

# Understanding **MACHINE LEARNING**

A Comic Beginner Series





# What is MACHINE LEARNING?

## Episode 1

Humans learn from experience.



So data is just examples?

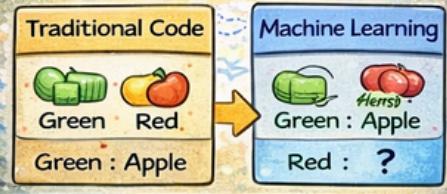


Exactly!

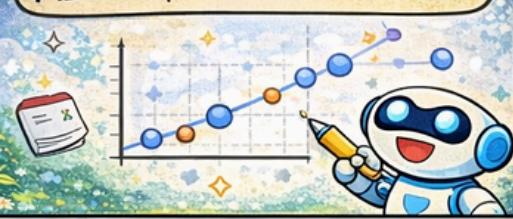
Machines learn from data!



With ML, we don't write rules --  
the computer finds them!



ML finds patterns hidden in the



ML helps us predict numbers  
and categories.



So ML is learning from data...

What's next?

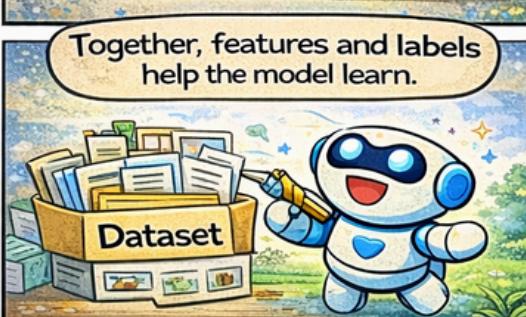
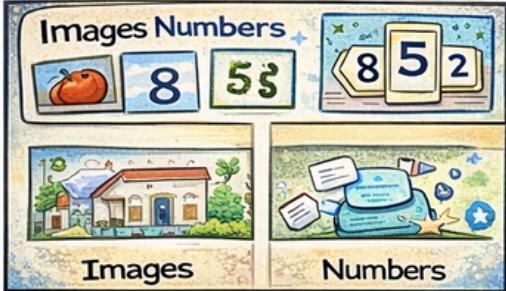


Next, we explore  
how data works!



# Data is the Fuel

## Episode 2



Good Data			Bad Data		
Size	Rooms	Location	?	?	?
1000 sq ft	3	New York	~?	?	?
1500 sq ft	4	Los Angeles	~?	?	?
2000 sq ft	5	Chicago	~?	?	?
2500 sq ft	6	Seattle	~?	??	?
3000 sq ft	7	Boston	~?	??	?



# Supervised vs Unsupervised Learning

Episode 3

## Supervised Examples

In supervised learning, we give the machine examples **with** labels.



## Unsupervised Learning

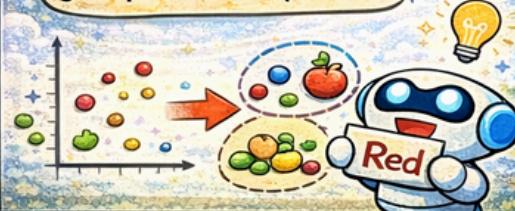
In unsupervised learning, the machine gets only the data... no labels.



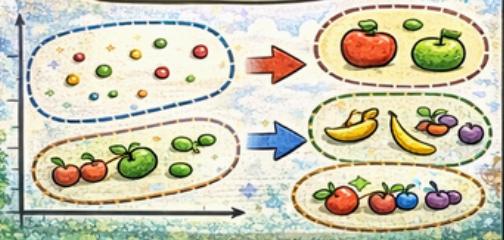
It groups similar things together automatically!



So it's like finding groups in messy data!



So it's like finding groups in messy data!



Oh! Labels decide everything!



See you in Episode 4:  
Training, Testing & Accuracy!

# Training, Testing & Accuracy

Episode 4

## Training Data



## Testing Data



Training is when the machine learns from examples.



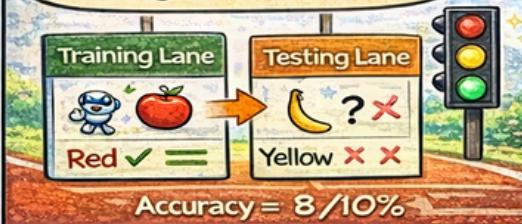
And testing checks how well it learned!



