DECISION TREES

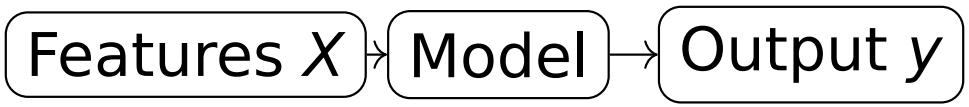
Intuition and Applications

April 20, 2026

Aryan Parikh

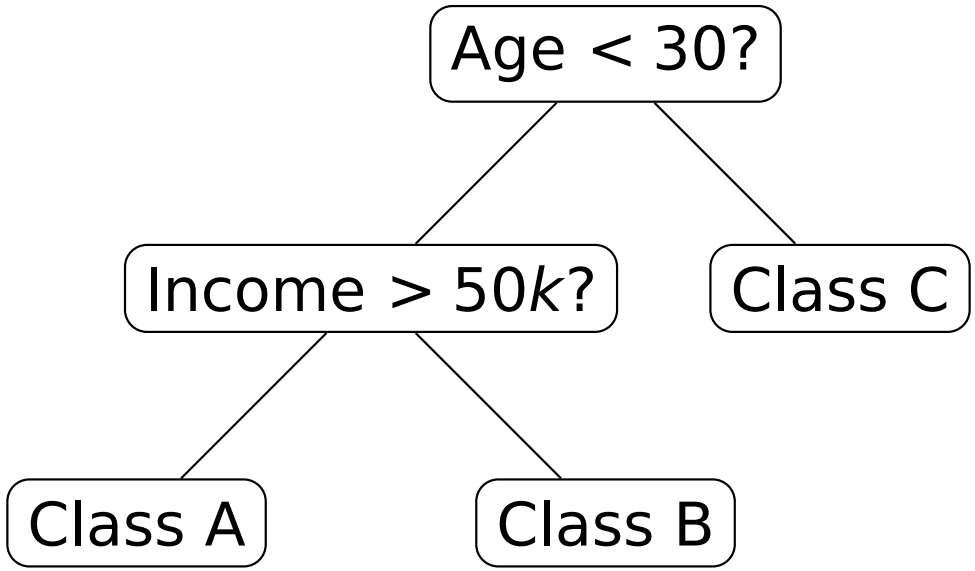
MOTIVATION

- Supervised learning: mapping from features X to outputs y using labeled data
- Training data consists of input-output pairs (x_i, y_i)
- Two main tasks:
  - Classification (discrete labels)
  - Regression (continuous values)
- Goal: build models that generalize well and are interpretable



WHY DECISION TREES?

- → Linear models struggle with nonlinear relationships
- → Hard to capture feature interactions
- → Decision Trees:
  - → Recursive partitioning of feature space
  - → Handle nonlinearity and interactions
  - → Work with mixed data types



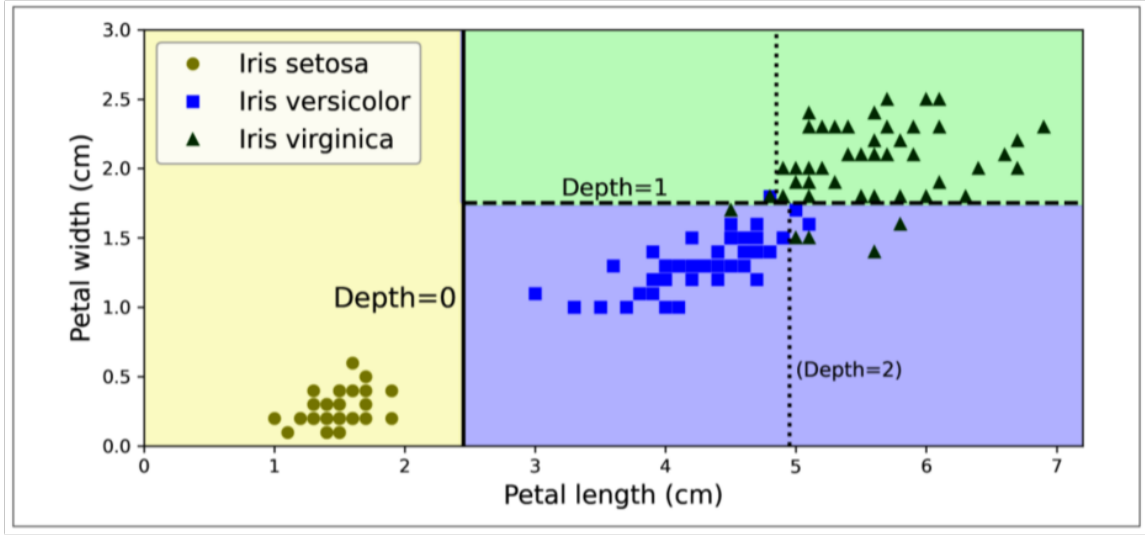
DECISION TREES FOR CLASSIFICATION

→ Goal: split data to maximize purity

→ Common criteria: Gini impurity

$$\text{Gini} = 1 - \sum_{k=1}^{K} p_k^2, \quad p_k : \text{proportion of class } k \text{ in node}$$

