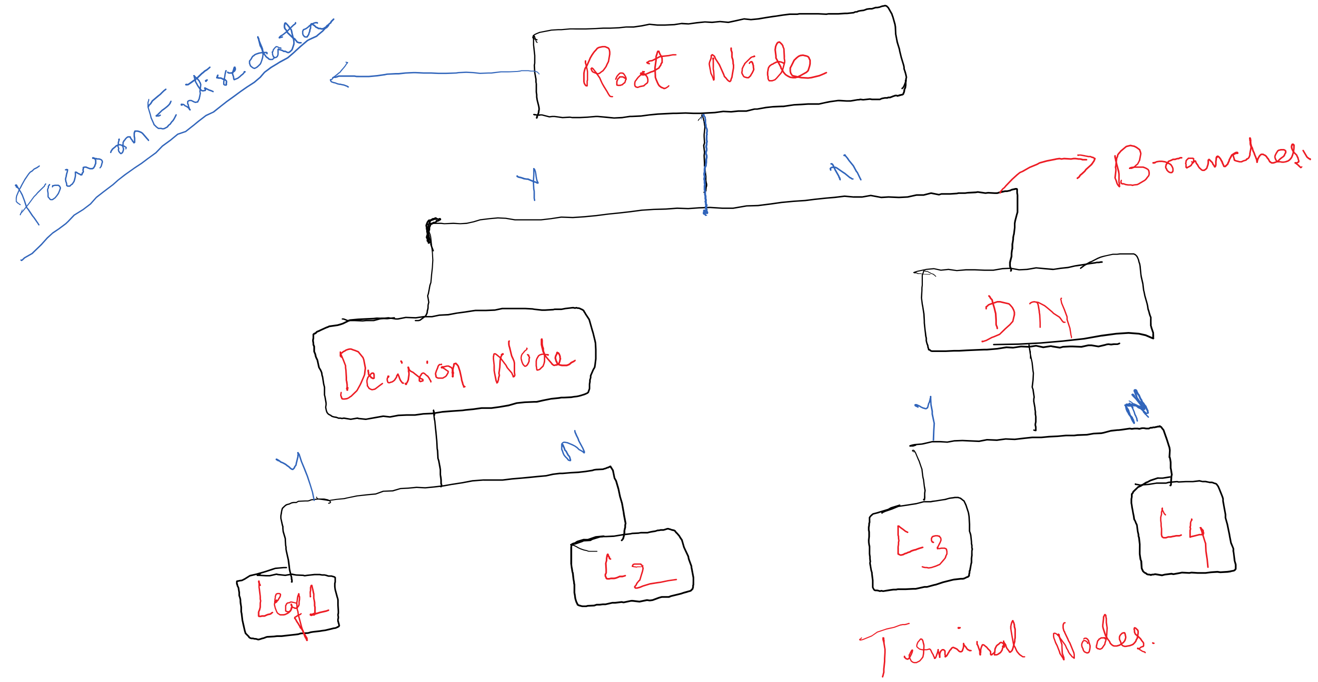
Decision Trees → Divide & Conquer → Most widely used

→ Makes Complex choices from simple set of choices & represent their learning in logical form



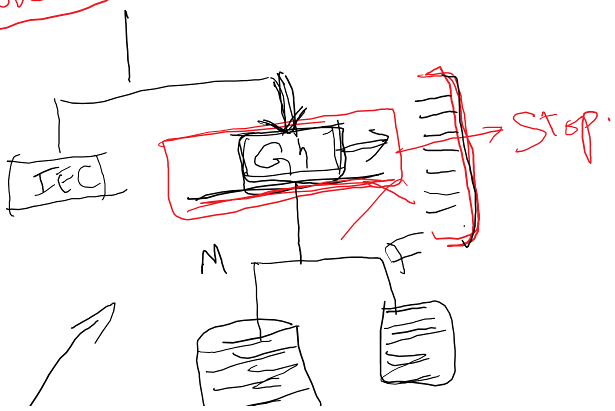
Recursive Patitioning → Enabling a machine to discover/learn something on its own

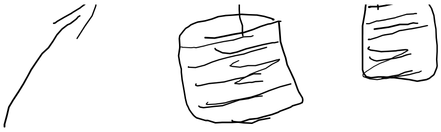
→ Splits → Data into Homogenious parts → IN/US

→ Terminates → All nodes have same class.