If it sees the answers again, the test isn't fair!



So training is practice..., testing is the real race!



More correct answers = better

High Accuracy	
Red	✓ ✓ 8
Yellow	✓ ✓
Green	✓ ✓

Low Accuracy	
Red	✗ ✗
Yellow	✗ ✗
Green	✗ ✗

$$\text{Accuracy} = 8 / 10 = 80\%$$

See you in Episode 5 !  
**Features & Labels !**



# Features & Labels

## Episode 5

### Features

- ✓ Size (1200 sq ft)
- ✓ Rooms (3)
- ✓ Location



### Labels



Features describe the data.



Good features help the model learn accurately.

#### Good

- ✓ Size
- ✓ Rooms
- ✓ Age of building

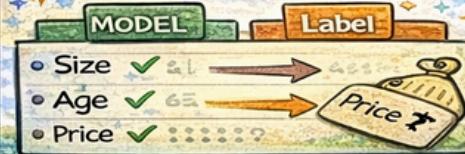
#### Bad

- ✗ House color
- ✗ Owner's name
- ✗ Random ID

Some features make no sense!



Meaningful features → meaningful predictions.



We use both kinds in ML!

#### Numericals

- Size
- Age
- Price

#### Categories

- City name
- color
- Yes/ No

Feature extraction helps organize data for learning.

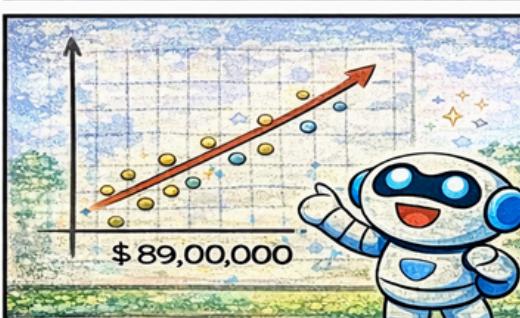
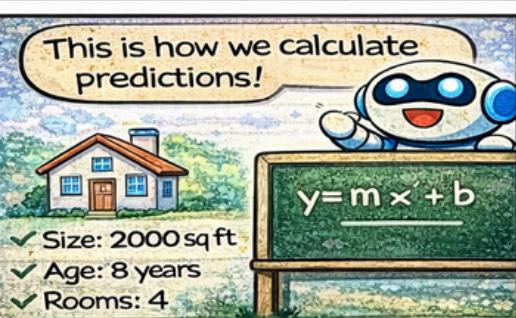
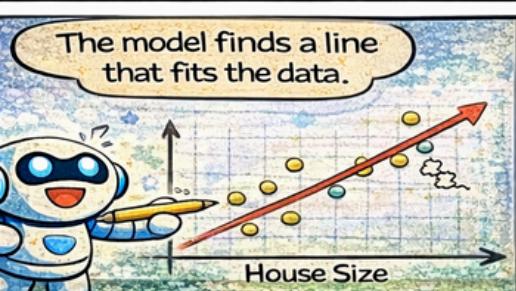
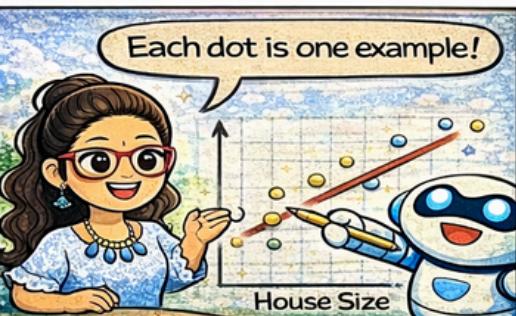
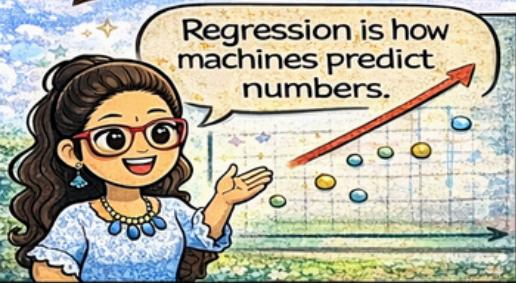


See you in Episode 6:

Regression—Predicting Numbers!

