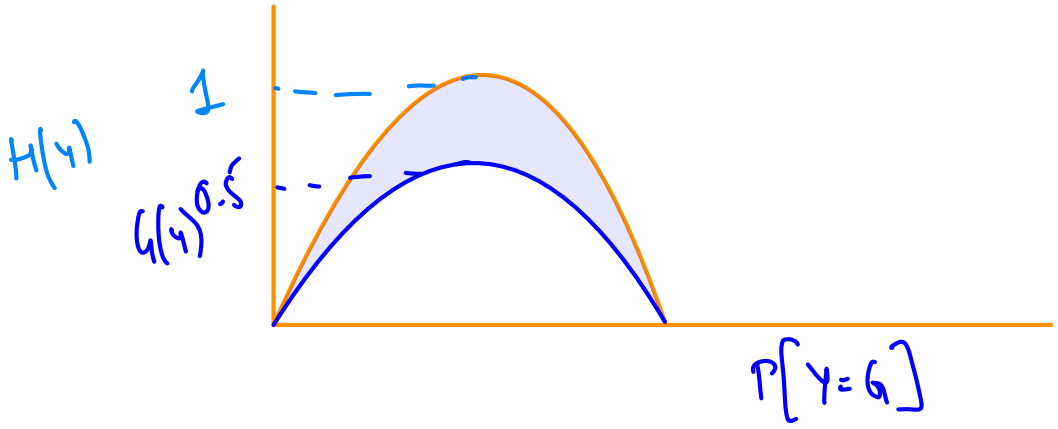
BAGGING & RANDOM FOREST

DT → Impurity → Entropy $H(Y)$ $[0,1]$ $\sum P(y_i) \log_2 [P(y_i)]$

Gini Impurity $G(Y)$ $1 - \sum [P(y_i)]^2$ $[0, 0.5]$



$Y = [0,1,2]$

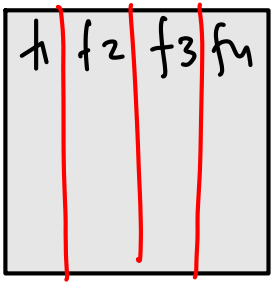
$1 - [P(Y=0)^2 + P(Y=1)^2 + P(Y=2)^2]$

Nomeical featun

f_1	f_2	f_3	f_4
C	C	N	N

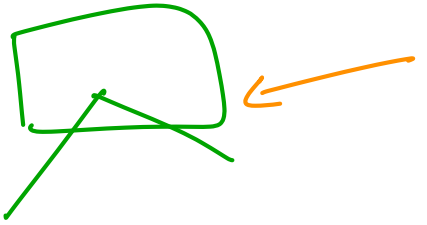
$n = 100$ (unique)

