Random Forest

We build Random Forest with multiple decision trees. If one classifier (Random forest here) is made of multiple classifier (Decision tree), this method is called

Ensemble Learning

The concept is that, when multiple weak learners works together, it works better than one perfect learner. So, we will make multiple weak decision trees, whose combined decision will give us more accurate prediction.

Steps in Random Forest:

1. Create multiple bootstrapped data set.

- Loop → for each bootstrapped dataset → (Bagging)
  - (Create decision trees)
    - Loop → for each node in the decision tree
      - 2. Randomly select a few features
      - 3. compute Gini Impurity and decide the node
    - Store the tree.

4. Majority voting for classification decision.

Let us ~~first~~ Build Random Forest on the given Dataset.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
Day1	Sunny	Hot	High	Weak	~~No~~ NO
Day2	Sunny	Hot	High	Strong	~~No~~ NO
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	NO
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	NO
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	overcast	Mild	High	Strong	Yes
Day13	overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	NO

(1) Bootstrapping

We will create multiple bootstrapped dataset at first. Bootstrapping is a method in statistics and ML where a random number of data/~~n~~ or instances gets selected. The word "Random" ~~Forest~~ in random forest comes from this random selecteection of data/instances.

Let us create 3 bootstrapped dataset ~~eg~~ with 6 instances each →

Bootstrapped dataset 1 –

Let's assume Day 10, Day 11, Day 12, Day 13, ~~an~~ Day 14 and Day 2 gets ~~selected~~ → copied →

Day	outlook	temperature	Humidity	Wind	Play Tennis
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	overcast	Mild	High	Strong	Yes
Day13	~~Rain~~ overcast	Hot	Normal	Weak	Yes
Day14	~~Sunny~~ Rain	Mild	High	Strong	NO
Day 2	Sunny	Hot	High	Strong	NO

Bootstrapped dataset 2 –

We will randomly copy Day 1, Day 2, Day 3, Day 4, Day 5 and Day 2 Again : 30 in Bootstrapped dataset, Instances can be repeated.

Day	outlook	temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	NO
Day2	Sunny	Hot	High	Strong	NO
Day3	overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day2	Sunny	Hot	High	Strong	NO

Repeated

Bootstrapped Dataset 3:-

We will select/copy Day6, Day7, Day8, Day9, Day10, Day13 randomly.

Day	outlook	Temperature	Humidity	wind	Play Tennis
Day6	Rain	Cool	Normal	Strong	NO
Day7	overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	weak	NO
Day9	Sunny	Cool	Normal	weak	Yes
Day10	Rain	Mild	Normal	weak	Yes
Day13	overcast	Hot	Normal	weak	Yes

Bagging:

Now we have to create Decision trees, for each Bootstrapped dataset. This technique is called Bagging, where we create several models on different random samples (bootstrapped) dataset

We will now create Full Decision Trees from all Bootstrapped dataset Let us create the first Decision Tree from the Bootstrapped dataset-1. For this at first we need to decide, what will be the root node.

[2] Feature selection -

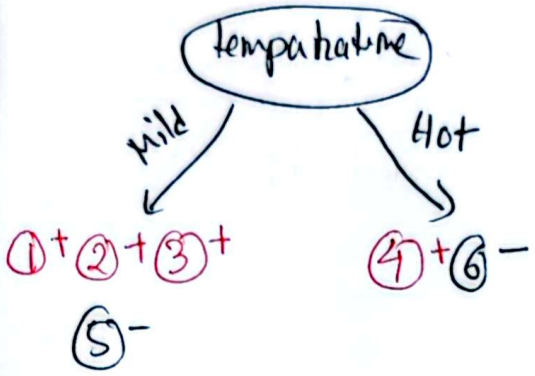
While creating the decision tree, whenever we are deciding a node, we will consider only a subset of features instead of all. Let us select Tempakature and Humidity for root node from bootstrapped dataset 1.
So the training set will look like -

	Day	Temperature	Humidity	Play Tennis
(1)	Day 10	Mild	Normal	Yes
(2)	Day 11	Mild	Normal	Yes
(3)	Day 12	Mild	High	Yes
(4)	Day 13	Hot	Normal	Yes
(5)	Day 14	Mild	High	No
(6)	Day 2	Hot	High	No

To create Decision Trees, we will use the Gini Impurity Formula. Let us calculate Gini Impurity of Temperature and Humidity.

