

MACHINE LEARNING SERIES!





Real-World Datasets

Episode 16

Wow! Real ML uses
all kinds of data!



CSV is the most popular format
for machine learning.



Excel files can have many
sheets. ML reads them too!



APIs send data from websites
and apps directly to us.



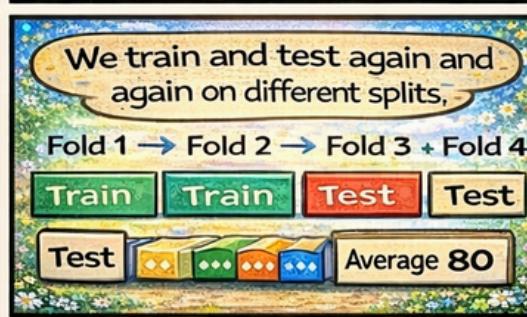
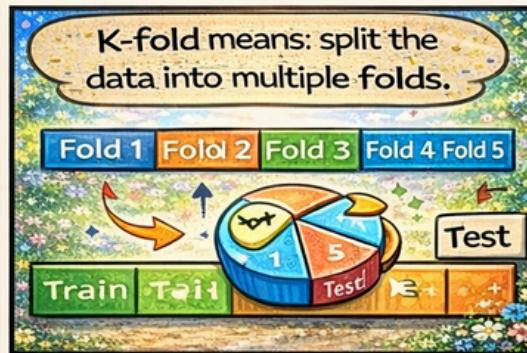
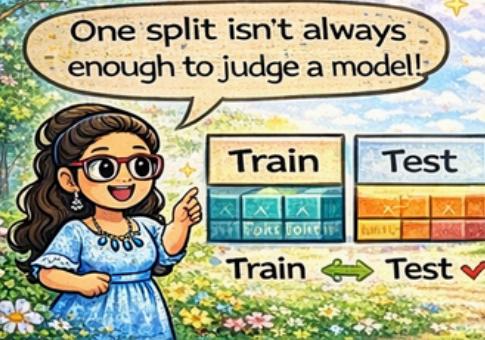
Kaggle is a huge library of
public datasets for ML practice!



See you in Episode 17 --
Splitting Data Properly (Cross-Validation) !

Episode 17 - Splitting Data Properly

(Cross-Validation)



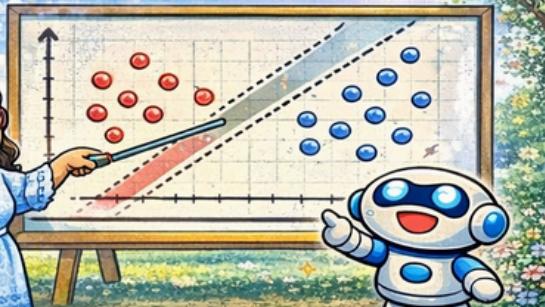
Episode 18 - Naive Bayes

Probability made simple!



Support Vector Machines (SVM)

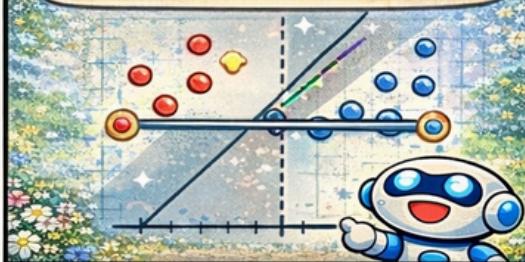
SVM finds the line that best splits the data!



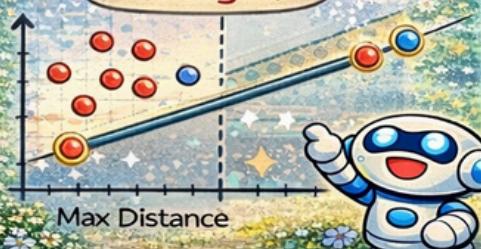
There are many ways to separate the points...



