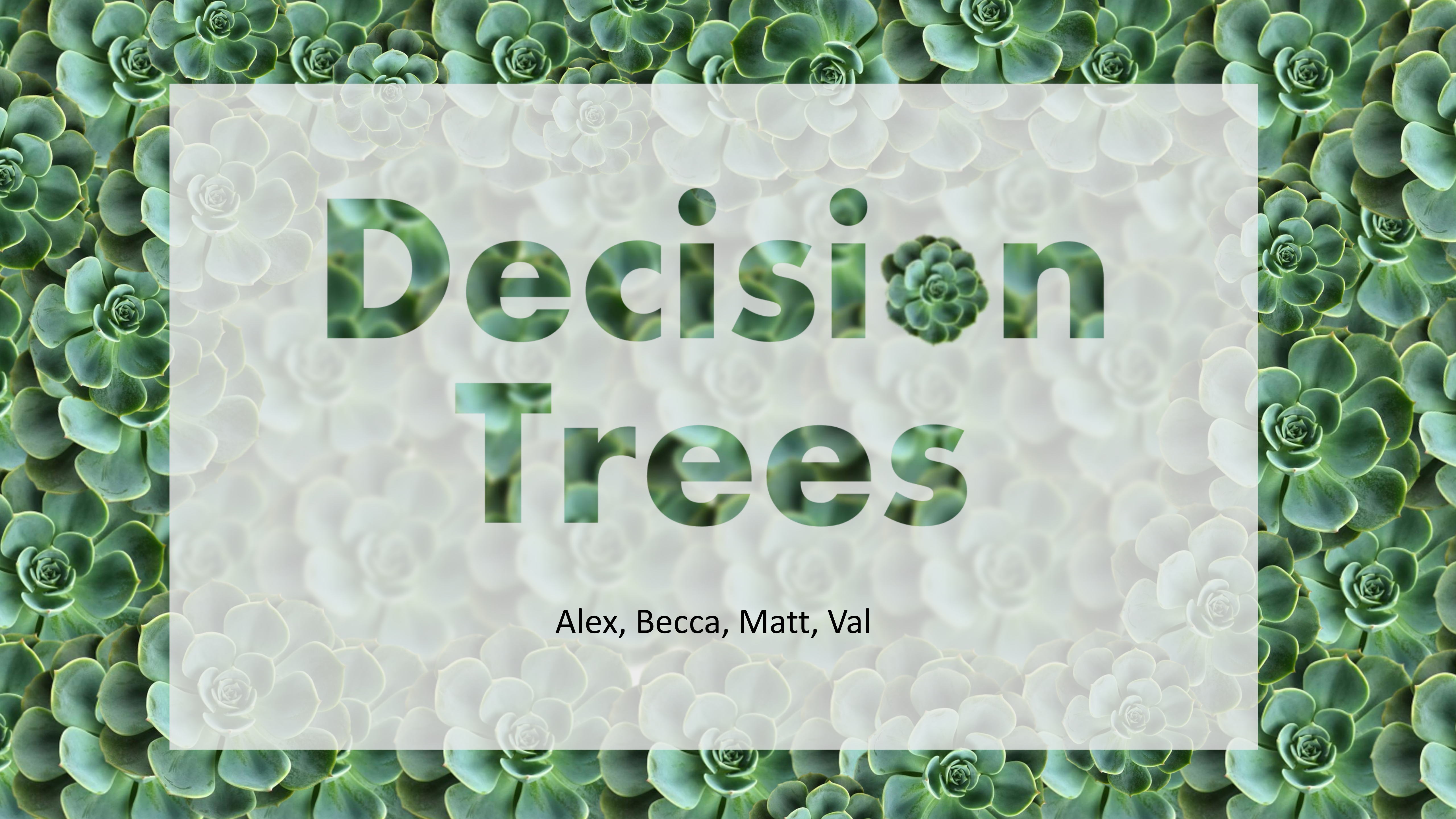


Decision Trees

Alex, Becca, Matt, Val

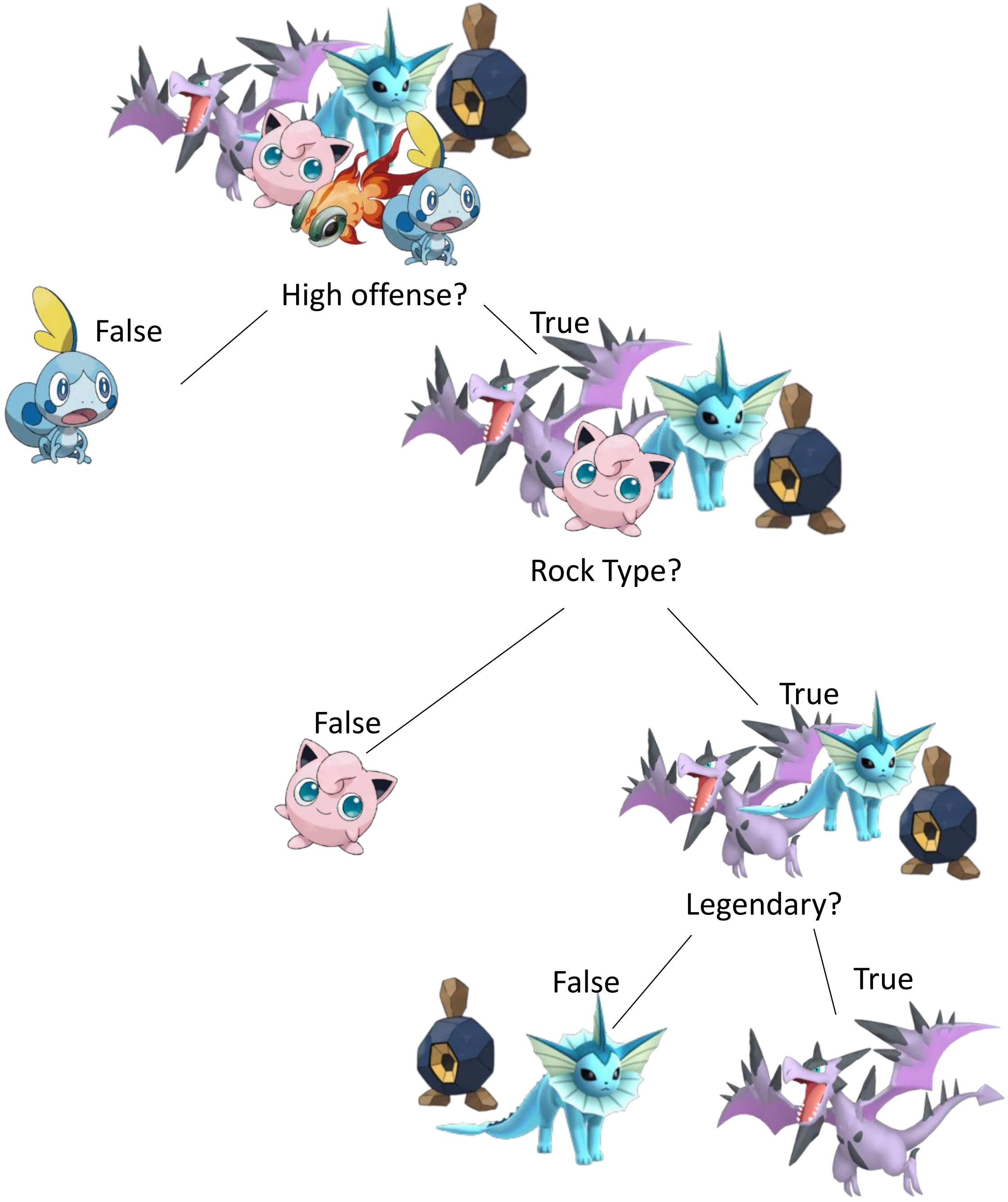
Overview

Decision Trees

- Nonparametric supervised learning algorithm
- Best for decision making and prediction
- Bases decisions on how a previous set of questions were answered