3] Calculating Gini Impurity:

Temparature



⇒ Gini (Temparature = Mild)

$= 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$

⇒ Gini (Temparature = Hot)

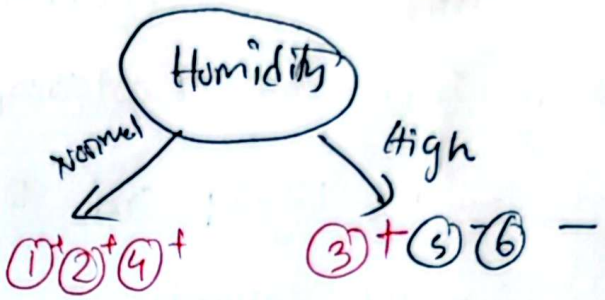
$= 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$

⇒ Gini (Temparature)

$= \frac{4}{6} \times 0.375 + \frac{2}{6} \times 0.5$

$= 0.417$

Humidity



⇒ Gini (Humidity = Normal)

$= 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$

⇒ Gini (Humidity = High)

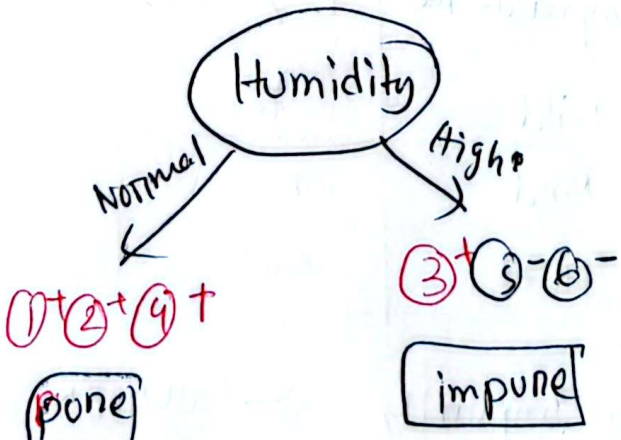
$= 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.44$

⇒ Gini (Humidity) =

$= \frac{3}{6} \times 0 + \frac{3}{6} \times 0.44$

$= 0.2223$

As Gini (Temparature) > Gini (Humidity), we will consider Humidity as root node.



Decision = Yes

we need to again decide a feature for this node as split.

For the Impure Node, we have to do the process Again.

- At first we will separate the datapoints of the impure node from the Bootstrapped Dataset-1.

~~Data~~ Day 12, Day 14 and Day 2 have ended up in the impure branch.

Data Set for the next level Node –

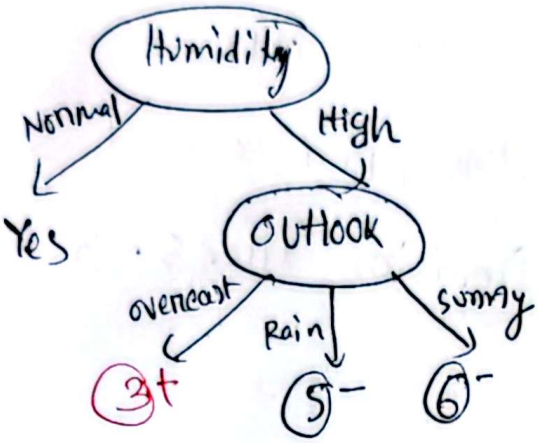
Day	~~tempa~~ outlook ~~feature~~ temparature	Humidity	Wind	Play Tennis?
Day12	overcast	Mild	High	Strong	Yes
Day14	Rain	Mild	High	Strong	No
Day2	Sunny	Hot	High	Strong	No

Again we will select two features to decide the split for the Node. Let us select outlook and Temparature →

	Day	outlook	Temparature	Play Tennis
(3)	Day 12	overcast	Mild	Yes
(5)	Day 14	Rain	Mild	No
(6)	Day 2	Sunny	Hot	No.

