Random Forest CLASSIFICATION AND REGRESSION

$d' < d$

$f_1 \, f_2 \, f_3 \, f_4 \text{ -- } f_m$

$\leftarrow \text{ --- } m \text{ --- } \rightarrow$

$\{f_1, f_2, f_3, f_4\}$

d' + FS

RS

Row SAMPLING + Feature Sampling

Dataset

d

200

New Test Data

RS replacement + FS

RS + FS

RS + FS

M1   DT1 $\longrightarrow$ O

M2   DT2 $\longrightarrow$ O

M3   DT3 $\longrightarrow$ O

$\vdots$

M_n $DT_n \rightarrow 1$

CLASSIFICATION
$\Downarrow$
Majority Voting
Classifier

Regression
$\Downarrow$
Average

Time Complexity $\uparrow$
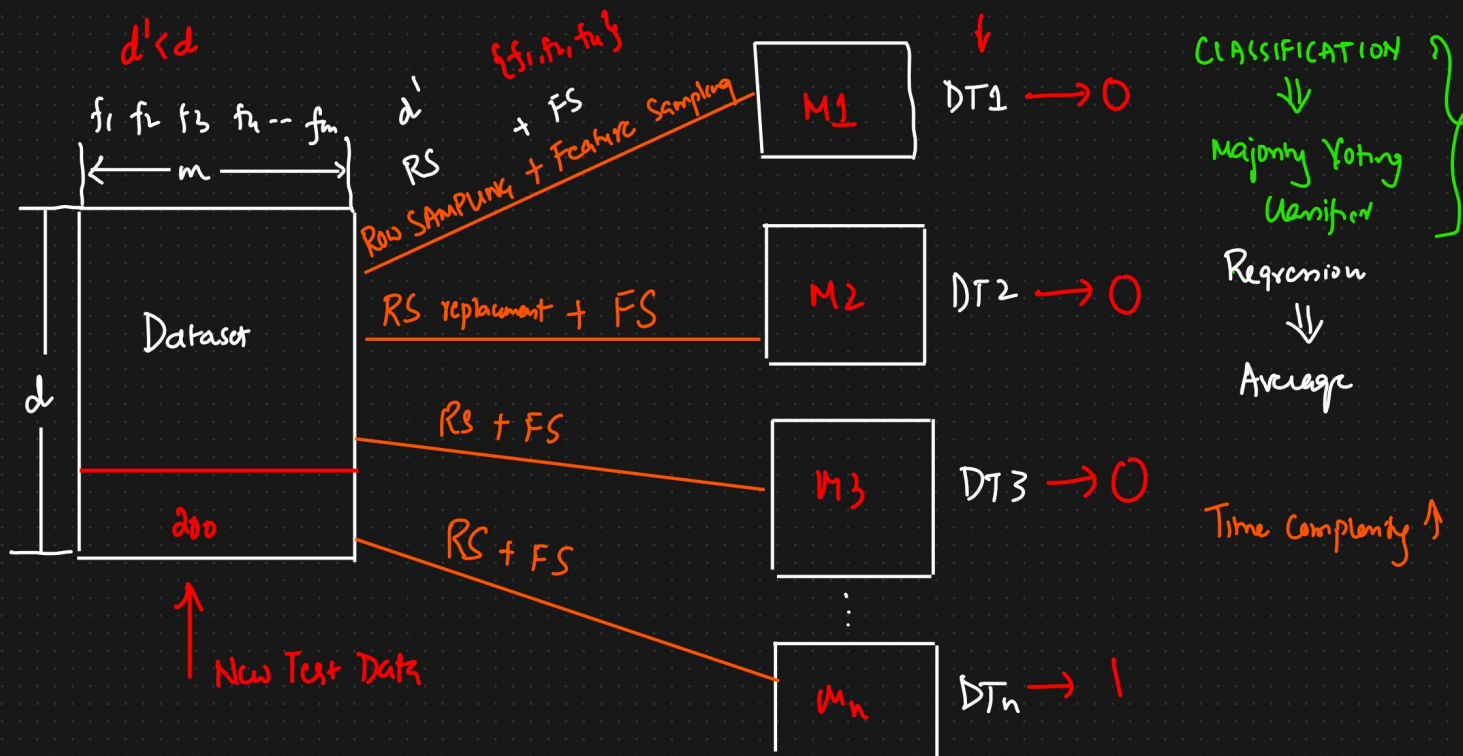
CLASSIFICATION - MAJORITY VOTING CLASSIFIER