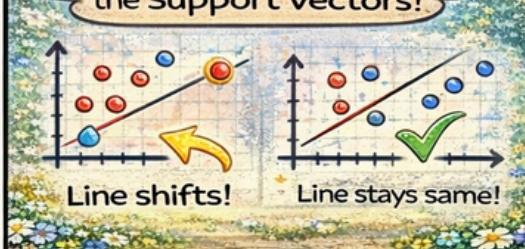
The best line keeps the widest space from both classes.



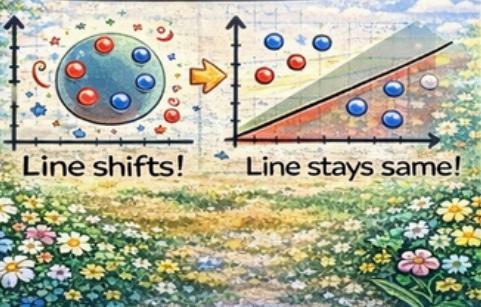
This space is called the **margin**.



These important points are the **Support Vectors**!



Sometimes data isn't linearly separable...



Kernels help SVM separate complex shapes!

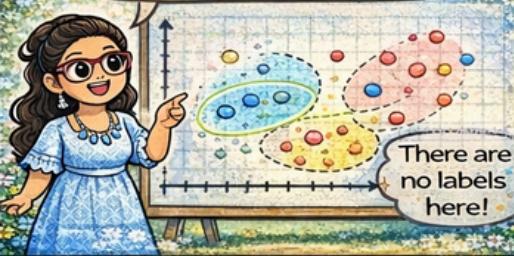


Clustering: k-Means (Unsupervised Learning)

Clustering groups similar items together - without labels!



k-Means discovers patterns we can't easily see.



Step 1: Choose k (Number of Clusters)

We choose how many clusters.



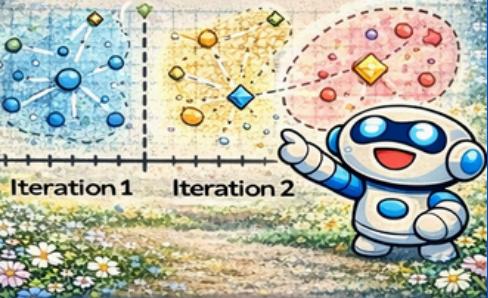
Step 2: Drop Random Centroids

Are these the leaders?



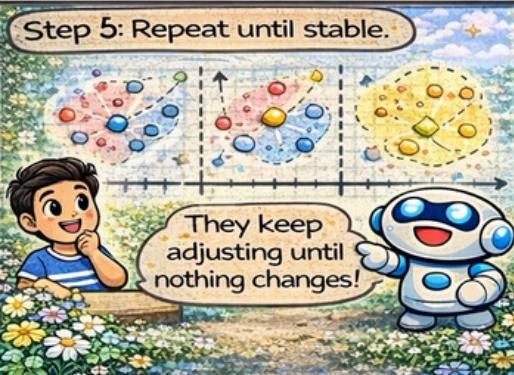
Yes! These starting points are called centroids.

Step 3: Assign Points to nearest centroid!



Step 5: Repeat until stable.

They keep adjusting until nothing changes!



And that's your final clustering!



Dimensionality Reduction (PCA)



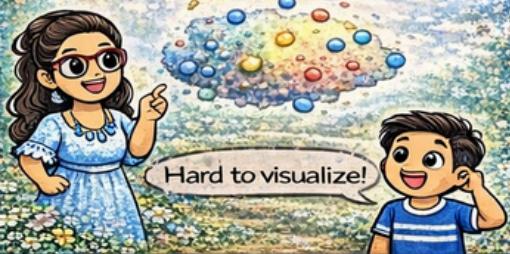
TOO MANY FEATURES!

Keep important patterns

Reduce noise



TOO MANY FEATURES!



