DECISION TREE

MOHAMMAD GHODDOSI

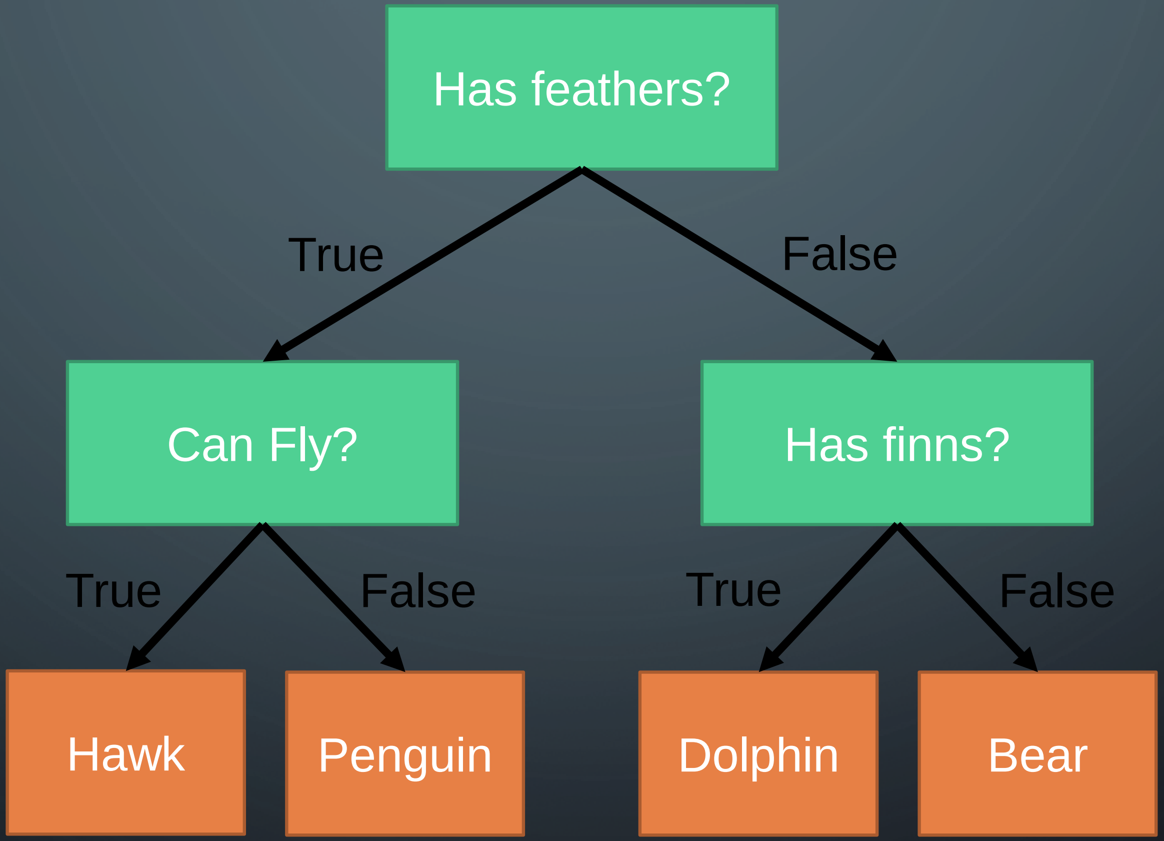
WHAT IS DECISION TREE

- Flowchart-like structure
  - Each internal node represents a test on an attribute
  - Each branch represents the outcome of the test
  - Each leaf node represents a class label

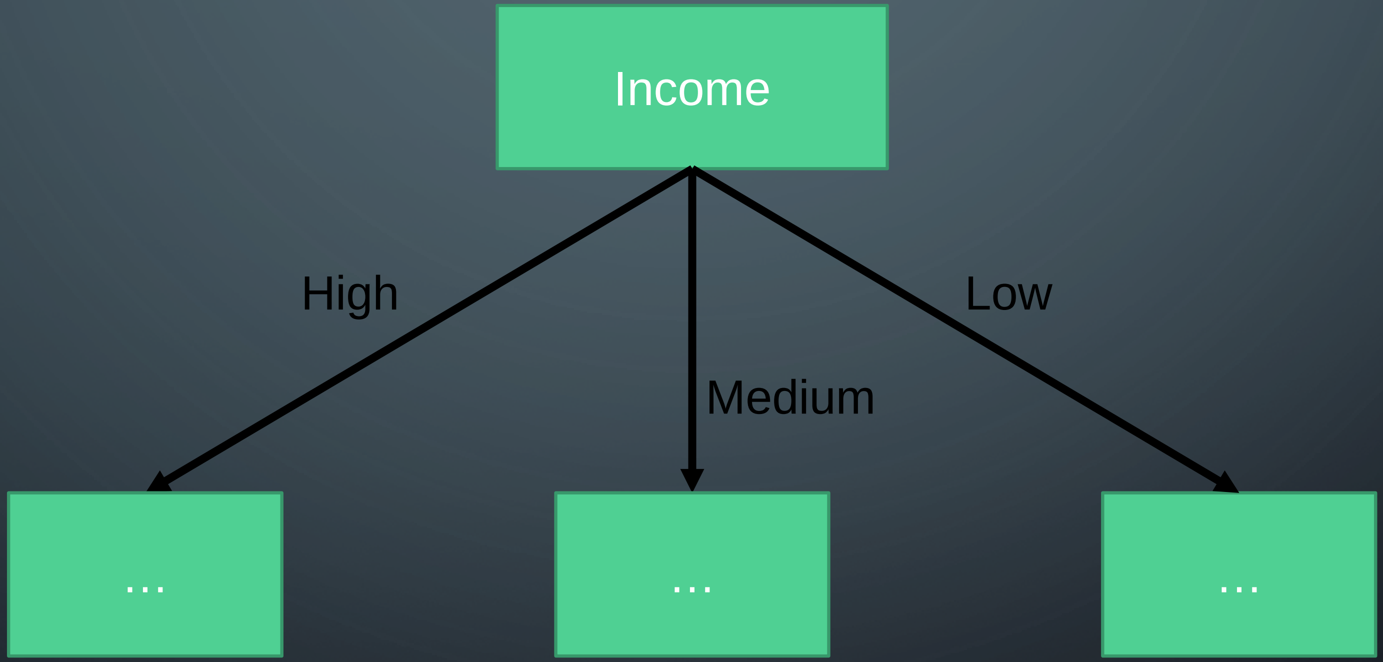
LEAF NODES

- Max depth
- Min Samples
- Output is most frequent class in leaf
- Output probability is proportion of most frequent class in leaf to others