$d = 4$



$\Rightarrow$ DT

IG



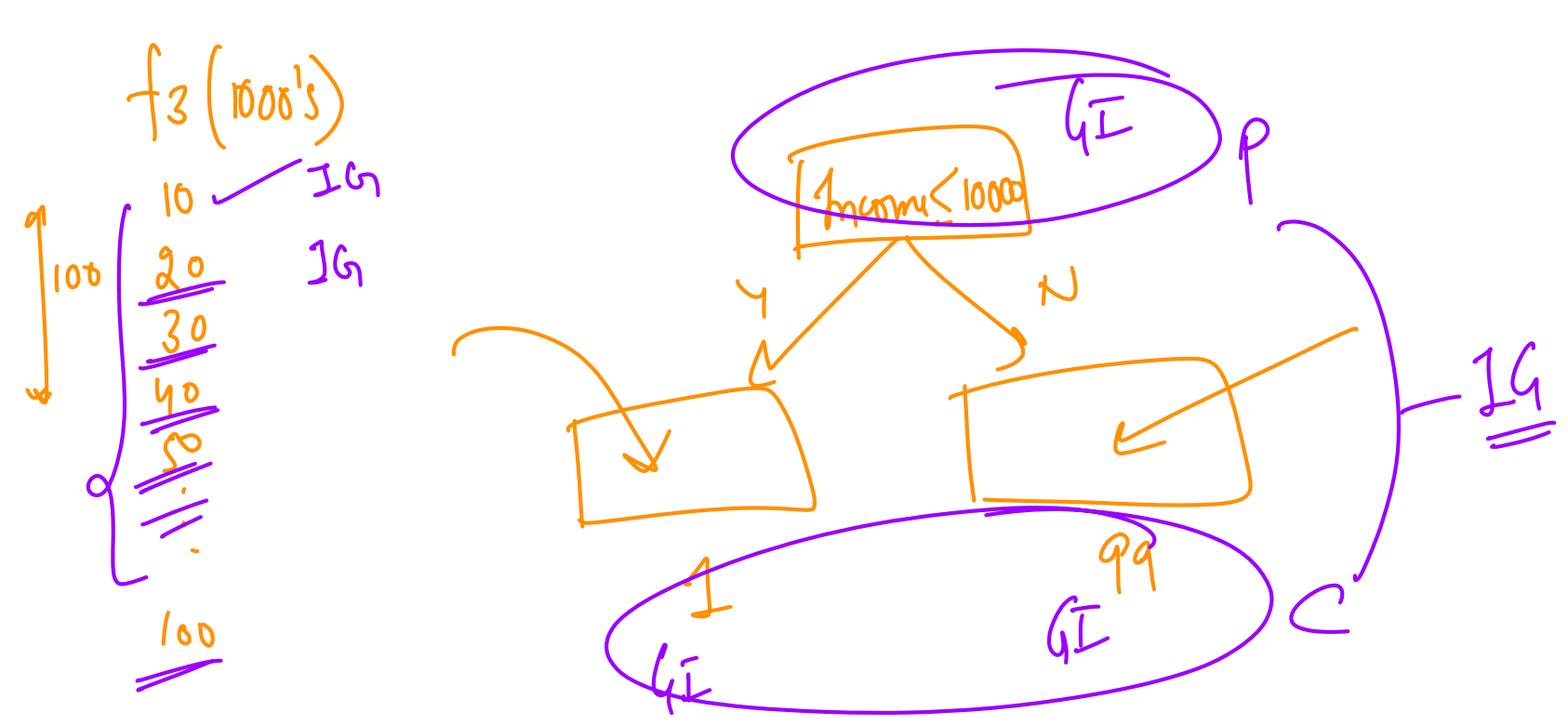
$f_3 \rightarrow$ IG $\rightarrow$ 100

$f_4 \rightarrow$ IG $\rightarrow$ 100

$f_1 \rightarrow$ IG - 1 ✓

$f_2 \rightarrow$ IG - 1 ✓

202

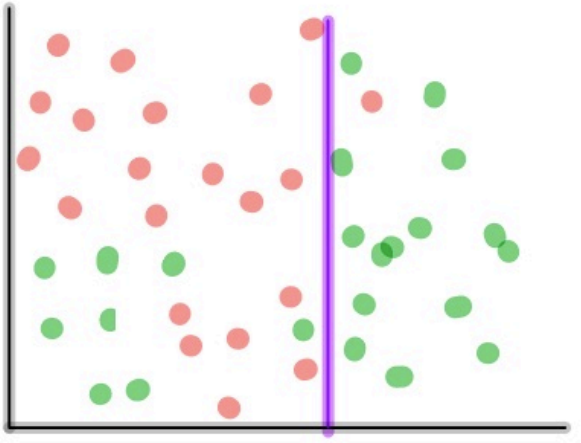

f3 (1000's)
10
20
30
40
50
100
100
IG
IG
GI
P
Income < 10000
Y
N
IG
I
GI
99
GI
C

DT

Overfit ⇒ Too Complex ← d ↑

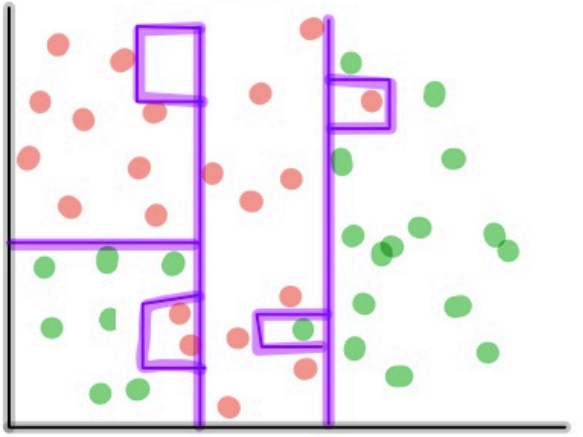
Underfit ⇒ Too simple. ← d ↓

Underfitting



Shallow tree

Overfitting



Deep Tree.

Impact of outlier

$d \uparrow$ Impact of outlier is significantly observed.

$d \downarrow$ Impact of outlier is either not observed or minimal.

Q DO WE REQUIRE feature Scaling

$$\text{Entropy} = -\sum_{i=1}^{k} P(y_i) \log P[y_i]$$

$$\text{Gini} = 1 - \sum_{i=1}^{k} [P(y_i)]^2$$

Count of Occurence → calculating based on frequency

lot of feature → high Dimensional data

f → 1000's → 400 C, 600 N } Computational requirement

Q Is DT a good idea?

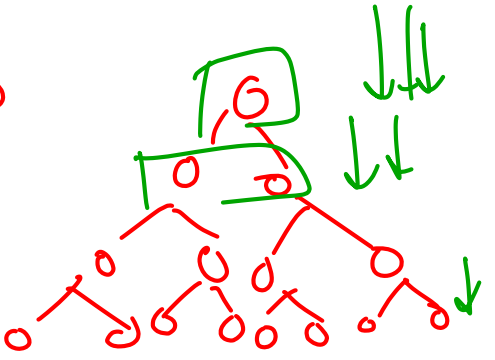
↳ Slow

↳ Lots of Computational resources required } →

Imbalanced data

↳ Should I use DT ?

→ Rebalance.



Feature Importances

How do you find out feature importance ?

