



# DECISION TREE

MOHAMMAD GHODDOSI

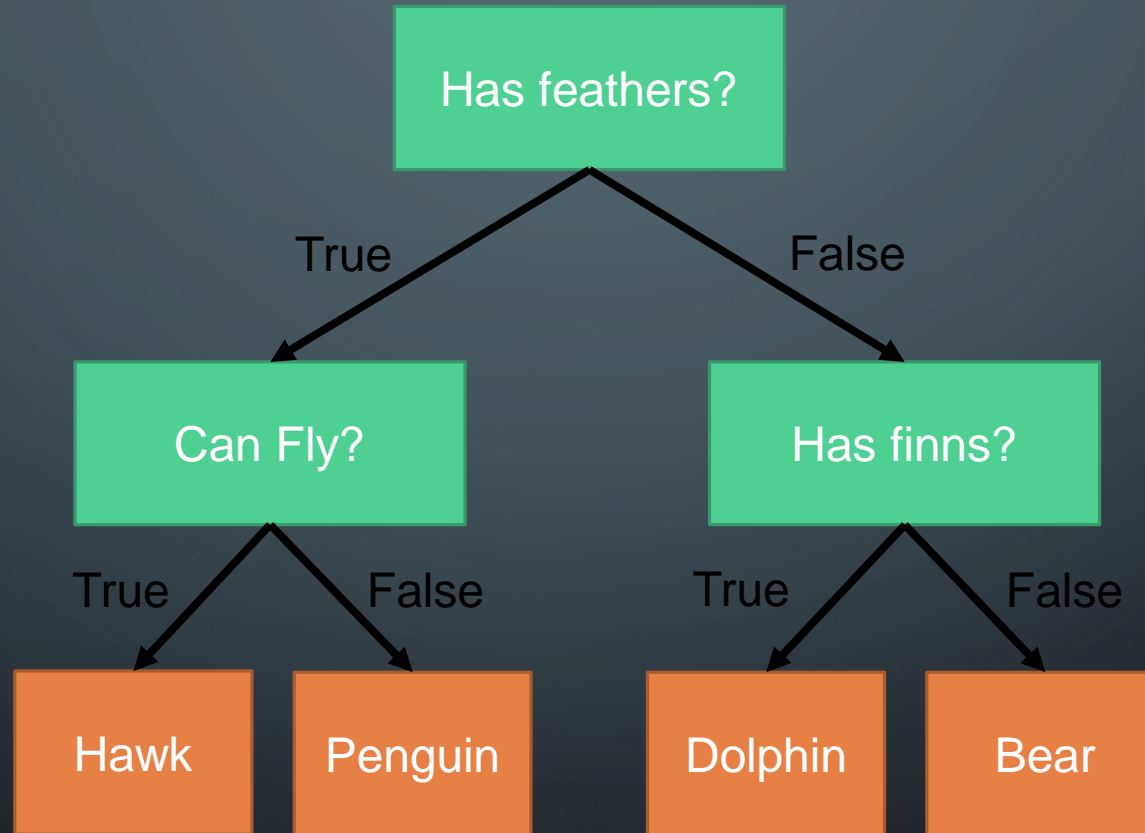
# WHAT IS DECISION TREE

- 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