Now we will calculate Gini Impurity of outlook and Temparature to decide the split.

Outlook



Gini (Outlook = overcast)

$= 1 - (\frac{1}{1})^2 - (\frac{0}{1})^2 = 0$

Gini (Outlook = Rain)

$= 1 - (\frac{0}{1})^2 - (\frac{1}{1})^2 = 0$

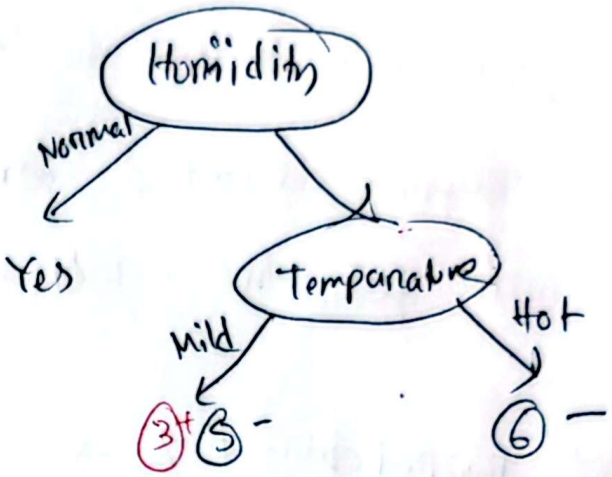
Gini (Outlook = Sunny)

$= 1 - (\frac{0}{1})^2 - (\frac{1}{1})^2 = 0$

∴ Gini (overcast) $= \frac{1}{3} \times 0 + \frac{1}{3} \times 0 + \frac{1}{3} \times 0$

$= 0$

Temparature



Gini (Temparature = Mild)

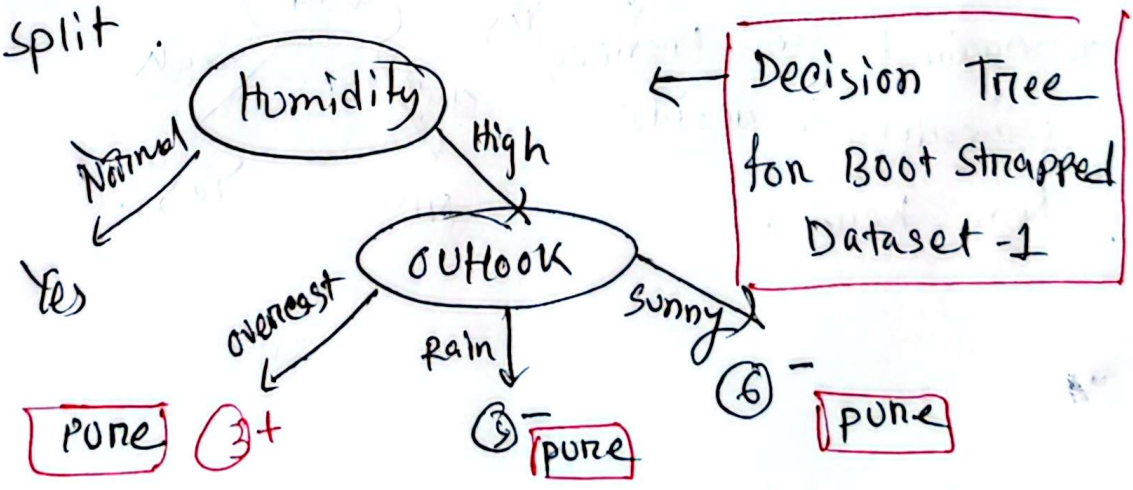
$= 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0.5$

Gini (Temparature = Hot)

$= 1 - (\frac{0}{1})^2 - (\frac{1}{1})^2 = 0$

∴ Gini (Temparature) $= \frac{2}{3} \times 0.5 + \frac{1}{3} \times 0 = 0.33$

As Gini (Outlook) < Gini (Temparature) we will use Outlook as the split.

