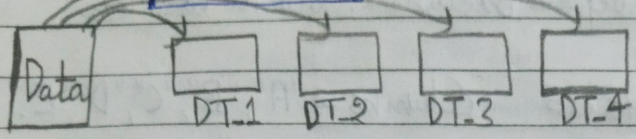
Bagging.....

Date.....

Q What is the difference between the Bagging and Random Forest...

Ans (1) Bagging



As, per the Sampling, the data will be shared with all the Models.

In Bagging, the "Sampling" is done.

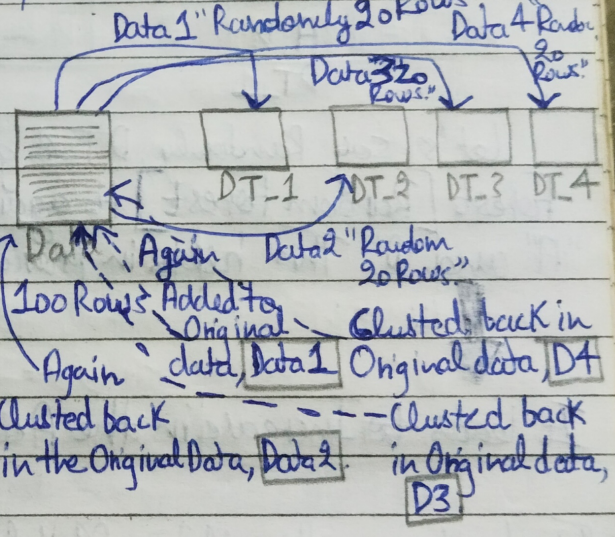
- Working Columns Sampling
- Rows Sampling
- Combination of both Rows as well as Column Sampling.

Let's say, My Data ⇒ "df" includes 100 Rows, 5 Columns. and We have 4 Decision_Tree Models to train. Columns ⇒ "A", "B", "C", "D", "E" of data "df".

So, our data ⇒ "df" includes 100 Rows.

Working of this Sampling

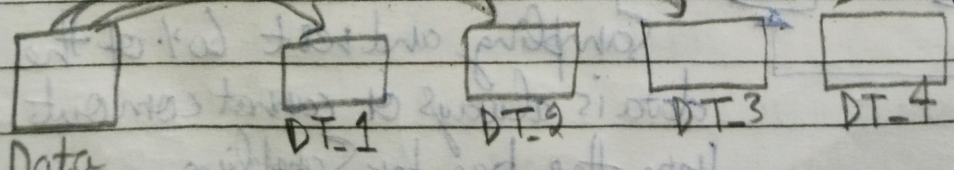
A,B	DT_1	⇒ Randomly (Any 2 Columns)
D,A	DT_2	⇒ Randomly (Any 2 Columns)
C,B	DT_3	⇒ Randomly (2 Col.)
D,B	DT_4	⇒ Randomly (2 Col.)



*So, before Execution, the Columns were randomly provided to the Models [DT_1, DT_2, DT_3, DT_4, DT_5].

(2) Decision Forest

Sampling is same i.e. data is transfered either in Row Sampling Method or Column Sampling Method....



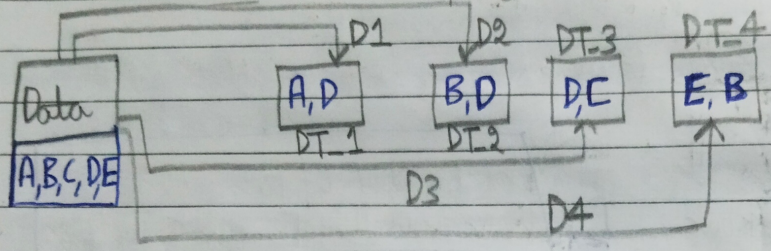
Spiral

Random Forest....

Date.....

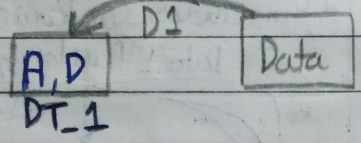
But, In Decision Forest, the Sampling of data is same as in the Case of Bagging but, we have see that in Bagging, the data is provided to the Models [DT_1, DT_2, DT_3, DT_4] before Execution. But in Random Forest this happens:-

"Column Sampling is done".... Columns ⇒ "A", "B", "C", "D", "E".



* Talking about the DT_1

"Assuming that 2 Columns Randomly will be provided to each Column."



Let's say Randomly, DT_1 got columns "A", "D". But, Here in decision Forest [Random Forest] it again randomly selects from the columns "A" and "D". This helps in providing more Randomness to the Models.