C.4.5 → Single thread → open Source

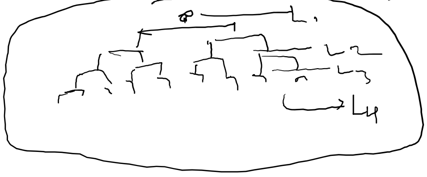
C.5.0 → Multi thread → Premium

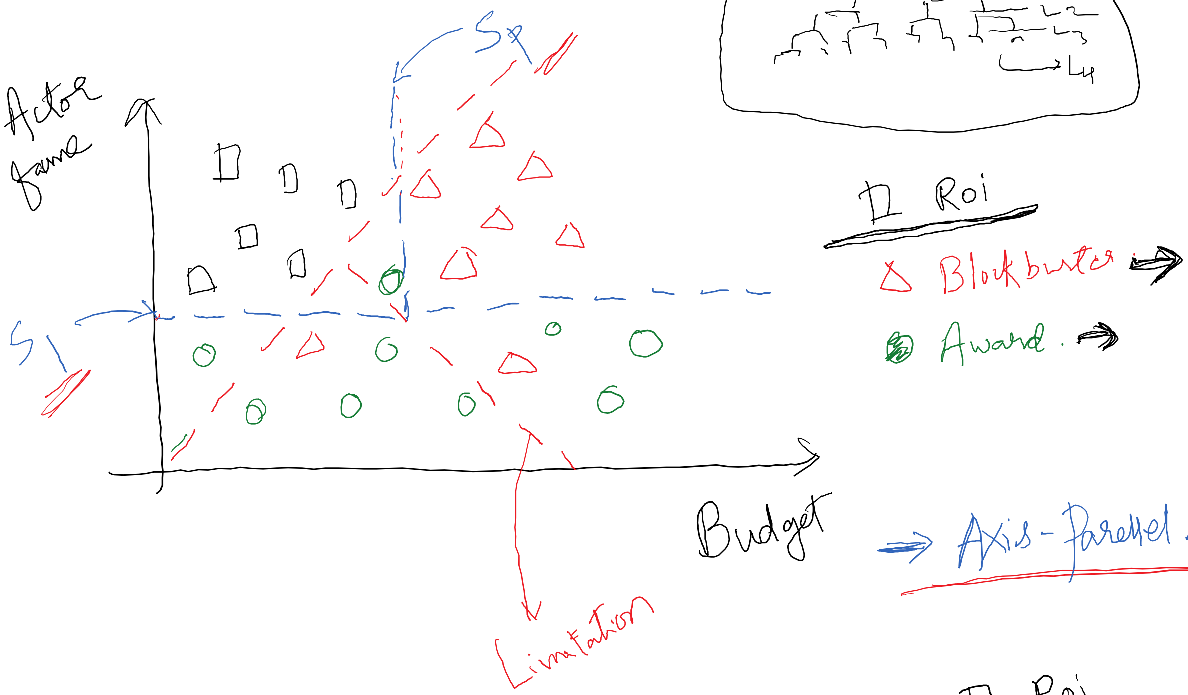




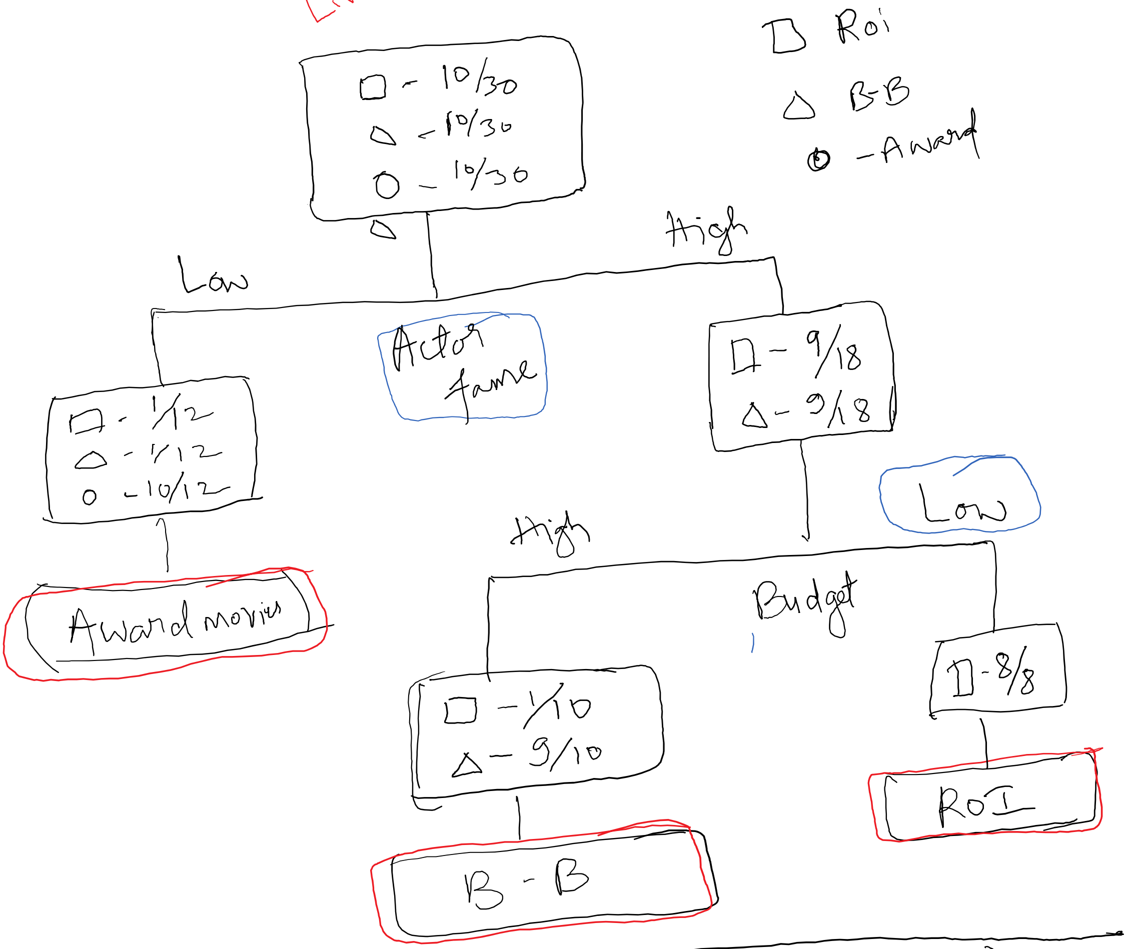
→ No features remaining

→ Has grown to a predicted limit → 4 levels





⇒ Axis-Parallel split



( C4.5 → Java based ( R- Package RWeka )

C.5.0 Algorithm / C.4.5 → Java based (R-package RWeka)

→ Degree to which the subset contains Homogenous elements, "PURITY"

→ All are homogenous → "PURE" class

→ C.5.0 uses $\boxed{\text{Entropy}}$ → Quantifies the Randomness of a set

Entropy ↑ → Very Diverse

Entropy ↓ → Very "PURE"

2 possible → Entropy btn 0 & 1

n — " — " — " — 0 to $\log_2(n)$

$$\text{Entropy}(S) = \sum_{i=1}^{c} -P_i \log_2(P_i)$$

S → Data segment

C → no of class levels.

$P_i$ → Proportion of values falling in a class.

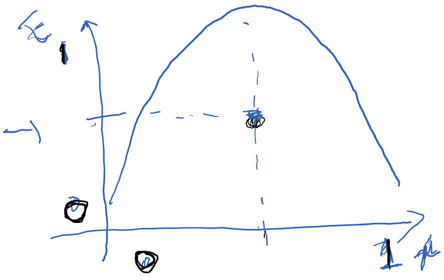
Ex. 2 classes → Red (60%), White (40%)

$$\text{Entropy} = -0.6 \times \log_2(0.6) - 0.4 \times \log_2(0.4) = 0.97 \uparrow$$

Not PURE

$\boxed{\text{Information gain} =}$ $E(S_1) - E(S_2)$

Info gain ↑ & Entropy ↓ →



→ Other forms of splits → Gini index, $\text{Chi}^2$, gain Ratio.