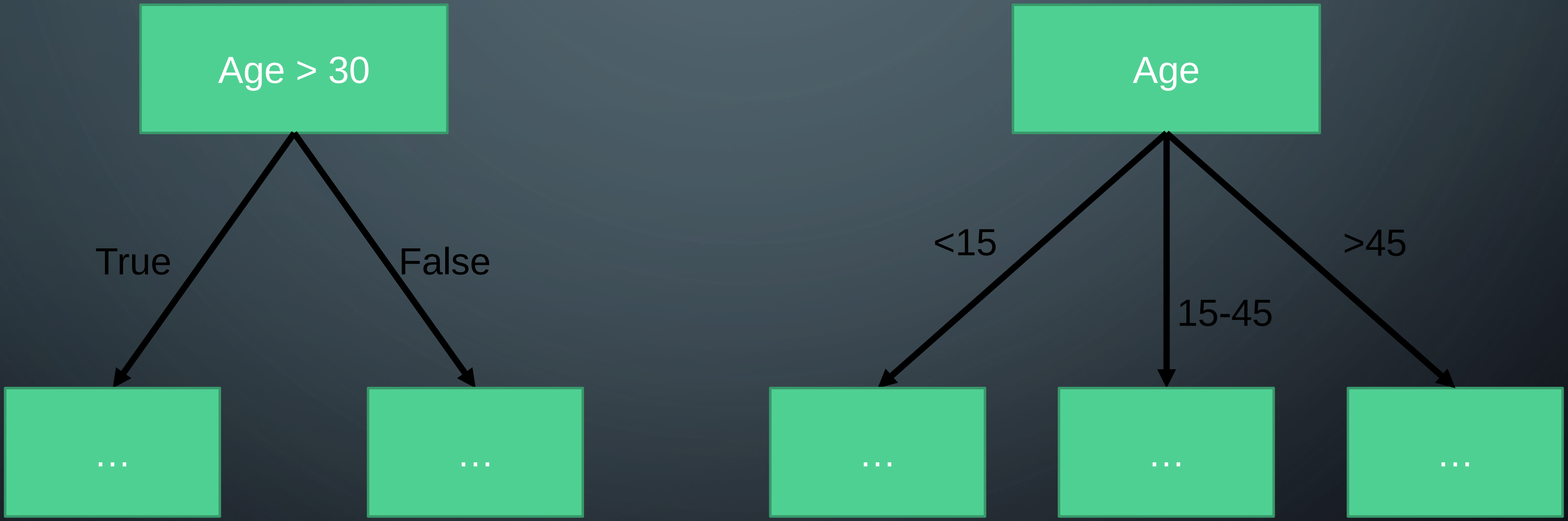
EXAMPLE - BINARY



EXAMPLE - CATEGORICAL



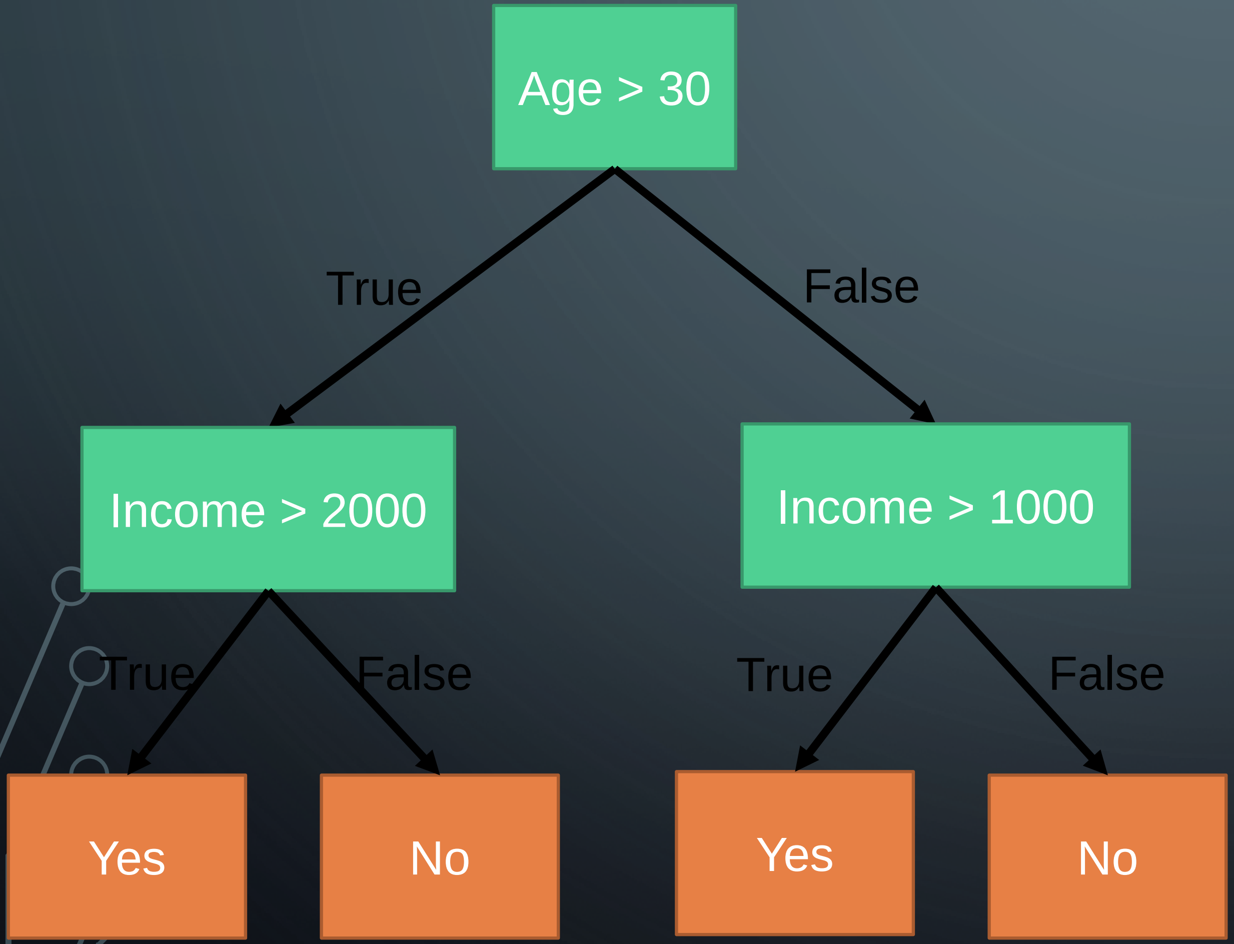
EXAMPLE - CONTINUOUS



DECISION TREE – PROS AND CONS

- Pros:
    - Understandable rules
    - Low computation
    - Both continuous and categorical variables
    - Shows more important features
- Cons:
    - Not powerful in regression
    - computationally expensive to train