Example

- Ash wants to choose one of his Pokechu to battle a Charizard:
    - High offense
    - Rock
    - Legendary

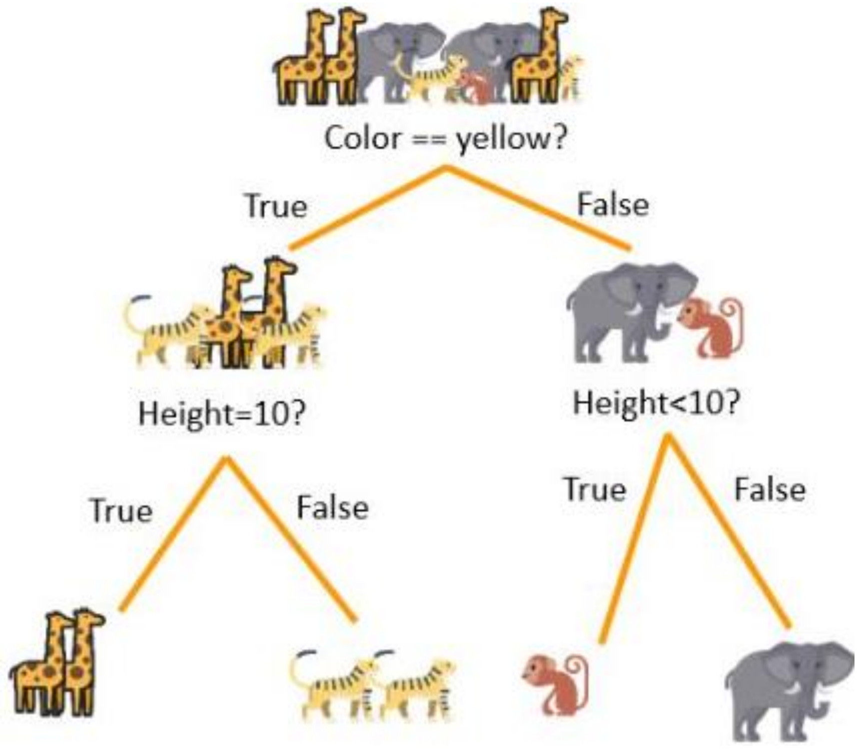


Terms to Know

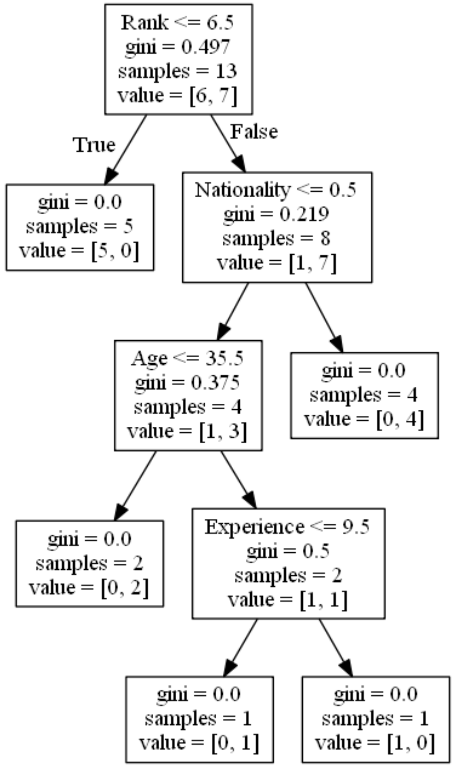
- Tree
  - Hierarchical structure mapping the possible outcomes of choices
- Nodes
  - Root node: the starting point of the tree
  - Decision node: Point where a decision must be made
  - End node: final outcomes of a decision path
- Entropy
  - Measure of disorder/randomness in a data set
- Information gain
  - Measure of how much information a feature provides about a class
    - Used to decide whether a feature should be used to split a node or not
- Greedy algorithm
  - Finds optimal solution at local level, but without “big picture” view

