Decision Tree : > Supervised

Day	Outlook ①	Temp ②	Humidity ③	Wind ④	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No.
3	Overcast	Hot	High	Weak	Yes.
4	Rain	mild	—	—	Yes
5.	Rain	—	—	—	Yes
6.	Rain	—	—	—	No
7.	Overcast	—	—		
8.	Sunny				
9.	Sunny	—	—		
10.	Rain	—	—		
12.	Overcast				

Step-1:

p = no. of yes values

n = no. of no. values.

p + n = total

p = 9

n = 5

p + n = 14

Information Gain: [step-1]

$$I(p,n) = \frac{-p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

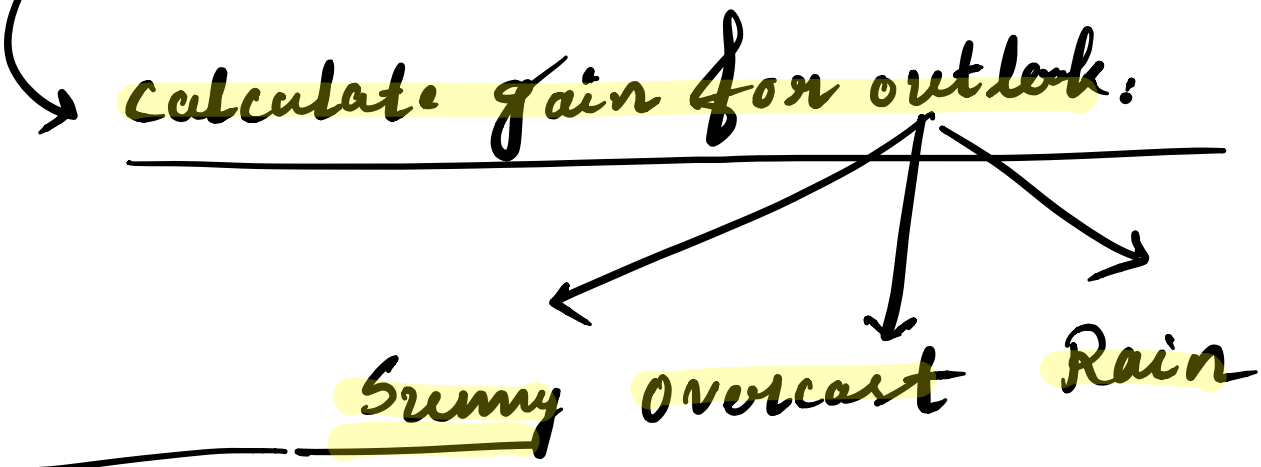
(yes, no's, total)

$$I(p,n) = \frac{-9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.9403$$

Calculate gain for each attribute [step-2]

Calculate gain for outlook:



Sunny

$p_i = 2$ ✓

$n_i = 3$ ✓

$$I(p,n) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= 0.971$$

Repeat the same for overcast and Rain.

outlook	p_i	n_i	$I(p,n)$
Sunny	2	3	0.971
Overcast	4	0	0
Rain	3	2	0.971

Calculate the entropy (step-3)

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{outlook}) = \frac{5}{14} \times 0.971 + \frac{5}{14} \times 0.971$$

$$= \text{org } 0.694$$

Step-4: Gain of outlook

$$\text{gain}(\text{outlook}) = I(p,n) - E(\text{outlook})$$

(step-1)

$$= 0.9403 - 0.694$$

$$= 0.246$$

Repeat the same steps for Temp, Humidity and wind.

Step-5:

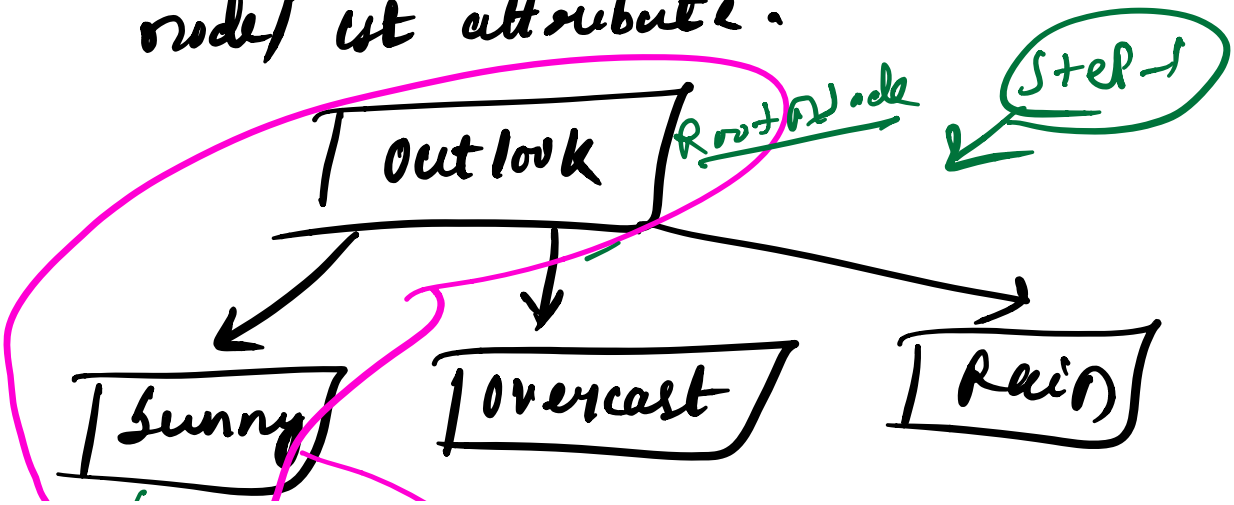
Gain(Temp) = 0.029

Gain(Humidity) = 0.151

Gain(wind) = 0.048

Gain(Outlook) = 0.246 → highest value among all

Hence choose outlook as the root node/ 1st attribute.



Iteration-2:

Now its turn to calculate the decision tree for this part.

→ Create a subset from the main dataset containing all values for outlook and sunny.

Day	outlook	Temp	H	W	P
8	Sunny	mild	High	W	No
9	Sunny	Cold	N	W	Yes
11	Sunny	Hot	N	S	Yes

↳ Repeat all the above steps for this new table

Random Forest :

why not only decision tree?

↓

overfitting may be a prob.

Solution ⇒ Create a bunch of decision tree based on random sample of data.

→ Define some set of subsample
→ For each subsample create a decision tree.
→ Then when we have a bunch of decision tree that we

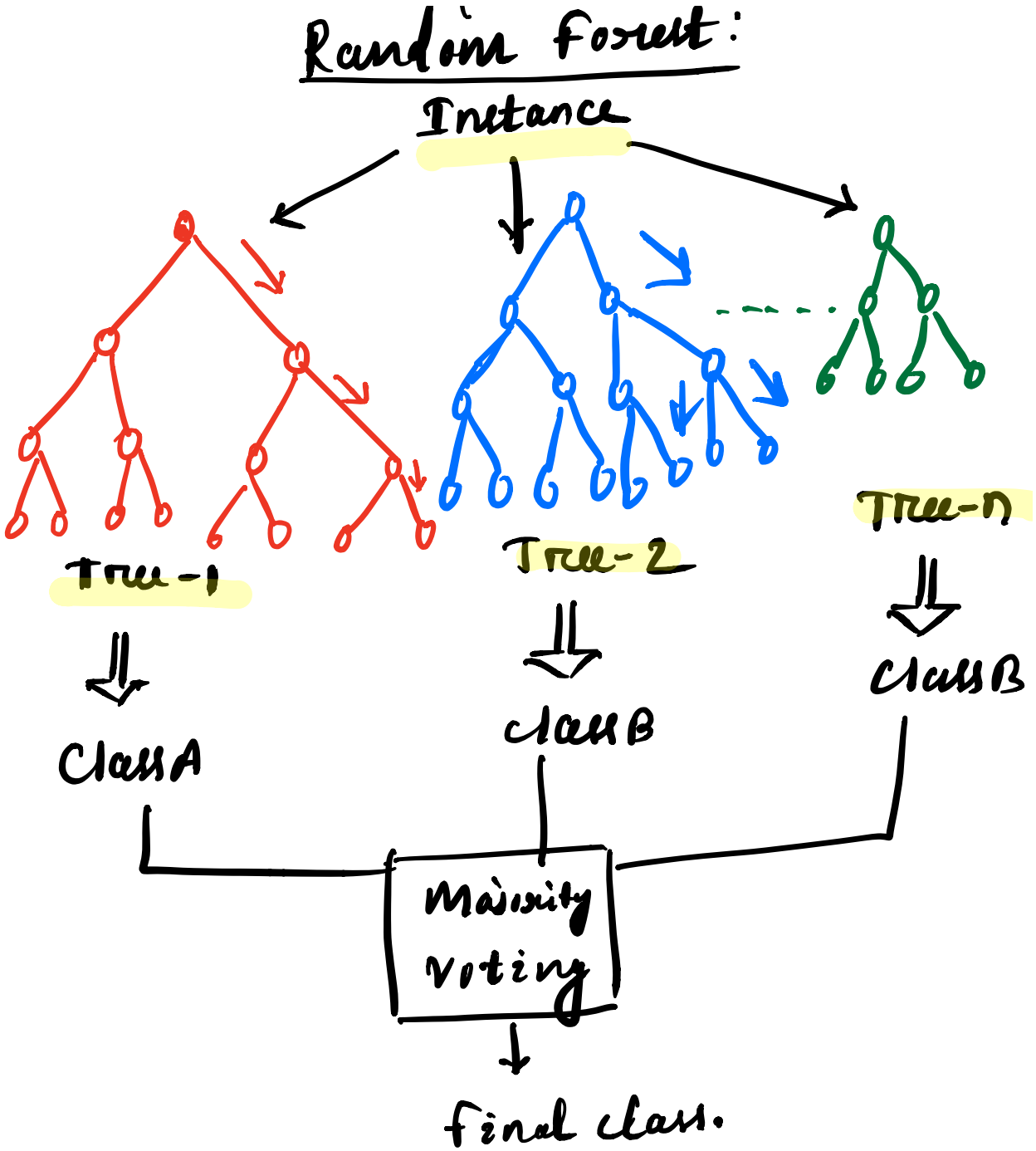
generated, when we have a new data point we will use all these decision tree to predict its class and then we take majority of vote.