DECISION TREE IN 2D SPACE



Age > 30

True

False

Income > 2000

Income > 1000

True

False

True

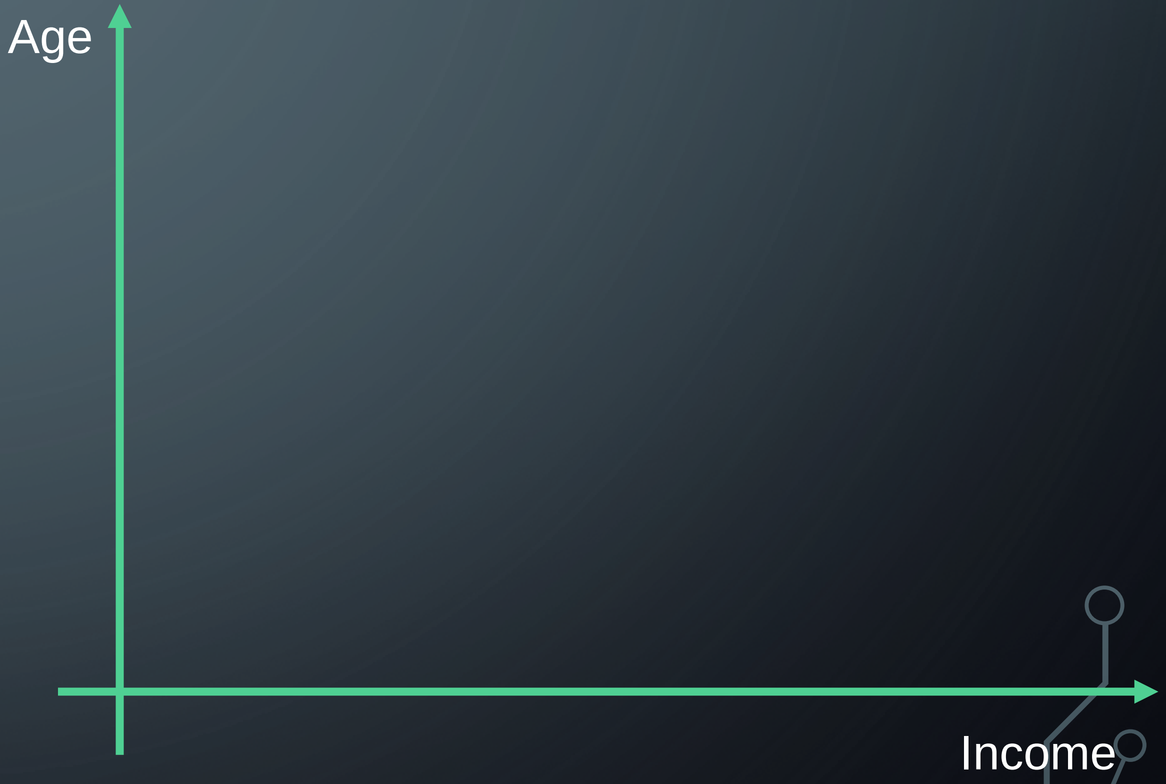
False

Yes

No

Yes

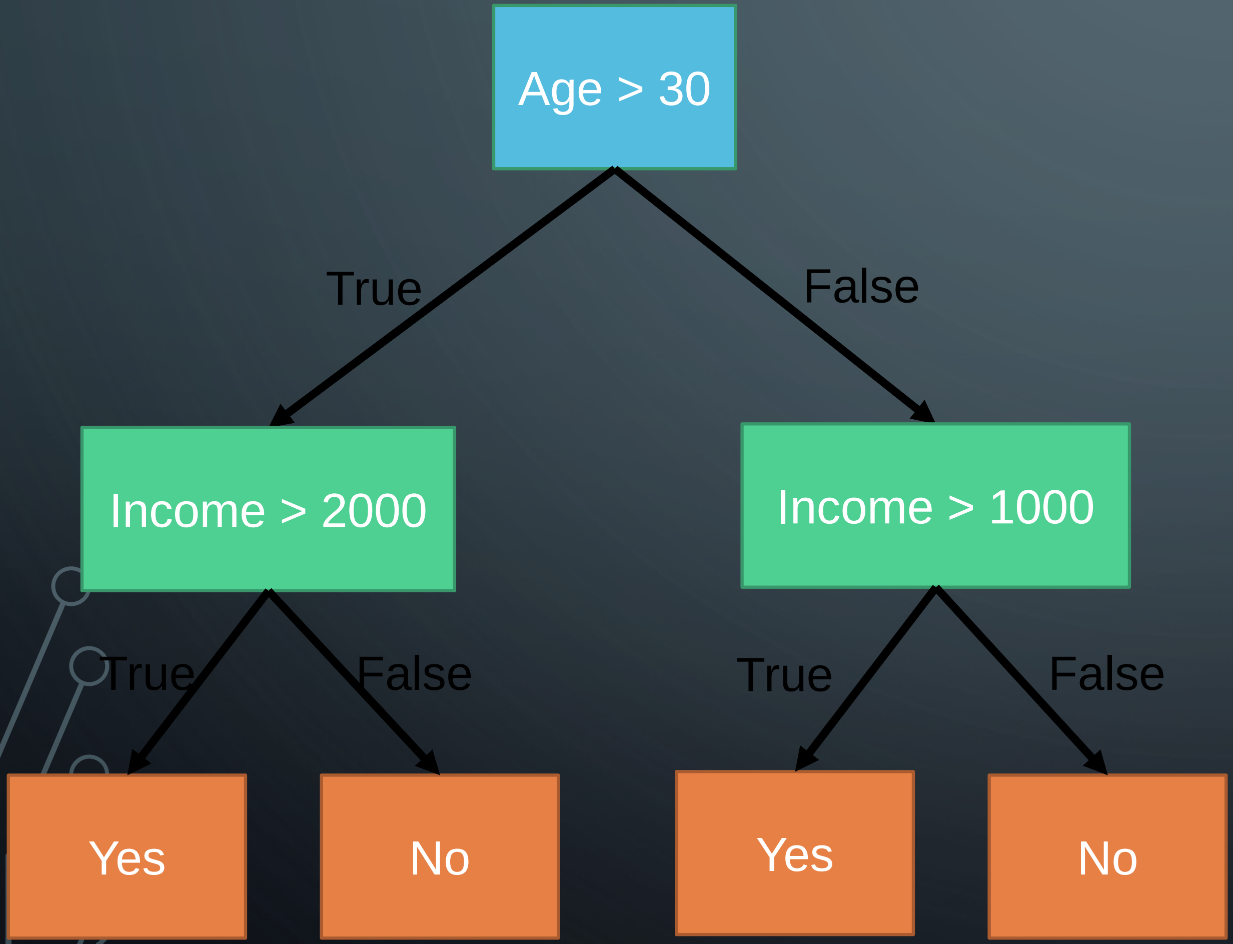
No

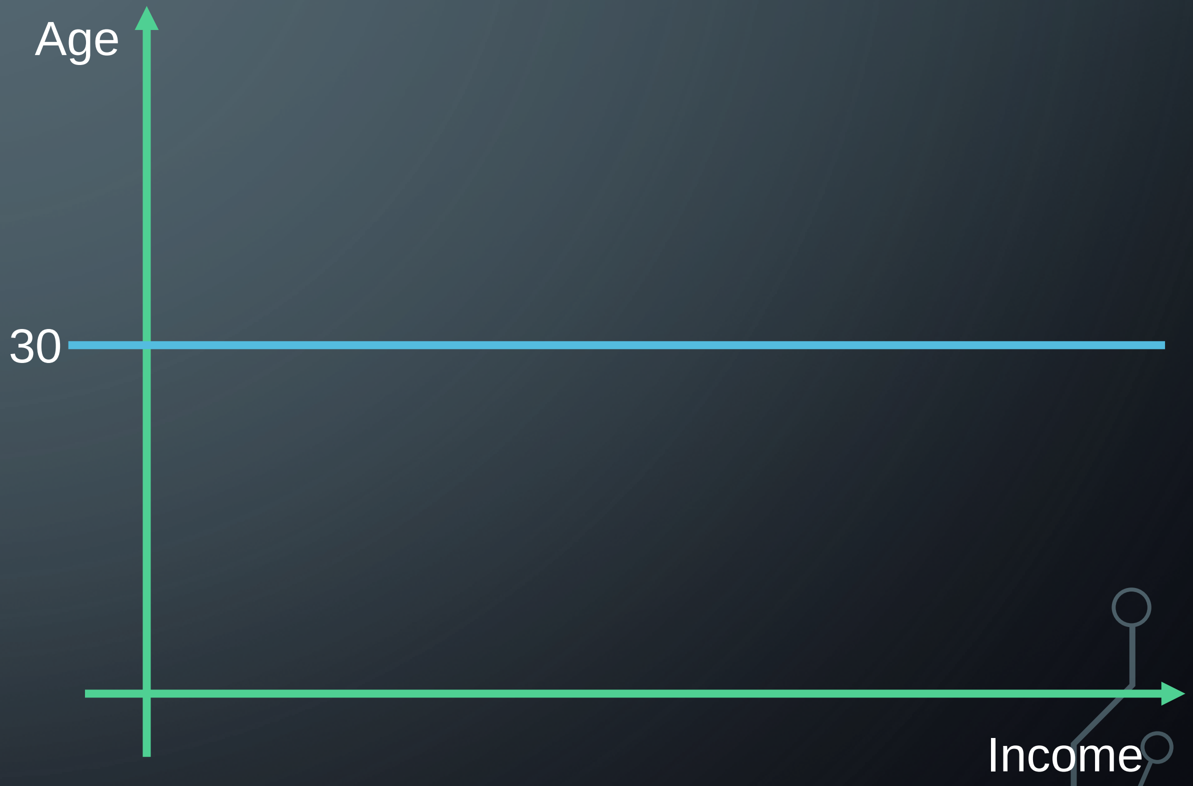


Age

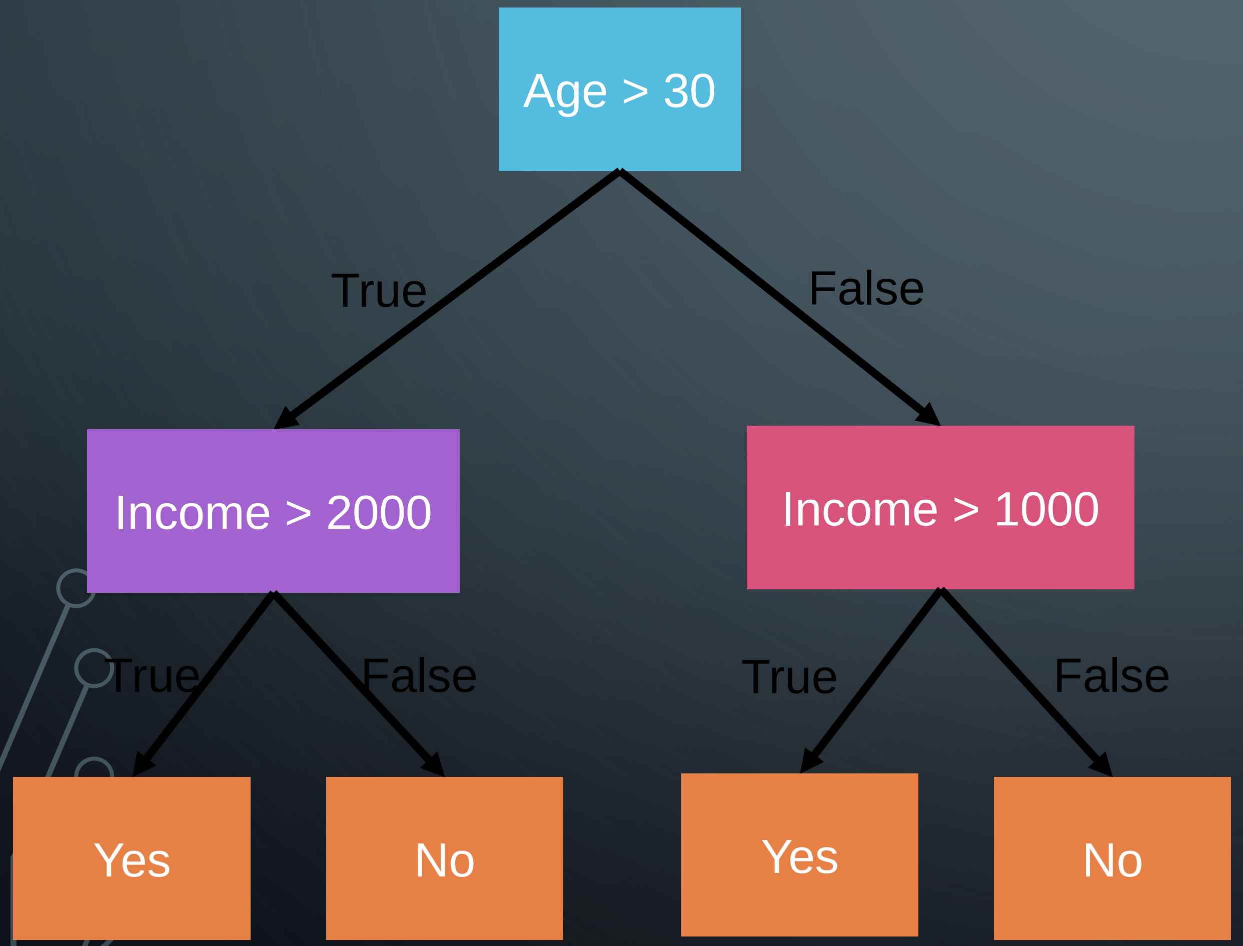
Income

