Decision Trees

~Abhishek Kumar

Decision Tree

- Supervised
- Classification algorithm

Example

- Predict if John will play tennis

- Training set:

  9 days = played (Yes)

  5 days = didn't play (No)

- Test

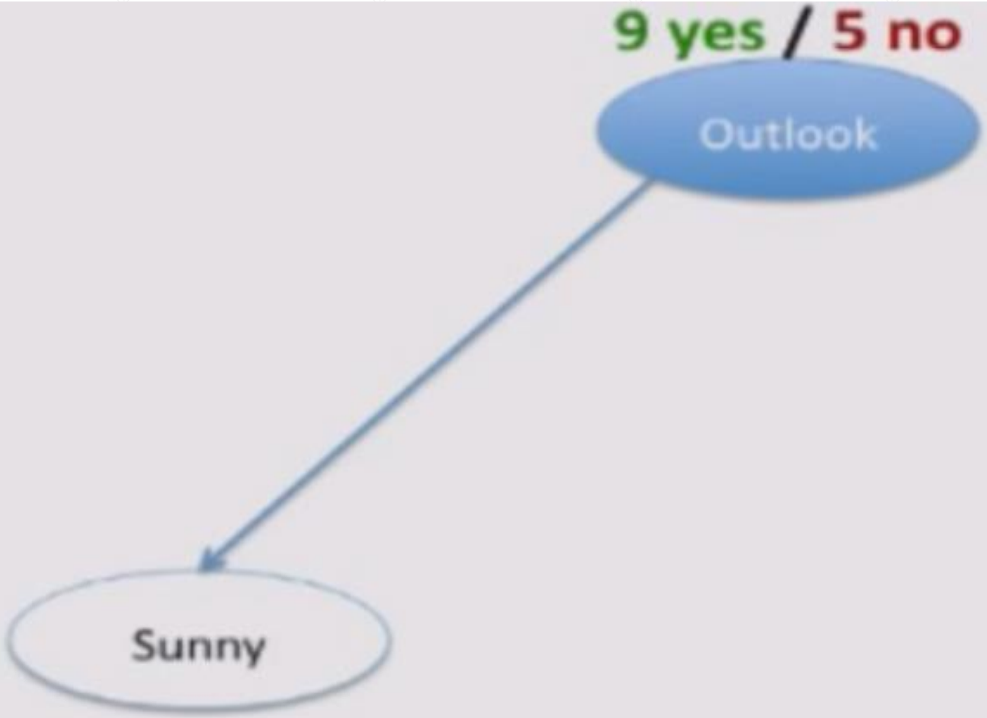
  D15, Rain, High, Weak, ?

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Idea

- Split the dataset into subset
- Check if they are pure
- If yes = Stop
- else keep splitting

- And check which subset our test data point fall into.



Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

9 yes / 5 no

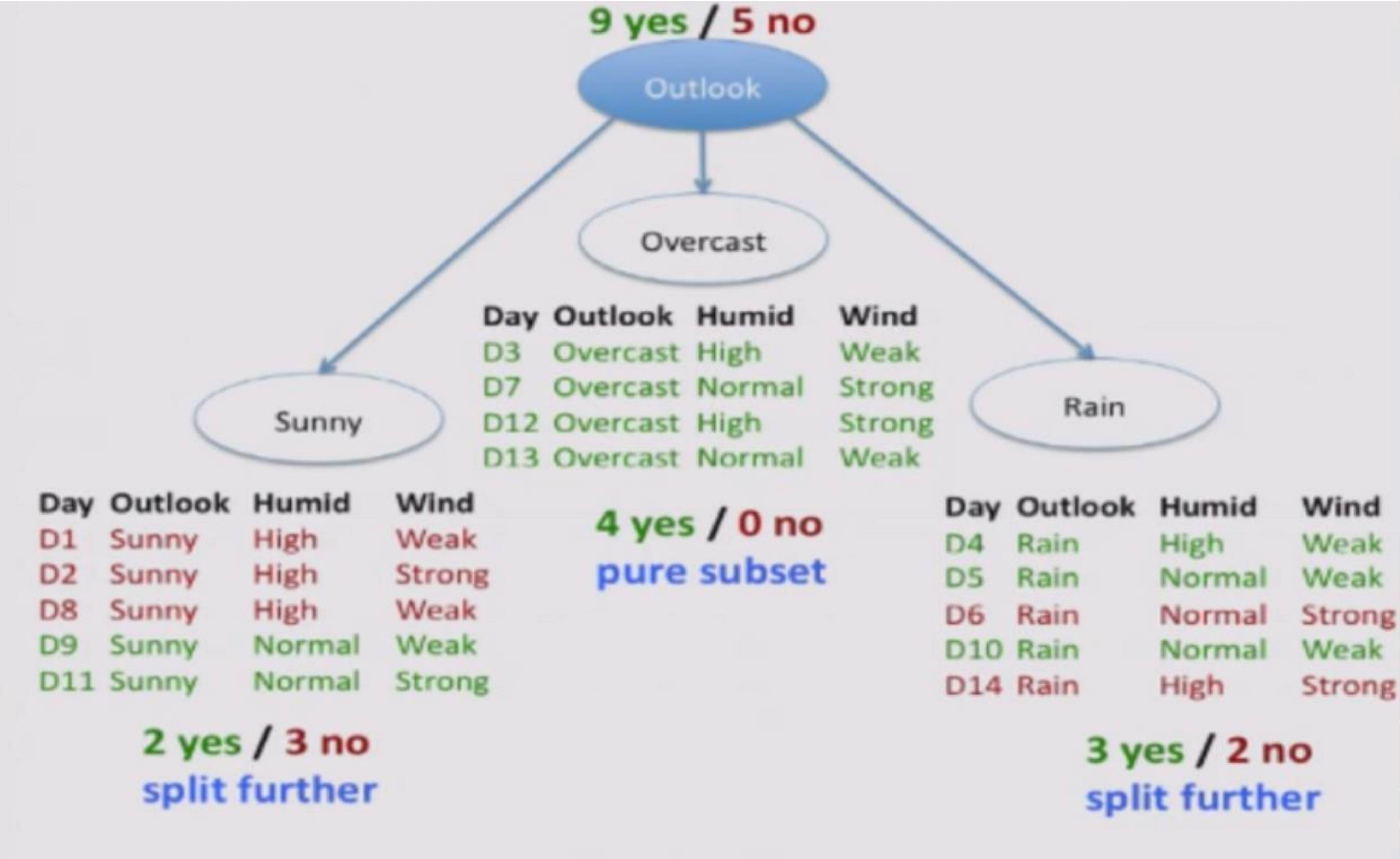
Outlook

Overcast

Day	Outlook	Humid	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

Sunny

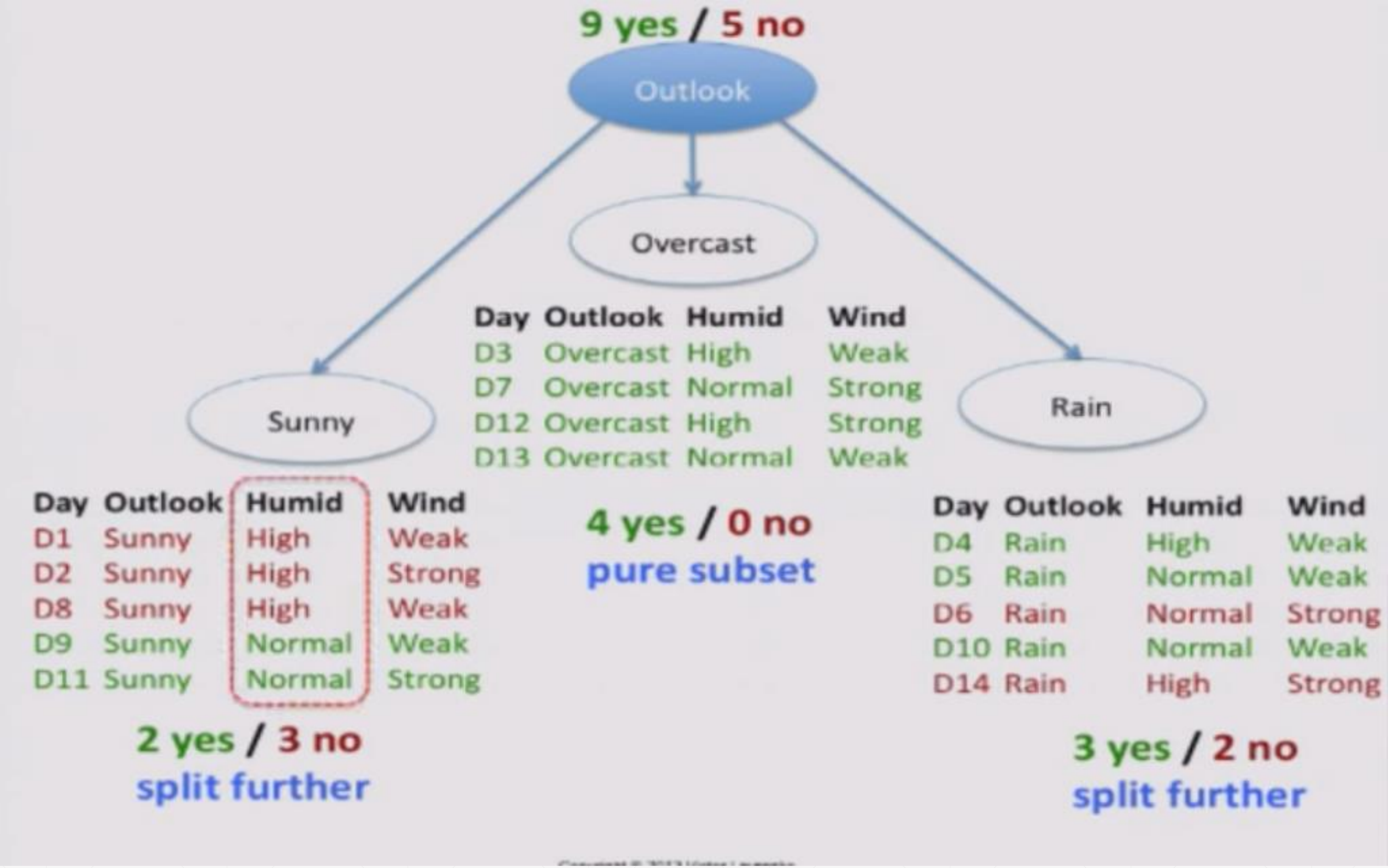
Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