Total IG → ✓

Regression

↑ DT

MSE

Gini

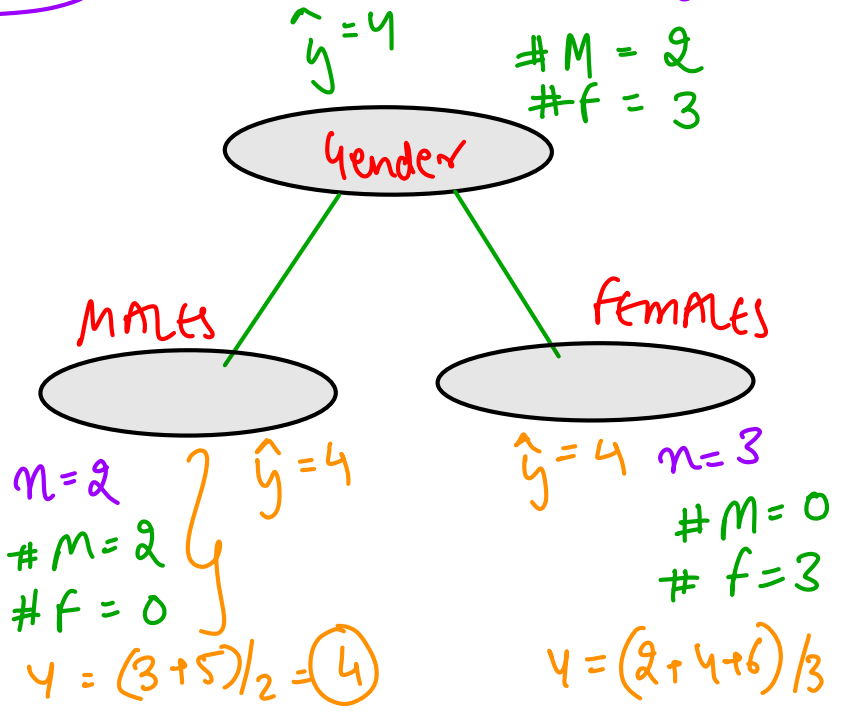
Entropy

$$\frac{1}{n}\sum (y-\hat{y})^2$$

Decision Trees for Regression

f1	f2	f3	y	
F	Y		2	D
M	N		3	D-
F	Y		4	D
M	N		5	D-
F	N		6	D

$$\frac{1}{n}\sum (y-\hat{y})^2$$

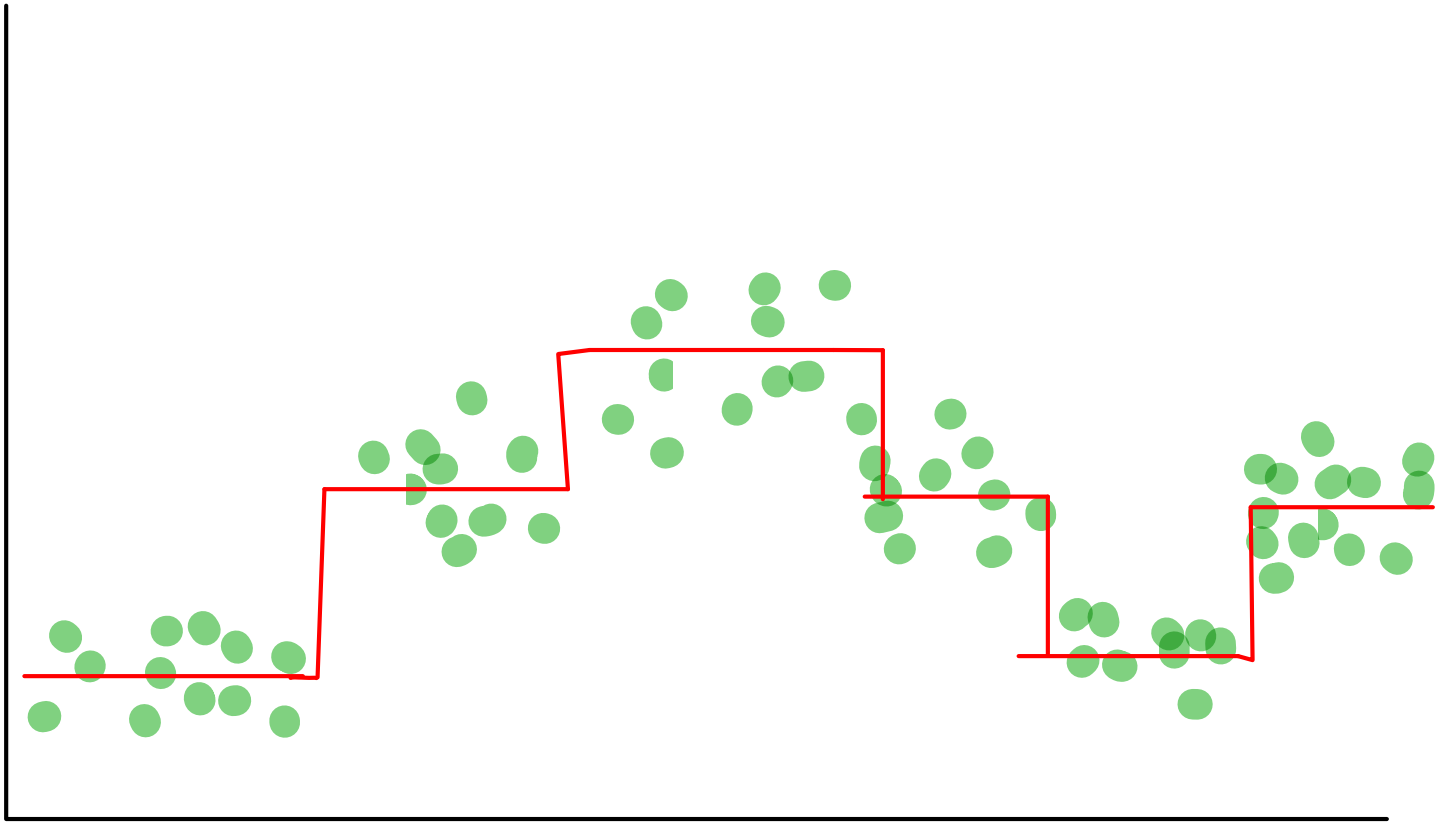


C_1 MSE_{Males} $= \frac{1}{2}\left[(3-4)^2+(5-4)^2\right] =$

C_2 $MSE_{Females}$ $\frac{1}{3}\left[(2-4)^2+(4-4)^2+(6-4)^2\right] =$

$MSE_{children} \Rightarrow \frac{2}{5} MSE_m + \frac{3}{5} MSE_f$

$IG = MSE_{parent} - MSE_{children\ (weighted)}$



Ensemble

↳ Group of things || → Models