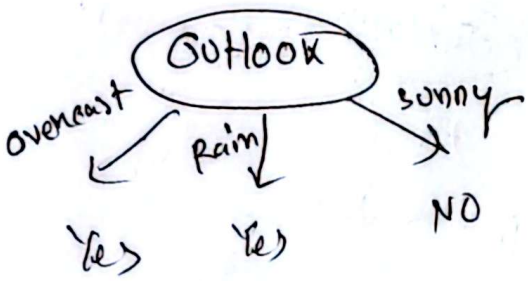


Decision. Yes NO NO

we will stop growing the tree as all leaves are pure.

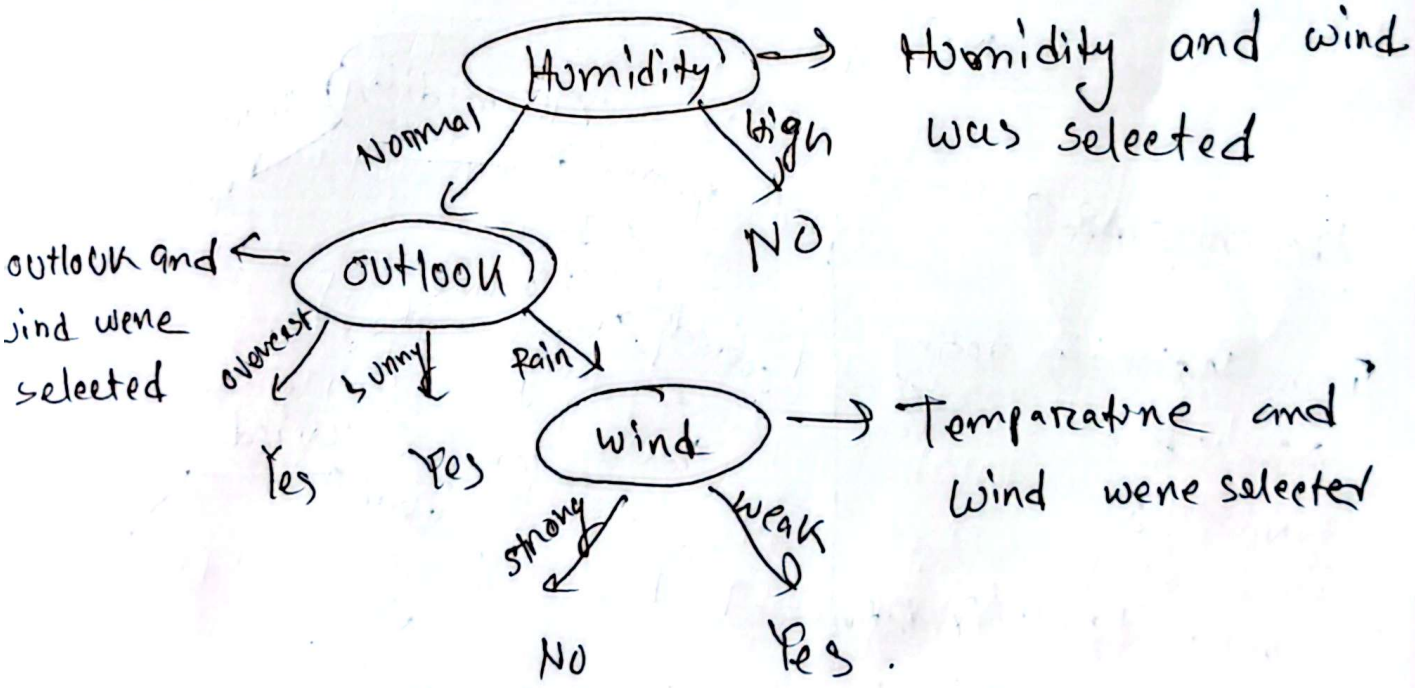
Similarly if we find Decision Tree for ~~Dataset~~ for Bootstrapped dataset-1 and Bootstrapped dataset-2, we will get the following Decision Trees.