Types

Categorical



Regression


Rank <= 6.5
gini = 0.497
samples = 13
value = [6, 7]

True

False

gini = 0.0
samples = 5
value = [5, 0]

Nationality <= 0.5
gini = 0.219
samples = 8
value = [1, 7]

Age <= 35.5
gini = 0.375
samples = 4
value = [1, 3]

gini = 0.0
samples = 4
value = [0, 4]

gini = 0.0
samples = 2
value = [0, 2]

Experience <= 9.5
gini = 0.5
samples = 2
value = [1, 1]

gini = 0.0
samples = 1
value = [0, 1]

gini = 0.0
samples = 1
value = [1, 0]

Test your Knowledge

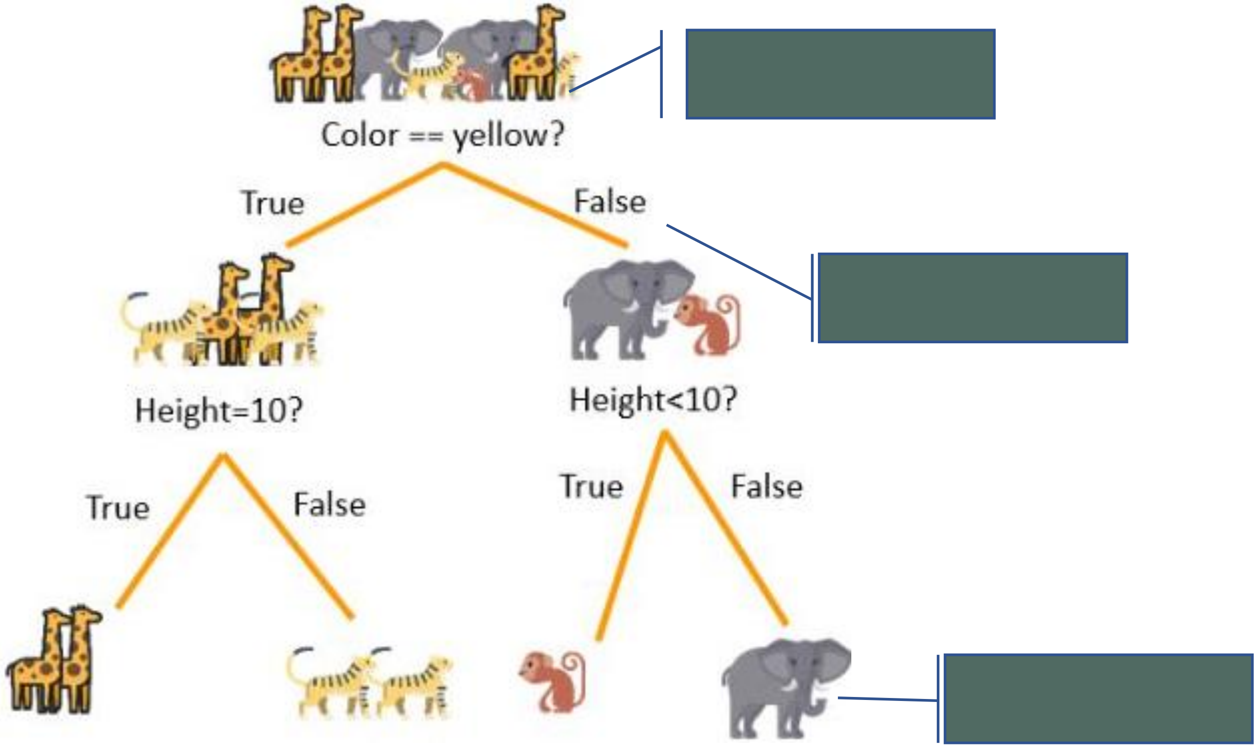
Tree

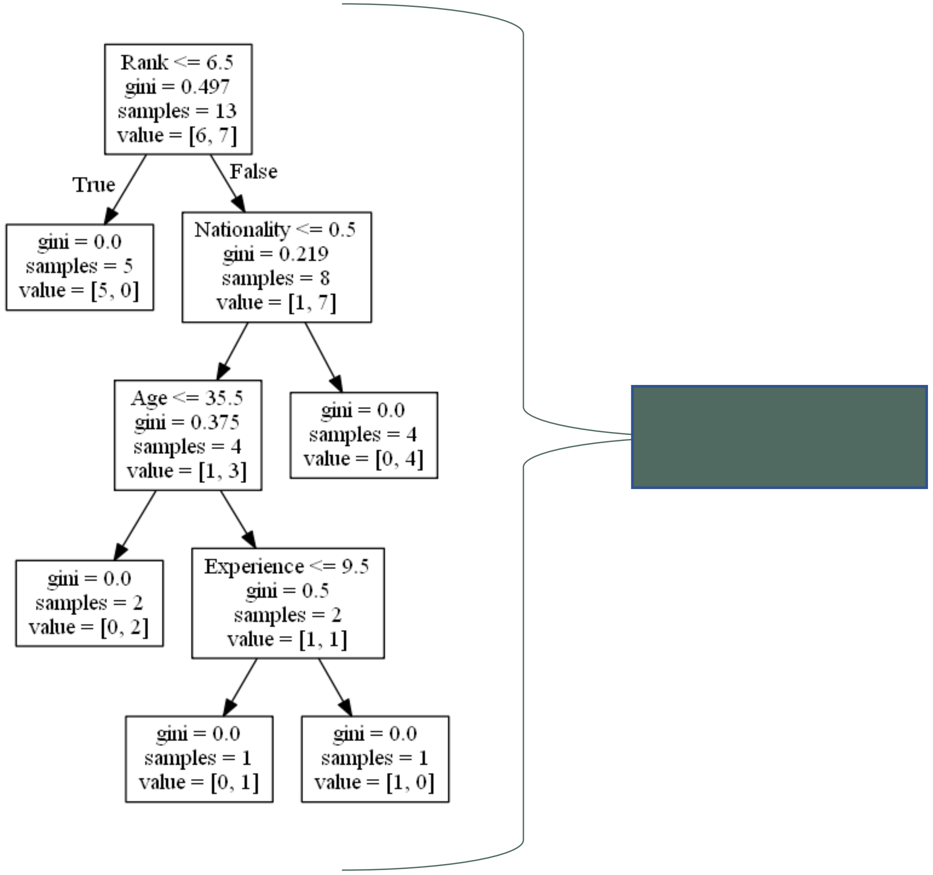
Root node

Decision node

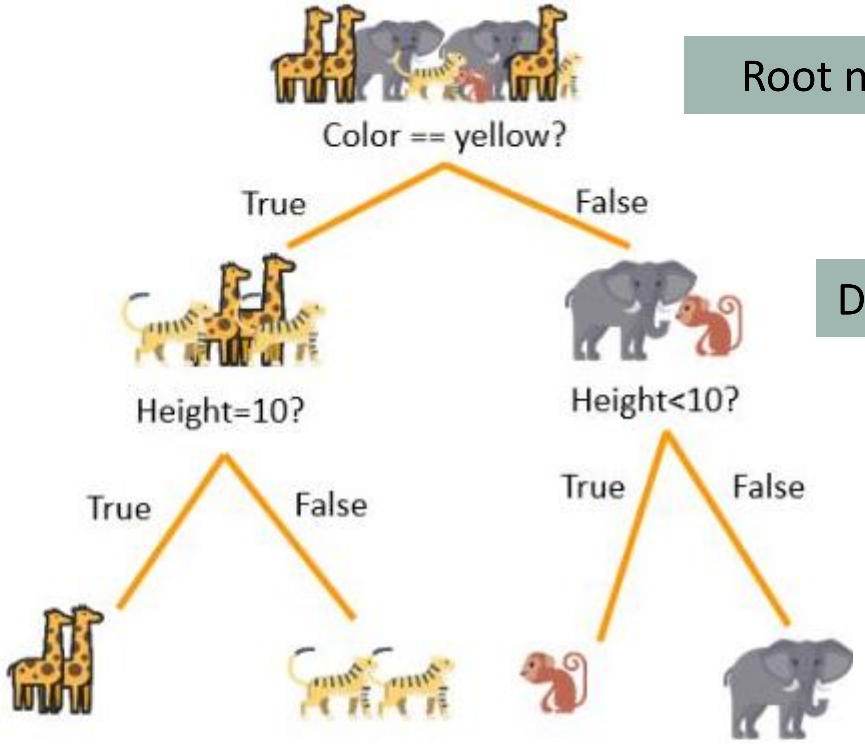
End node

Entropy