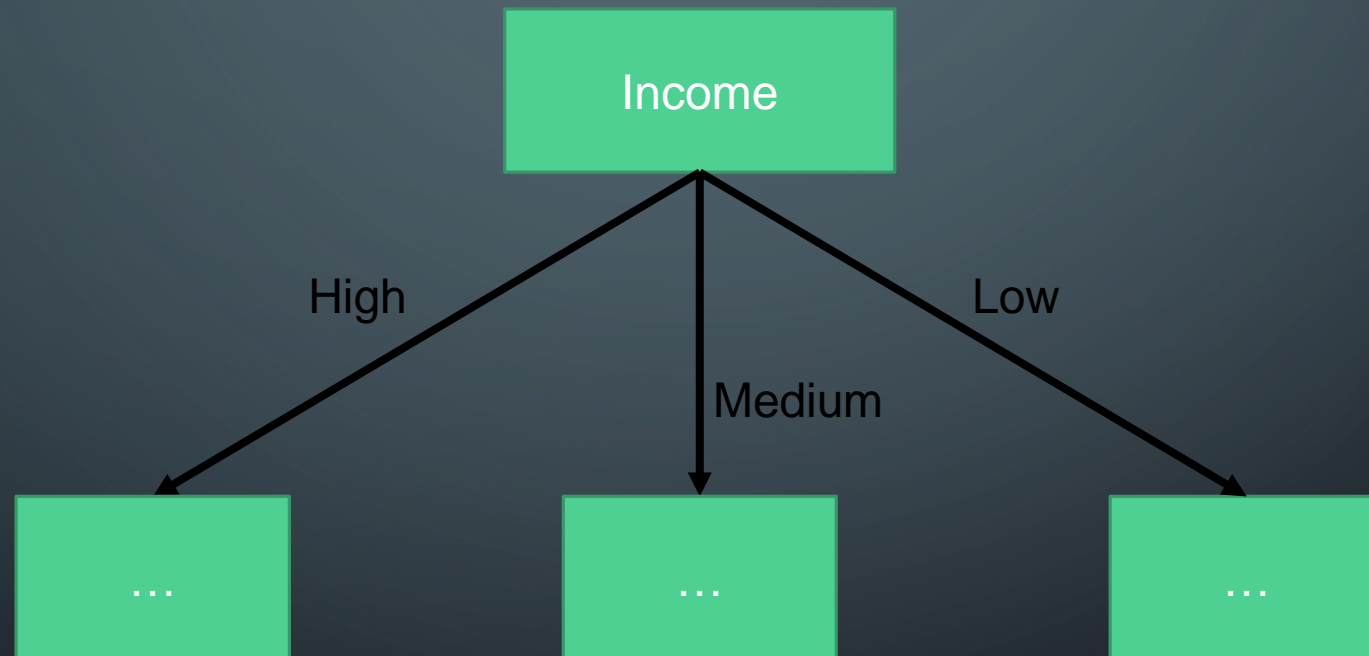
# LEAF NODES

- 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

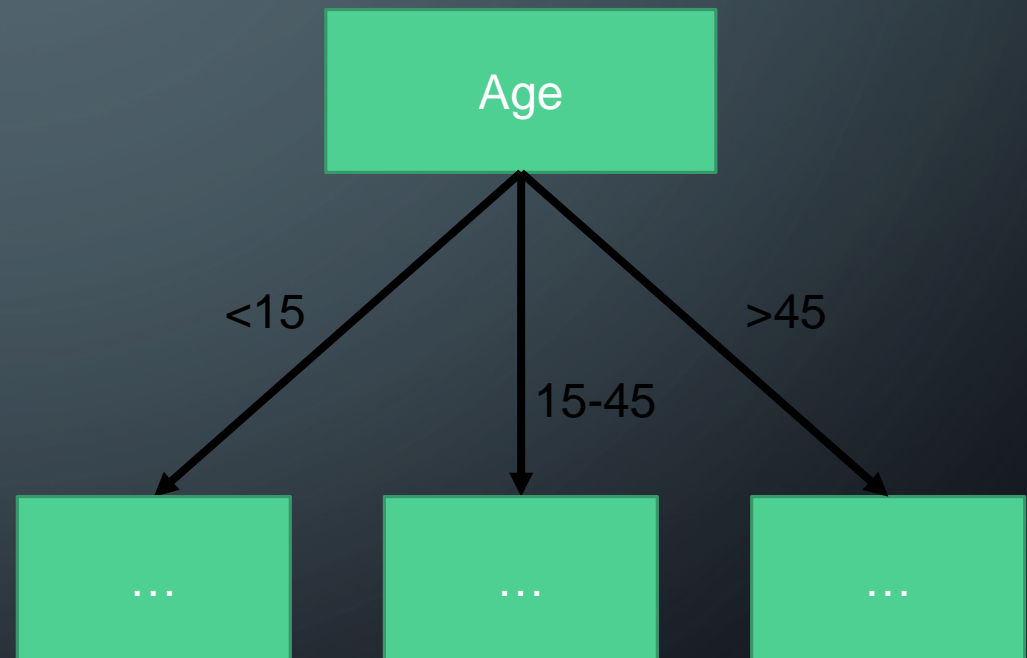
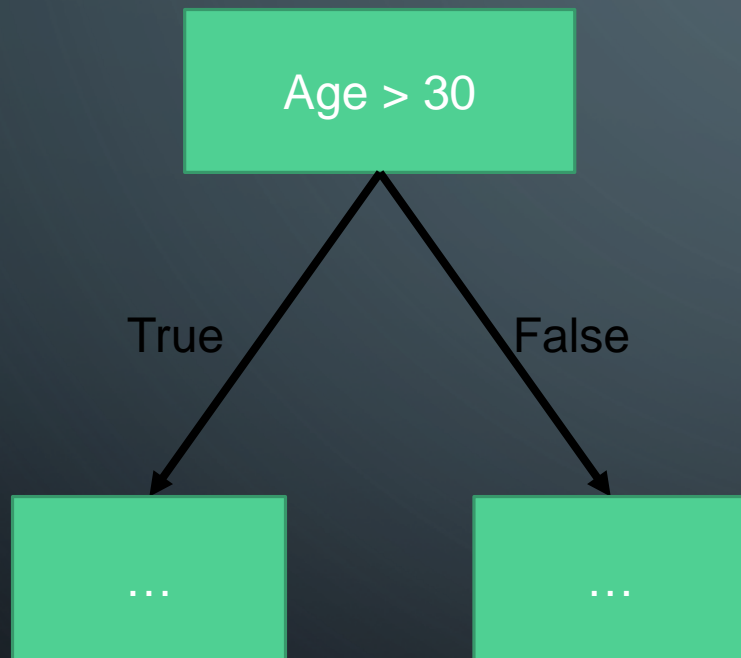