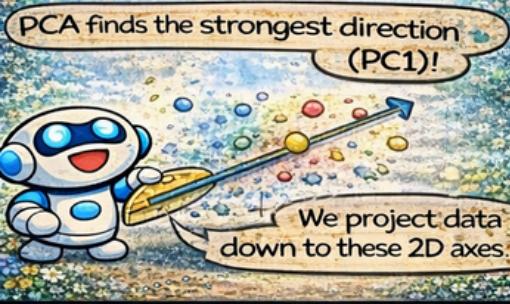
Hard to visualize!

Data is like a cloud of points in 3D!

Hard to visualize!

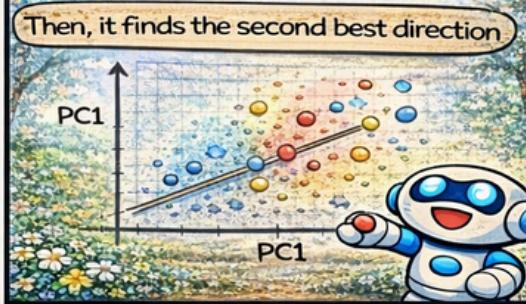


PC1

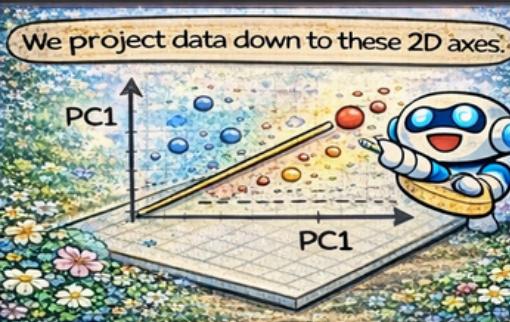


PCA finds the strongest direction (PC1)!

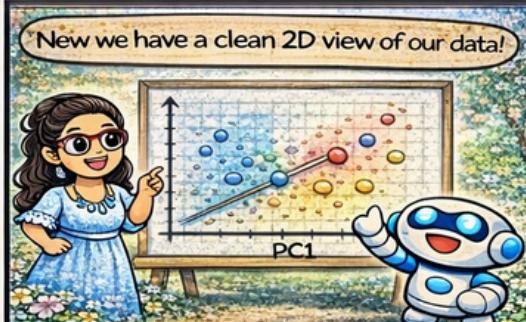
We project data down to these 2D axes.



PC1



We project data down to these 2D axes.



New we have a clean 2D view of our data!

See you in Episode 22 – Feature Engineering!



Episode 22 – Feature Engineering

| Height | Weight | Age | Curc | Wesinger |
|---------|--------|-----|------|----------|
| Fificht | 75 kg | 20 | 20.0 | 20.0 |
| Eldeter | 70 kg | 28 | 20.9 | 20.9 |
| Etemi | 29 kg | 28 | 20.8 | 20.8 |
| Sulonis | 22 kg | 22 | 20.0 | 22.0 |
| Fleight | 17 kg | 32 | 20.9 | 20.9 |
| Fleight | 10 kg | 93 | 20.0 | 20.0 |



We create better input features to help ML learn better!

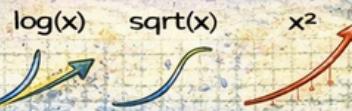


Creating new features.

| Height | Weight | Age | Weather |
|--------|--------|-----|---------|
| 170 cm | 70 kg | 25 | Sunny |
| 180 cm | 70 kg | 25 | Sunny |



Transformations help reduce skew and highlight patterns!

