Department of Computer Science and Engineering

ID3 ALGORITHM

By,

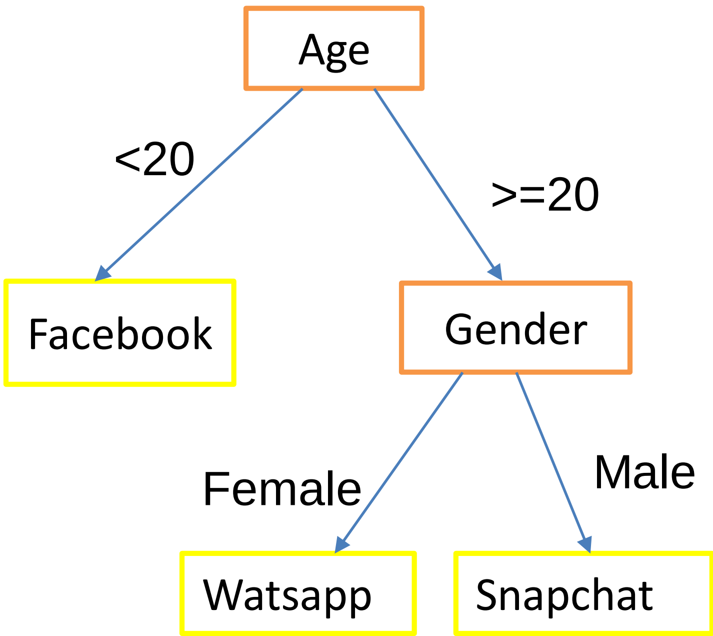
Mrs. Shruthi.J
Assistant Professor
Dept. of Computer Science & Engineering
BMS Institute of Technology, Bengaluru.

Agenda

- ✓ Overview of Decision Tree Learning
- ✓ ID3 algorithm
- ✓ Example
- ✓ Advantages and Disadvantage
- ✓ Exercise

Between Gender and age, which one seems more decisive for predicting what app the user download

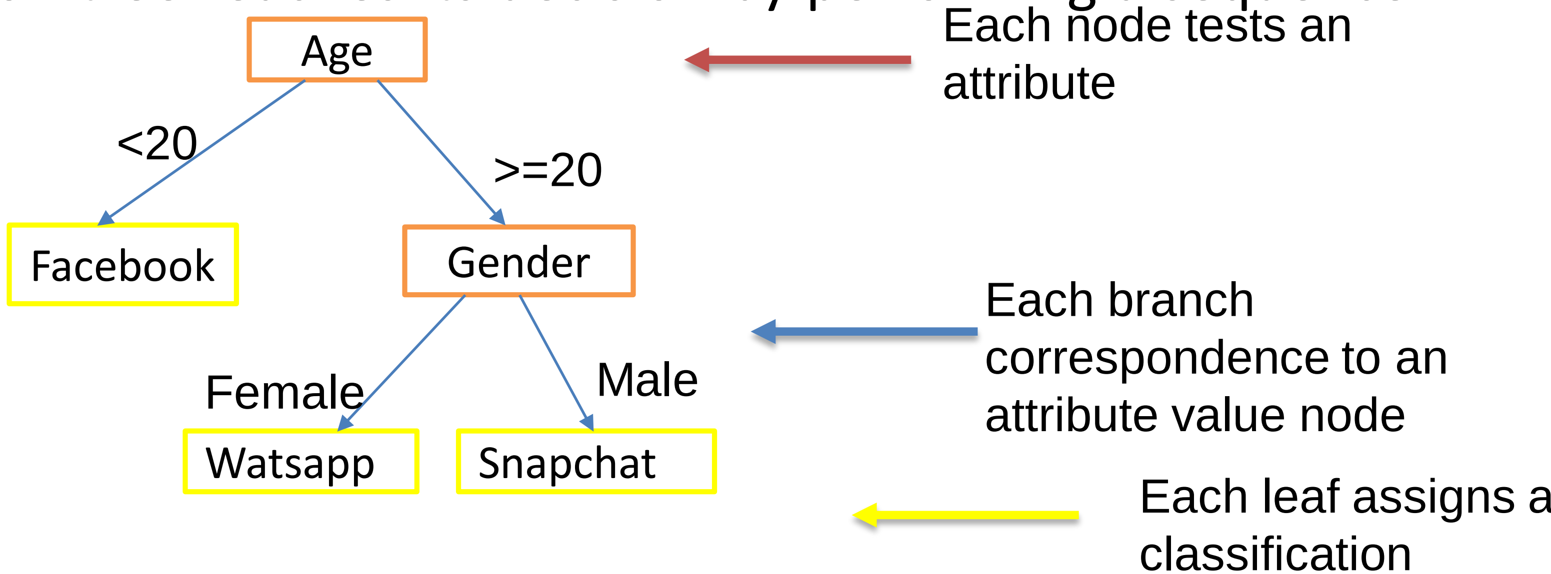
Gender	Age	App used
Female	15	Facebook
Female	25	Watsapp
Male	32	Snapchat
Female	40	Watsapp
Male	12	Facebook
Male	14	Facebook