# EXAMPLE - CATEGORICAL





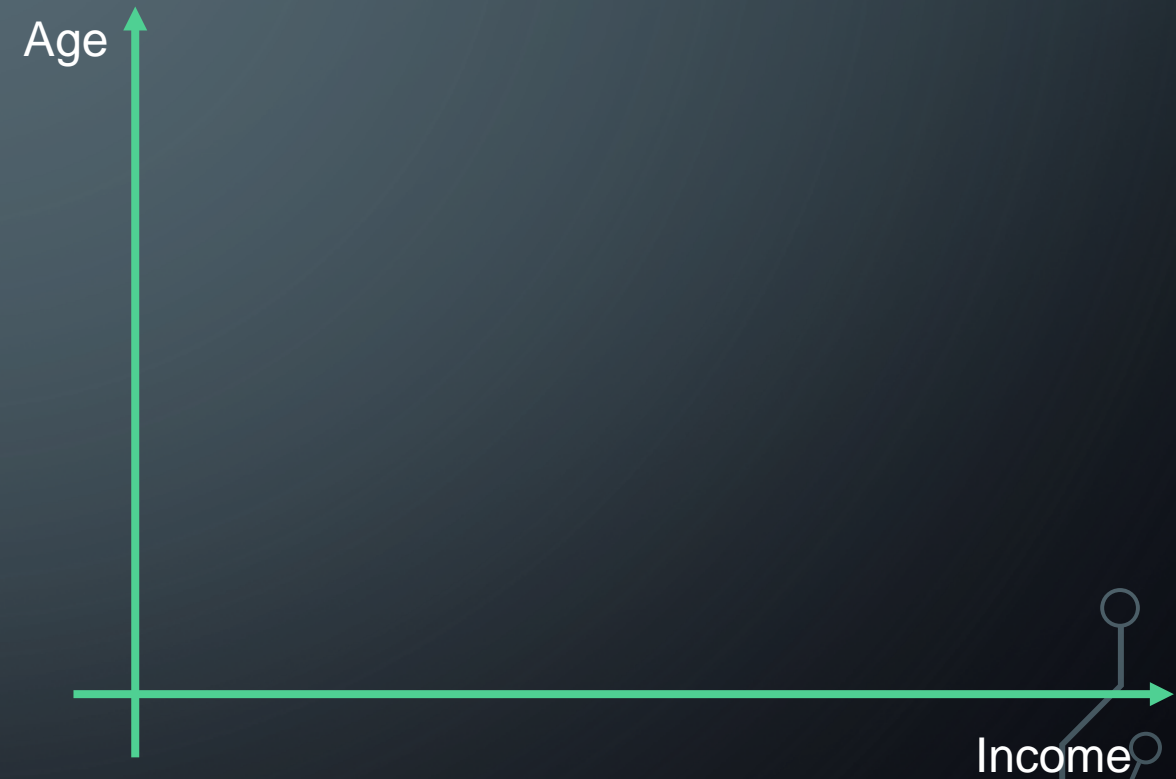
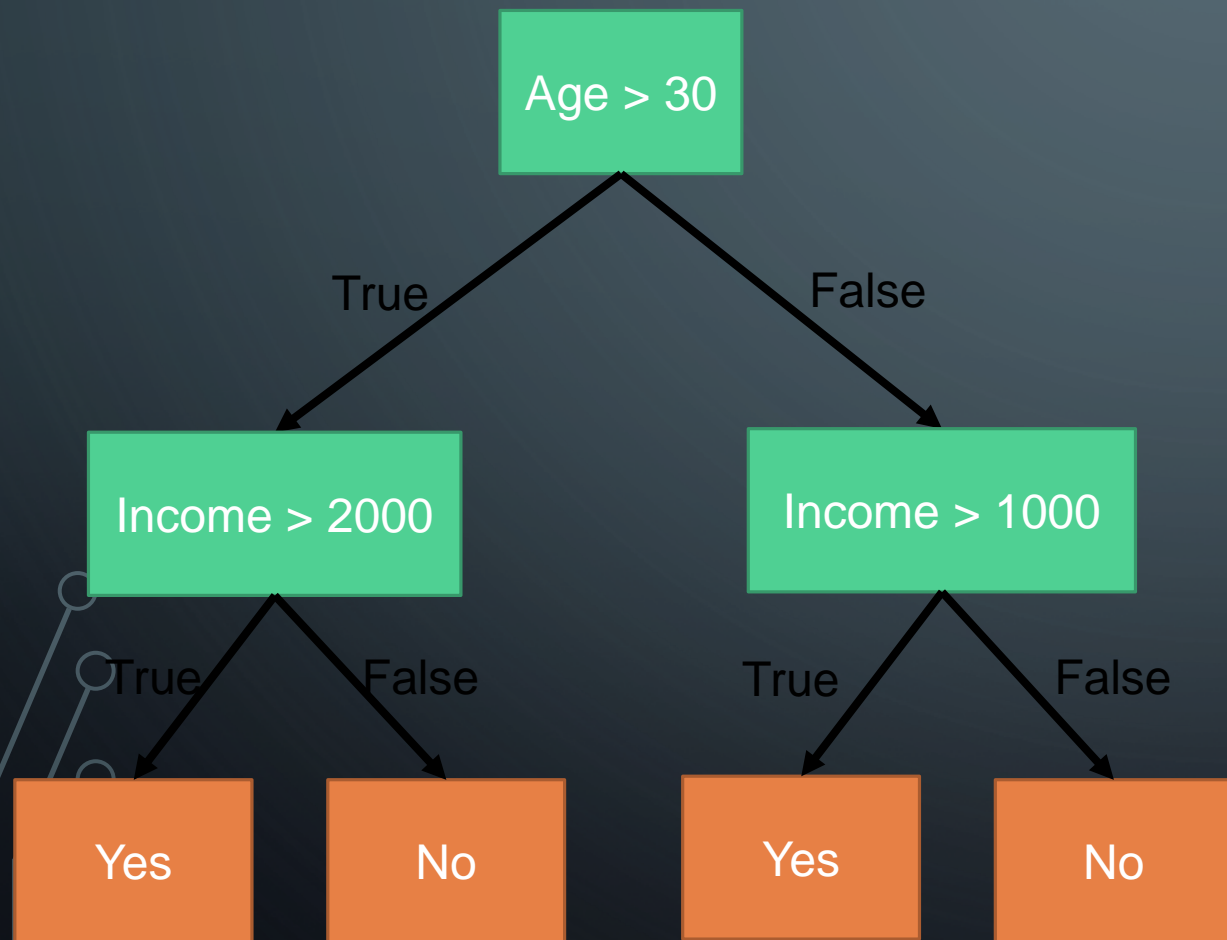
# EXAMPLE - CONTINUOUS