→ Hence Random forest gives us a higher accuracy then using a decision tree alone.

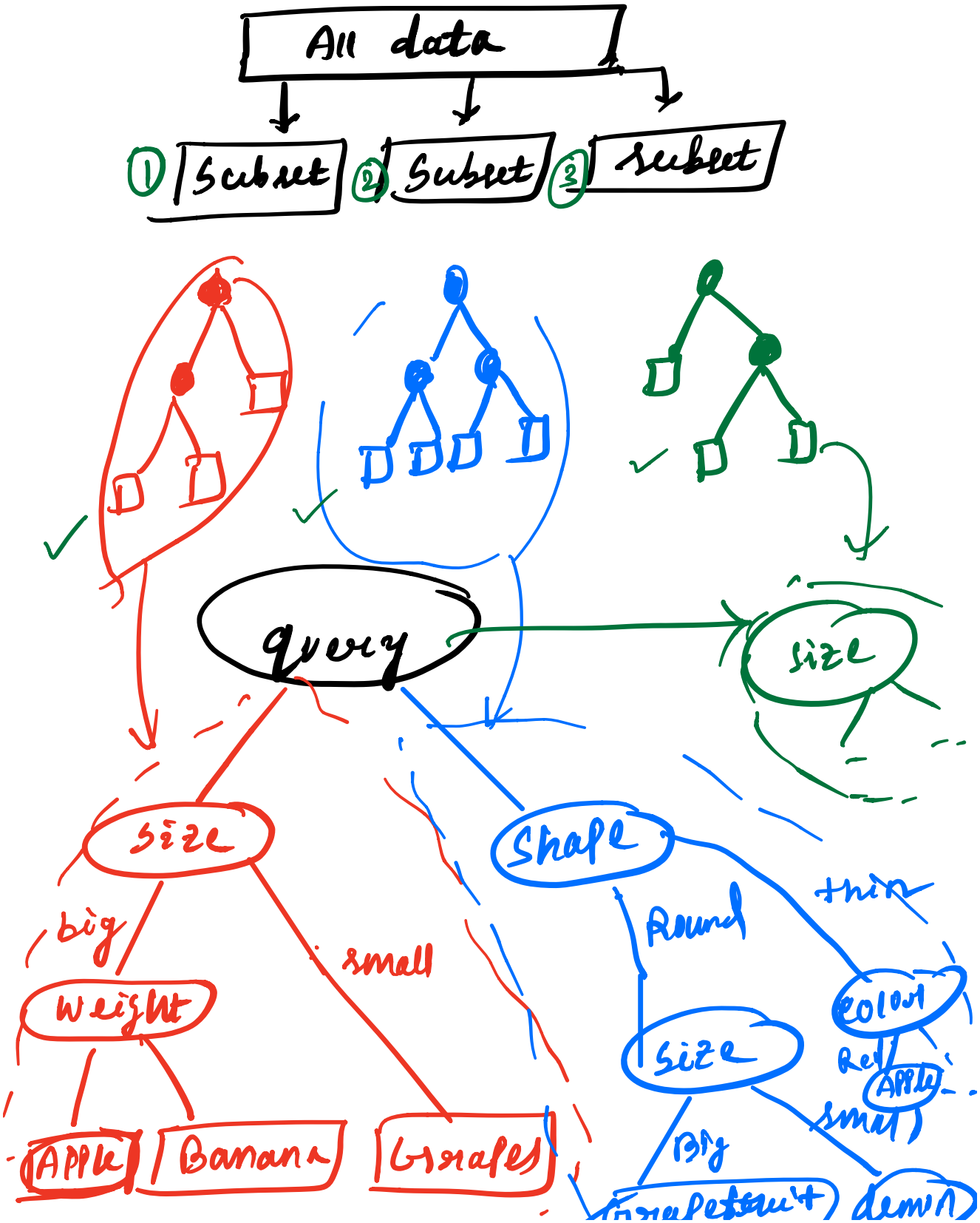
→ Both Decision Tree and Random Forest can be used for both classification as well as for regression.

