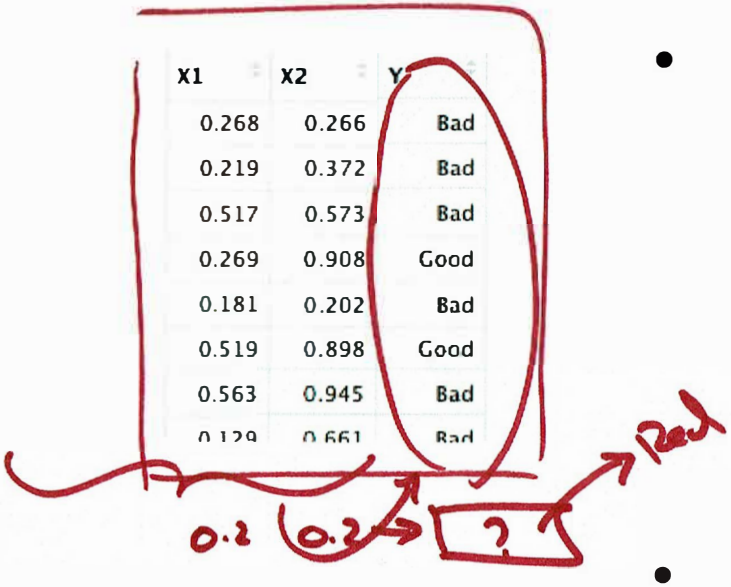
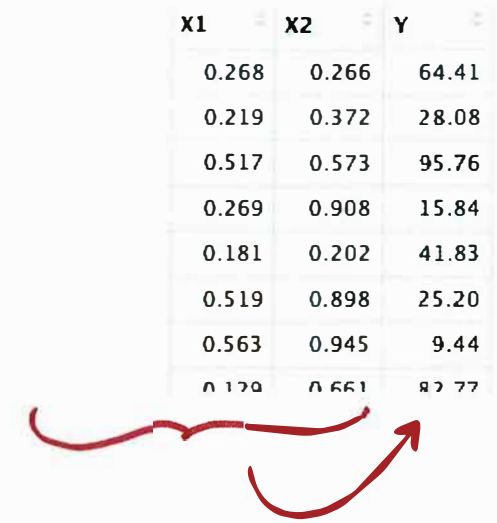
Classification and Regression

greatlearning
Learning for Life

- Decision Trees can be used for both

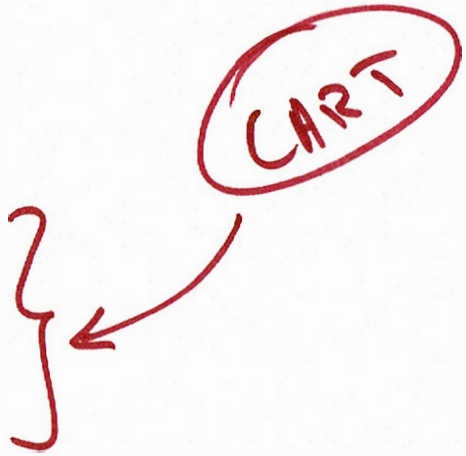


X1	X2	Y
0.268	0.266	Bad
0.219	0.372	Bad
0.517	0.573	Bad
0.269	0.908	Good
0.181	0.202	Bad
0.519	0.898	Good
0.563	0.945	Bad
0.129	0.661	Bad