So we all know what is Decision Tree!!!!

Decision Tree Learning

- Uses decision tree to go from **observations** to **conclusion** about the items target values.
- Predictive modelling approach used in statistics, **machine learning** and data mining.
- The input and output values can be discrete or continuous.
- A decision tree reaches its decision by performing a sequence of tests.



Decision Tree Learning Algorithm

- ID3 (Iterative Dichotomiser 3)
- C4.5 (successor of ID3)
- CART (Classification and Regression tree)

ID3Algorithm

- Invented by Ross Quinlan in 1975.
- Used to generate a decision tree from a given data set by employing a **top down** to test each attribute at every node of the tree.
- No back tracking
- The resulting tree used to classify the future samples.

ID3 Algorithm

- Dichotomisation means dividing to two completely opposite things
- Algorithm iteratively divides into two groups which are the most dominant attribute and other to construct a tree.
- Most dominant attribute can be found by calculating the Entropy and Information Gains of each attribute.
- Most dominant one is put on the tree as decision node.
- Entropy and Gain scores would be calculated again among the other attributes.
- Procedure continues until reaching a decision for that branch.

Entropy

- A formula to calculate the homogeneity of a sample
- A completely homogeneous sample has entropy of 0 (Leaf node).
- An equally divided sampled has entropy of 1.
- The formula for entropy is:

$$\text{Entropy}(S) = \sum -p(I)\log_2 p(I)$$

-p/(p+n)log2(p/(p/n))
-n/(p+n)log2(n/(p/n))

- Where p(I) is the proportion of S belonging to class I. ∑ is over total outcomes .
- Example: If S is a collection of 14 examples with 9 YES and 5 NO examples then
- Entropy(S)=-(9/14)log2(9/14)-(5/14)log2(5/14)=0.949

Information Gain (IG)

- The information gain is based on the decrease in entropy after a dataset is split on an attribute.
- The formula for calculating information gain is:

$$Gain(S,A)= Entropy(S)-\sum[p(S|A)*Entropy(S|A)]$$

- Make a decision tree node containing that attribute
- Recurse on subsets using remaining attributes.

Steps in ID3 Algorithm:

1. **Compute the entropy for the data-set**
2. **For every attribute/feature**
    i. Calculate **entropy** for all categorical values
    ii. Take **average information entropy** for the current attribute

$$\sum(pi+ni/(p+n)\ I(pi,ni))$$

    iii. Calculate **gain** for the current attribute
3. **Pick the highest gain attribute**
4. **Repeat until we get the tree we desired.**

Example to play tennis or not

Training Examples

Examples,
minterms,
cases, objects,
test cases,

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

Calculate Entropy(Step1)

Target/Decision

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

Decision column consists of 14 instances and includes two labels: YES(9) and NO(5)

Entropy(Decision) = -p(yes)*log2p(yes)-p(no)*log2p(no)

Entropy(Decision) =-(9/14)*log2(9/14)-5/14)*log2(5/14)=0.94

Wind Factor on decision (step2)

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

- Wind attributes has two labels: weak and strong.
- Calculate (D|w=weak) and (D|w=strong)