# Regression – Predicting Numbers

## Episode 6



See you in Episode 7:  
Classification – Predicting Categories!

# Classification – Predicting Categories

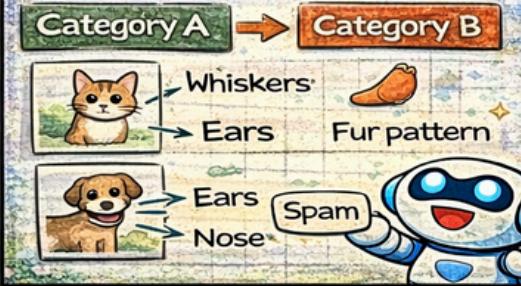
Episode 7



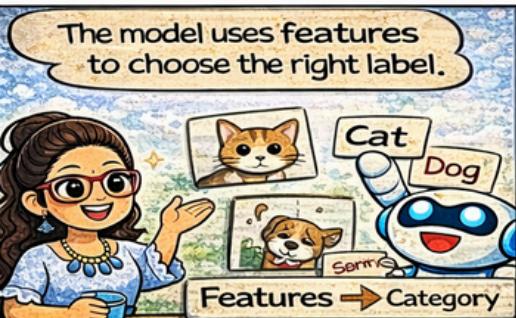
Classification is how we predict “categories, not numbers!”



We train using labeled examples.



The model uses features to choose the right label.



That line helps classify new points!



Classification can predict more than two categories too!

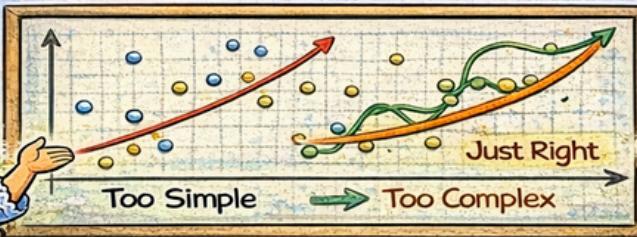


See you in Episode 8:  
Underfitting & Overfitting!

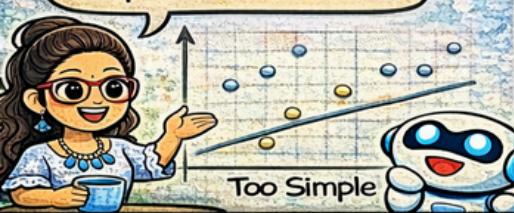


# Underfitting & Overfitting

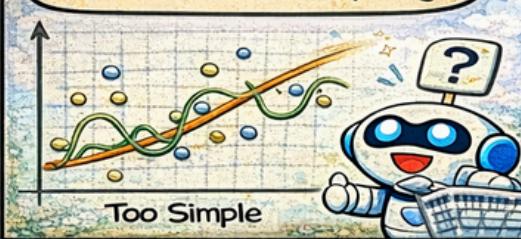
Episode 8



A good model learns the pattern in the data.



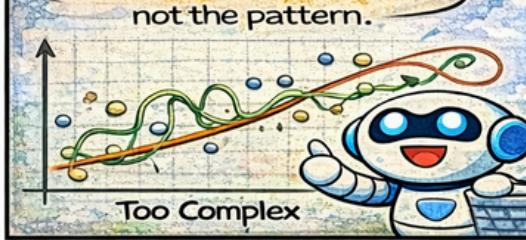
This line misses everything!



The model is not learning enough.



It learns even the noise, not the pattern.



Training Accuracy: 100%

Testing Accuracy: 40%



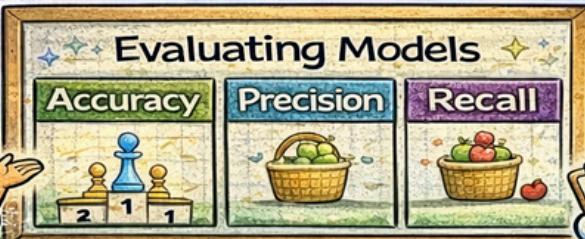
A balanced model captures the pattern without memorizing.