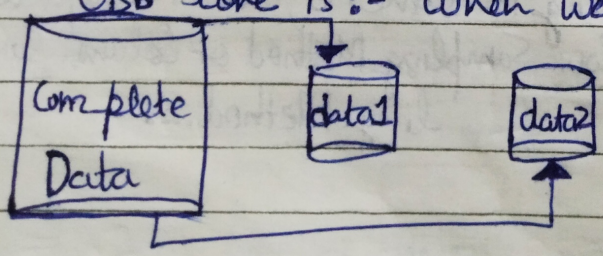
It leads to Increase in the Accuracy of the Prediction.....

This happens with all the Models available in the Random Forest...

Q What is OOB in Bagging?

Ans Out of Bag Score or OBB Score is:- When we take the data after Sampling i.e; data 1 or data 2 only 30% of the actual data is used in Sampling and rest 60% of the data is always or cannot come out from the bag for Sampling...

Out of Bag Score [OOB_Score]



Spiral

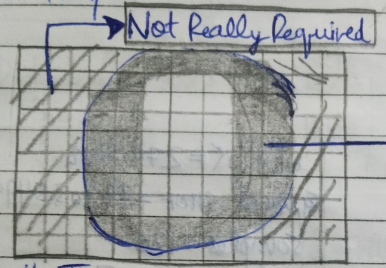
Feature Importance....

Date.....

Q) Why we use the Feature Importance?

Ans Feature Importance ⇒ Feature Importance is used to take the important features [columns] from the complete data. These important features are important because their availability is majorly affecting the Prediction [Output].

For example:-



Handwritten Image of Zero "0".

Let's say I have the "MNIST" Data.

Shape ⇒ "MNIST" Data.

Important Columns ⇒ 28X28 [28-Rows, 28-Columns] ⇒ 784

As per the image, I can see that the Important Features [Columns] are marked as 0.

So, the "Random Forest" have the availability to calculate the Feature Importance within the data...

Calculating Feature Importance

(i) Calculating the Feature Importance of the Decision Trees!!

FORMULA....

Where:- N_t ⇒ No. of Rows in the Node.
N ⇒ Total No. of Rows in the table.
impurity ⇒ gini impurity.

$$Node_{Impurity} = \frac{N\_t}{N}\left[impurity - \left(\frac{N\_t\_r}{N} \times right\ impurity\right) - \left(\frac{N\_t\_l}{N} \times left\ impurity\right)\right]$$

N_t_r ⇒ No. of Rows in the right Node.

right impurity ⇒ gini impurity of right node.

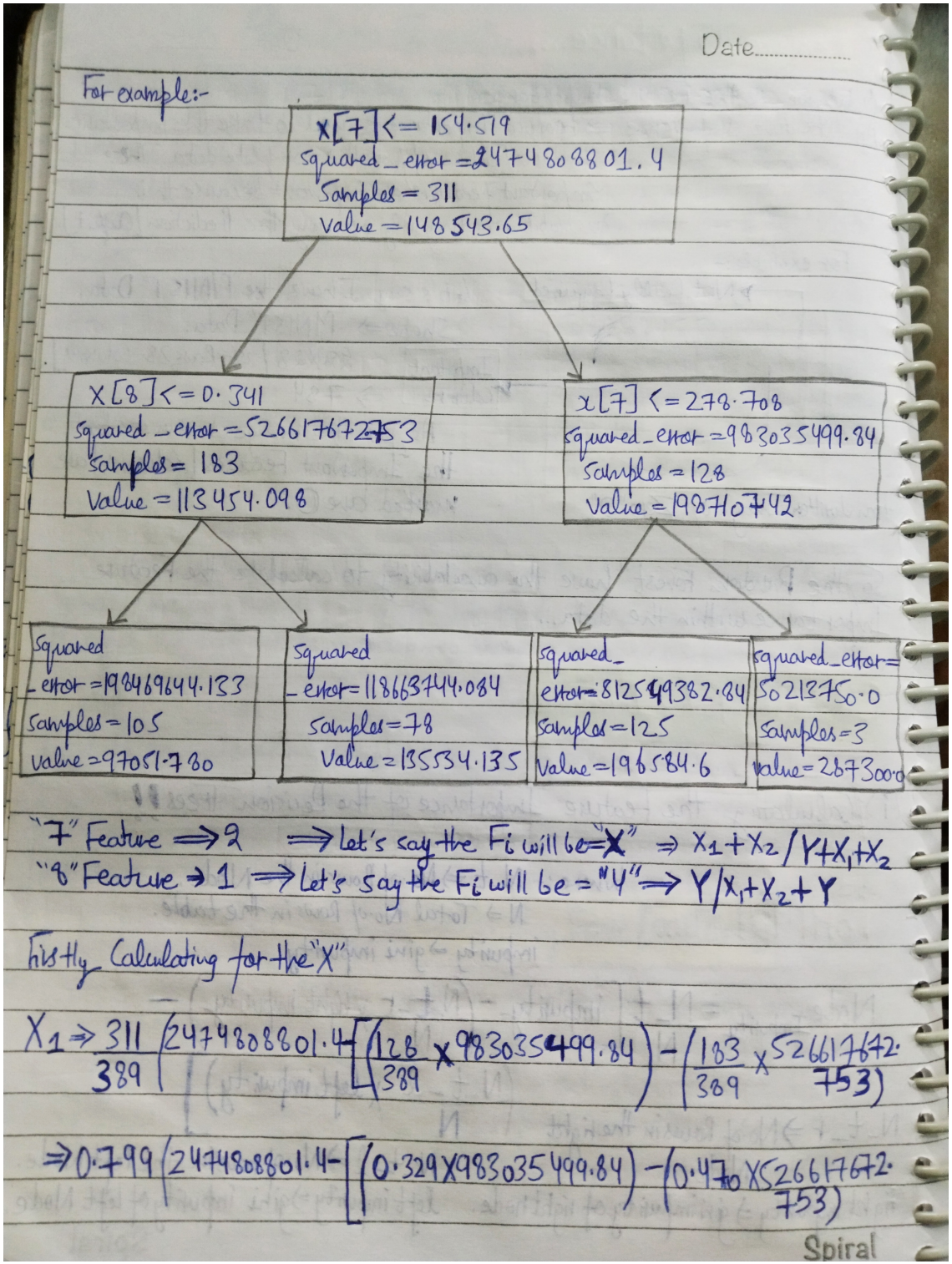
N_t_l ⇒ No. of rows in the left Node.