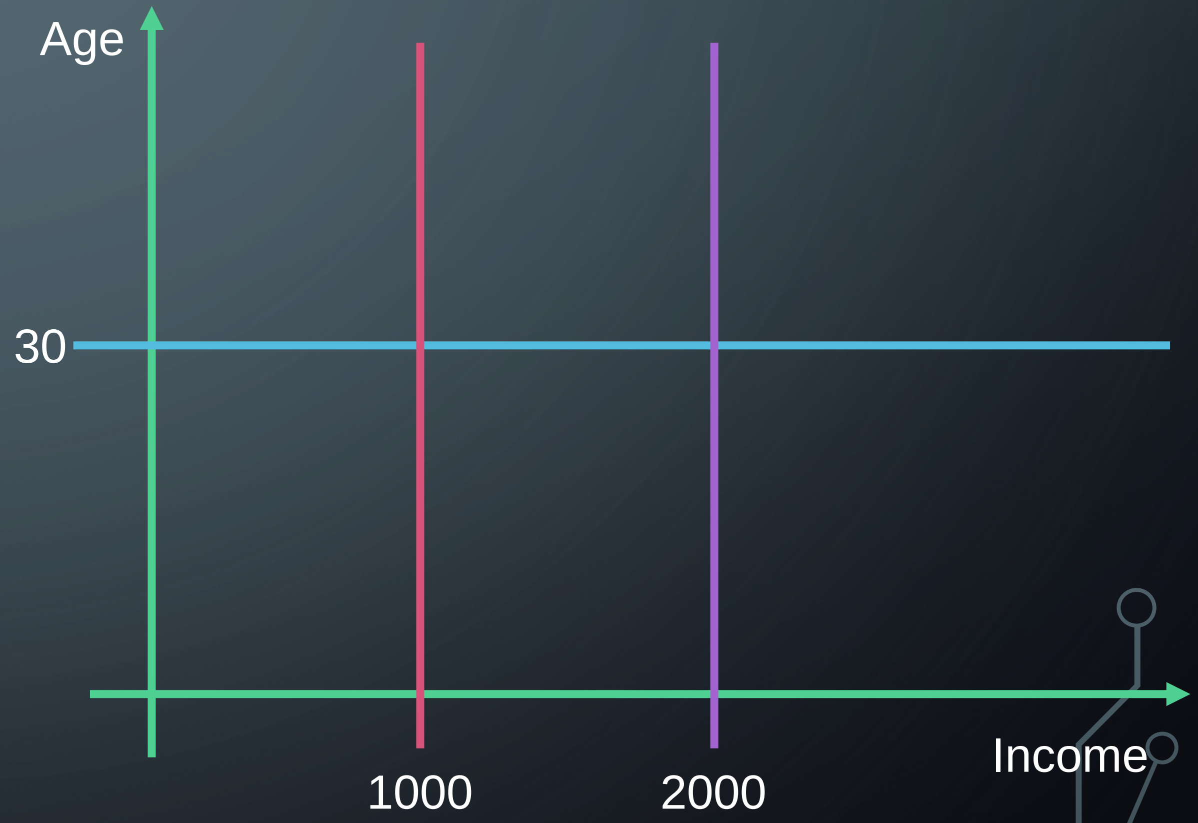
DECISION TREE IN 2D SPACE



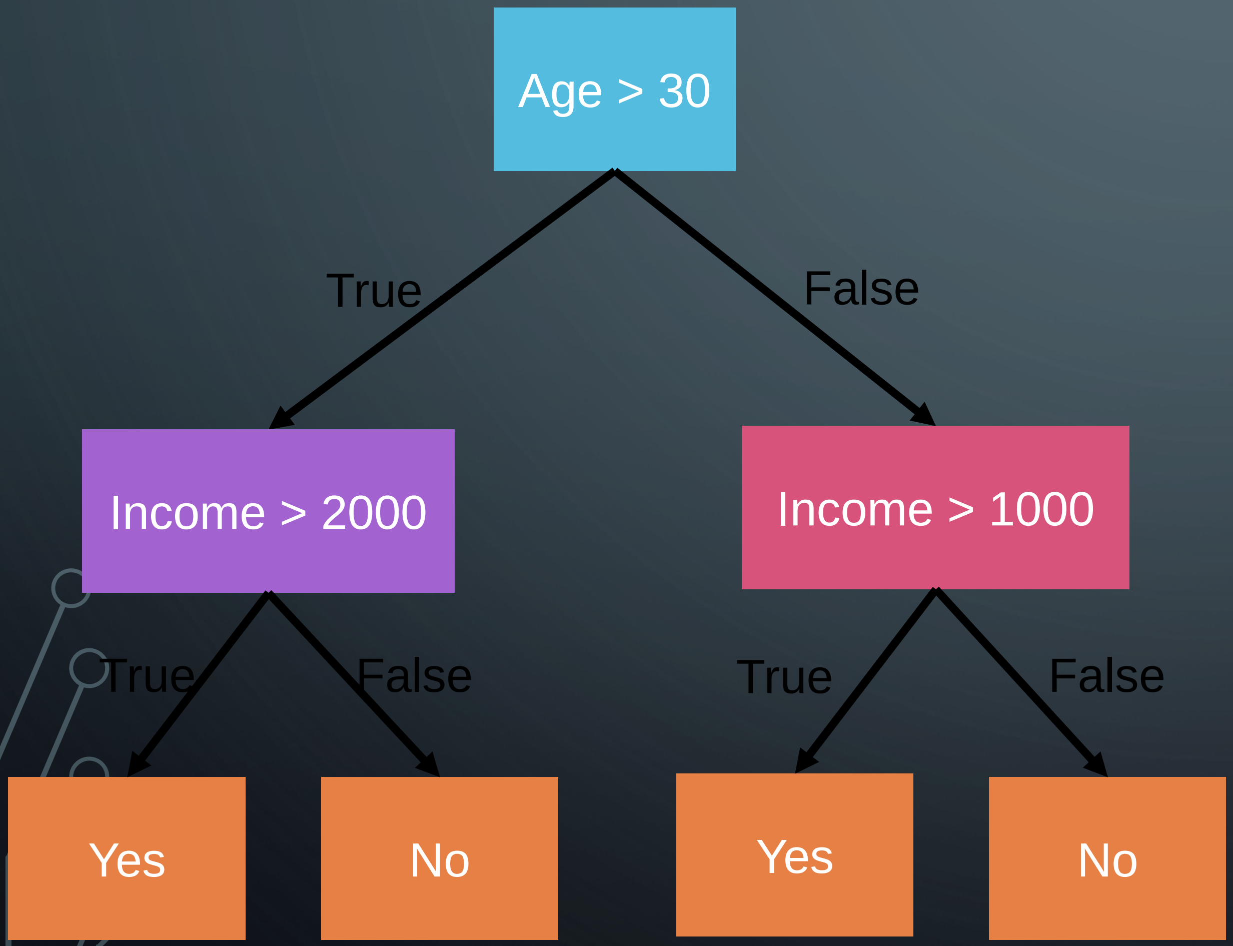


DECISION TREE IN 2D SPACE





DECISION TREE IN 2D SPACE



Age > 30

True

False

Income > 2000

Income > 1000

True

False

True

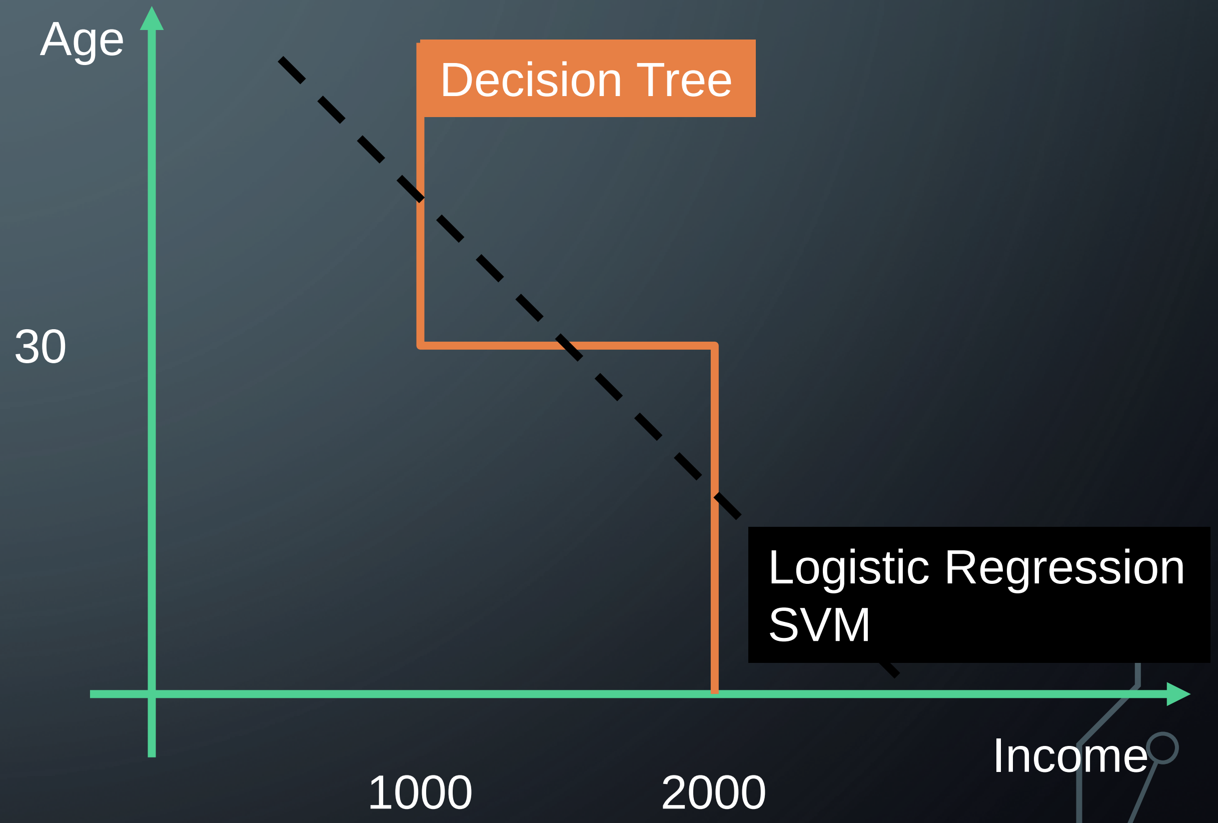
False

Yes