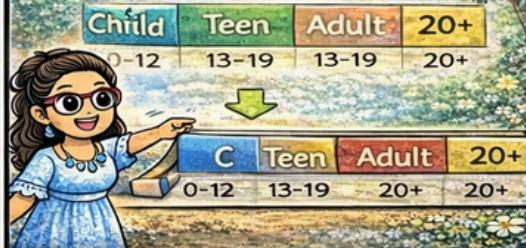


| | | | | |
|-----|-------|-------|-------|-------|
| Rod | 1,0,0 | Grech | 0,1,0 | 0,0,1 |
|-----|-------|-------|-------|-------|

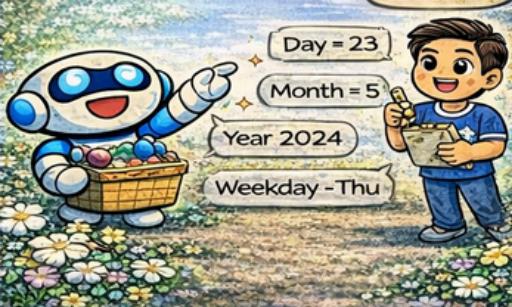
Models need numbers—so we encode categories.



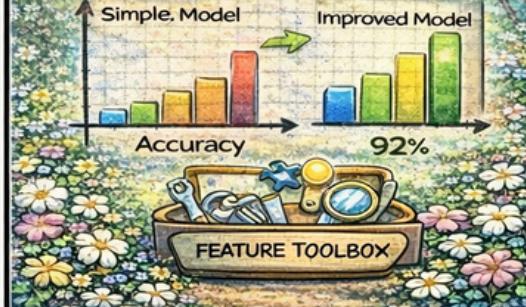
We can group values into buckets!



A single column can give many useful features!

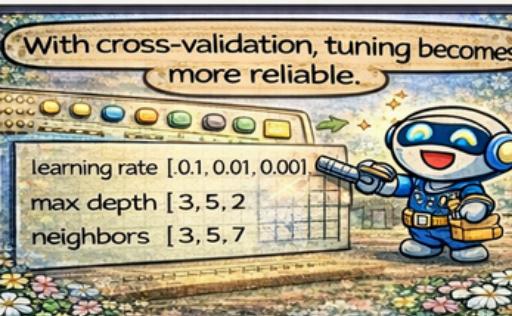
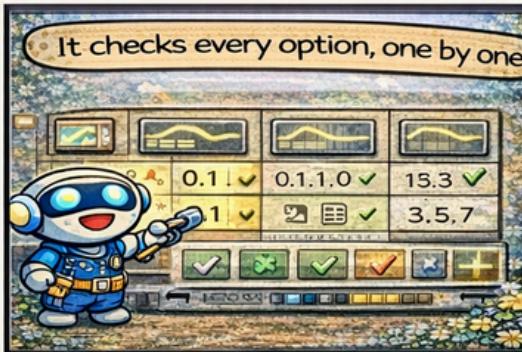
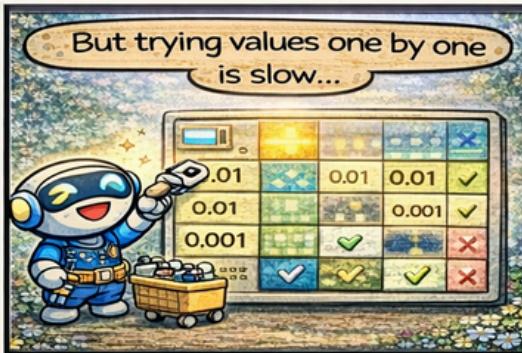
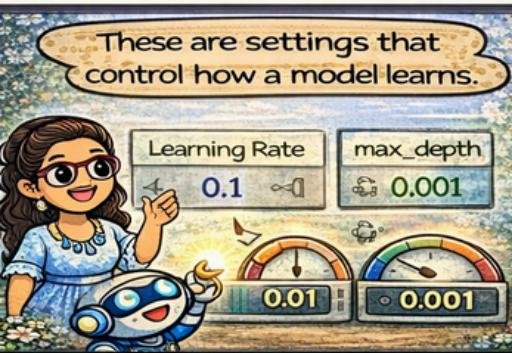
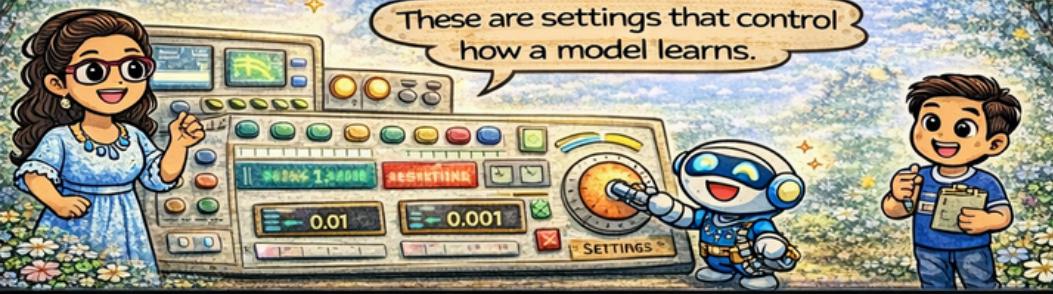