# DECISION TREE – PROS AND CONS

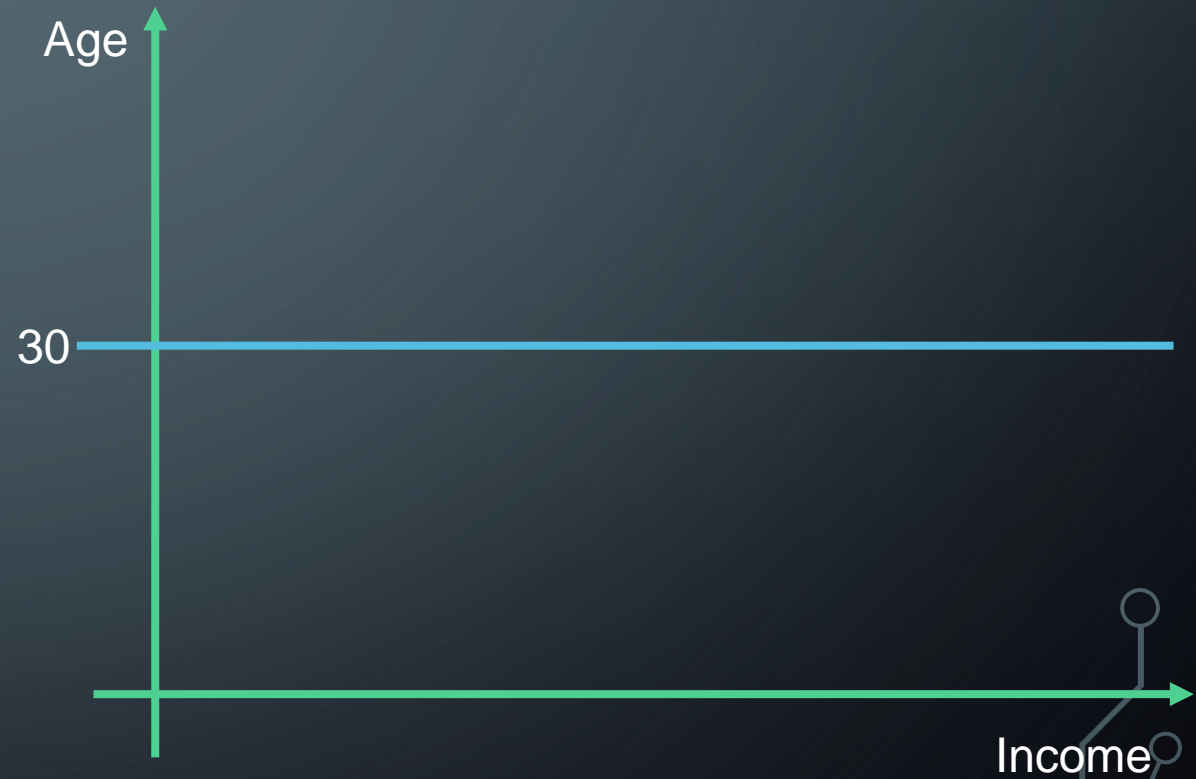
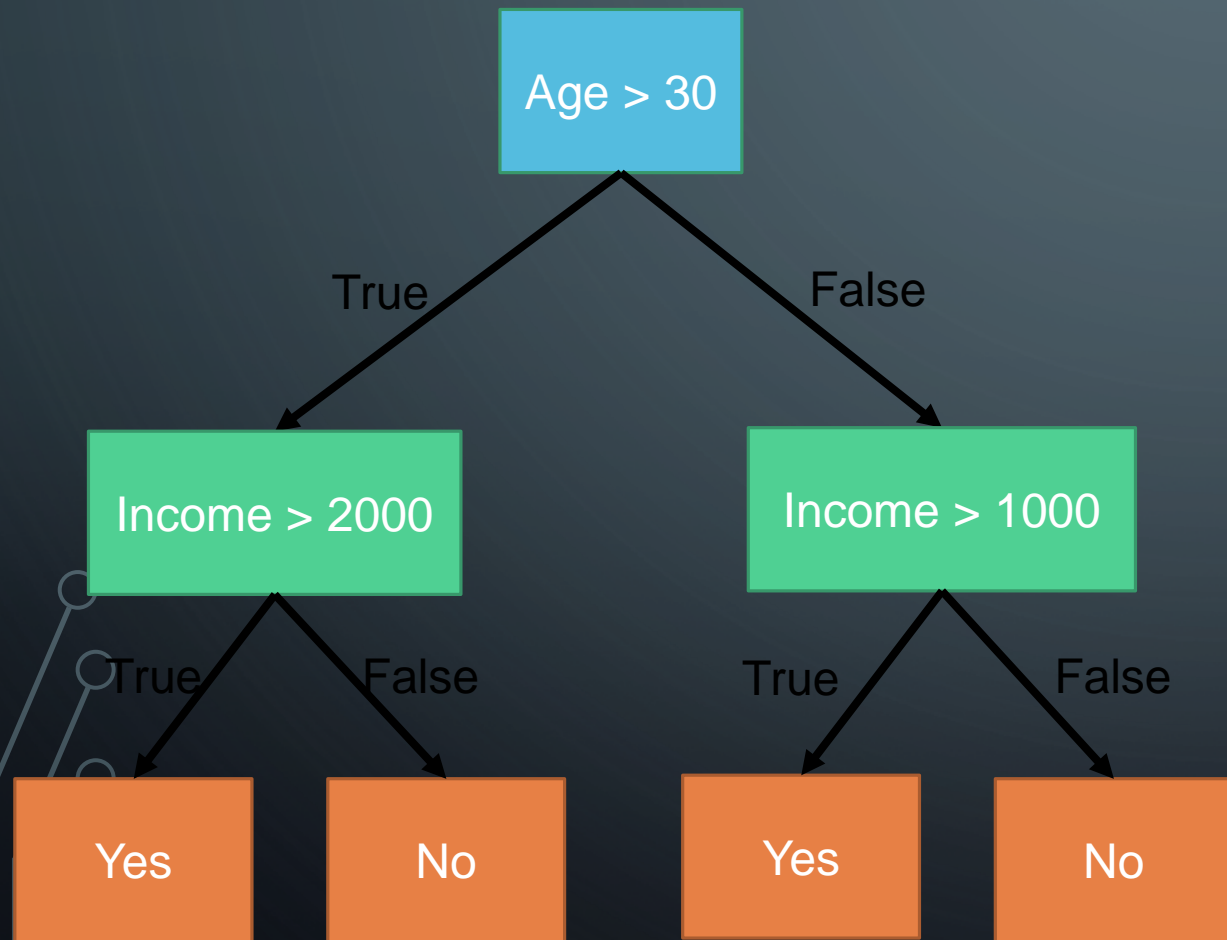
- Pros:
  - Understandable rules
  - Low computation
  - Both continuous and categorical variables
  - Shows more important features
- Cons:
  - Not powerful in regression
  - computationally expensive to train