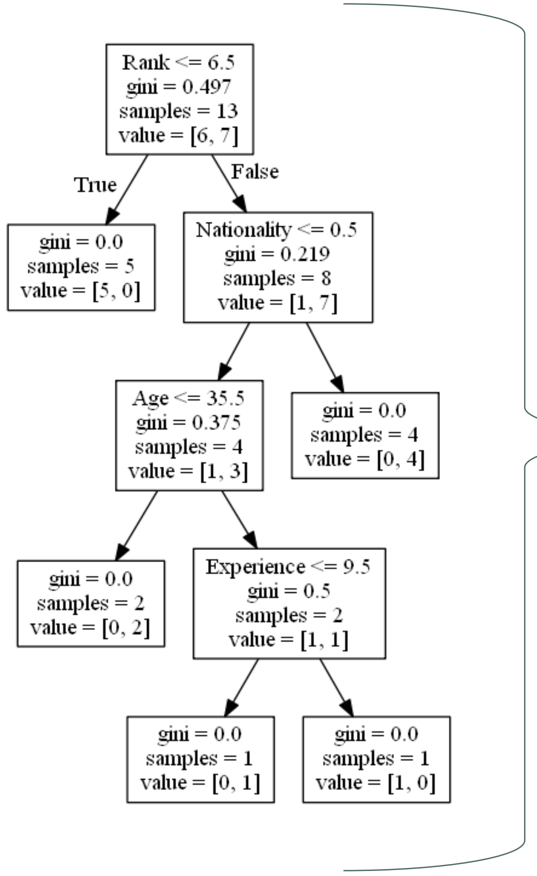
Answers



Root node

Decision node

End node



Tree

Advantages and Disadvantages

- Advantages
  - High accuracy
  - Less up-front effort
    - Normalization not necessary
  - Easy to depict and explain
  - Flexible model, no prerequisites

- Disadvantages
  - Does not support missing values (NaN)
  - Can have high variance
  - Takes more processing power/time
  - Inadequate for continuous variables

Hyperparameters

- Maximum depth: determines the maximum number of levels in the decision tree
- Minimum samples split: sets the minimum number of samples required to split an internal node
- Minimum samples leaf: sets the minimum number of samples required to be at a leaf node
- Maximum features: sets the maximum number of features considered for splitting a node
- Criterion: determines the metric used for evaluating the quality of a split (e.g. gini impurity, entropy)
- Splitter: determines the strategy used for splitting at each node (e.g. best, random)
- Class weight: determines the weights of classes in case of imbalanced data
- Random state: sets the random seed for reproducibility of results.

Resources

- https://www.geeksforgeeks.org/decision-tree/
- Video Tutorial
- Documentation (Scikit Learn)

Further Reading

- Fisher Yates Algorithm (article)
- Simpson Index (scientific paper)

Questions?