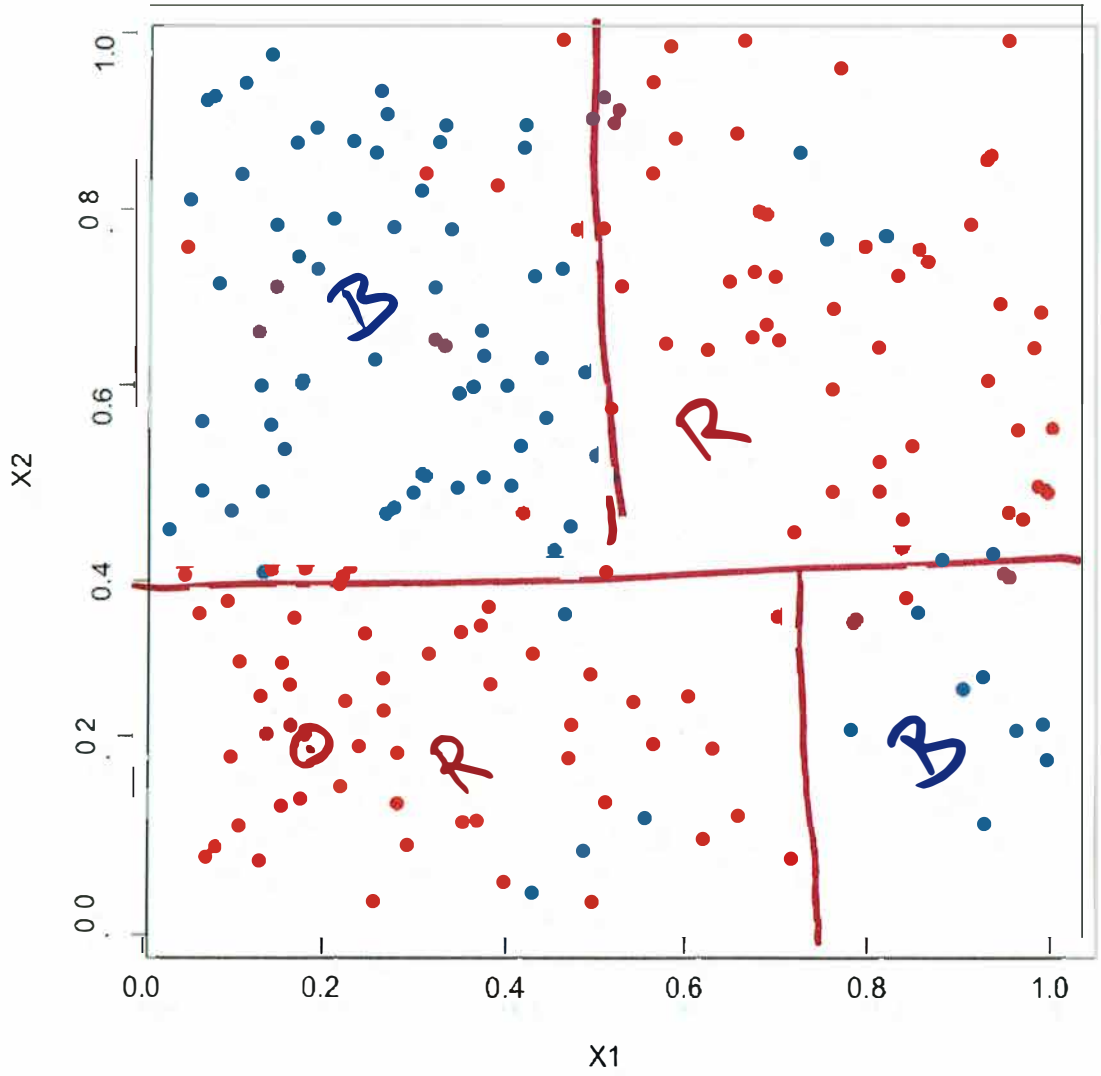
X1	X2	Y
0.268	0.266	64.41
0.219	0.372	28.08
0.517	0.573	95.76
0.269	0.908	15.84
0.181	0.202	41.83
0.519	0.898	25.20
0.563	0.945	9.44
0.129	0.661	82.77

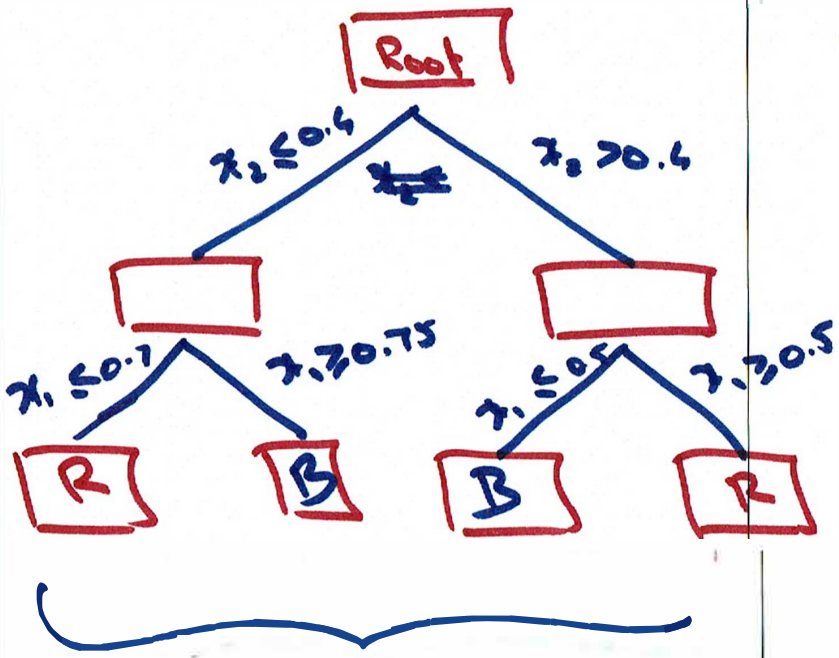
- Classification
  - Spam / not Spam
  - Admit to ICU /not
  - Lend money / deny
  - Intrusion detections