left impurity ⇒ gini impurity of left Node.

Spiral

Date.....

For example:-



"7" Feature ⟹ 2 ⟹ Let's say the Fi will be "X" ⇒ $X_1 + X_2 / Y + X_1 + X_2$

"8" Feature → 1 ⟹ Let's say the Fi will be = "Y" ⟹ $Y / X_1 + X_2 + Y$

Firstly Calculating for the "X"

$$X_1 \Rightarrow \frac{311}{389}\left(247480880 1.4 - \left[\left(\frac{128}{389} \times 983035499.84\right) - \left(\frac{183}{389} \times 526617672753\right)\right]\right.$$

$$\Rightarrow 0.799\left(247480880 1.4 - \left[\left(0.329 \times 983035499.84\right) - \left(0.470 \times 526617672753\right)\right]\right.$$

Spiral

Date..................

X_1

$\Rightarrow 0.799(2474808801.4 - [323418679.4 - 247510306.2])$

$\Rightarrow 0.799(2474808801.4 - 75908373.2)$

$\Rightarrow 0.799(2398900428)$

$\Rightarrow \boxed{1916721442}$ — (1)

$X_2 \Rightarrow \frac{128}{389}\left(983035499.84 - \left[\left(\frac{3}{389} \times 50213750.0\right) - \left(\frac{125}{389} \times 812549382.84\right)\right]\right)$

$\Rightarrow 0.329(983035499.84 - [(7.71 \times 50213750.0) - (0.32 \times 812549382.84)])$

$\Rightarrow 0.329(983035499.84 - [387148012.5 - 260015802.5])$

$\Rightarrow 0.329(983035499.84 - [127132210])$

$\Rightarrow 0.329[855903289.8]$

$\Rightarrow \boxed{281592182.3}$ — (2)

Calculating for the Y

$Y \Rightarrow \frac{183}{329}\left(526617672.753 - \left[\left(\frac{78}{329} \times 118663744.084\right) - \left(\frac{105}{329} \times 198469644.133\right)\right]\right)$

$\Rightarrow 0.55(526617672.753 - [(0.23 \times 118663744.084) - (0.31 \times 198469644.133)])$

$\Rightarrow 0.55(526617672.753 - [27292661.14 - 61525589.68])$

$\Rightarrow 0.55(526617672.753 - [-34232928.54])$

$\Rightarrow 0.55(560850601.3)$

$\Rightarrow \boxed{308467830.7}$ — (3)

Date.....

Feature Importance of "7" Feature will be:- $\frac{191672144+28159218203}{2198313624+4}$

$\Rightarrow \frac{2198313624}{2506781455}$

$\Rightarrow \boxed{0.87694}$

Feature Importance of "8" Feature will be:- $\frac{308467830.7}{2506781455}$

$\Rightarrow \boxed{0.12305}$