Pruning → Trim → Reduce the size to generalise the data.

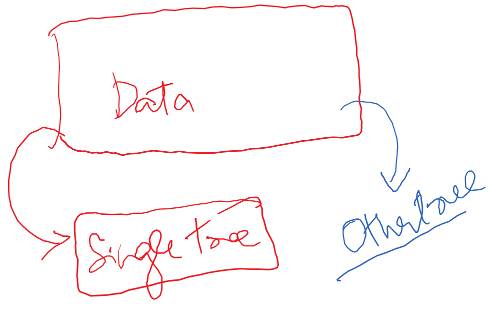
Pre pruning → Early Pruning → get result → stop.

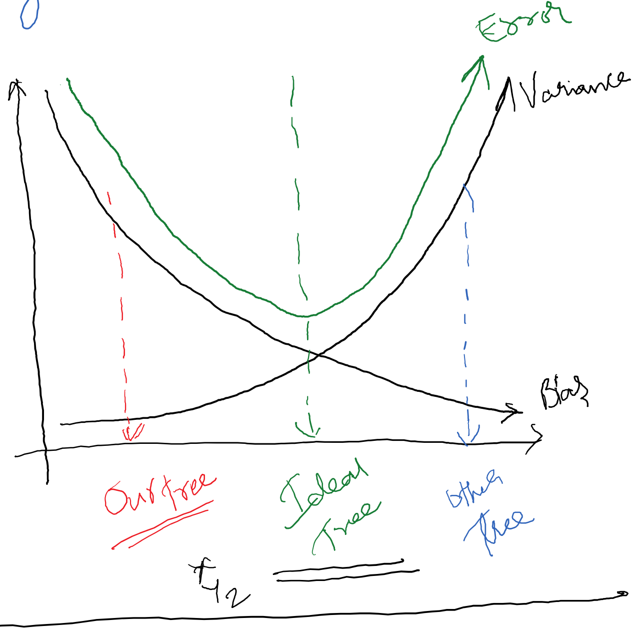
Post pruning → Allow the tree to grow → Based on need → Trim

## Bagging, Boosting & Random Forests

Ensemble method → Group methods.

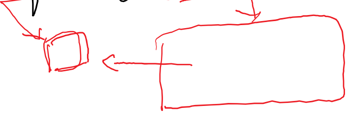
⊛ → Group of predictive models to achieve better Accuracy & Model stability → It will give a boost to tree models.