Weak attribute	Yes	No	Entropy
Weak	6	2	0.811
Strong	3	3	1

Weak wind factor (step 2.i)

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D13	Overcast	Hot	Normal	Weak	Yes

- Entropy(D|W=weak)=-p(no)*log2p(no)-p(yes)*log2p(yes)
- =-(2/8)*log2(2/8)-(6/8)*log2(6/8)=0.811

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

D2	Sunny	Hot	High	Strong	No
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D14	Rain	Mild	High	Strong	No

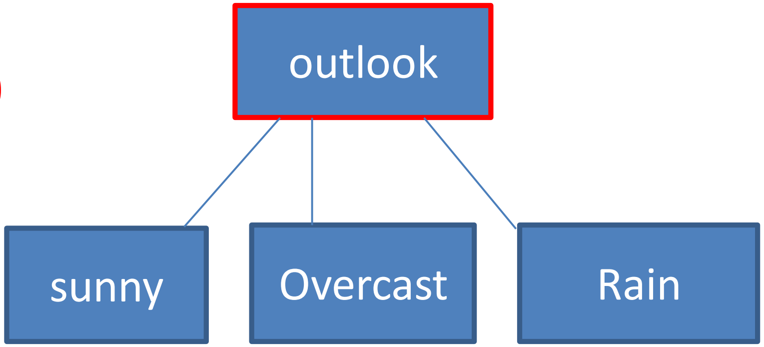
Strong Wind factor (step 2.i)

- Entropy(D|W=strong)=-p(no)*log2p(no)-p(yes)*log2p(yes)
- =-(3/6)*log2(3/6)-(3/6)*log2(3/6)=1

Gain of wind (step 2.ii and 2.iii)

- Gain(D,W)=Entropy(D)-[p(D|W=weak)*Entropy(D|w=weak)]-[p(D|W=strong)*Entropy(D|w=strong)]
- =0.940-[(8/14)*0.811]-[(6/14)*1]=0.048

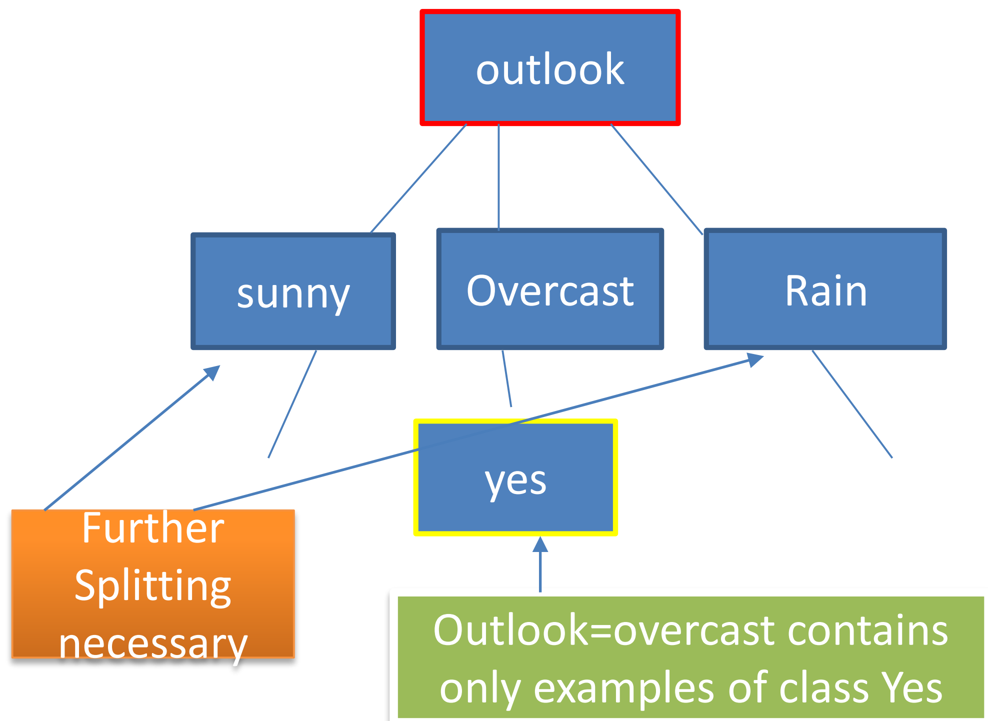
- Other factors on decision (step 3)
- Applied similar calculation on the other columns.
- Gain (Decision, outlook)= 0.246
- Gain (Decision, temperature)= 0.029
- Gain (Decision, humidity)= 0.151
- Gain (Decision, wind)= 0.048



- Outlook factor produce the highest score so appear as root node.