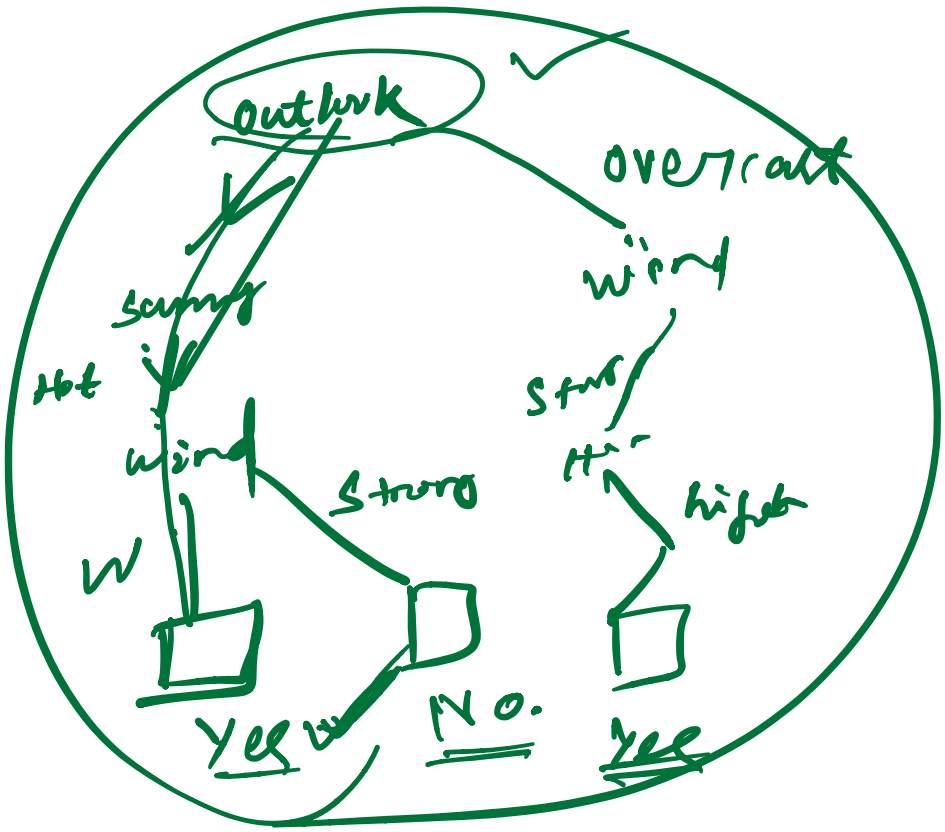
→ most used machine learning technique.

Random Forest:



→ The more trees we add the more accuracy & core we gets.


All data
1 Subset
2 Subset
3 Subset
query
size
size
big
small
weight
Apple
Banana
Grapes
Shape
Round
thin
size
color
Red
Apple
small
Big
demin


Outlook
Overcast
Sunny
Wind
Hot
Strong
Yes
No.
Yes