See you in Episode 9:  
Evaluating Models!

# Evaluating Models

Episode 9



But accuracy can trick us when classes are imbalanced!



Model predicts "Green" every time accuracy looks high..



Confusion matrix shows all prediction outcomes.



Precision asks:

- Of what we predicted positive, how many were right?



Confusion matrix shows ALL prediction outcomes.

Recall asks:

- Of all real positives, how many did we catch?



Precision avoids mistakes.  
Recall avoids missing important cases.

Now I know how to measure a model properly!

Accuracy



Accuracy

Precision



Precision

Recall



Confusion matrix



See you in Episode 10:  
**Data Cleaning Basics!**

# Data Cleaning Basics

Episode 10



V	S	D	B	I	Z	...
FILY	?			?	ELLO	LRY M
X	?			?	OL	SIMA
3-5	?			?	LO	LO



Clean data helps models learn better!



We can fill missing values or remove them.

Some data points are just... empty!



Duplicates confuse the model!



We keep only one copy.

Before

a	✓	?
a	✓	?

After

o	✓	?
o	✓	✓



Outliers can be fixed—or removed



Remove outlier

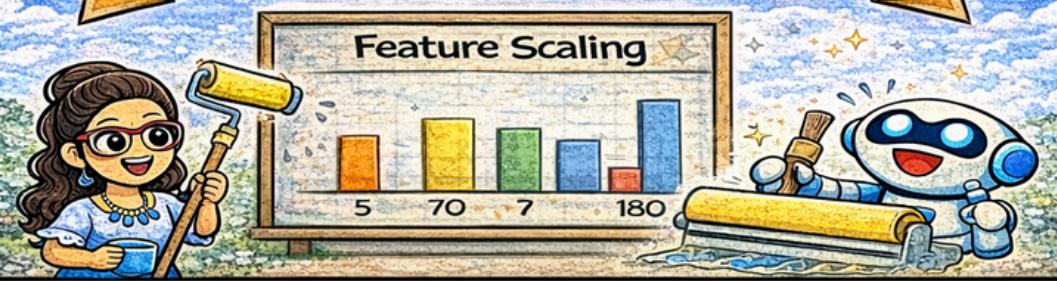
Clean and ready for training!



See you in Episode 11:  
Feature Scaling!

# Feature Scaling

Episode 11



When features are on different scales, models get confused.

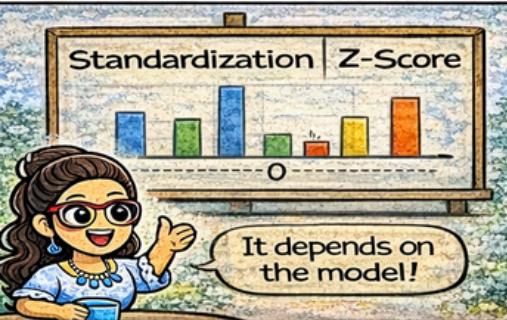


Standardization shifts data to mean 0 and equal spread.

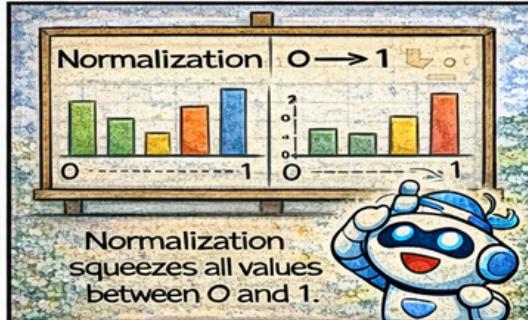
Scaling makes features easier for models to compare.



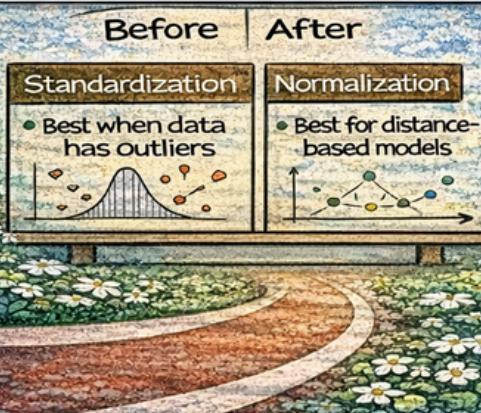
Normalization squeezes all values between 0 and 1.