- Overcast outlook on decision (step 4)
- Decision will always be yes if outlook were overcast.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D3	Overcast	Hot	High	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes



Calculate (outlook=sunny|temperature)gain

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes

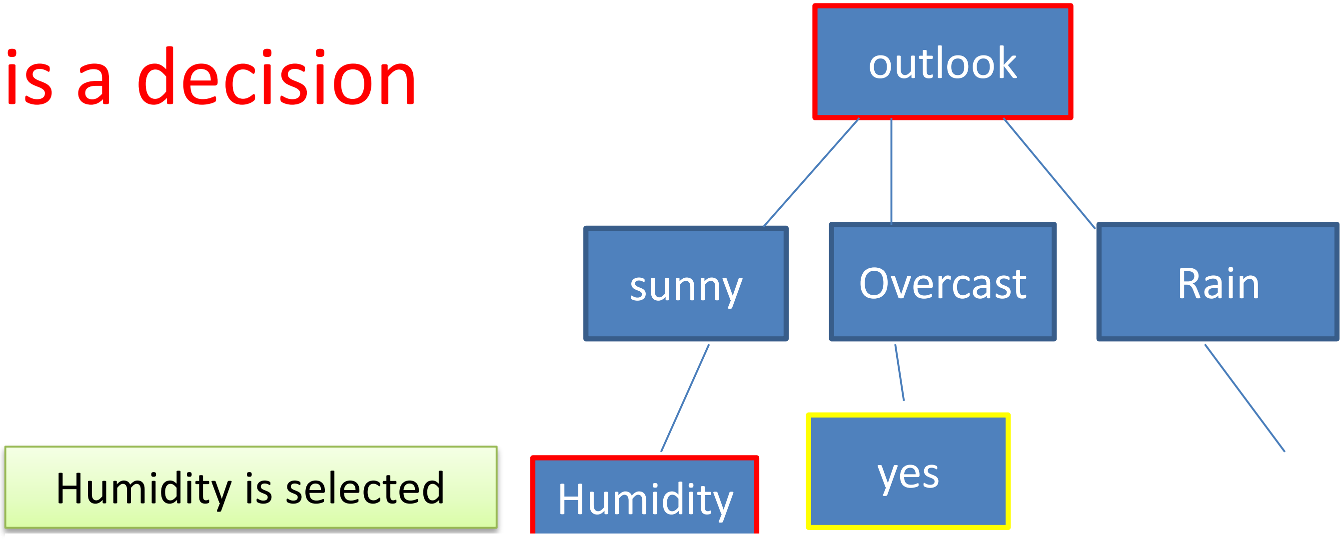
Sunny |temp	yes	No	Information
Hot	0	2	0
Mild	1	1	1
Cool	1	0	0