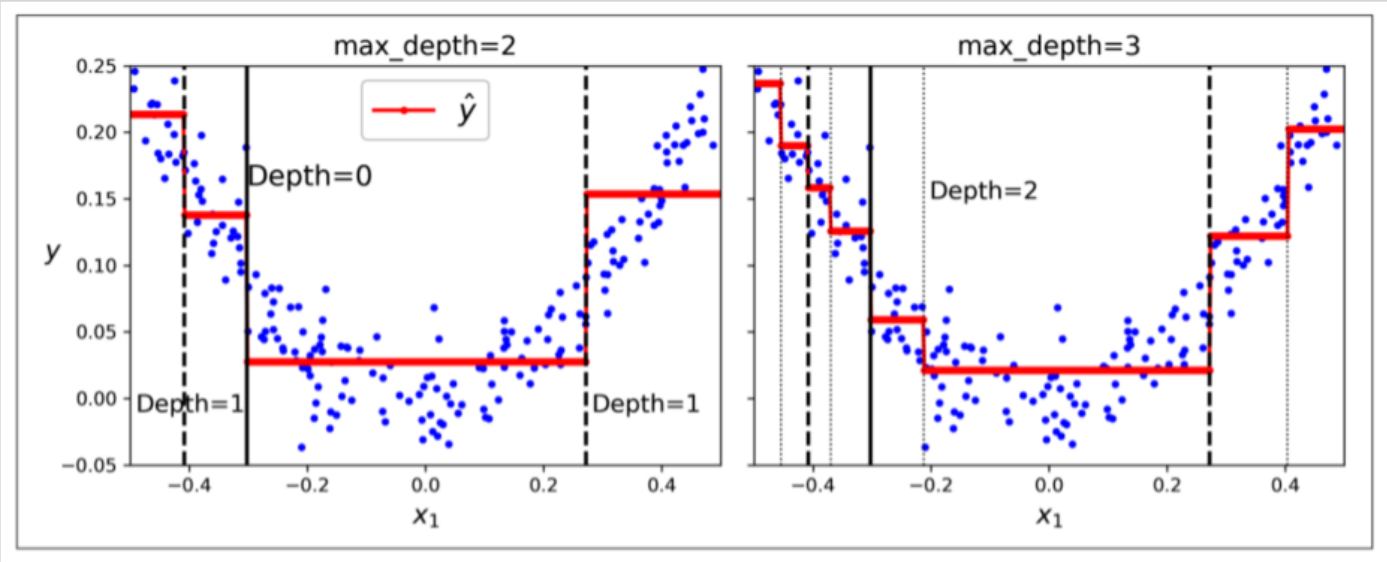
$$k : \text{class index}, \quad K : \text{total number of classes}$$



DECISION TREES FOR REGRESSION

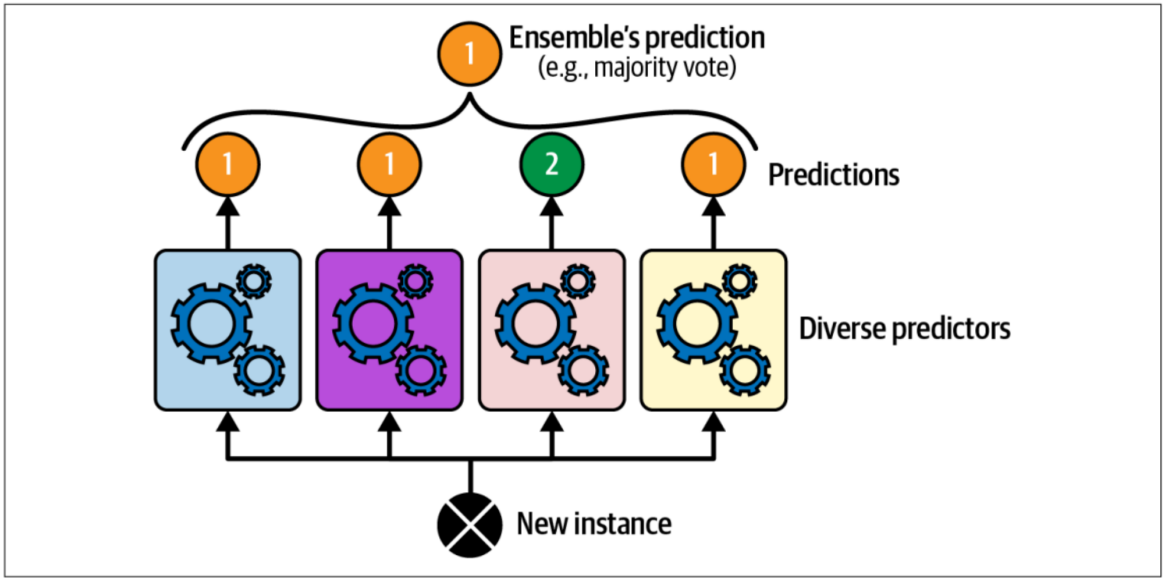
- → Goal: minimize prediction error (MSE)
- → Prediction: average value in each leaf

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$



RANDOM FORESTS

- Combine many decision trees
- Use bagging (bootstrap aggregation)
- Random feature selection
- Predictions:
  - Classification: majority vote
  - Regression: average



TAKEAWAYS

- → Decision trees are interpretable and flexible
- → Handle nonlinear relationships naturally
- → Random forests improve accuracy via ensembles

Decision trees are the foundation of many modern ML methods

Thank you!

WORKS CITED

Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 3rd ed. Sebastopol, CA: O'Reilly Media, 2025.