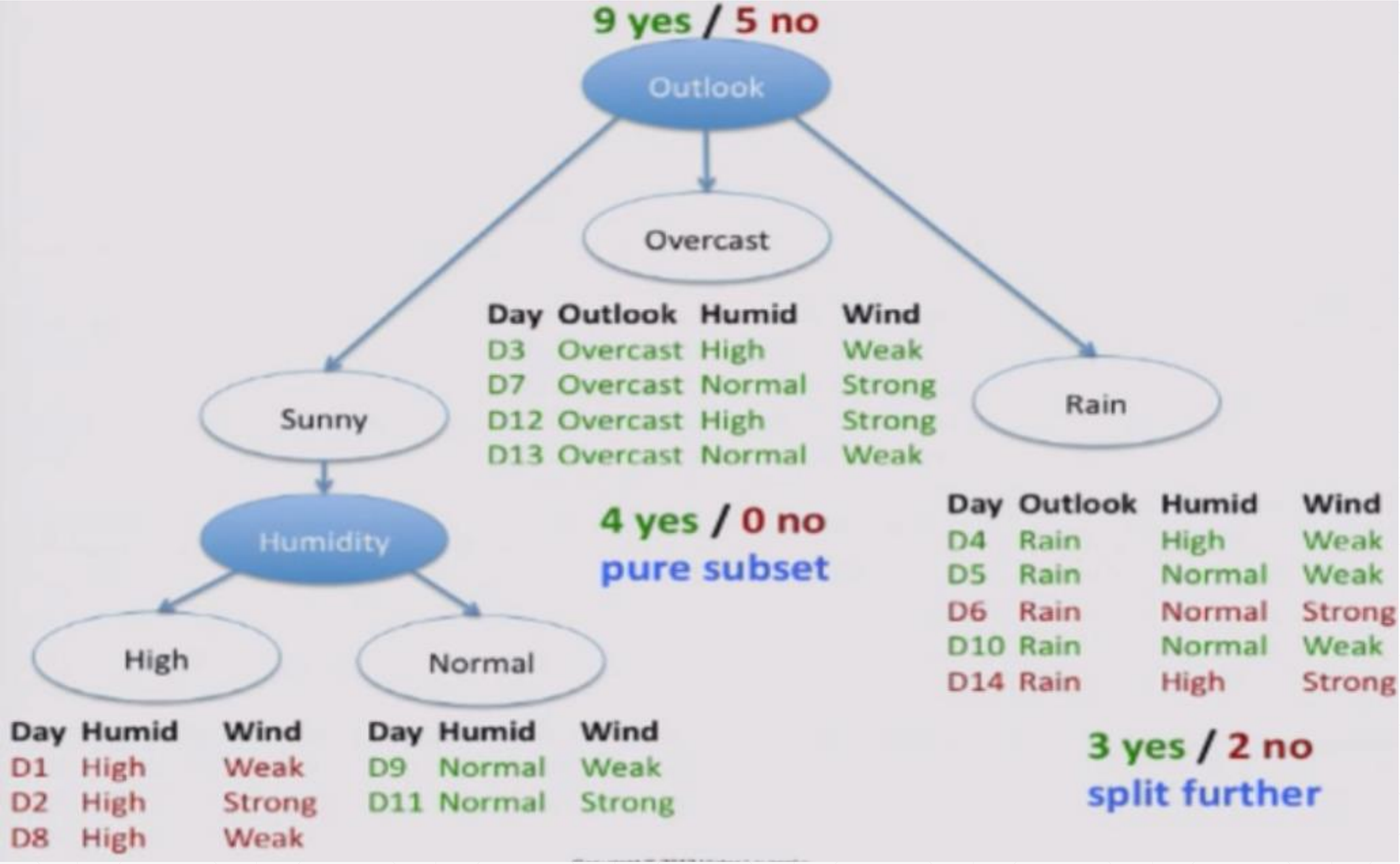


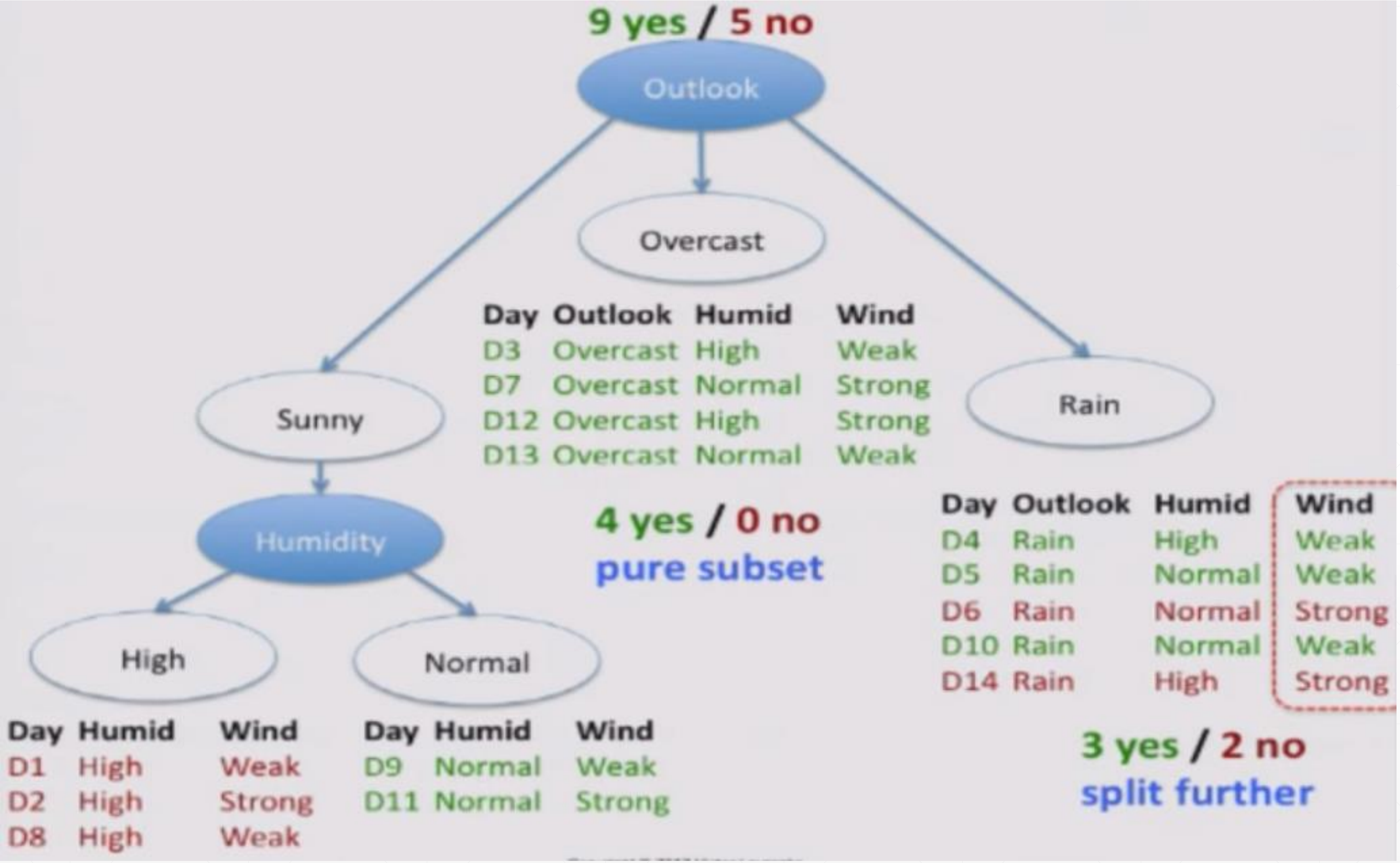
Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