Entropy(outlook=sunny|temperature(hot))=

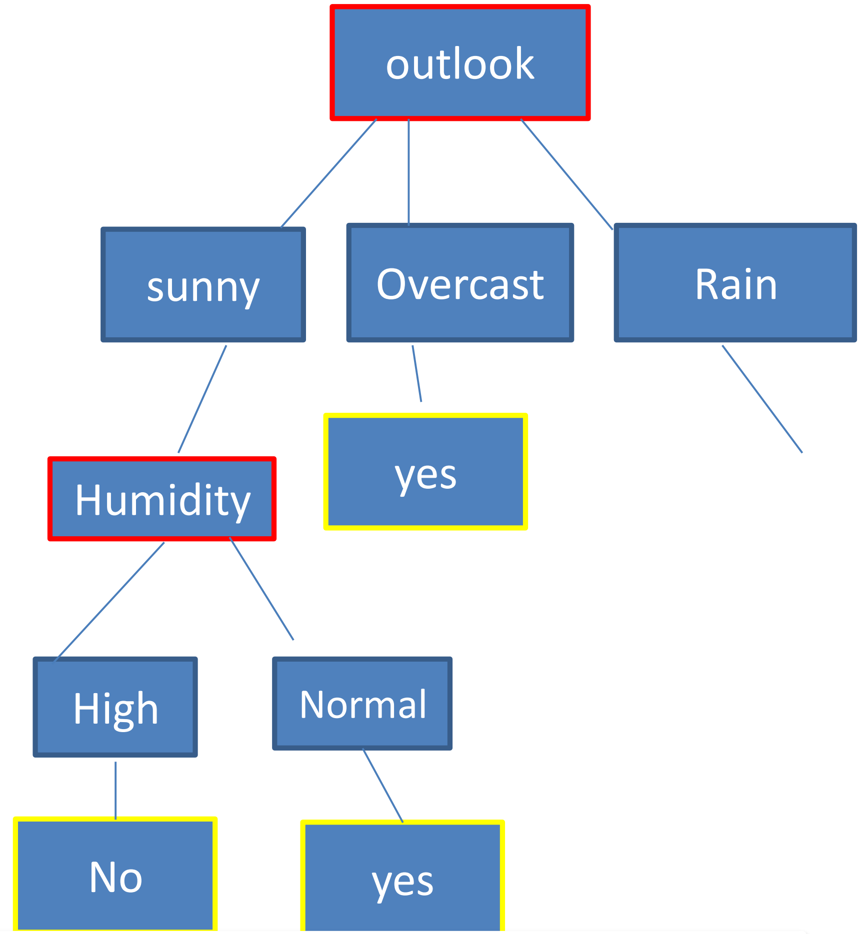
-p(yes)*log2p(yes)-p(no)*log2p(no)

-(0/2)log2(0/2)-(2/2)log2(2/2)=**0**

- Sunny outlook on decision
- (outlook=sunny|temperature)gain=0.570
- (outlook=sunny|Humidity)gain=0.970
  (outlook=sunny|wind)gain=0.019
- Humidity is a decision



Day	Outlook	Temp	Humidity	Wind	Play tennis
D1	Sunny	Hot	High	Weak	NO
D2	Sunny	Hot	High	Strong	NO
D3	Overcast	Hot	High	Weak	YES
D4	Rain	Mild	High	Weak	YES
D5	Rain	Cool	Normal	Weak	YES
D6	Rain	Cool	Normal	Strong	NO
D7	Overcast	Cool	Normal	Strong	YES
D8	Sunny	Mild	High	Weak	NO
D9	Sunny	Cool	Normal	Weak	YES
D10	Rain	Mild	Normal	Weak	YES
D11	Sunny	Mild	Normal	Strong	YES
D12	Overcast	Mild	High	Strong	YES
D13	Overcast	Hot	Normal	Weak	YES
D14	Rain	Mild	High	Strong	NO