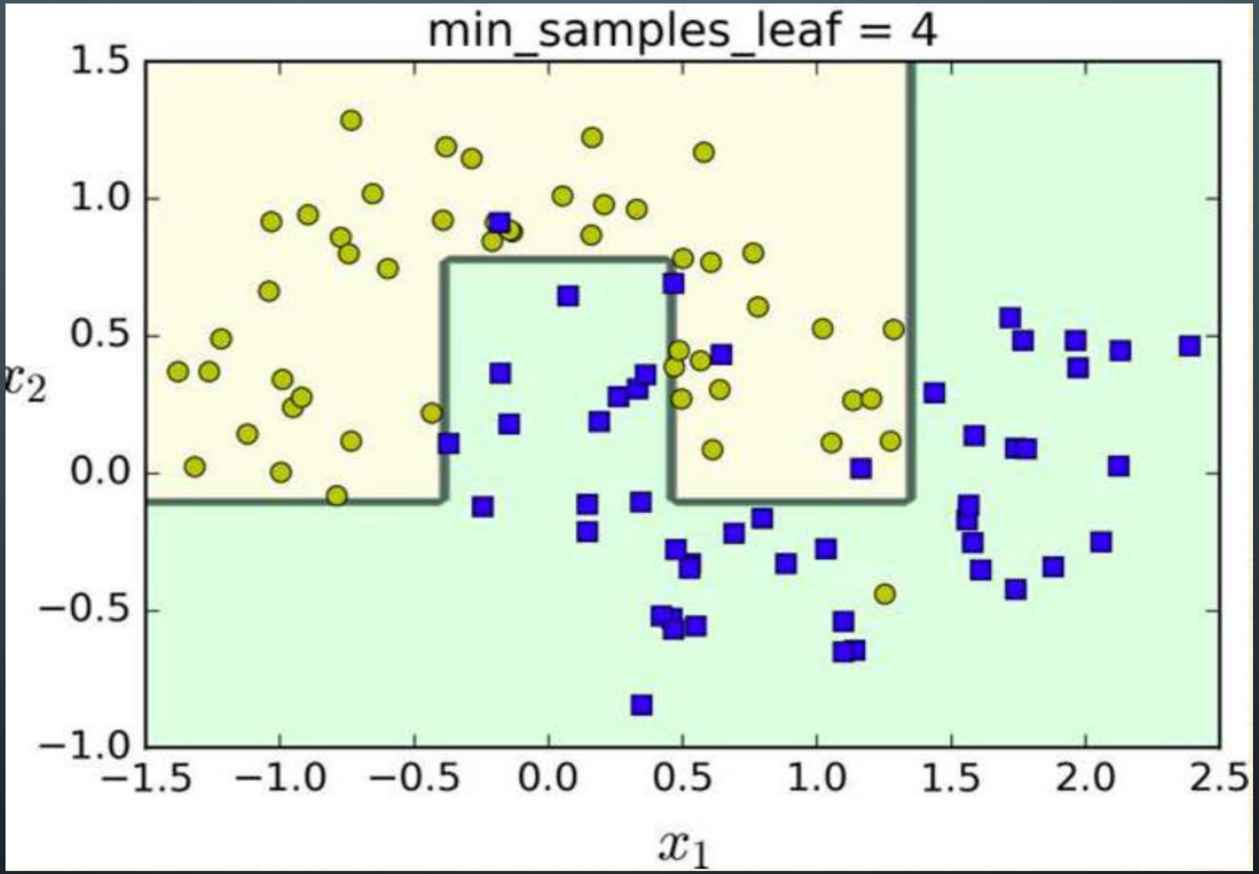
No

Yes

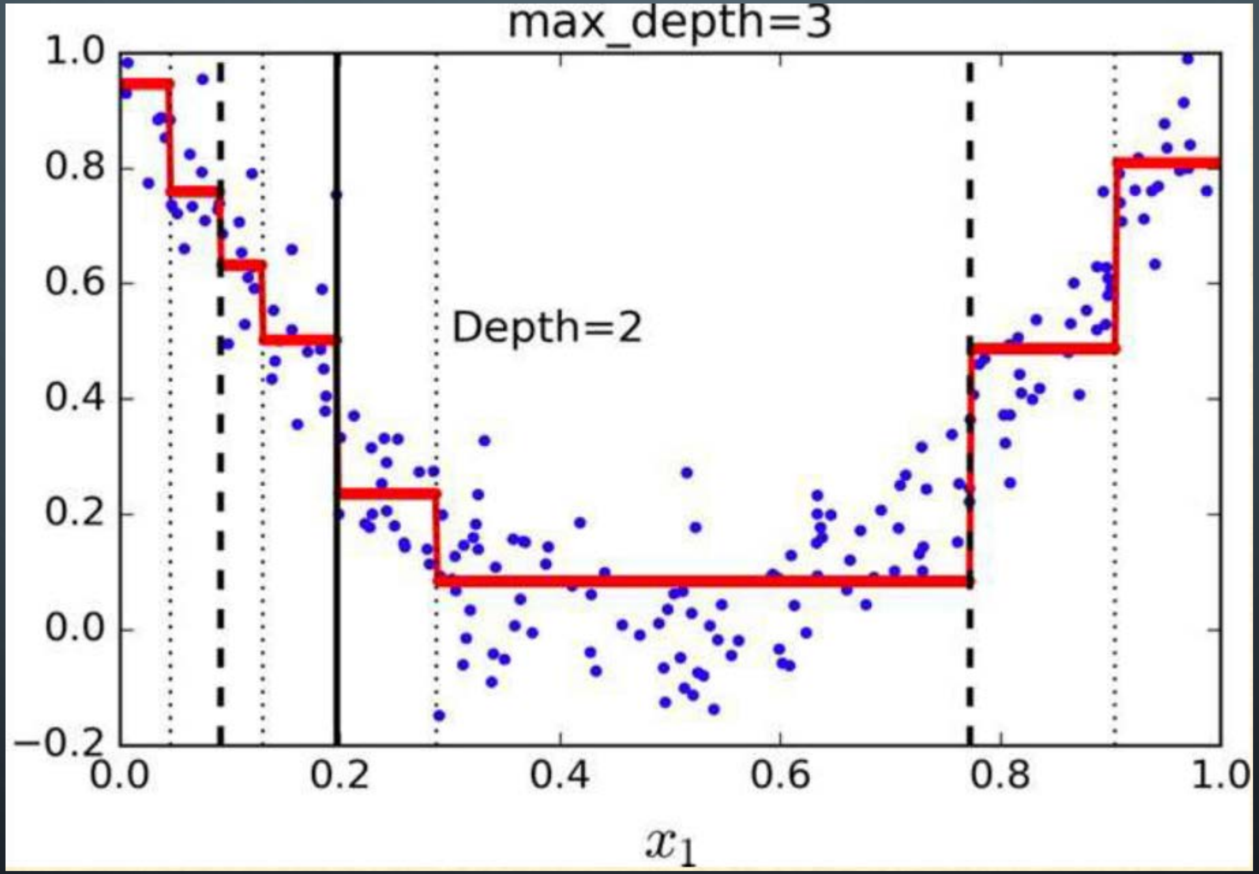
No



DECISION TREE FOR CLASSIFICATION



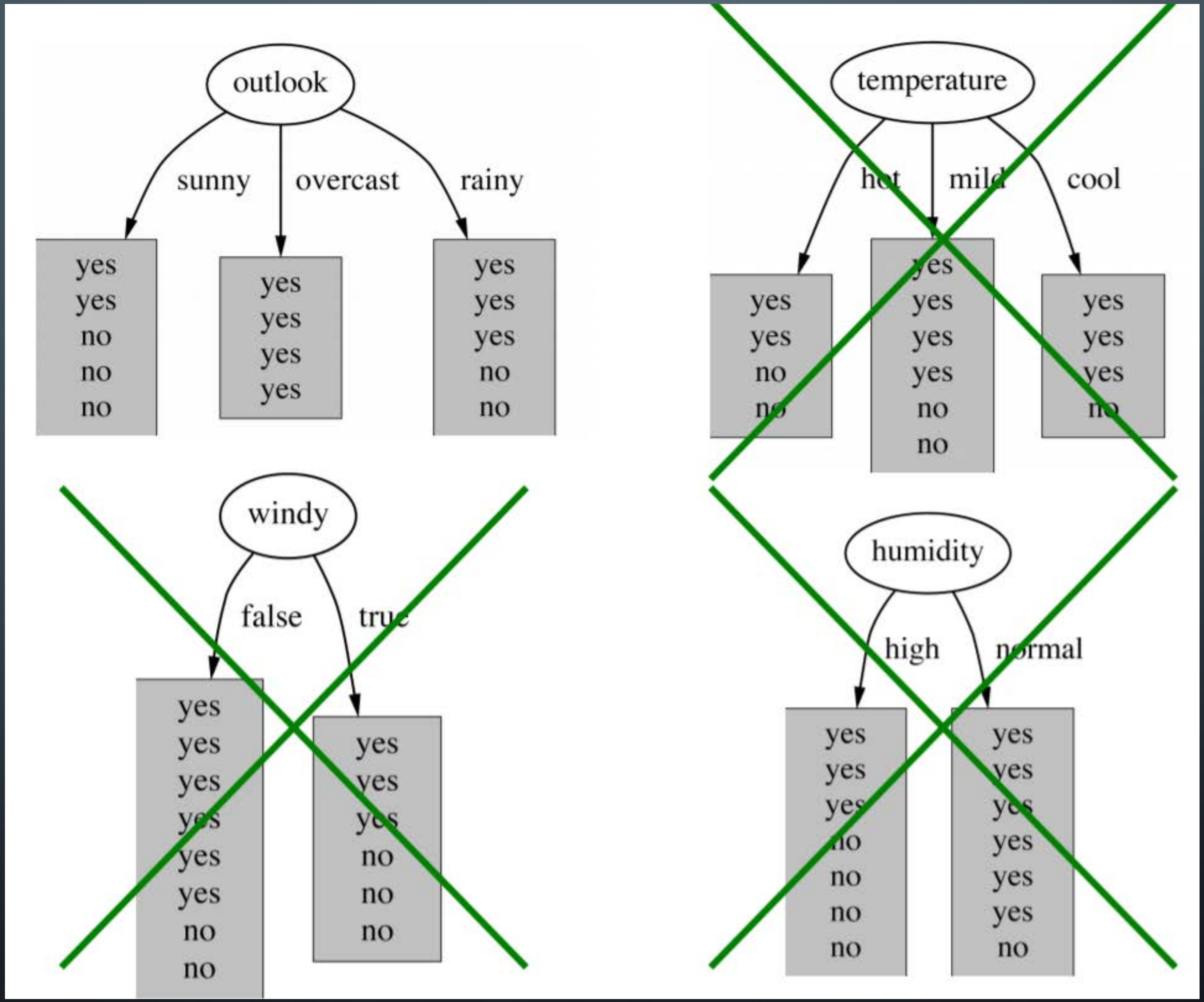
DECISION TREE FOR REGRESSION



CART ALGORITHM

- Split dataset into 2 subsets using a single feature k and threshold t_k
- Search for best (k, t_k) that produces the purest subsets.
- Do same algorithm recursively for subsets …

PURITY


outlook
sunny
overcast
rainy
yes yes no no no
yes yes yes yes
yes yes yes no no
temperature
hot
mild
cool
yes yes no no
yes yes yes yes no no
yes yes yes no
windy
false
true
yes yes yes yes yes yes no no
yes yes yes no no no
humidity
high
normal
yes yes yes no no no no
yes yes yes yes yes yes no

PURITY

- Entropy
- Average Entropy / Information
- Information Gain
- Gain ratio
- Gini Index

$$Entropy: E(S) = -\sum_{i=0}^{c} p_i \log p_i$$

$$Information = I(S, A) = \sum_{i} \frac{|S_i|}{S} . E(S_i)$$

$$Gini: G(S\ ) = 1\ - \sum_{i=0}^{c} p_i^2$$

$$Gini: G(S, A) = \sum_{i} \frac{|S_i|}{S} . G(S_i)$$

OVERFITTING AND PRUNING

- Pre-pruning
  - Stop growing a branch when information becomes unreliable
- Post-pruning
  - simplify tree after training by replacing some nodes with leafs
- Post-pruning preferred in practice

jupyter