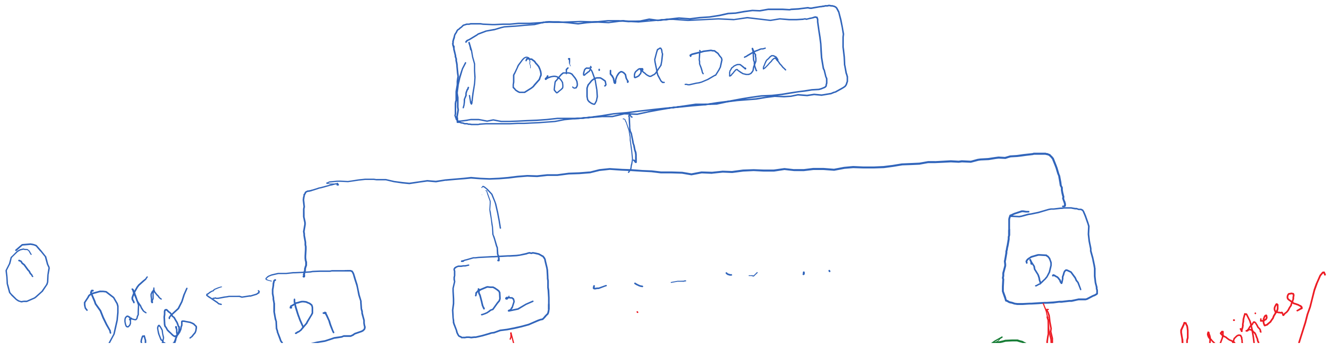


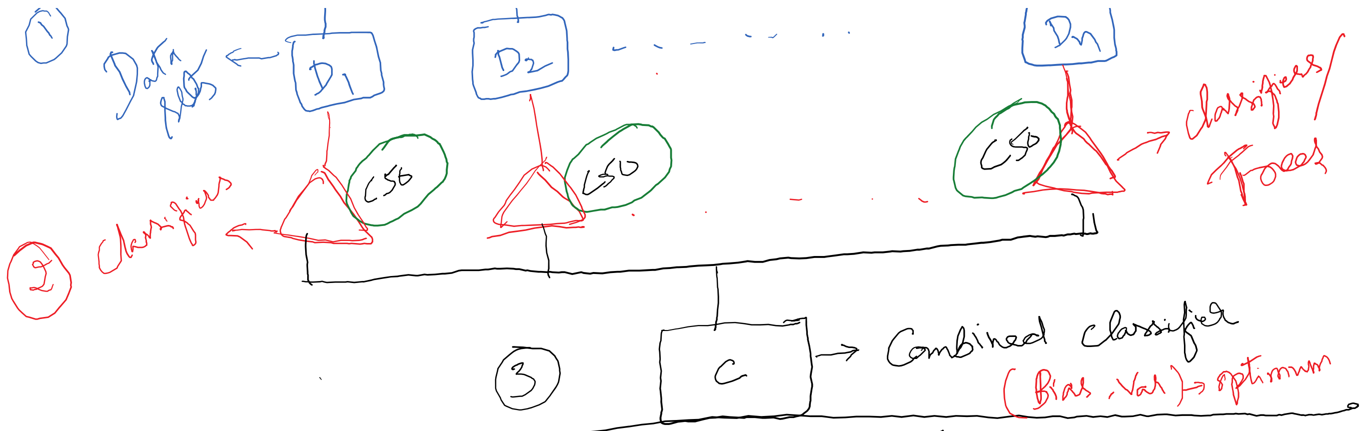


$E$ = Noise + Bias + Variance

Bagging Reduce Variance of the predictors by Combining results of Multiple classifiers on different Samples of same data.







① → Sample with Replacement

→ Subset selection, Shrinkage

② → Use same classifier.

C50, RPart, CART

③ → Combine

→ Mean — Mostly

→ Median

→ Modes

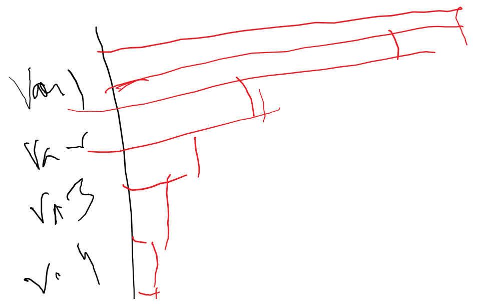
## Random Forest

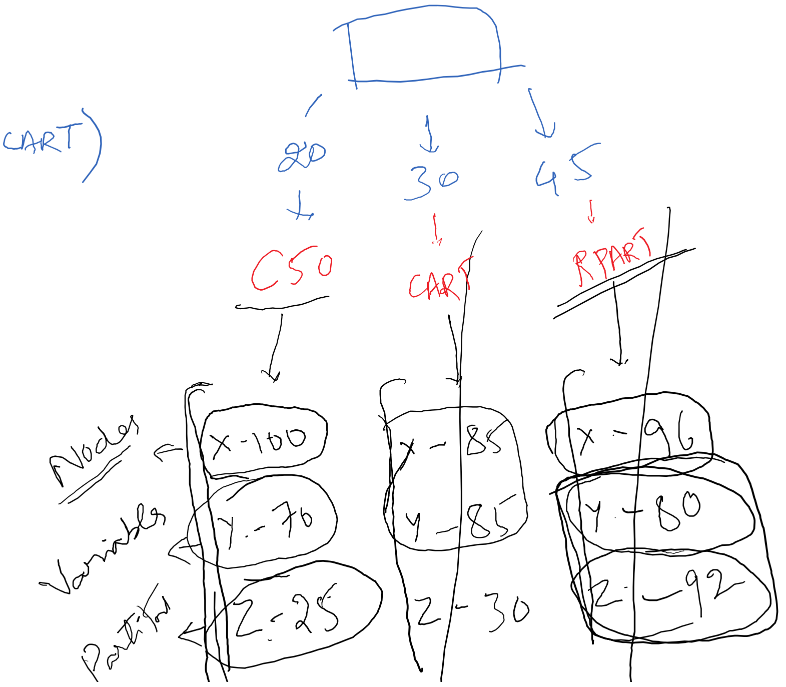
- → Apply → Regression & also classification
- → Undertakes → Dimensionality Reduction
- → It Treats Missing values, Outliers, Variable transformation
- → Gives the list of Important variables → Contributing %P.
- → CANNOT Control the mechanism