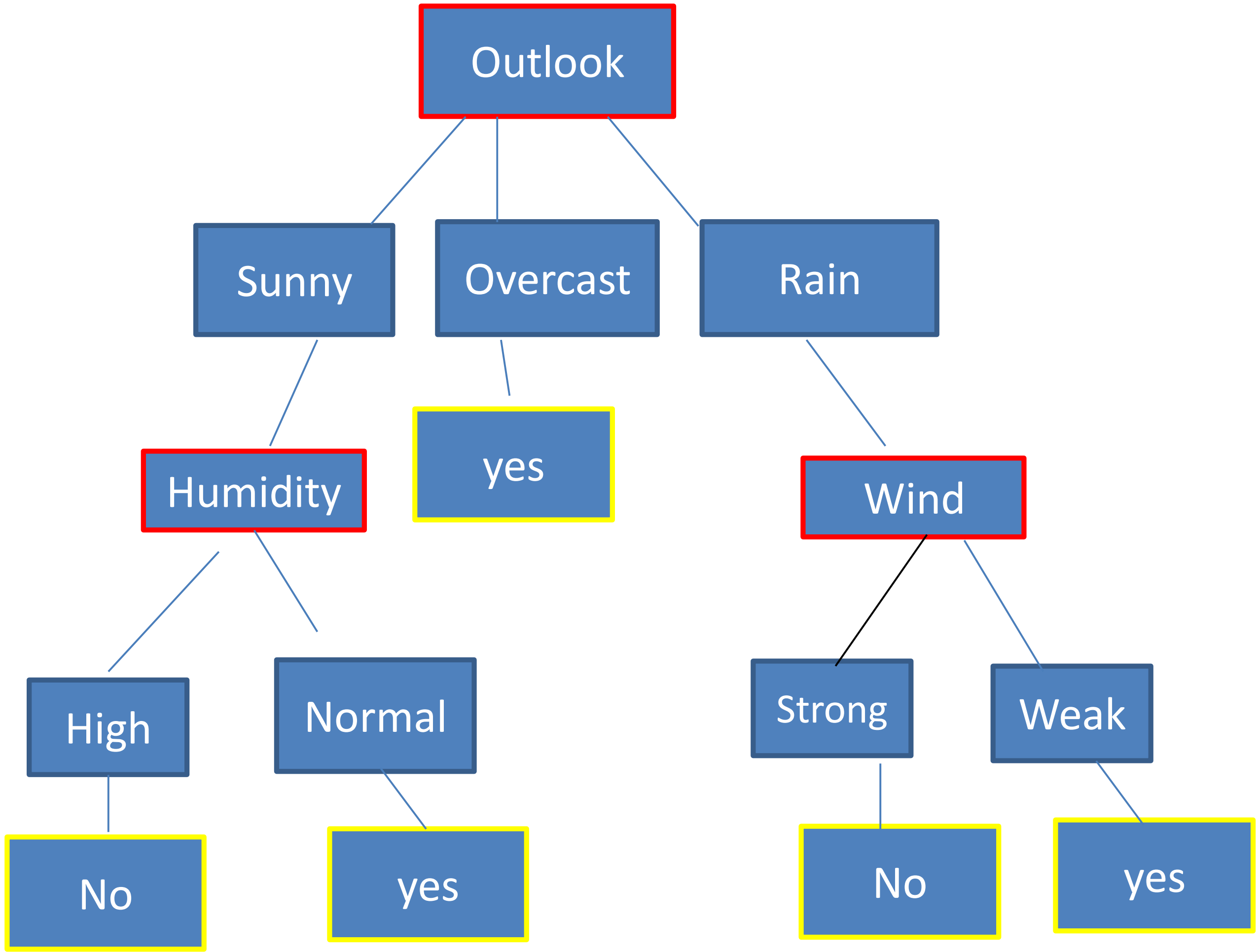
Pure leaves No further expansion necessary

Day	Outlook	Temp	Humidity	Wind	Play tennis
D1	Sunny	Hot	High	Weak	NO
D2	Sunny	Hot	High	Strong	NO
D3	Overcast	Hot	High	Weak	YES
D4	Rain	Mild	High	Weak	YES
D5	Rain	Cool	Normal	Weak	YES
D6	Rain	Cool	Normal	Strong	NO
D7	Overcast	Cool	Normal	Strong	YES
D8	Sunny	Mild	High	Weak	NO
D9	Sunny	Cool	Normal	Weak	YES
D10	Rain	Mild	Normal	Weak	YES
D11	Sunny	Mild	Normal	Strong	YES
D12	Overcast	Mild	High	Strong	YES
D13	Overcast	Hot	Normal	Weak	YES
D14	Rain	Mild	High	Strong	NO

- Rain outlook on decision
- Gain(outlook=rain|temperature)
- Gain(outlook=rain|humidity)
- Gain(outlook=rain|wind)

- Wind produce highest score

Complete Decision Tree



Points to remember

- Entropy value should be between **0 to 1**.

