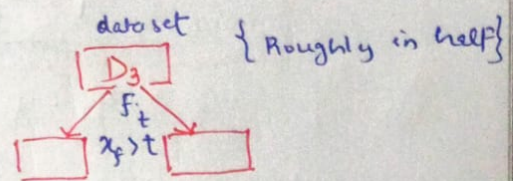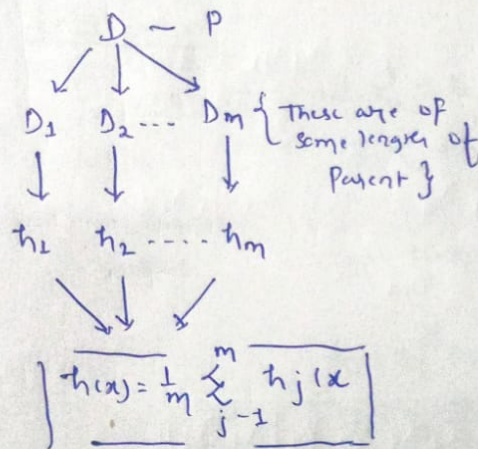✏️

# Random Forest

A
**Random Forest** is a machine learning algorithm used for both classification and regression tasks. It works by creating multiple decision trees during training and then combining their results to make a final prediction.

- **Decision Tree**: This is like a flowchart where data is split into smaller and smaller groups based on certain rules (e.g., "Is this value greater than X?"). Each split helps make a decision.

- **Random**: In Random Forest, we randomly choose a subset of data points and a subset of features (columns) for each decision tree. This randomness helps create trees that are different from each other.

- **Forest**: Instead of relying on just one decision tree, the algorithm creates many decision trees (a forest). The final prediction is made by taking the average of all the decision trees' predictions (for regression) or by a majority vote (for classification).

## Day 4: Random Forest

Random Forest is a commonly-used Machine learning algorithm Trademarked by Leo Breiman and Adele Cutler, that combines the output of Multiple decision trees to reach a single Result. Its ease to of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

Random forest: { feature Selection }

$$D \sim P$$

$D_1 \quad D_2 \cdots \quad D_m$ { These are of some length of Parent }

$h_1 \quad h_2 \cdots h_m$

$$h(x) = \frac{1}{m} \sum_{j-1}^{m} h_j(x)$$

dataset { Roughly in half }

$D_3$
$f_t$
$x_f > t$

$k << $ dimension

only split on $k << d$ features

$k = \lceil \sqrt{d} \rceil$
↳ hyper parameter

[ training ]  [ Validation ]

→ Here we don't need to do this we can directly calculate test error
↳ training or Validation split

$$\underset{loss}{L(x)} = \frac{1}{m} \sum_{j=1}^{m} \ell(h_j(x), y)$$

ex. case
$D = [a, b, c]$
$D_1 = [b, c, b]$

So we are calculating loss other dataset didn't contain a.

Out of Bag Error

$$R_{OOB} = \underbrace{\frac{1}{n} \sum_{i=1}^{n}}_{\text{Average}} \underbrace{\frac{1}{z_i} \sum_{j}}_{\substack{(x_i, y_i) \neq \emptyset}} \ell \left( h_j(x_i), y_i \right)$$

<span style="color:red">That are not trained in dataset</span>

$$\boxed{\begin{array}{l} z_i = \sum 1 \\ (x_i, y_i) \neq \emptyset \end{array}}$$