Better features = smarter models!





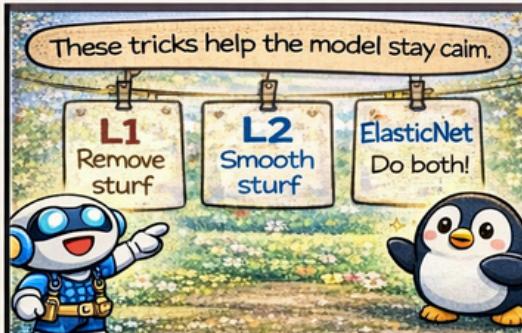
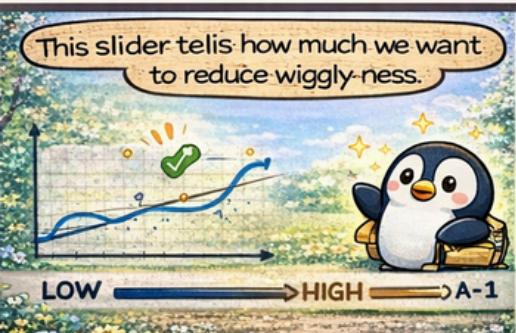
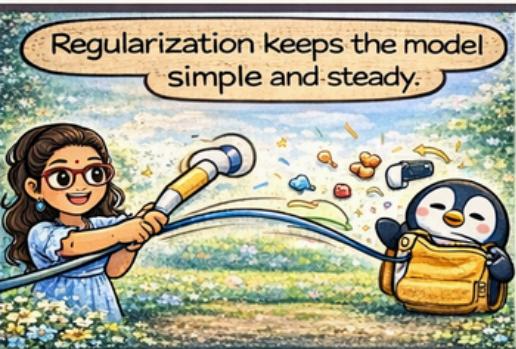
Episode 23 – Hyperparameter Tuning



See you in Episode 24 – Regularization (Avoiding Overfitting)!

