QUANZETTA

# Did the Model Learn the Training Data?

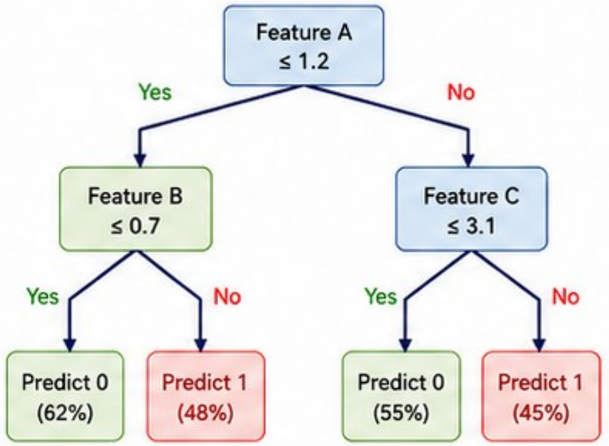
Let's compare the same model on the data it studied and the data it never saw.

**1. Training Set**



**Data the model studied**
(Practice)

**2. Same Trained Model**




Trained Decision Tree Model

**3. Test Set**



**Data the model never saw**
(Exam)

`model.predict(x_train)` → **train_accuracy**

`model.predict(x_test)` → **test_accuracy**

**To diagnose the model, we should compare training performance with test performance.**

QUANZETTA

# A Surprising Result

The same model behaves very differently.

## Training Accuracy



**100%**

On the training data the model has seen before.

?

**Why such a big gap?**

## Test Accuracy



**59.5%**

On new data the model has never seen.

**Excellent training performance does not guarantee strong performance on new data.**

QUANZETTA

# Memorizing vs Understanding

Being perfect on familiar examples is not the same as generalizing well.

## Practice Questions




Aces the practice questions
(Training Data)



## New Exam






Struggles on the new exam
(Unseen Data)

**Training Set**
= practice questions

=

**Test Set**
= new exam

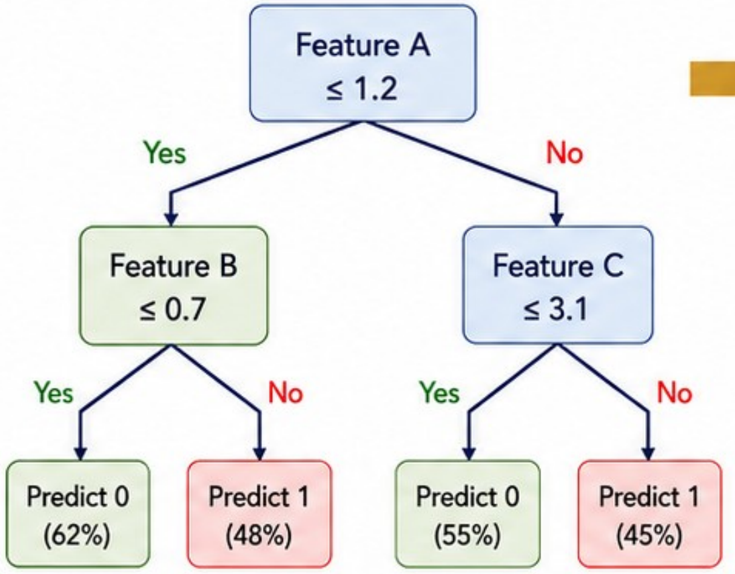
**A good model should generalize, not just memorize.**

QUANZETTA

# Look Inside Our Decision Tree

How complex did the tree become?

**Small, Shallow Tree**



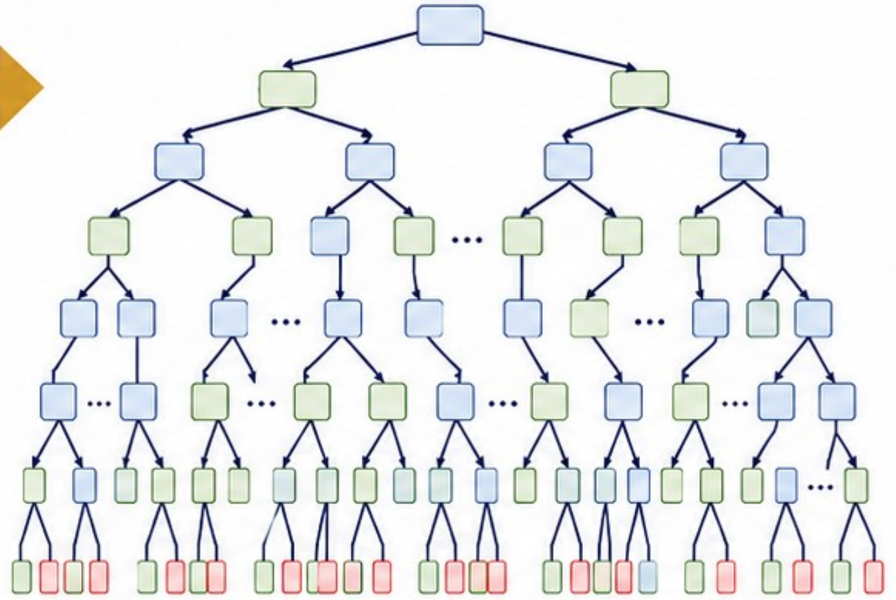

**Inspect the Tree**

```
1  model.get_depth()
2  model.get_n_leaves()
```

Depth = **28**

Leaves = **1,396**

**Large, Deep Tree**



**An unrestricted tree can keep growing until it learns very specific rules.**

QUANZETTA

# Now We Can Name the Pattern

Three common situations when we compare training and test performance.