↳ Multiple model → Used together ⇒ Answer.

iphone

Father ⇒ NO

Mother ⇒ YES

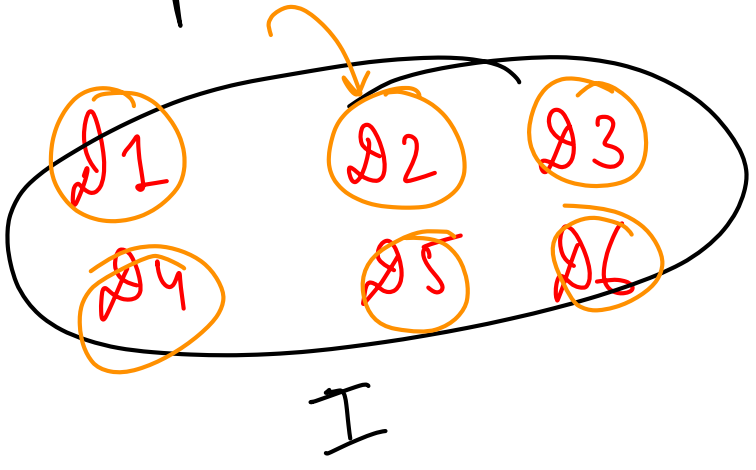
Sistu ⇒ YES

F1 ⇒ Y

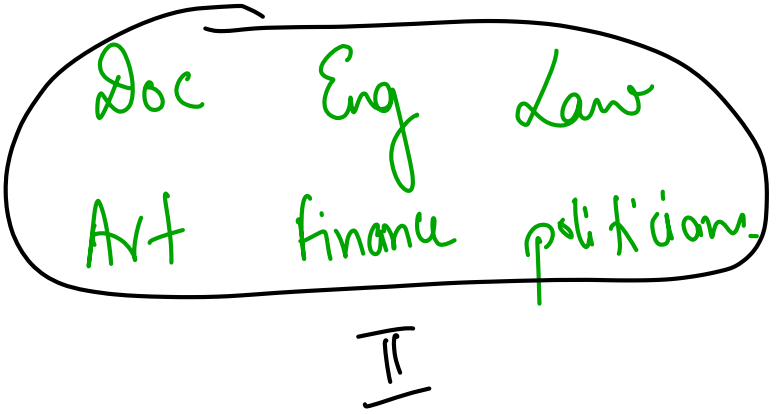
F2 ⇒ Y

F3 ⇒ Y

Implement a new cell related to medicine



I



II

$[m_1 \quad m_2 \quad m_3 \quad \ldots \quad m_k] \Rightarrow$ Base learner.

Combine $\Rightarrow$ final answer.

1. BAGGING
2. BOOSTING
3. CASCADING
4. STACKING

Quiz time!

Quiz Ended!

Which one of these is a type of ensemble learning technique?

29 users have participated

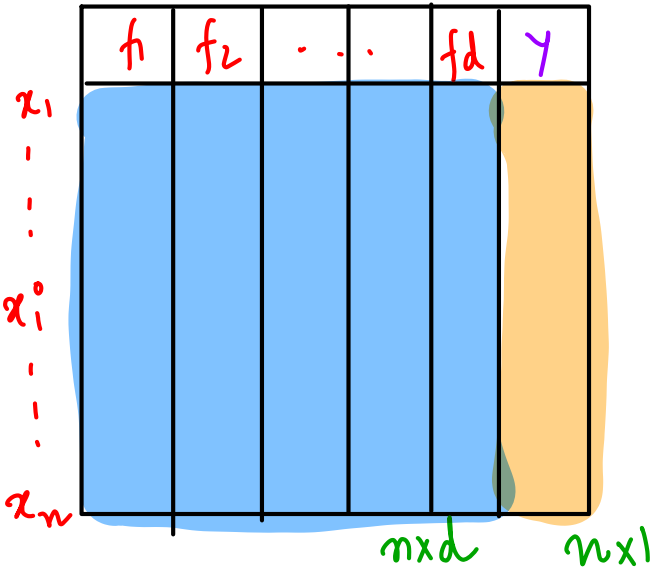
A	Selecting	0%
B	Collecting	7%
C ✓	Bagging	93%
D	Randomizing	0%