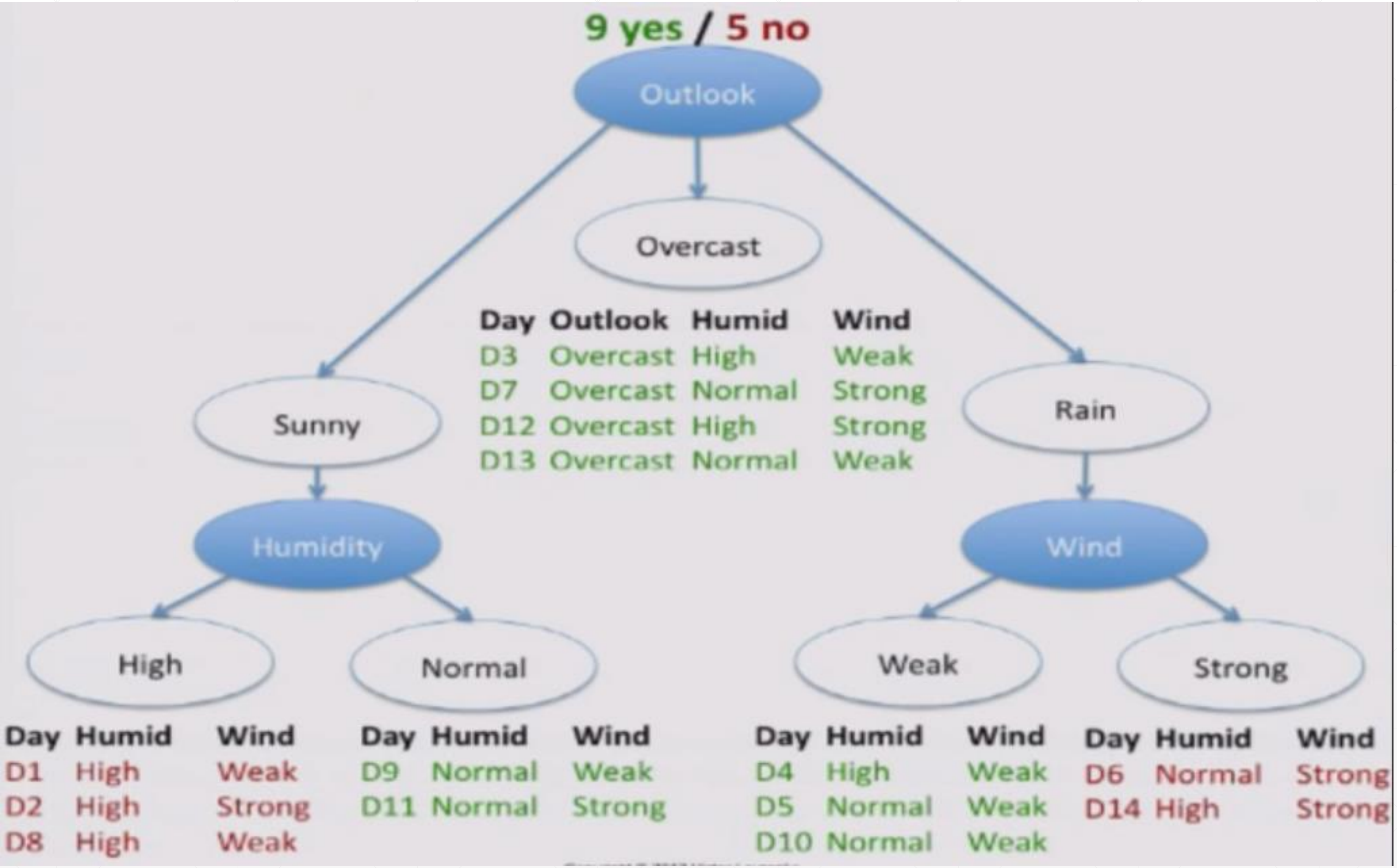
Day	Outlook	Humid	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

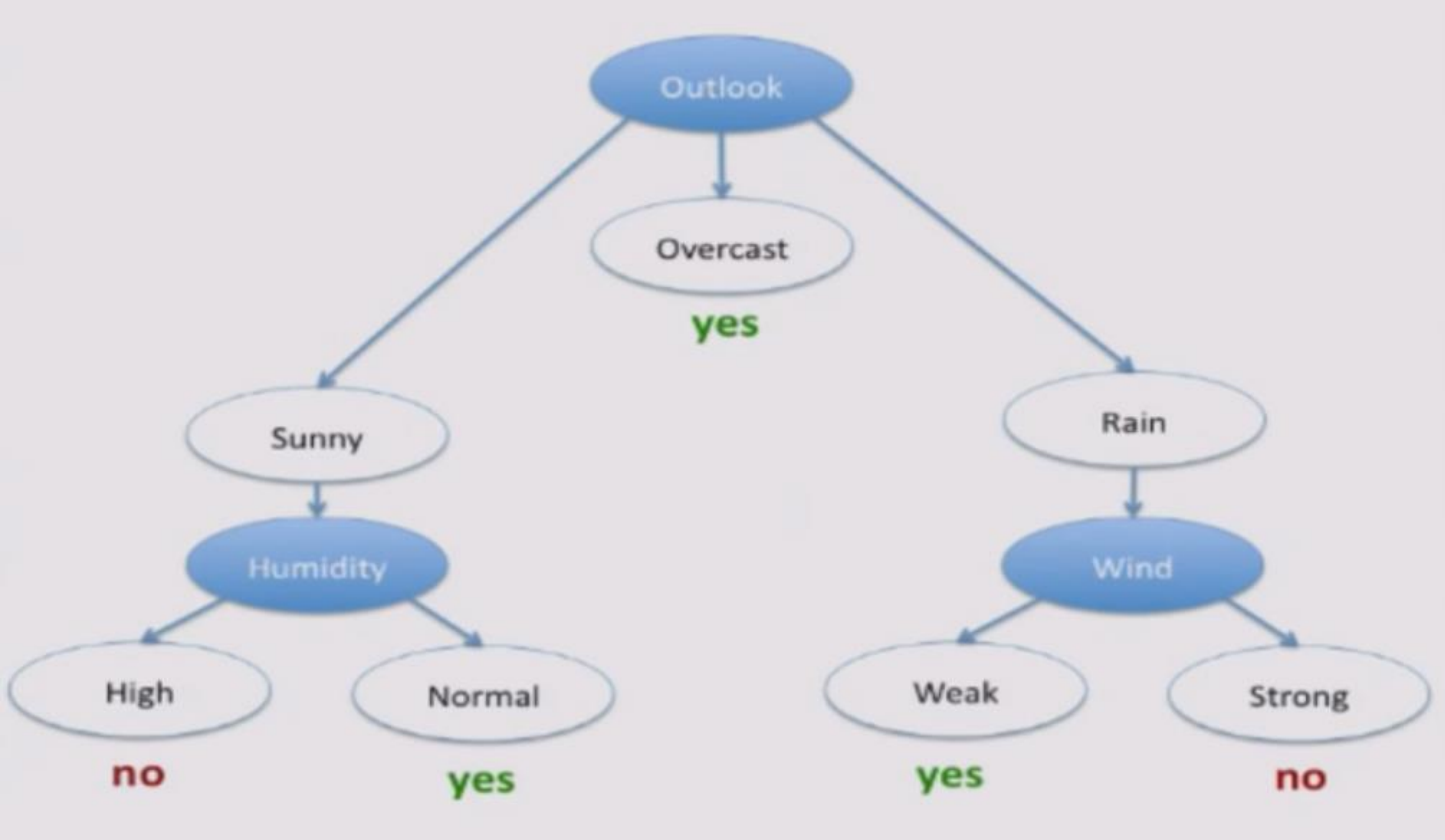
Day	Outlook	Humid	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