## Underfitting

Both train and test are low.




## Good Fit

Both train and test are good and close.




Our model

## Overfitting

Train is very high, test is much lower.

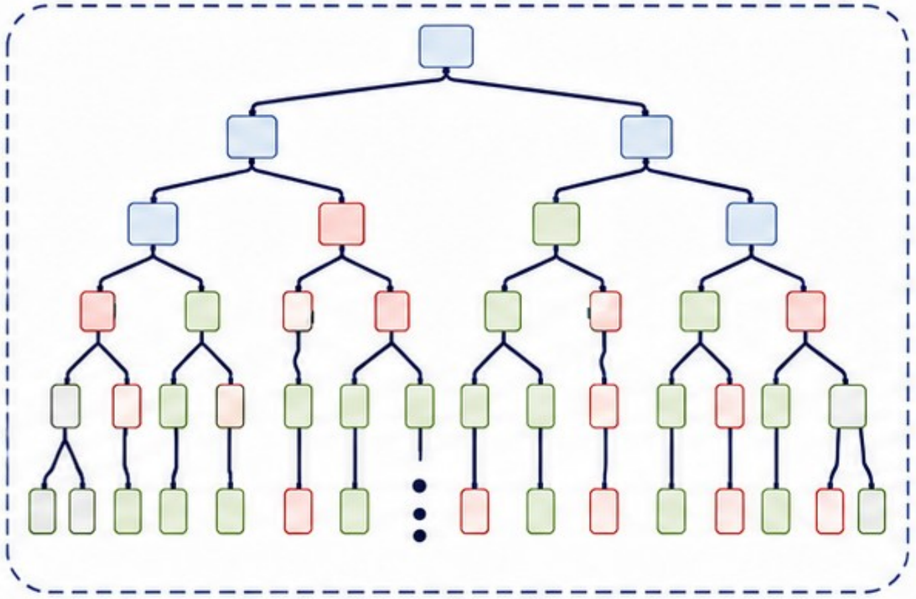



**Our current model matches the overfitting pattern.**

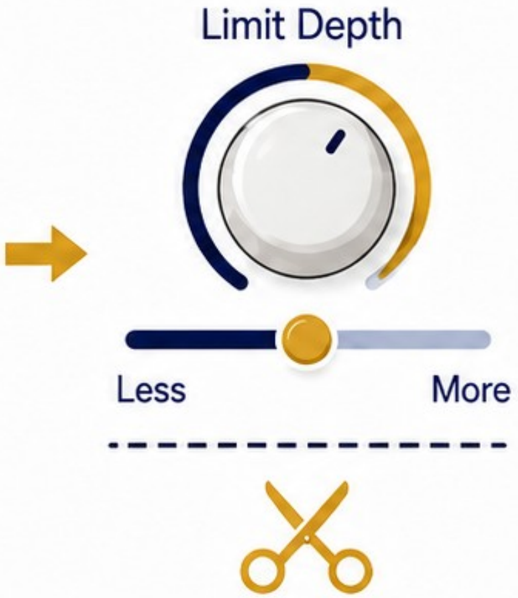
QUANZETTA

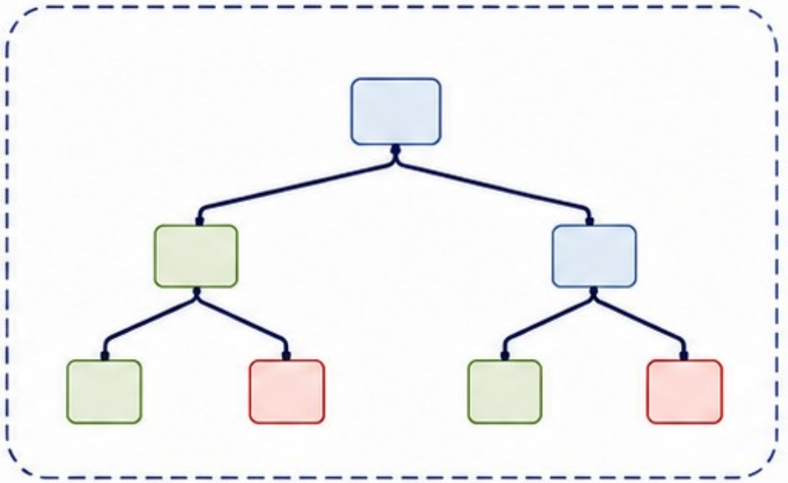
# What Should We Do Next?

The tree may be learning too much detail.



Too deep. May overfit.



Simpler. More balanced. Better generalization.

**Can we stop the tree from growing too much?**

**Next lesson:** control model complexity to improve generalization.