BAGGING = BOOTSTRAPPED AGGREGATION

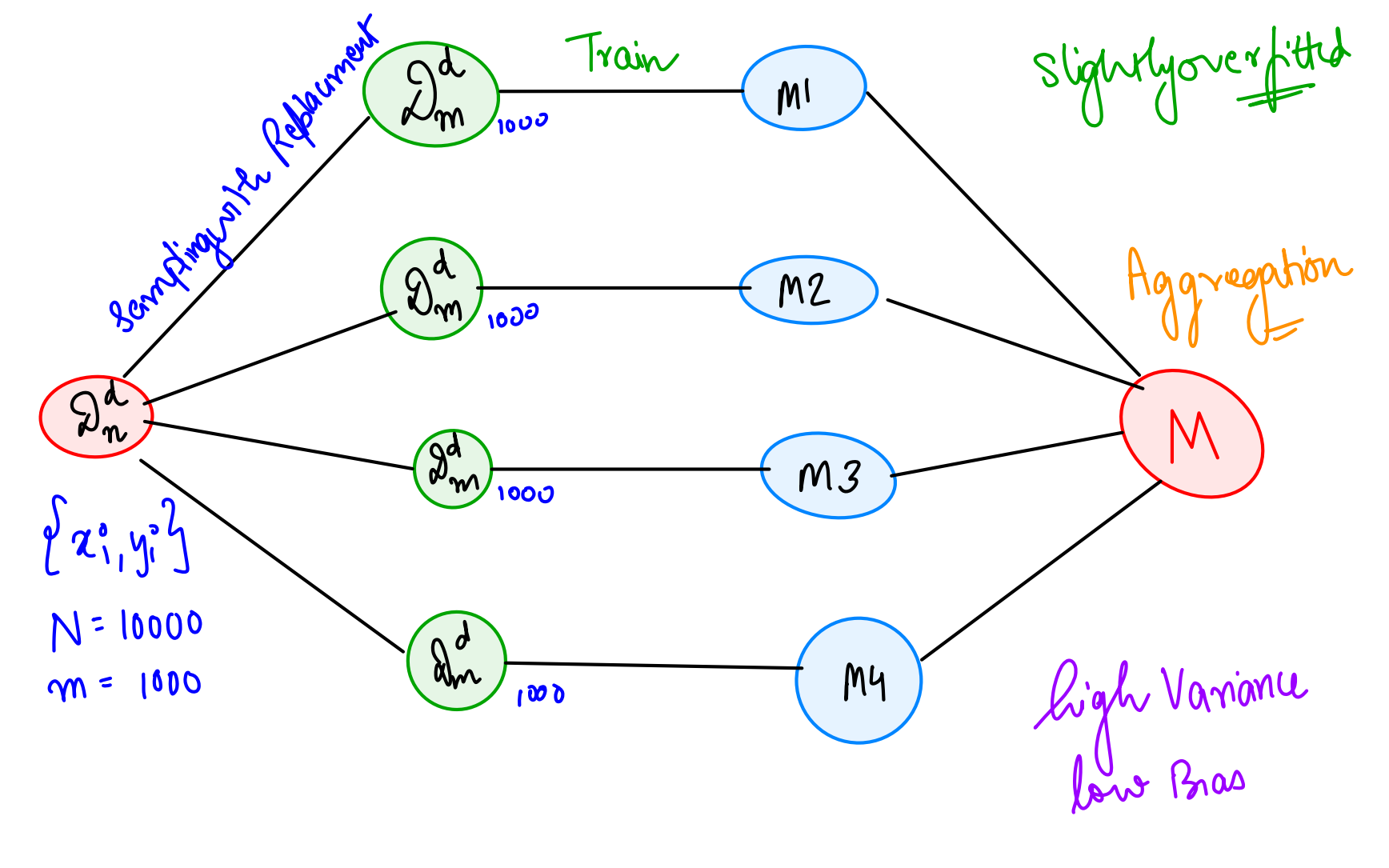
Bootstrapped Sampling → Sampling with replacement

Aggregation → Mean, Max, Min.