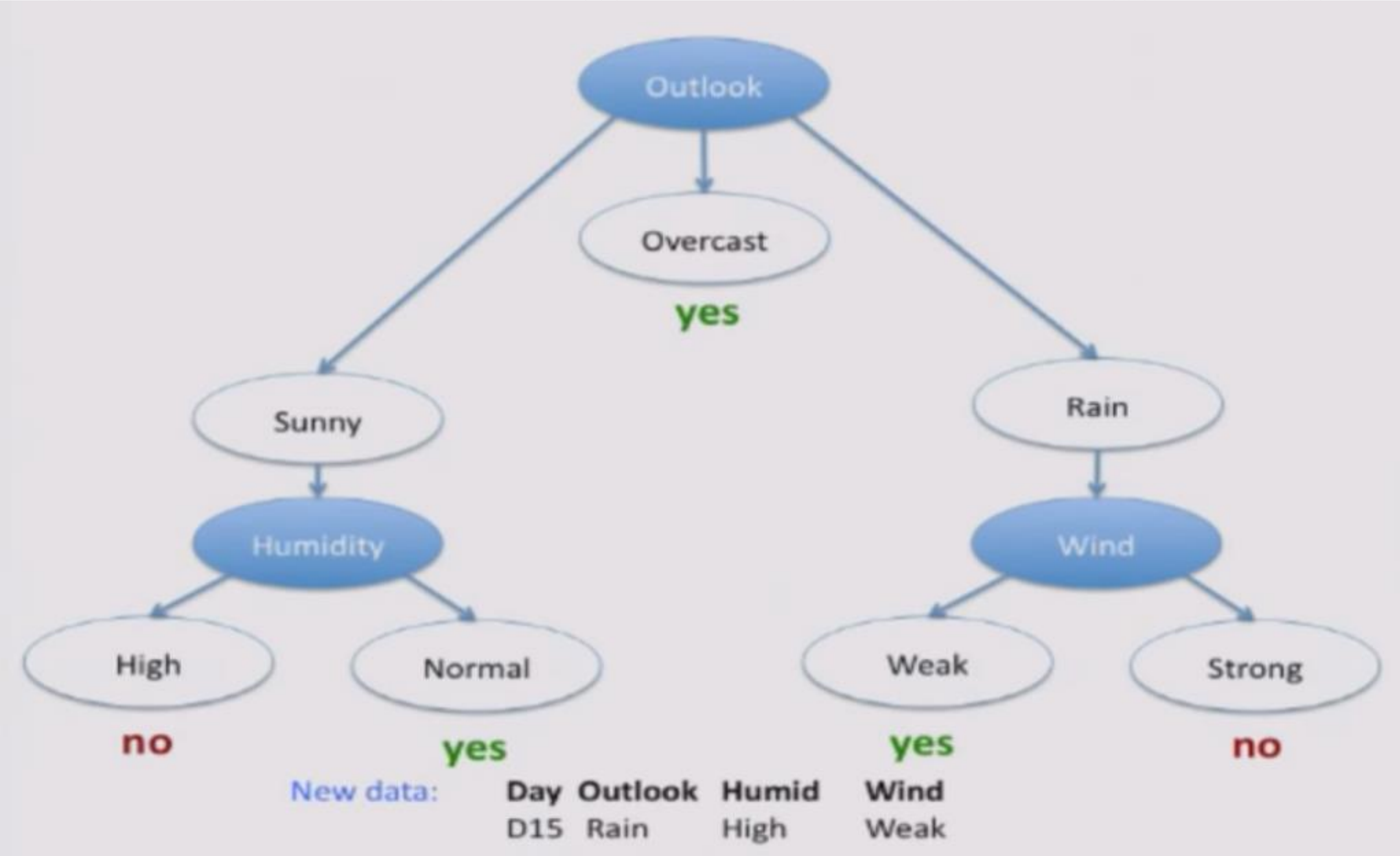

9 yes / 5 no
Outlook
Overcast
Sunny
Rain
Day Outlook Humid Wind
D3 Overcast High Weak
D7 Overcast Normal Strong
D12 Overcast High Strong
D13 Overcast Normal Weak
4 yes / 0 no
pure subset
Day Outlook Humid Wind
D1 Sunny High Weak
D2 Sunny High Strong
D8 Sunny High Weak
D9 Sunny Normal Weak
D11 Sunny Normal Strong
2 yes / 3 no
split further
Day Outlook Humid Wind
D4 Rain High Weak
D5 Rain Normal Weak
D6 Rain Normal Strong
D10 Rain Normal Weak
D14 Rain High Strong
3 yes / 2 no
split further

