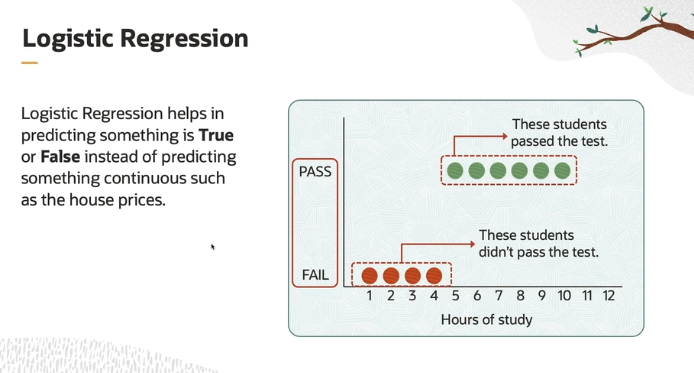
Classification in machine learning is like sorting objects into different boxes based on certain characteristics. Imagine you have a bunch of emails, and you want to separate them into two boxes: "Spam" and "Not Spam." This is what classification does.

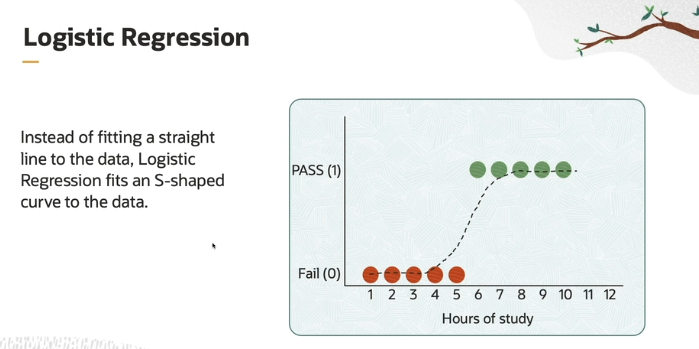
Here’s how it works:

1. **Labeled Data:** First, you give the machine a bunch of examples that are already labeled. For instance, you might show it 100 emails, where some are labeled as "Spam" and some as "Not Spam." These labels are like hints to help the machine learn.
2. **Training a Classifier:** The machine looks at the features (the words, subject line, sender, etc.) of these labeled emails. It tries to find patterns in the data that tell it what makes an email "Spam" or "Not Spam."
3. **Predicting:** After learning from these examples, the machine can now look at a new, unlabeled email and decide whether it should go into the "Spam" or "Not Spam" category.

So, classification is all about teaching the machine to put new data into the correct category based on what it learned from the past examples.

In your example of detecting spam mail, the machine is trained on emails labeled as either "Spam" or "Not Spam." When a new email arrives, the machine uses what it learned to guess if it’s spam or not, making this a **binary classification** because there are only two outcomes: "Spam" or "Not Spam."





**What is Logistic Regression?**

Logistic regression is a machine learning algorithm used when you want to predict whether something is **true** or **false**. It helps answer questions like: “Will this student pass or fail the test?” or “Is this email spam or not?”

**The Example: Pass or Fail**

Imagine you have a group of students, and you want to predict whether they will pass or fail an exam based on how many hours they studied. Here’s how it works:

* **Input (Independent Feature):** In this case, the **hours of study** are the input. It’s the factor that influences whether a student will pass or fail.
* **Output (Binary):** The output is simple: **pass** or **fail**. It’s binary because there are only two possible outcomes.

**How Does It Work?**

Unlike linear regression, which uses a straight line to predict results (like predicting house prices), **logistic regression** uses an **S-shaped curve** called the **sigmoid function**. The sigmoid curve is important because it turns any number into a value between **0 and 1**, which we can think of as a **probability**.

* **The Sigmoid Function:** This function transforms the hours of study into a probability between 0 and 1. The closer the value is to 1, the more likely the student is to pass. If the value is closer to 0, the student is more likely to fail.

**How Decisions Are Made:**

Let’s say a student studied for 5 hours. We input this number into the sigmoid function, and it gives us a probability, for example, 0.85 (85%) (0.85 isi liye aya bcuz sigmoid function 0 – 1 kay between value return krta hai, then phr usko multiply by 100 krkay final percentage nikal sktay hain) . This means there’s an 85% chance the student will pass. We can then set a threshold:

* If the probability is greater than 0.5 (50%), we predict the student will **pass**.
* If the probability is less than 0.5, we predict the student will **fail**.

**What does mean by setting Threshold ?**

In logistic regression, **setting a threshold** means deciding a cutoff point to classify an outcome into one of two categories (like pass/fail, spam/not spam).

The logistic regression model gives you a **probability** that a data point belongs to a certain class/category. This probability is a number between **0 and 1**. To make a decision, you need to set a **threshold**—a specific value where you draw the line between the two classes.

**Example:**

Let’s say you're predicting whether a student will pass or fail based on hours of study. The model gives you a probability, like 0.85 (85%).

* If the probability is **greater than the threshold**, we classify it as a **pass**.
* If the probability is **less than or equal to the threshold**, we classify it as a **fail**.

**Common Threshold:**

* The most common threshold is **0.5** (or 50%):
  + If the probability > 0.5, predict **pass**.
  + If the probability ≤ 0.5, predict **fail**.