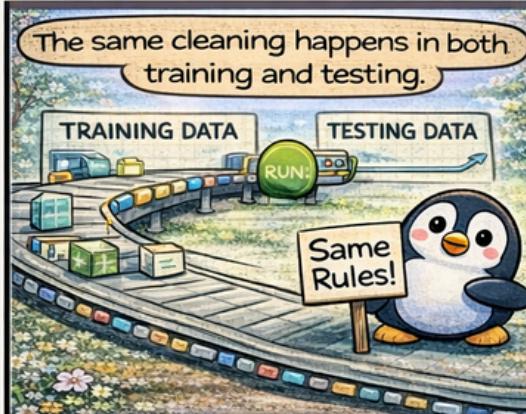
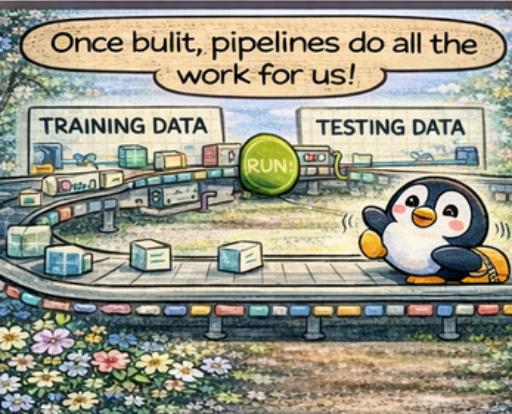
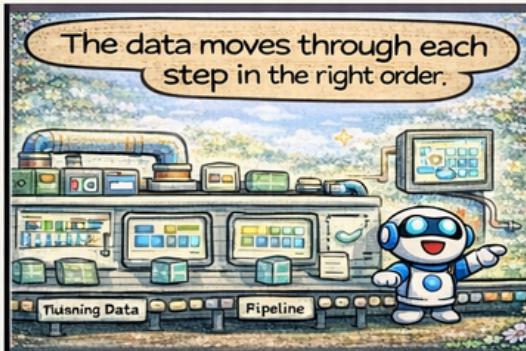
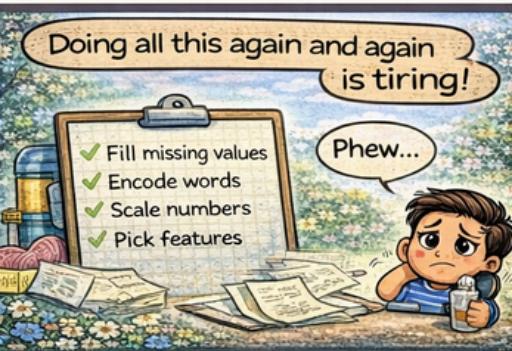


Episode 24 – Regularization (Keep It Simple!)



See you in Episode 25 – Decision Trees (Advanced Magic)!

Episode 25 - Pipelines & Preprocessing



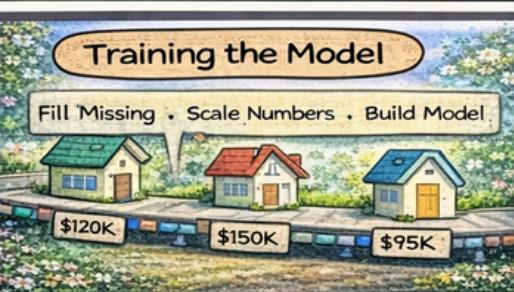
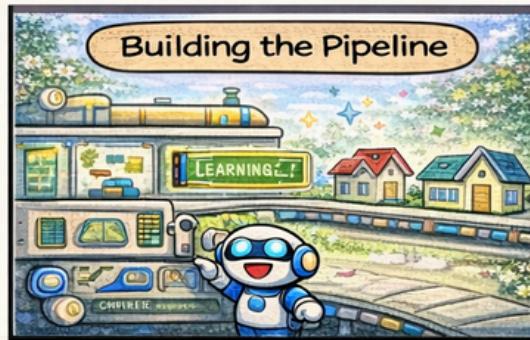