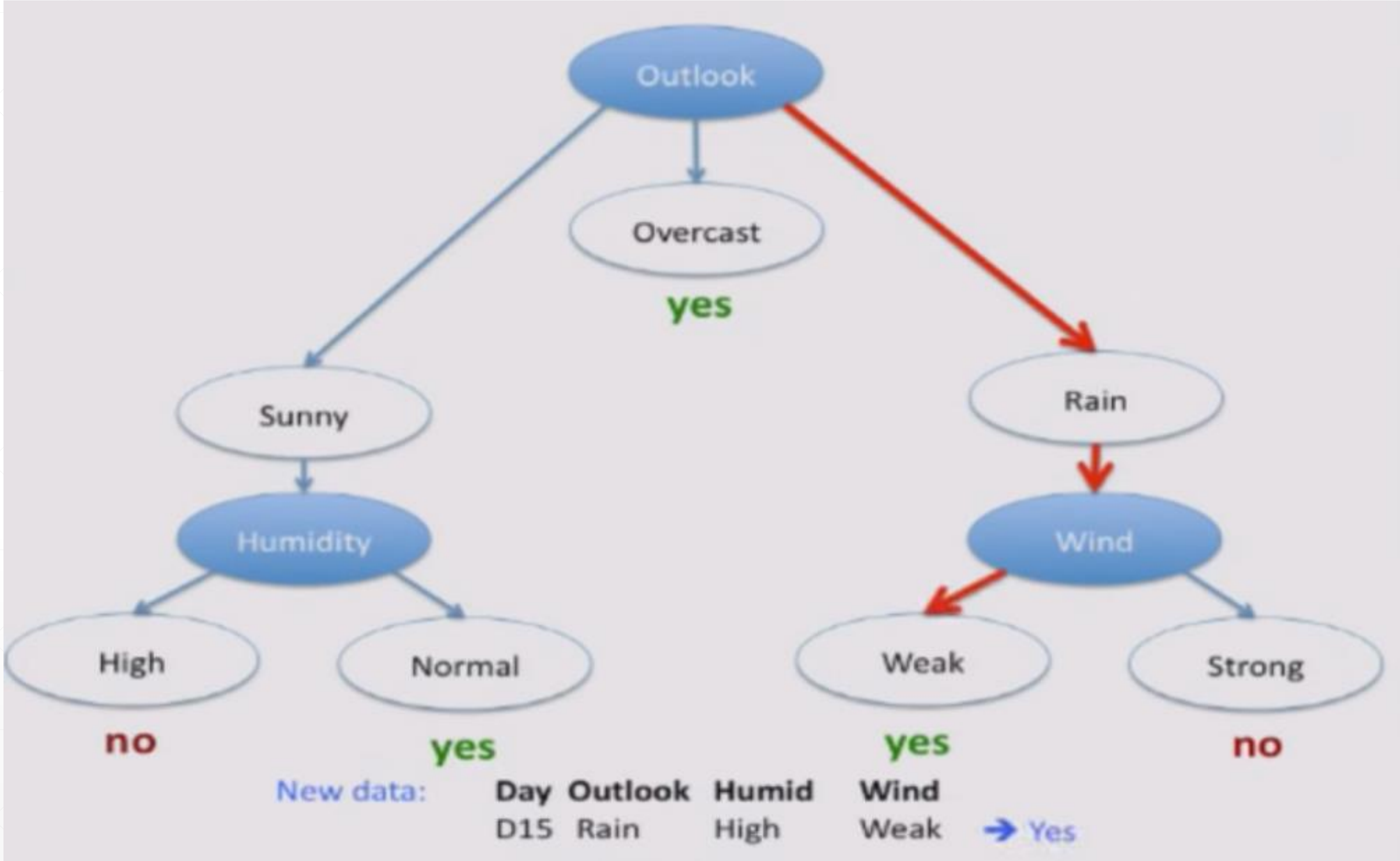

9 yes / 5 no
Outlook
Overcast
Sunny
Rain
Day Outlook Humid Wind
D3 Overcast High Weak
D7 Overcast Normal Strong
D12 Overcast High Strong
D13 Overcast Normal Weak
4 yes / 0 no
pure subset
Humidity
High
Normal
Day Humid Wind
D1 High Weak
D2 High Strong
D8 High Weak
Day Humid Wind
D9 Normal Weak
D11 Normal Strong
Day Outlook Humid Wind
D4 Rain High Weak
D5 Rain Normal Weak
D6 Rain Normal Strong
D10 Rain Normal Weak
D14 Rain High Strong
3 yes / 2 no
split further