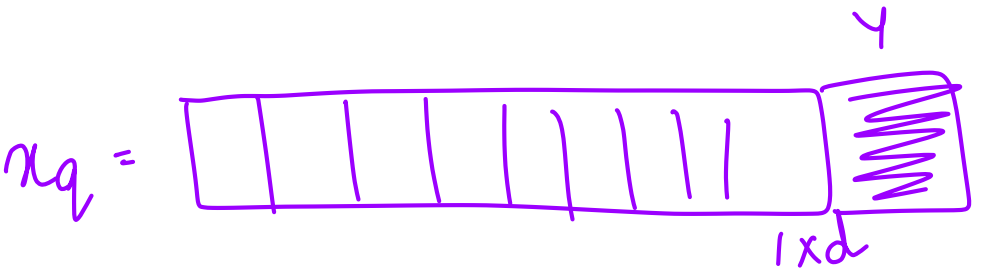
$\mathcal{D} =$

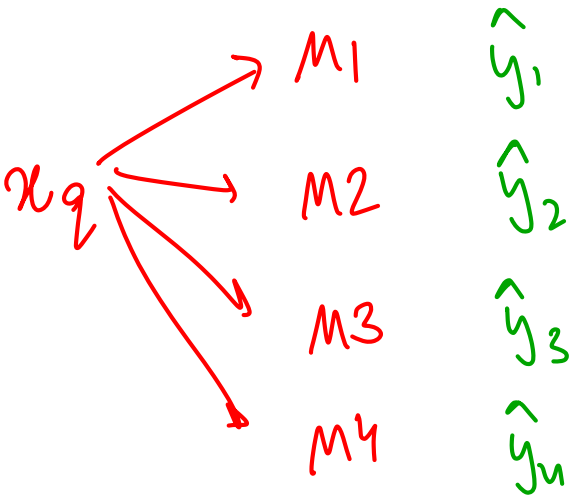


$\Rightarrow \mathcal{D}_n^d$


Sampling with Replacement
$\mathcal{D}_n^d$
$\mathcal{D}_m^d$
1000
Train
M1
M2
M3
M4
M
Aggregation
slightly overfitted
$\{x_i, y_i\}$
N = 10000
m = 1000
High Variance
low Bias

At Inference

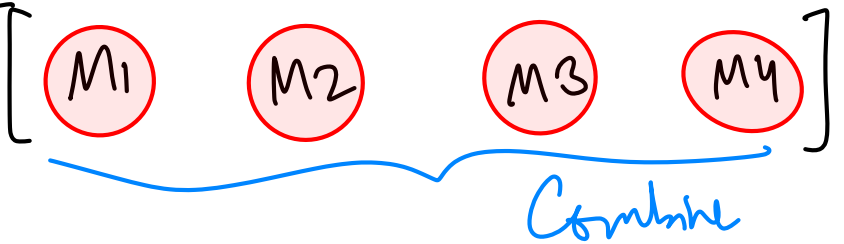




Classification ⇒ Majority Voting

Regression ⇒ Mean / Median.