Episode 26 - Mini Project: Predict House Prices



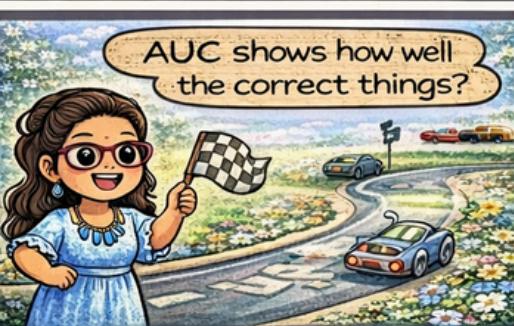
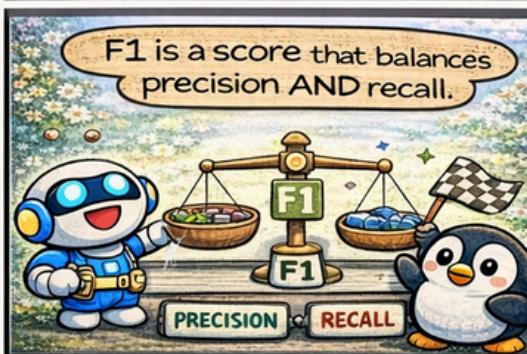
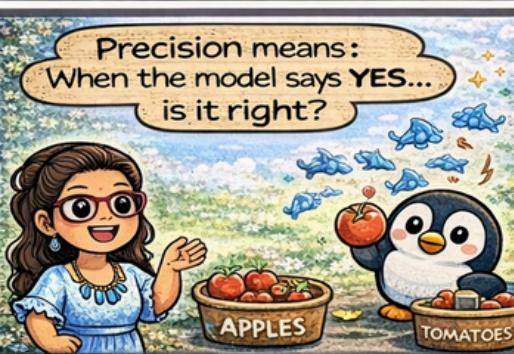
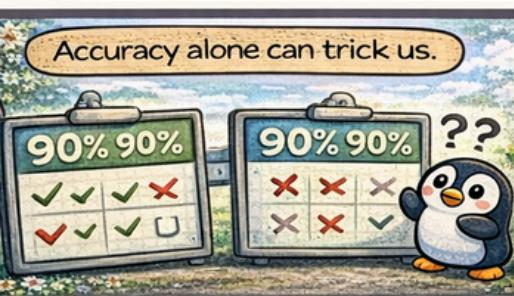
Our Dataset

| Size | ISQ FT | 8 | 3 | 6 | ✓ | Price |
|------|--------|---|---|-----|---|---------|
| 700 | 4,100 | 6 | 5 | 500 | 5 | \$1,200 |
| 850 | 6,500 | 6 | 3 | 500 | 4 | \$1,500 |
| 900 | 3,500 | 3 | 3 | Cat | 1 | \$1,500 |



See you in Episode 27 - Ensemble Learning (Teamwork of Models))

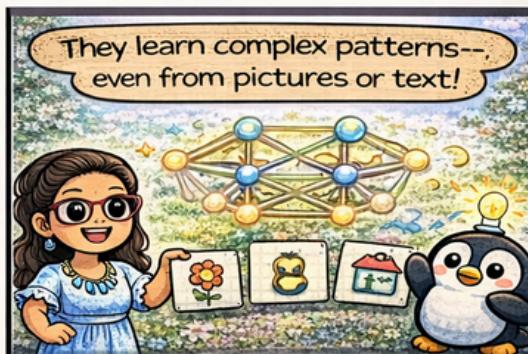
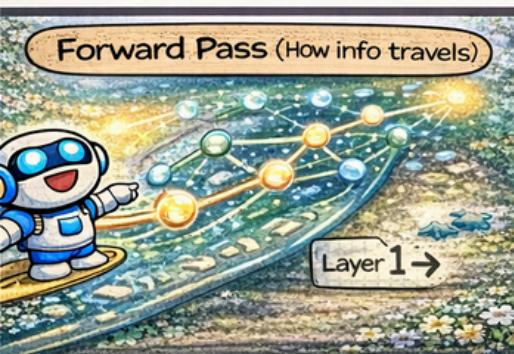
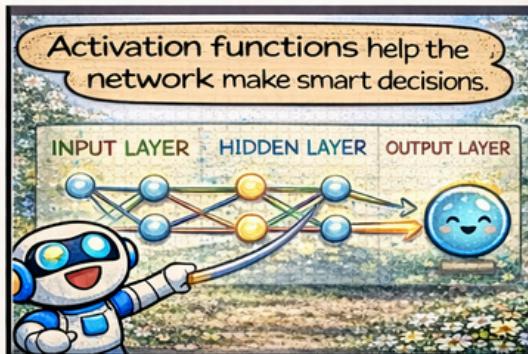
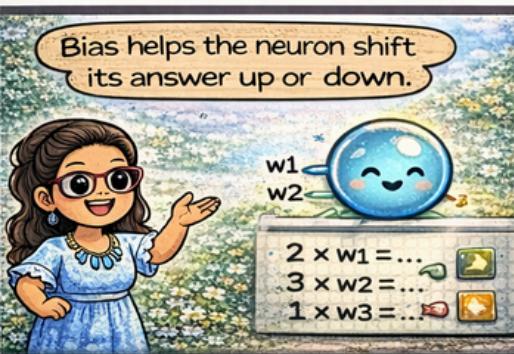
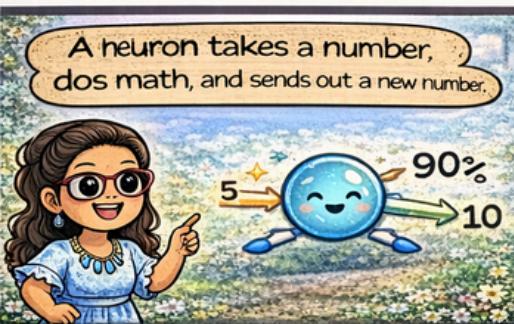
Episode 27 - Smarter Scores: AUC & F1