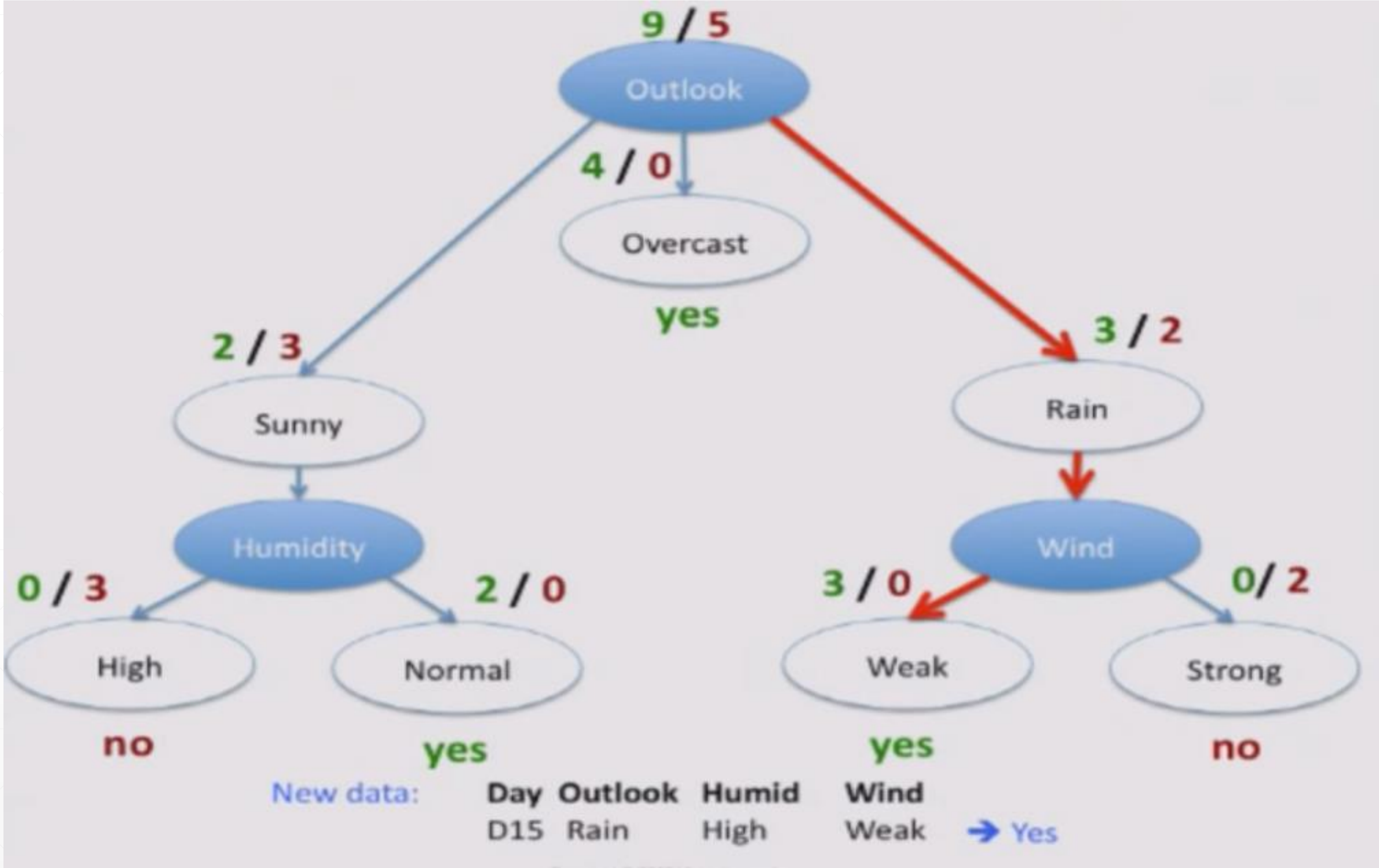

9 yes / 5 no
Outlook
Overcast
Sunny
Rain
Day Outlook Humid Wind
D3 Overcast High Weak
D7 Overcast Normal Strong
D12 Overcast High Strong
D13 Overcast Normal Weak
4 yes / 0 no
pure subset
Humidity
High
Normal
Day Outlook Humid Wind
D4 Rain High Weak
D5 Rain Normal Weak
D6 Rain Normal Strong
D10 Rain Normal Weak
D14 Rain High Strong
3 yes / 2 no
split further
Day Humid Wind
D1 High Weak
D2 High Strong
D8 High Weak
Day Humid Wind
D9 Normal Weak
D11 Normal Strong


9 yes / 5 no
Outlook
Overcast
Day Outlook Humid Wind
D3 Overcast High Weak
D7 Overcast Normal Strong
D12 Overcast High Strong
D13 Overcast Normal Weak
Sunny
Humidity
High
Day Humid Wind
D1 High Weak
D2 High Strong
D8 High Weak
Normal
Day Humid Wind
D9 Normal Weak
D11 Normal Strong
Rain
Wind
Weak
Day Humid Wind
D4 High Weak
D5 Normal Weak
D10 Normal Weak
Strong
Day Humid Wind
D6 Normal Strong
D14 High Strong


Outlook
Overcast
yes
Sunny
Rain
Humidity
Wind
High
Normal
Weak
Strong
no
yes
yes
no


Outlook
Overcast
yes
Sunny
Rain
Humidity
Wind
High
Normal
Weak
Strong
no
yes
yes
no
New data:
Day Outlook Humid Wind
D15 Rain High Weak


Outlook
Overcast
yes
Sunny
Rain
Humidity
Wind
High
Normal
Weak
Strong
no
yes
yes
no
New data:
Day Outlook Humid Wind
D15 Rain High Weak → Yes


9 / 5
Outlook
4 / 0
Overcast
yes
2 / 3
Sunny
3 / 2
Rain
Humidity
Wind
0 / 3
High
no
2 / 0
Normal
yes
3 / 0
Weak
yes
0/ 2
Strong
no
New data:
Day Outlook Humid Wind
D15 Rain High Weak → Yes

ID3 algorithm

- Split node (node, training set):
  - P = best node to split the training set
  - Split training example to child nodes
  - Each value of P, create a new child node
  - Split training example to child nodes
  - For each child node:
    - If subset is pure then stop
    - Else keep splitting (child_node, {subset})

Which is best attribute to split?

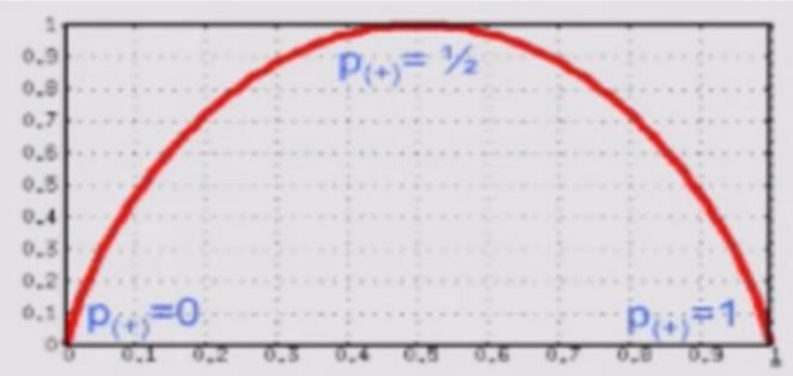
- Purity of subset is the determining factor
  - Pure (overcast) = completely certain vs Impure (Strong) = completely uncertain
  - Pure can be symmetric on both sides