BAGGING

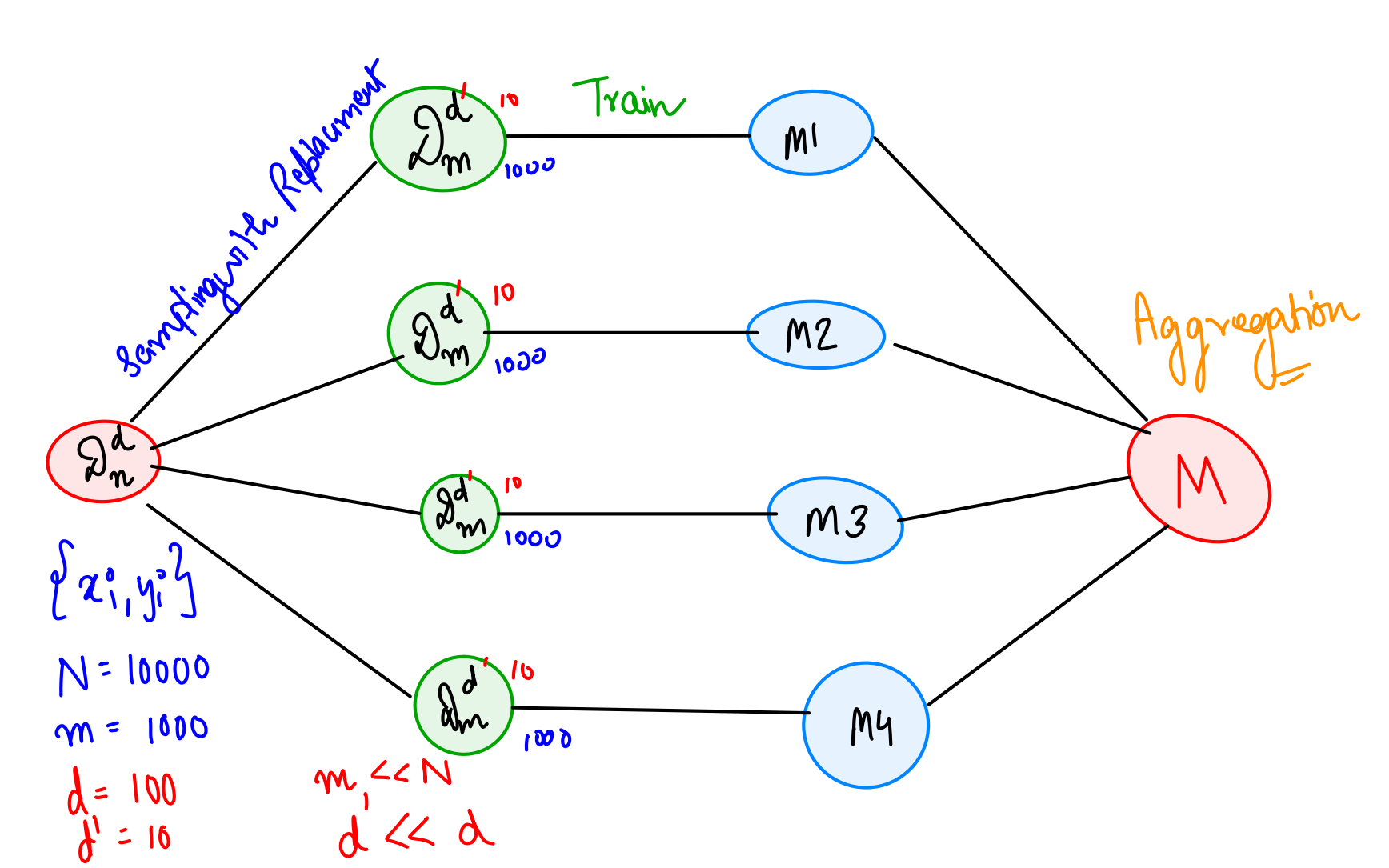


overfitted Variance ↑ Bias ↓

RANDOM FOREST ⇒ Collection of trees ⇒ decision trees.

① Randomly Select Row [Row Sampling]

② Randomly select Columns [Column Sampling]


Sampling with Replacement
Train
10
1000
M1
M2
M3
M4
M
Aggregation
$\{x_i, y_i\}$
N = 10000
m = 1000
d = 100
d' = 10
$m_1 << N$
$d << d$

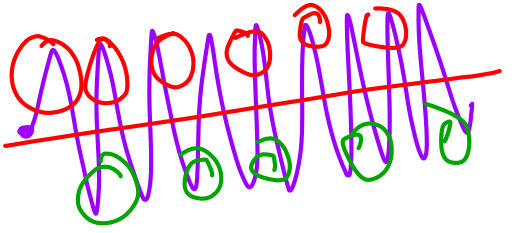
RF = DT + R.S + C.S + Aggregation

Base Learner

$m \ll N$ $\quad d' \ll d$.