Bootstrapped Dataset-2's Decision Tree



→ outlook and Temparature features were Randomly selected to check Gini

Bootstrapped Dataset-3's Decision Tree—

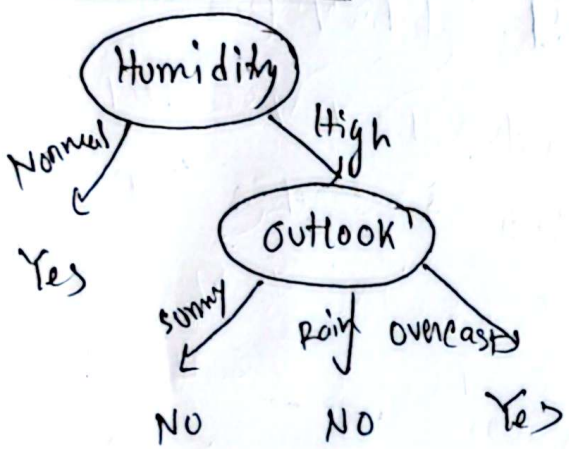


Now we have three Decision Trees of our Random Forest Model.

Let's predict class for the following Query.

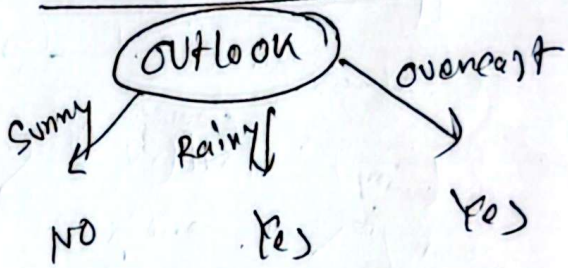
	Outlook	Temperature	Humidity	Wind	Play Tennis
Day 13 →	Rain	Hot	High	Weak	Yes

Decision Tree-1 →



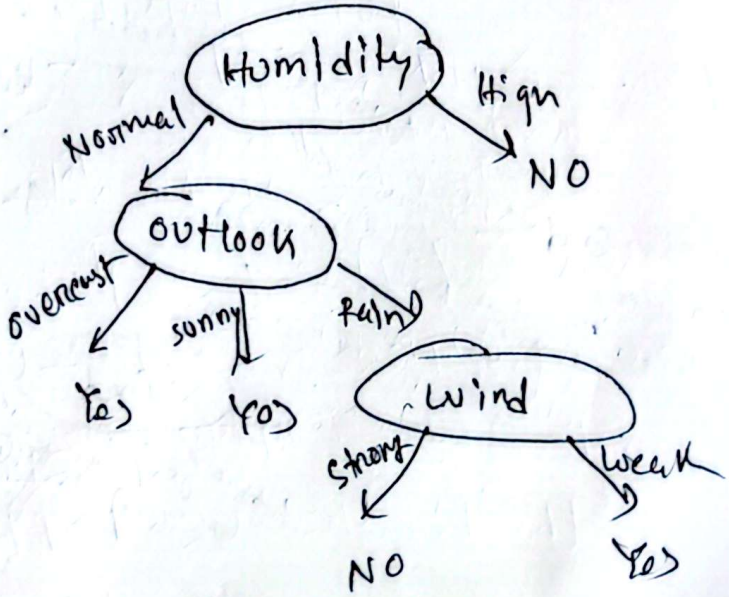
Query Result → NO

Decision Tree-2 →



Query Result → Yes

Decision Tree-3 →



Query Result → NO

~~Decision Tree-3~~

From Majority voting

Yes : 1

NO : 2

The final Result/Prediction of the Random Forest Mode is **NO**