Entropy

- $H(S) = -p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$

- Impure = 1

- Pure = 0



Information gain

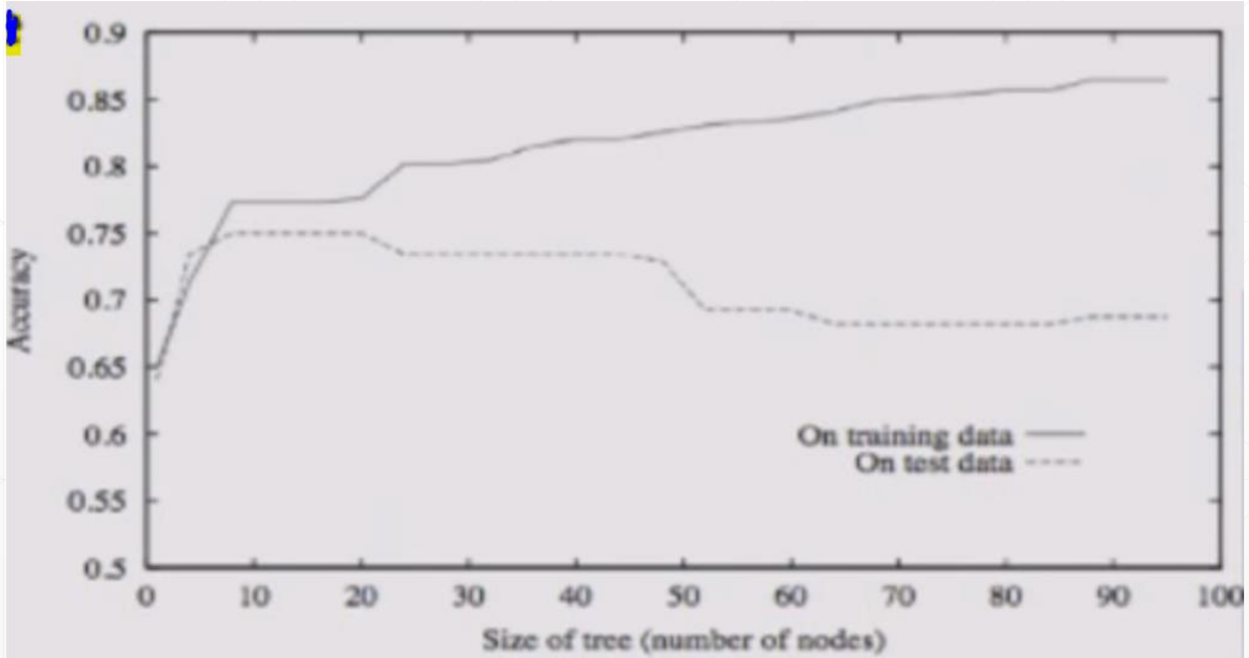
- Help us to find as many items in pure subsets

$$Gain(S,A) = H(S) - \sum_{V \in Values(A)} \frac{|S_V|}{|S|} H(S_V)$$

V ... possible values of A
S ... set of examples {X}
S_v ... subset where $X_A = V$

Overfitting

- Always classify training examples until we get singletons because singletons are pure.



How to Avoid overfitting?

- Pruning

Issue with Information gain

- Biased toward attributes with many values

- Use GainRatio:

$$SplitEntropy(S,A) = -\sum_{V \in Values(A)} \frac{|S_V|}{|S|} \log \frac{|S_V|}{|S|}$$

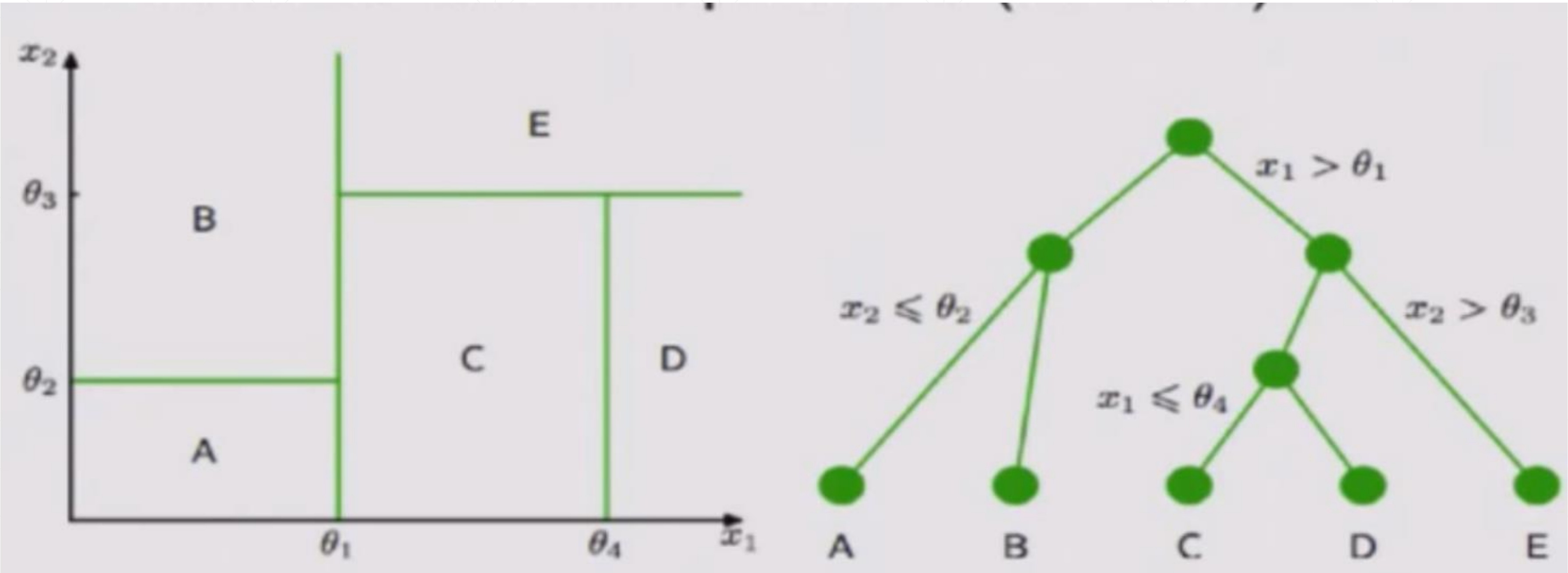
A … candidate attribute
V … possible values of A
S … set of examples {X}
S_V … subset where $X_A = V$

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitEntropy(S,A)}$$

penalizes attributes
with many values

Continuous Attributes

- Threshold is optimized



Muti-class classification

- Predict most frequent class in subset

Regression

- Predict average value from subset
- Linear regression at leaves.

Pros and cons

- Trees are interpretable (not a black box)
- Easily handle irrelevant features
- Can also handle missing values
- Compact --- very fast testing time O(depth)

- Only axis aligned decision boundaries
- Greedy alogorithm

Summary

- ID3 algorithm
- Greedily select next best attribute
- Entropy = certainty
- Information gain = reduction in uncertainty
- Gain ratio: favor big sets
- Prefer small trees higher gain at root
- If overfitting = post pruning
- Fast, compact and interpretable

Discussion

Thank you!