Decision Tree Split for numerical Features

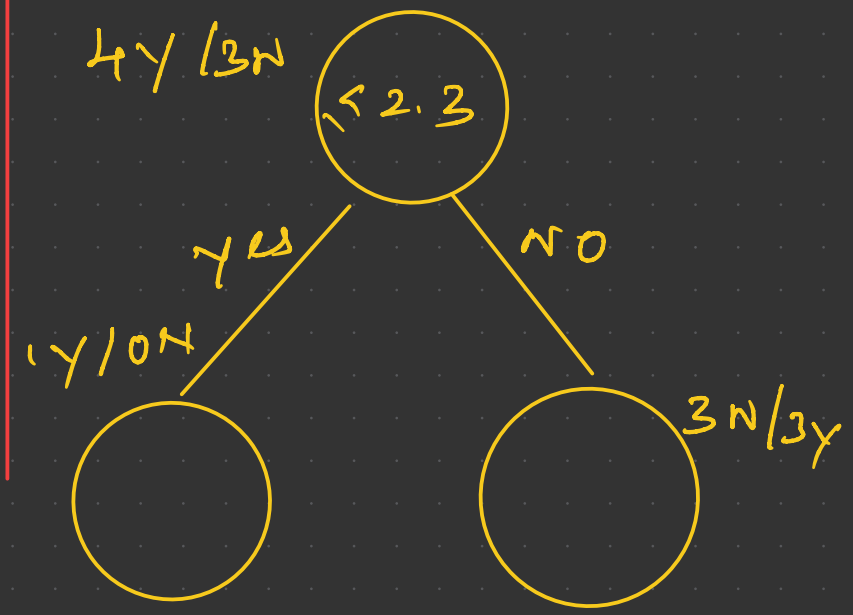
Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

If you look at the above dataset, it is of categorical feature. we could easily do the split on the categorical features, what if the features are numerical ?

F_1	O/P
2.3	yes
3.6	yes
4	NO
5.2	NO
6.7	yes
8.9	no
10.5	yes

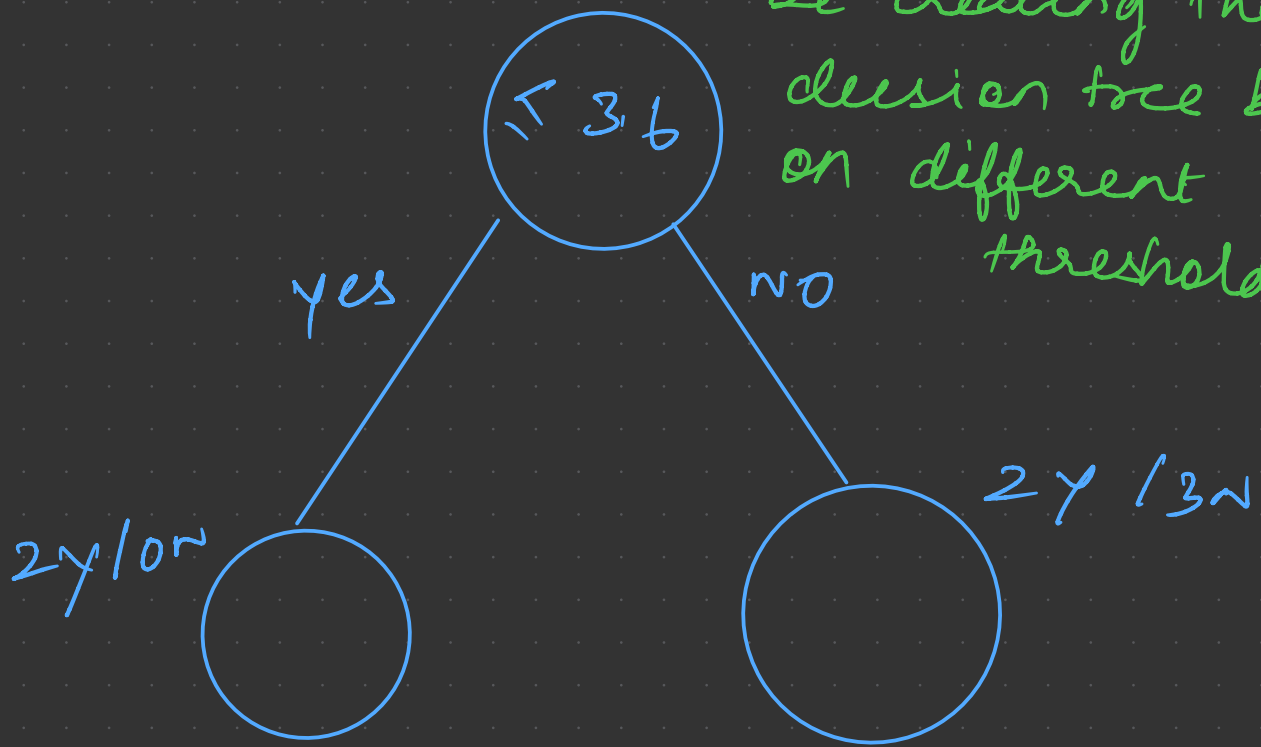
① Sort the features based on ascending order.

① Threshold = 2.3



② Threshold = 3.6

like this I will be creating the decision tree based on different threshold.



Now to understand better which split is better, we will be using Information Gain.

in a nutshell, whichever is having highest information gain we will be selecting that particular split.

There is a drawback in this method,

what if we have millions of records?

→ Time complexity is usually high.