BIAS ↓

VARIANCE ↑



'M' ⇒ Low BIAS + Low VARIANCE

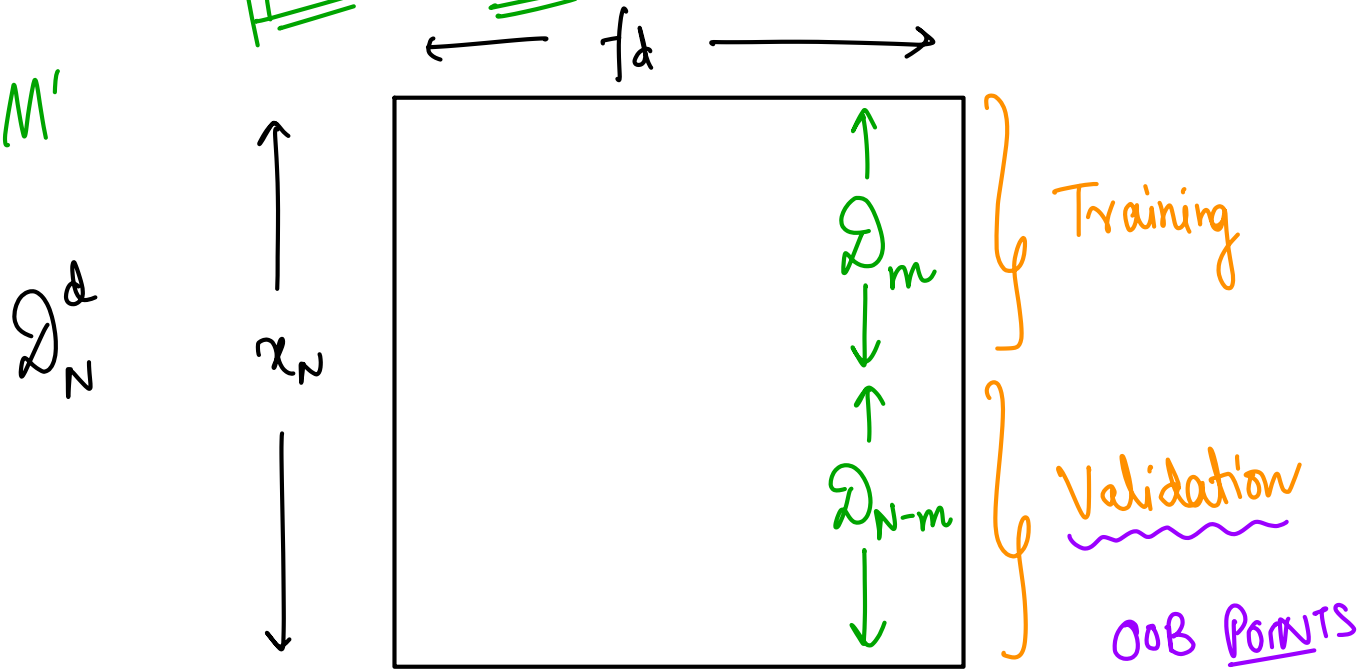
Aggregation + Randomisation

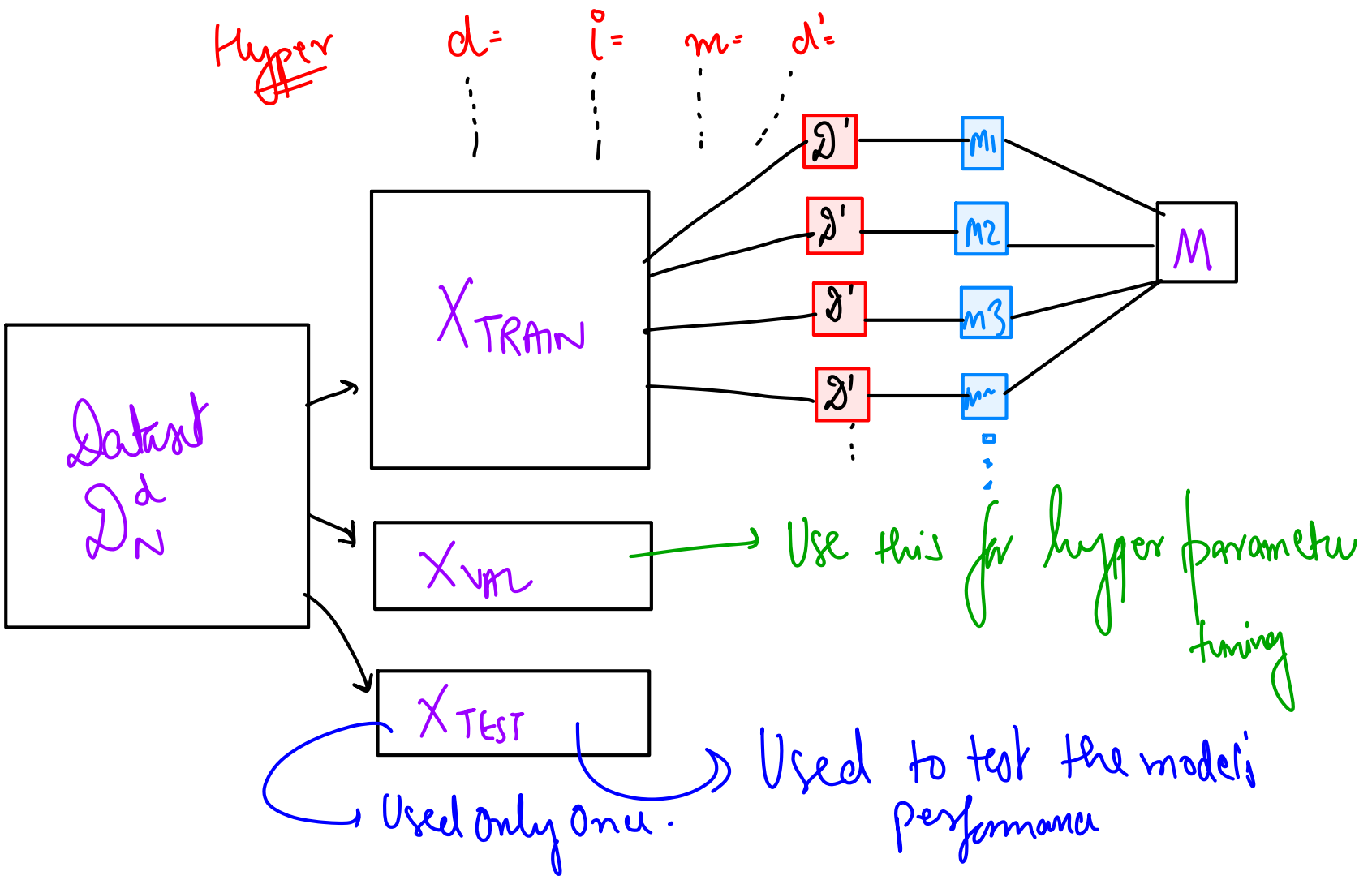
How do we validate RF ?

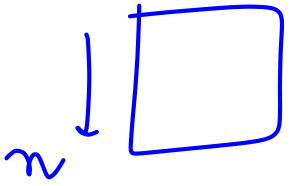
D = Tr (M) Val Test

⇒ K → different model

⇒ 'M'

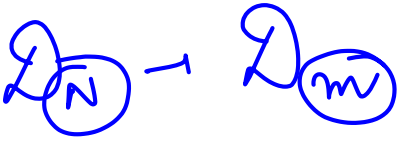



Hyper
d=
i=
m=
d'=
Dataset
D_N^d
X_{TRAIN}
X_{VAL}
X_{TEST}
D'
D'
D'
D'
m1
m2
m3
M
Use this for hyper parameter tuning
Used only once.
Used to test the model's performance



Quiz time!

Quiz Ended!



If a dataset contains "n" rows, and "m" of these rows are sampled to train the base learners in Random Forest, what will be the cross-validation data for each of the models?

21 users have participated

	A	Complete dataset with "n" rows	5%
	B	A part of "m" sampled rows	19%
✓	C	Remaining "n-m" rows after sampling	76%
	D	None of the above.	0%