See you in Episode 23 - Ensemble Learning (Teamwork of Models))!



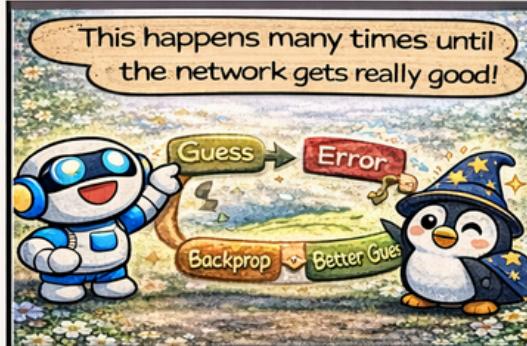
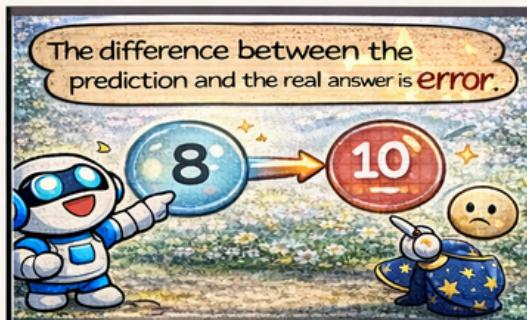
Episode 28 - Intro to Neural Networks



See you in Episode 29 - Training Neural Networks (Backpropagation Magic!)



Episode 29 - Training Neural Networks (Backpropagation Magic!)



See you in Episode 30 - Mini Project. Build a Neural Network!



Episode 30 – Responsible AI



AI must help people – not hurt or confuse them.



Be Fair

AI should treat everyone fairly.



Use Safe Data

Use data that is safe, respectful, and allowed.



Don't Let AI Be Mean

AI should never be harmful or unkind.



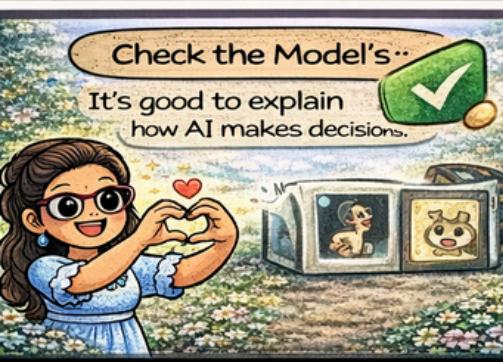
Check the Model's Mistakes

We check models often to keep them safe.



Check the Model's...

It's good to explain how AI makes decisions.



Humans in Control

AI helps humans... but humans stay in charge.



A Sriju Comic



Curiosity Lives Here!

Stories that smile back at you