- If all examples are positive or all are negative then entropy will be **zero(low)**

- If half of the examples are of positive and half are negative class then entropy is **one (high)**

Advantages

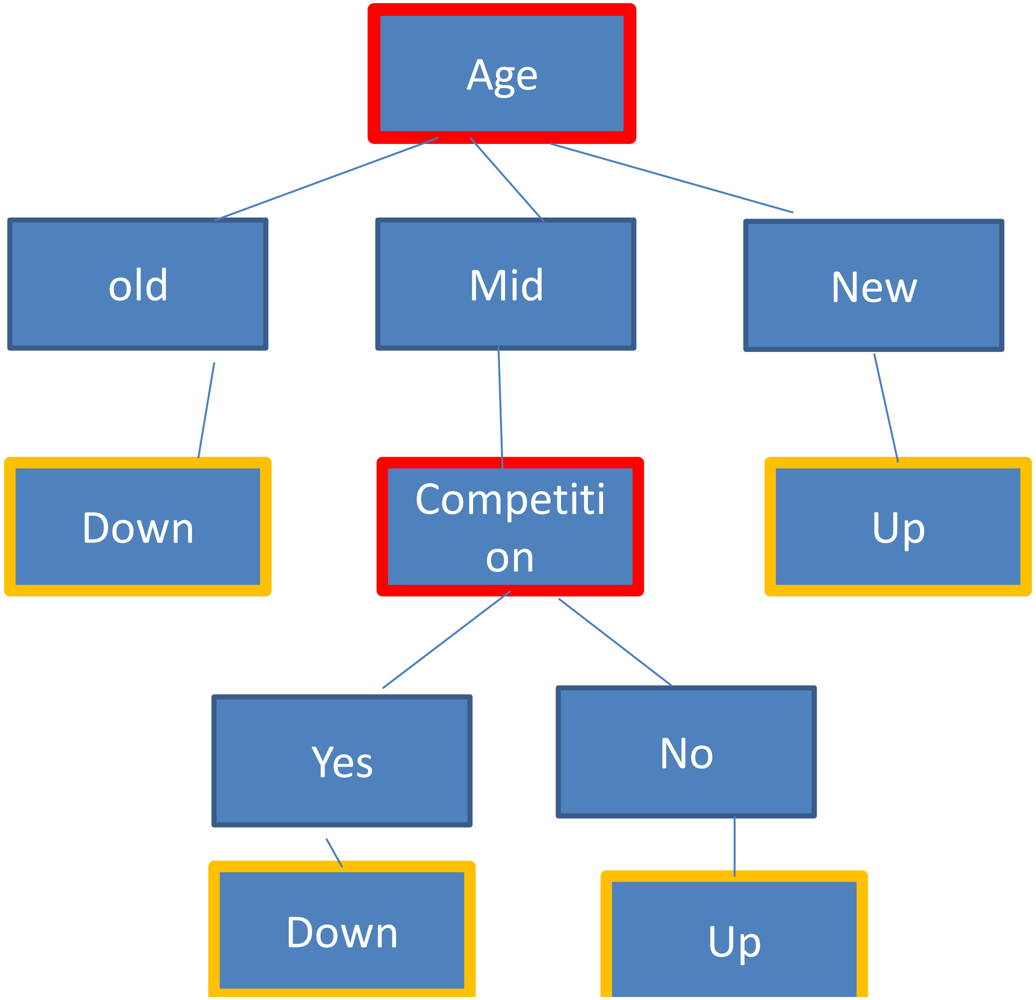
- Understandable **prediction rules** are created from the training data.
- Builds **fastest** and **short tree**
- Only need to test **enough attributes** until all data is classified.
- Finding leaf nodes enables test data to be **pruned**, reducing number of tests.
- **Whole dataset** is searched to create tree.

Disadvantages

- **Overfit** to the training data if small sample is tested.
- Only **one attribute at a time** is tested for making a decision.
- Does not handle numeric attributes and missing values.
- Classifying continuous data may be computationally **expensive**, as many trees must be generated to see where to break the continuum.

Let us solve

Age	Competition	Type	Profit
Old	YES	Software	Down
Old	NO	Software	Down
Old	NO	Hardware	Down
Mid	YES	Software	Down
Mid	YES	Hardware	Down
Mid	NO	Hardware	Up
Mid	NO	Software	Up
New	YES	Software	Up
New	NO	Hardware	Up
New	NO	Software	Up


Age
old
Mid
New
Down
Competiti
on
Up
Yes
No
Down
Up