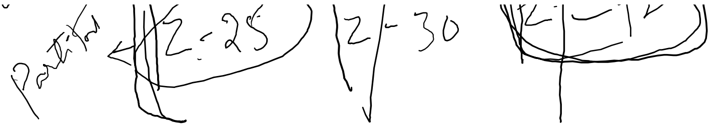
→ Multiple Trees → No pruning

→ Different methods → (C50, RPART, CART)

→ Voting →







→ N variables → M < N → M → Random variables

→ Allow The Trees to grow fully → No pruning

→ Aggregate data of those N Trees

**Boosting** → Refers To a family of Algorithms to make Weak learners To STrong learners.

(Iterative process)

XGBOOT
GBM

→ classify S/NS

→ Promotional Image → S

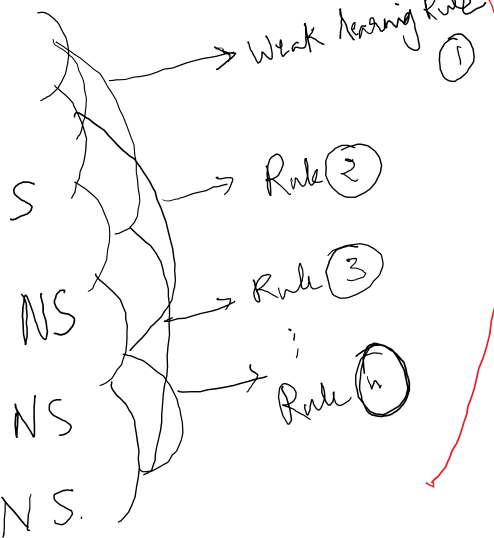
→ Only Hyperlink → S

→ You have won $ ... → S

→ IEC College → NS

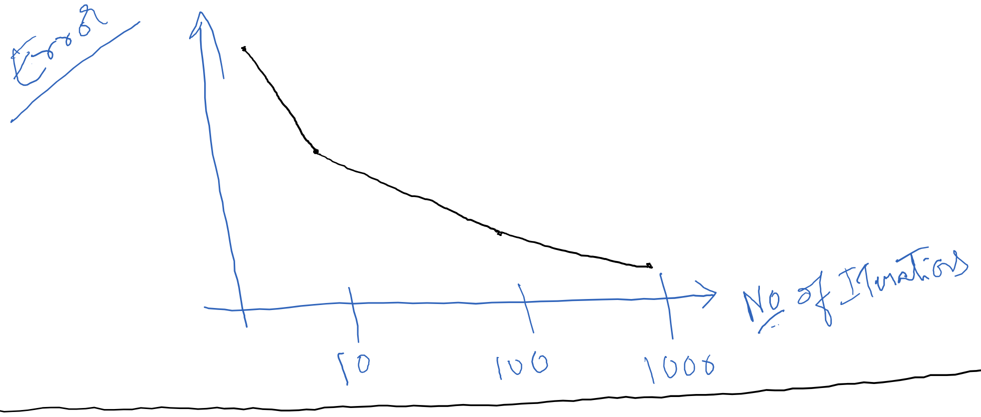
→ offer letter Company → NS

→ Bank e-statement → NS.



Strong Rule

No of Iterations


Error
No of Iterations