# DECISION TREE IN 2D SPACE

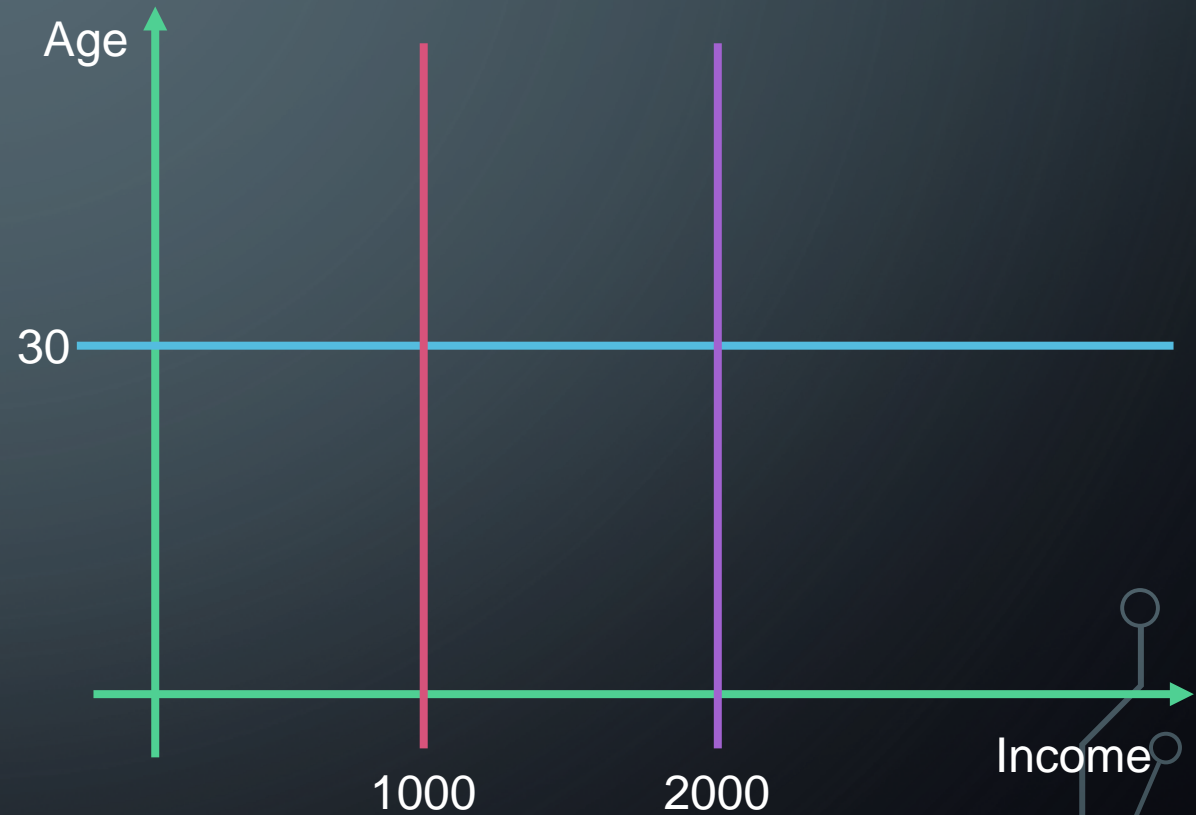
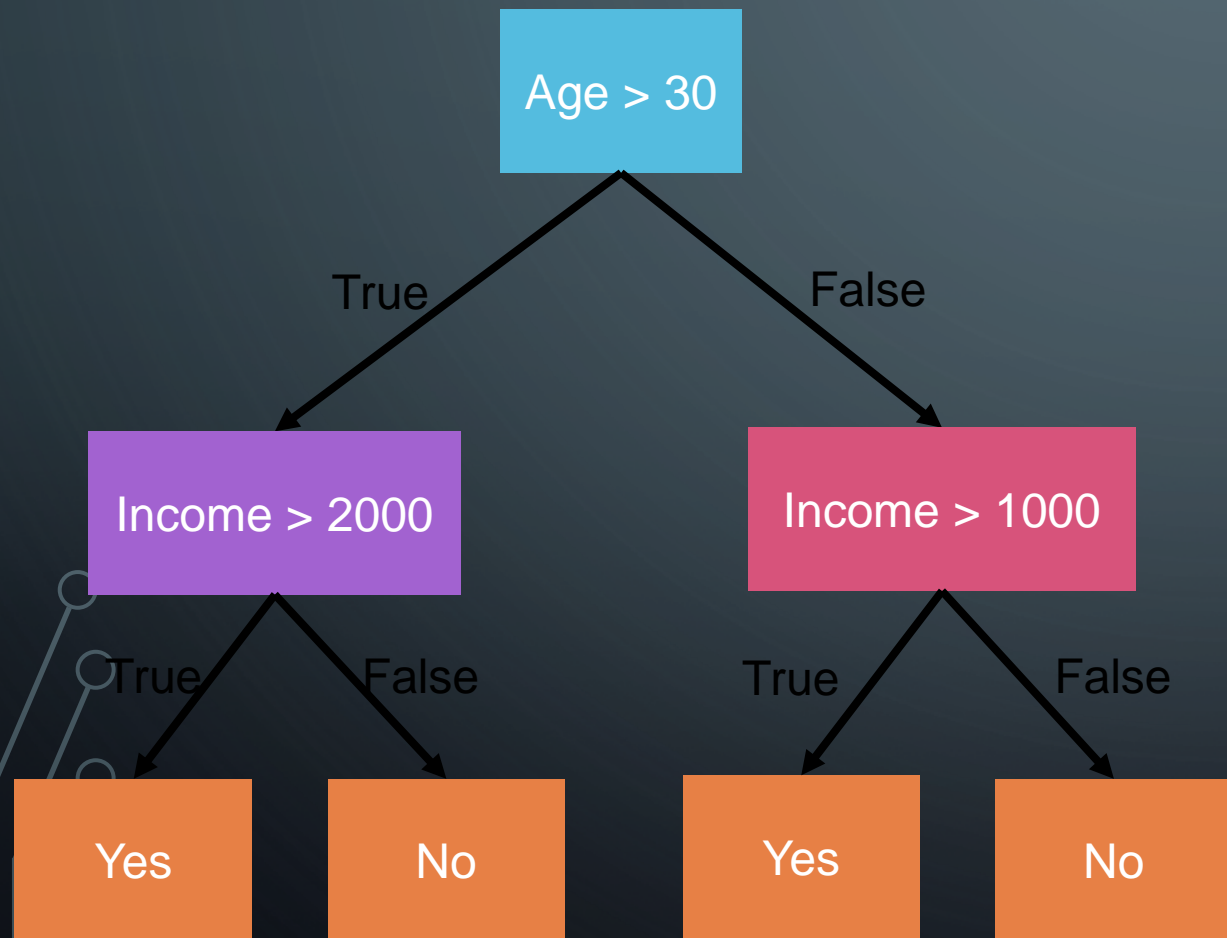