REGRESSION - Average O/P of the Models.

Why should we use Random Forest instead of DT ?

Decision Tree

Generalized Model $\begin{bmatrix} \rightarrow \text{low Bias} \\ \rightarrow \text{low Bias} \end{bmatrix}$

Overfitting

Train Acc $\uparrow$ $\longrightarrow$ Low Bias $\longrightarrow$ Low Bias

Test Acc $\downarrow$ $\longrightarrow$ High Variance $\longrightarrow$ Low Variance

Random Forest

Out of Bag Evaluation (Random Forest) Out of Bag Score?

Bagging

$d' < d$
$m' < m$

Row Sampling + FS
with Replacement
d', m'

DT1

DT2

DT3

DT4

n

d DATASET

$\frac{1}{3} * d$

Out of Bag

Validation dataset
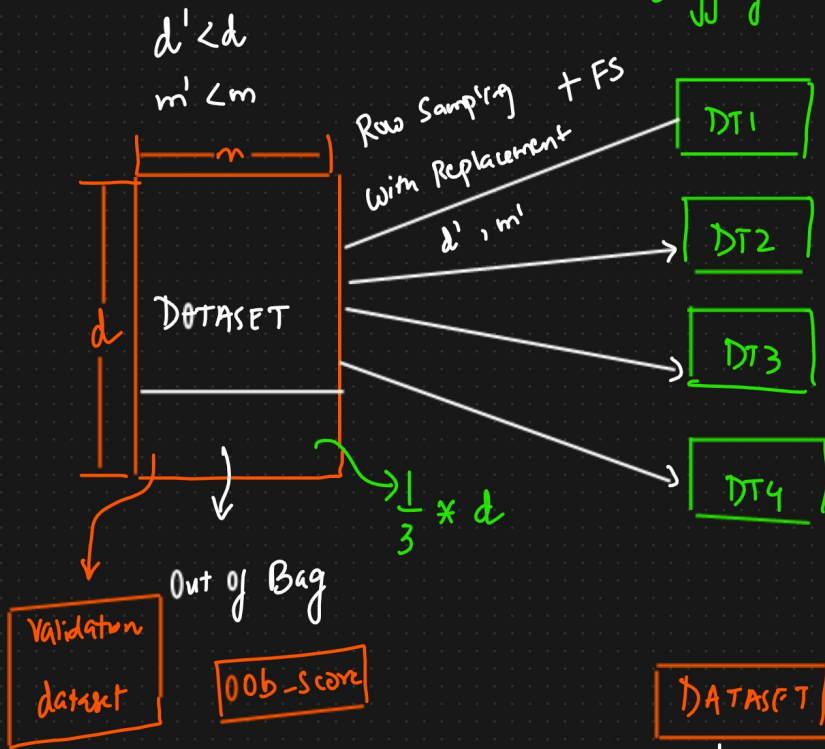
Oob-score

Sklearn

Oob = TRUE

Oob Score $\Rightarrow$ Percentage

Validation of Out of Bag DATA

DATASET

TRAIN

TEST

1000

TRAIN

Validation

$\Rightarrow \dfrac{2}{3} * d$

$\Rightarrow \dfrac{1}{3} \times d$

$\dfrac{2}{3} * 1000$

$\dfrac{1}{3} \times 1000$