You can adjust the threshold based on how you want the model to behave. If you want to be stricter about predicting "pass," you might set the threshold higher, like 0.7. This means:

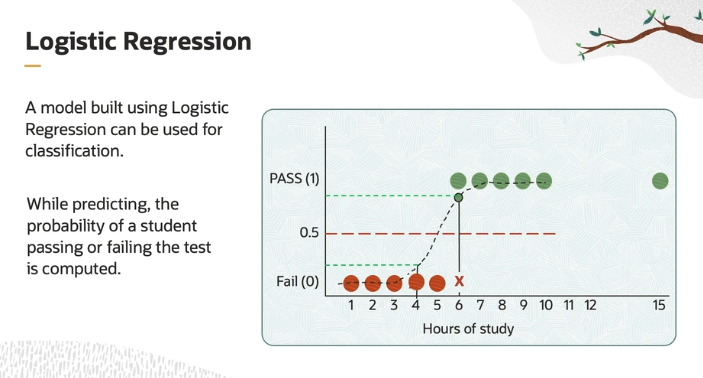
* Only if the probability is > 0.7 will the model predict **pass**.
* Otherwise, it will predict **fail**.

In short, the **threshold** is the cutoff that helps the model make a final yes/no (binary) decision based on the predicted probability.

**Summary:**

* **Logistic regression** helps us make a yes/no decision based on some input (like hours of study).
* It uses the **sigmoid function** to convert the input into a probability.
* Based on that probability, we can decide whether something will happen (pass/fail, spam/not spam, etc.).

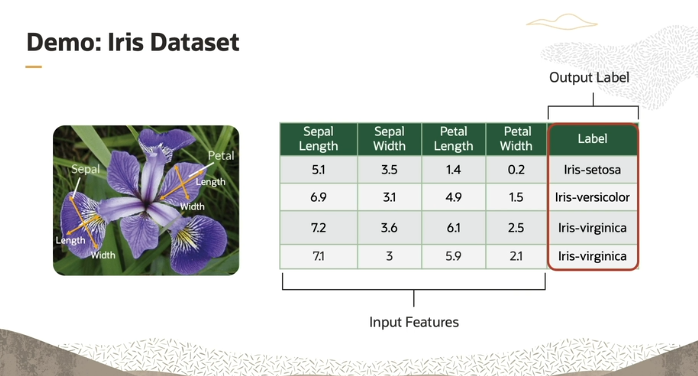
This way, logistic regression makes decisions based on probabilities and helps classify data into two groups.



**As we have already understood in above explanation that what is threshold ?**

But again here we will understand that why it is set mainly? So as we know that the sigmoid function returns value b/w 0 and 1, which is a probability so how our model will identify/classify on the basis of probability that the student is **pass or fail.** SO for that purpose we have to create a cutoff line between 0 and 1 at a **0.5 point** (you can also consider it as 50%) as you can see in the above pic and this line or cutoff point is known as **Threshold**. Now the model will classify that the student is fail or pass on the basis of if probability is greater than **0.5** so student can be considered pass or chance are of **pass** and if less than or equal to **0.5** so it can be considered fail or chance are of **fail.**

For example, a student who studies for six hours and having 80% probability to pass in the test is classified as "pass." A student who studies for four hours and have a 20% probability to pass the test is classified as "fail."



The explanation you're asking about is referring to how the **Iris dataset** is used in machine learning to perform **multi-class classification**. Let’s break it down step by step:

1. **Iris Dataset**:
   * This dataset is a well-known one in machine learning, containing **150 instances** (or examples) of **three different types of Iris flowers**:
     + **Iris setosa**
     + **Iris versicolor**
     + **Iris virginica**
2. **Attributes (Features)**:
   * Each flower in the dataset is described by **four characteristics** (also known as features):
     + **Sepal length**
     + **Sepal width**
     + **Petal length**
     + **Petal width**
   * These features are numerical values that describe the size of different parts of the flower, and they are used as the input for the machine learning model.
3. **Multi-class Classification**:
   * Since there are **three classes of flowers** (setosa, versicolor, virginica), the goal is to **classify** each flower into one of these three categories based on the features.
   * This is a case of **multi-class classification** because there are **more than two possible classes** to choose from (in contrast to binary classification, where there are only two outcomes).
4. **Logistic Regression**:
   * In the demo, a **logistic regression** model will be used. Even though logistic regression is often associated with binary classification, it can also be adapted to handle **multi-class classification** by using techniques like **one-vs-rest** or **softmax**.
   * The model will be trained to classify the flowers into one of the **three classes** based on the **four features** provided.
5. **Output Label**:
   * The **output label** in this case is the type of Iris flower. So, for each input (a flower’s sepal and petal measurements), the model will predict one of the three possible outputs: Iris setosa, Iris versicolor, or Iris virginica.

**In Simple Terms:**

* You have 150 examples of three different types of flowers, and you want to teach the model how to identify the flower type based on its size (sepal/petal length and width).
* Since there are three different types of flowers to predict (not just two), it's called **multi-class classification**.
* The model (logistic regression) will be trained on the data and will eventually be able to predict which of the three flowers a new example belongs to based on its measurements.

Top of Form

Bottom of Form