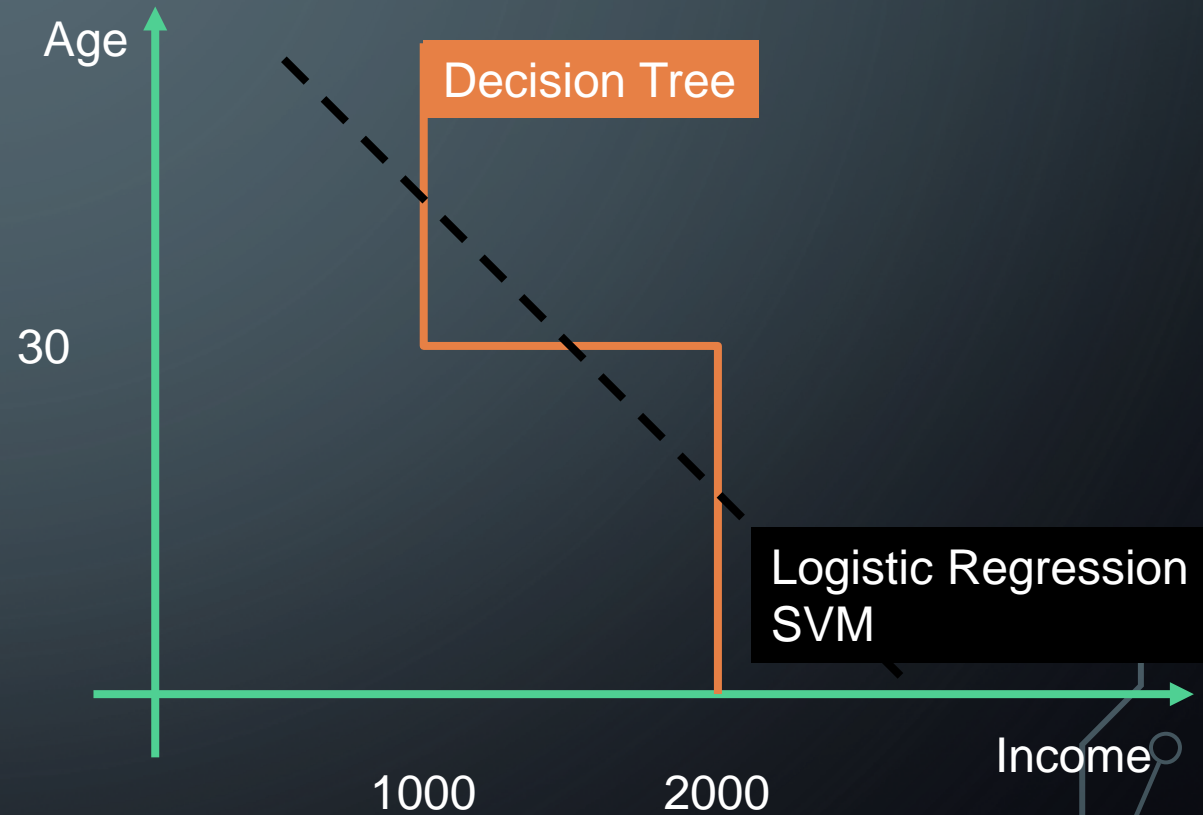
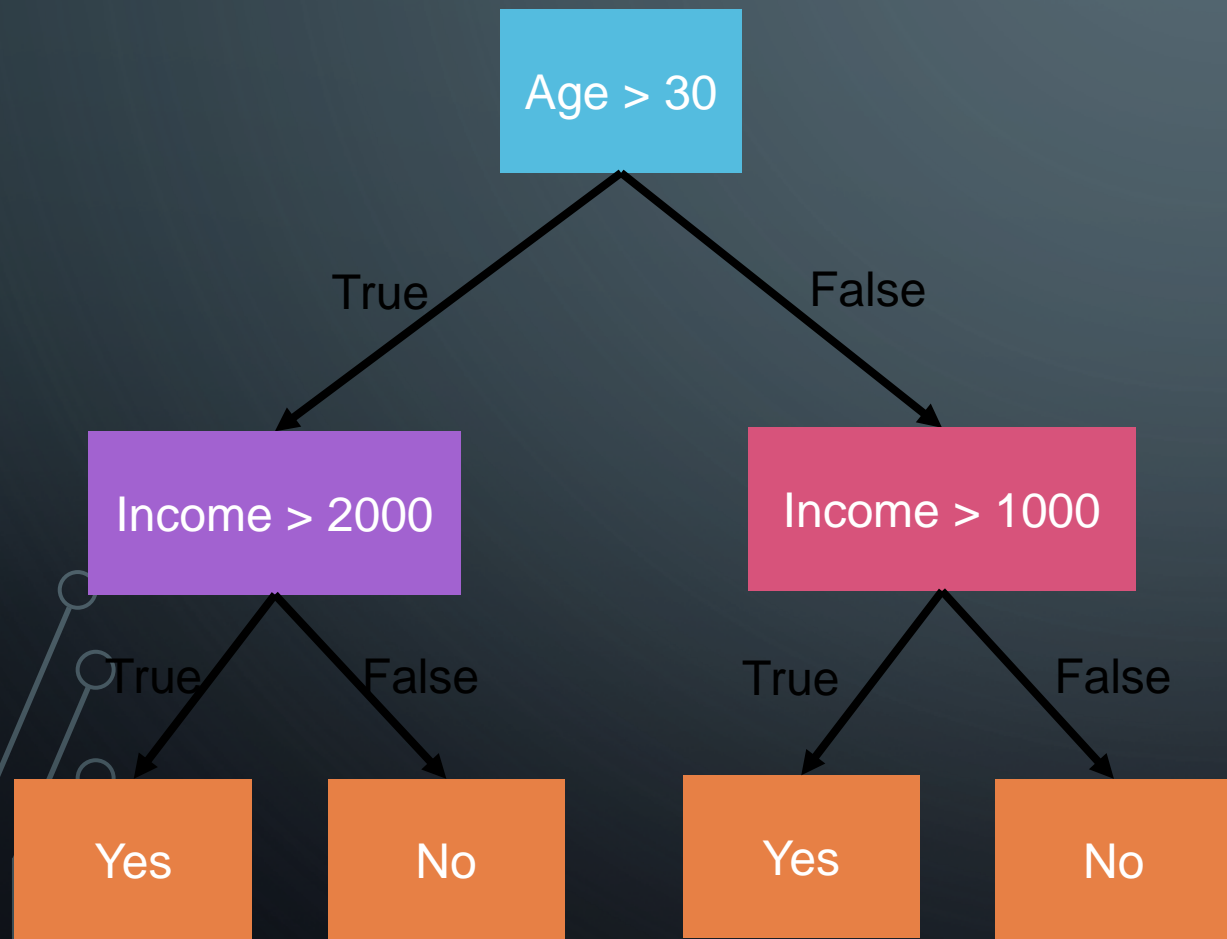
# DECISION TREE IN 2D SPACE