It depends on the model!



Normalization squeezes all values between 0 and 1.



# Train a Simple Model

## Episode 12

### Train a Simple Model (Linear Regression)

```
import pandas as pd  
from sklearn.linear_model import  
LinearRegression
```

Let's train our first ML model!



Here come the student scores!

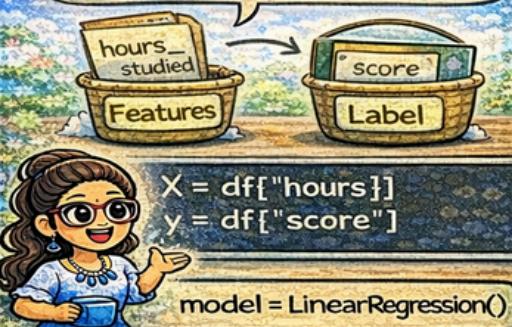
Here come the student scores!

```
df = pd.read_csv("scores.csv")
```

```
y = pd.read_csv("scores.csv")
```

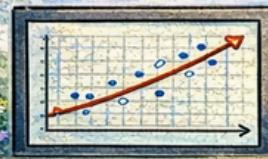


Features go in, labels come out.



We're teaching the model!

```
model = LinearRegression()  
model.fit(X, y)
```



So linear regression is just a straight line!



$$y = m \cdot x + b$$

slope      intercept





# Decision Trees Intuition

Episode 13



A decision tree decides by asking simple questions.

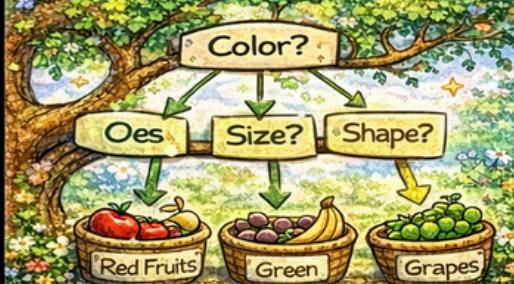


Oh! It's like a quiz with steps!



Is it furry? →  
No → Dog?

Each question splits the data.



We keep splitting until choices get clear!



?

?

We make decisions like this every day.



So leaves are the final predictions!



Is it raining?

```
graph TD; Rain[Is it raining?] --> GoOut[Go out.]; Rain --> StayHome[Stay home.]
```



# kNN & Distance-Based Learning

Episode 14



## KNN & Distance-Based Learning.



kNN predicts by looking at the closest neighbors.



Each dot belongs to a category.



What category does THIS



kNN checks who the nearest neighbors are!



Oh! Different k means different answers!



kNN is simple and powerful when data is not huge.

- ✓ Simple patterns
- ✓ Small to medium datasets
- ✗ Not great for very large data



See you in Episode 15:  
Mini Project - Predict  
Student Scores Again (with ML)

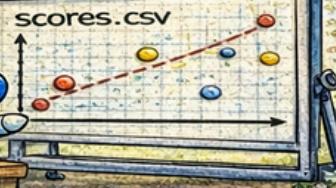
# Mini Project: Predict Student Scores!

Episode 15



## Student Score Predictor

Student	Hours	Score
1	40	40
3	77	?
2	?	?



First, load the raw data.

Student	Hours	Score
1	40	40
3	77	?
2	?	?



Train with one part., test with another.

Feature

Hours Studied

Missing

Duplicates



Let's clean the messy parts!



We'll use hours as the input,  
Score as the output.

model.fit(X\_train, y\_train)

Train with one part.

Dataset

1	3	2
2	6	2



Let's predict new student scores!

Input → hours = 5

Model → 78

Predictions → 78

SEVENTY,  
EIGHT!



Mini Project Complete –  
See you in Part 2!

**A Sriju Comic**



**Curiosity Lives Here!**

**Stories that smile back at you**