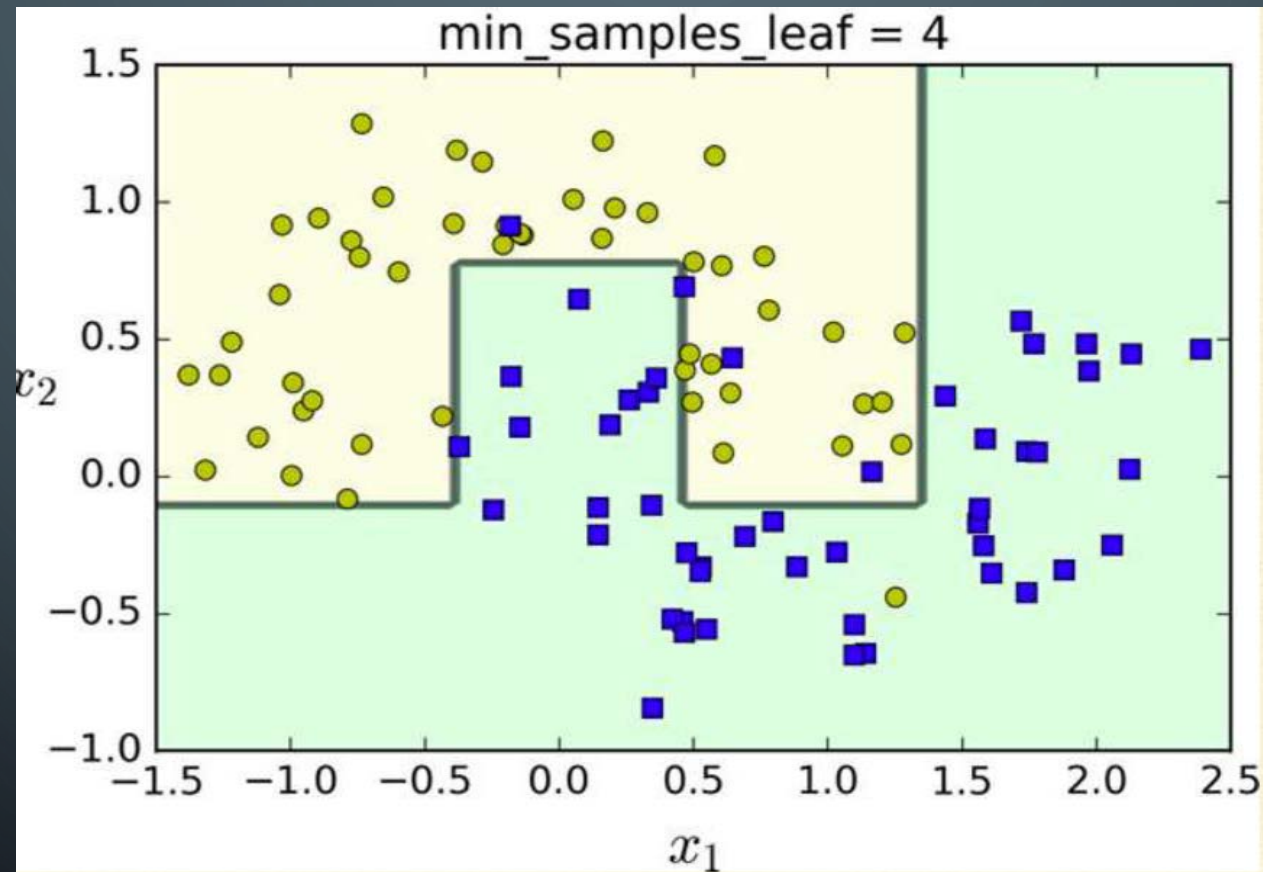
# DECISION TREE IN 2D SPACE



# DECISION TREE IN 2D SPACE

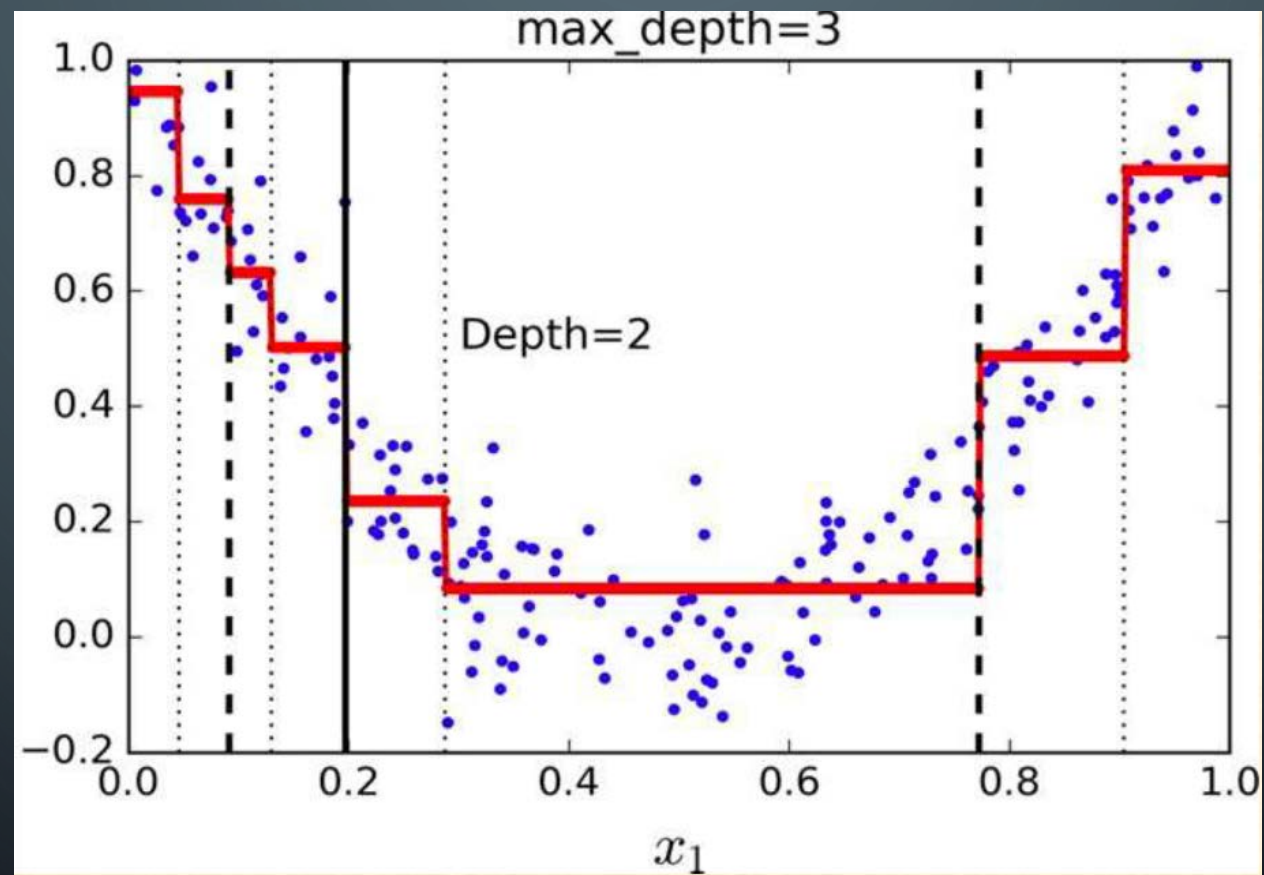


# DECISION TREE FOR CLASSIFICATION





# DECISION TREE FOR REGRESSION





# CART ALGORITHM

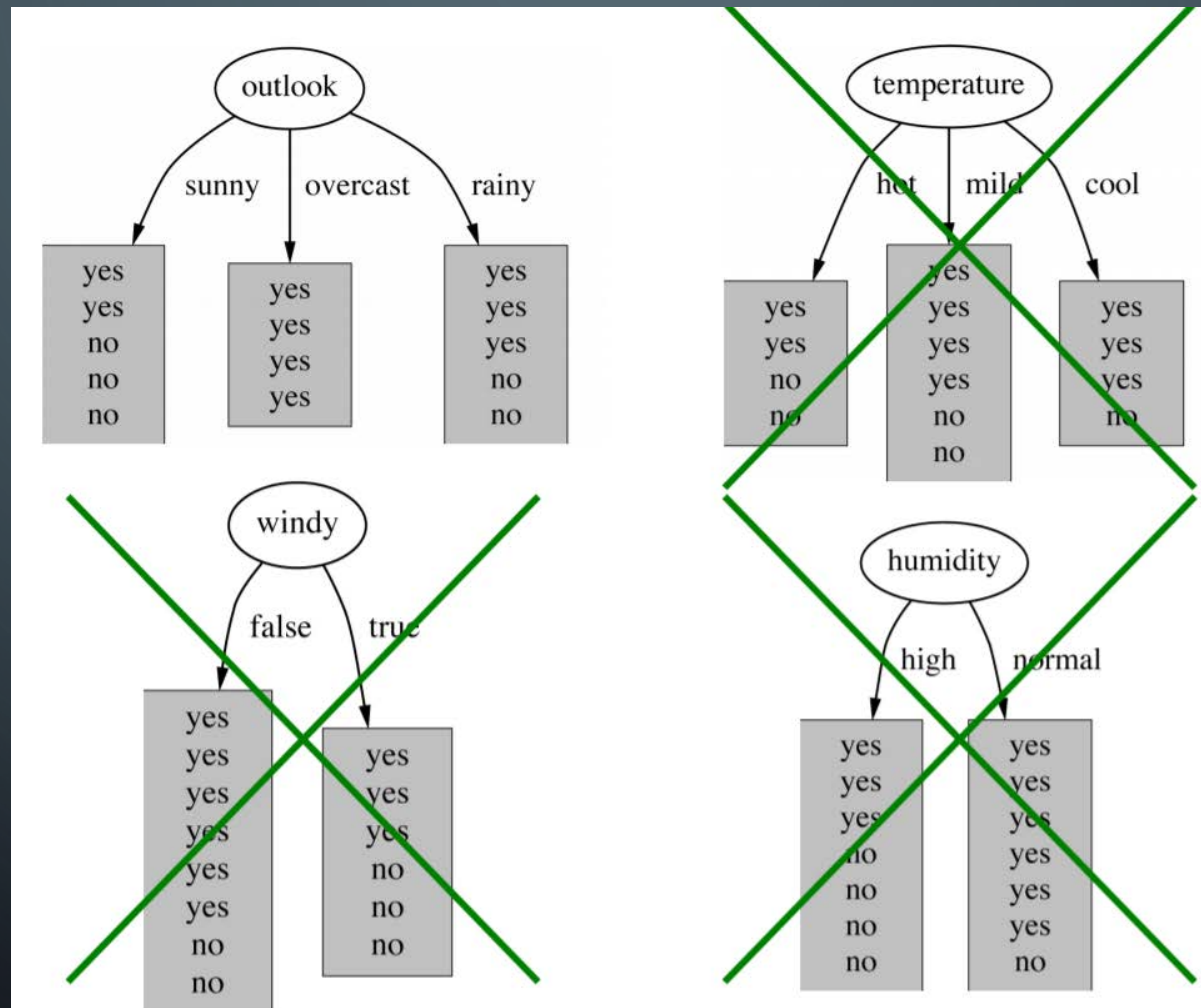
- Split dataset into 2 subsets using a single feature  $k$  and threshold  $t_k$
- Search for best  $(k, t_k)$  that produces the purest subsets.
- Do same algorithm recursively for subsets ...

# PURITY

The diagram illustrates the concept of Purity in decision trees. It shows four decision trees for the 'play' attribute, each with a root node and leaf nodes containing 'yes' or 'no' values.

- outlook:** A decision tree with three branches: sunny, overcast, and rainy. The sunny branch leads to a leaf node with 4 instances (3 yes, 1 no). The overcast branch leads to a leaf node with 4 instances (all yes). The rainy branch leads to a leaf node with 4 instances (3 yes, 1 no).
- temperature:** A decision tree with three branches: hot, mild, and cool. The hot branch leads to a leaf node with 4 instances (3 yes, 1 no). The mild branch leads to a leaf node with 4 instances (3 yes, 1 no). The cool branch leads to a leaf node with 4 instances (3 yes, 1 no).
- windy:** A decision tree with two branches: false and true. The false branch leads to a leaf node with 8 instances (6 yes, 2 no). The true branch leads to a leaf node with 4 instances (3 yes, 1 no).
- humidity:** A decision tree with two branches: high and normal. The high branch leads to a leaf node with 8 instances (6 yes, 2 no). The normal branch leads to a leaf node with 4 instances (3 yes, 1 no).

The trees for **temperature**, **windy**, and **humidity** are crossed out with a large red X, indicating they are impure. The tree for **outlook** is not crossed out, indicating it is pure.



# PURITY

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

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

$$\text{Information} = I(S, A) = \sum_i \frac{|S_i|}{S} \cdot E(S_i)$$

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

$$\text{Gini: } G(S, A) = \sum_i \frac{|S_i|}{S} \cdot G(S_i)$$



# 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



The Jupyter logo is centered in the image. It consists of two orange, curved, crescent-like shapes that form a circle around the word "jupyter". The word "jupyter" is written in a white, lowercase, sans-serif font. There are four white circles of varying sizes positioned around the logo: one at the top left, one at the top right, one at the bottom left, and one at the bottom right. The background is a dark blue gradient. In the corners, there are faint, light blue circuit-like patterns with lines and small circles.

jupyter