- Regression
  - Predict stock returns
  - Pricing a house or a car
  - Weather predictions (temp, rain fall etc)
  - Economic growth predictions
  - Predicting sports scores

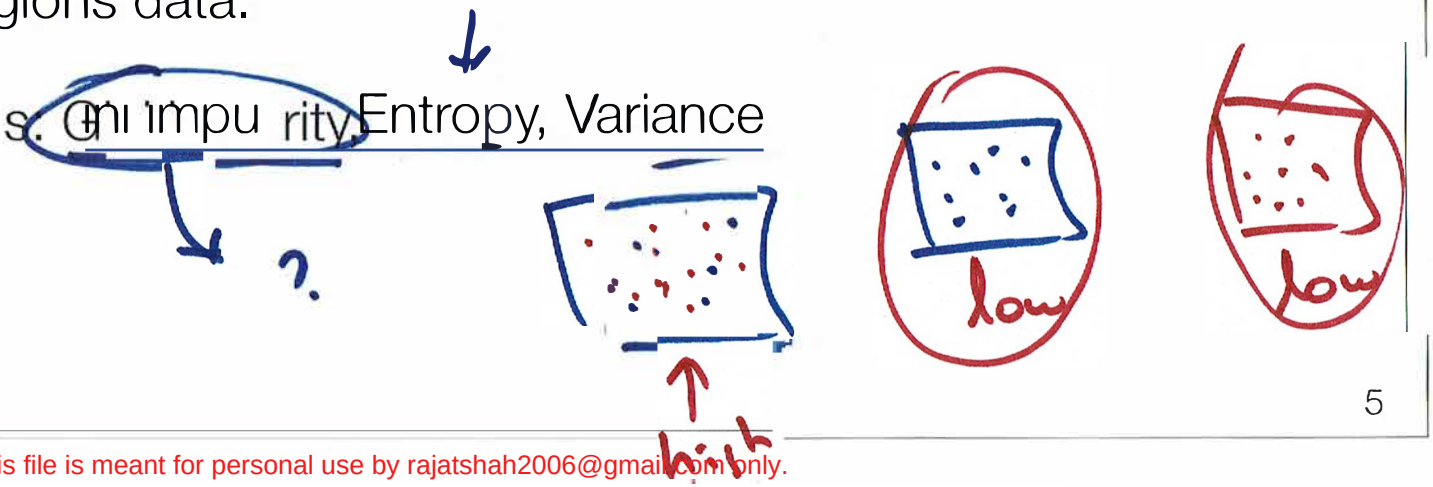
Visualizing Classification as a Tree





Metrics

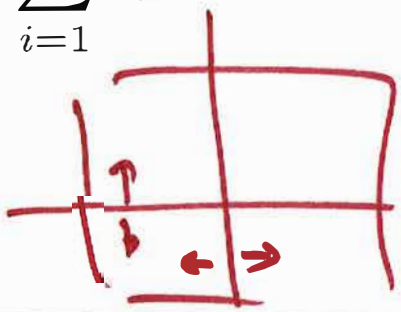
- Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items.
- Different algorithms use different metrics for measuring "best"
- These metrics measure how similar a region or a node is. They are said to measure the impurity of a region.
- Larger these impurity metrics the larger the "dissimilarity" of a nodes/regions data.
- Example: Gini impurity, Entropy, Variance

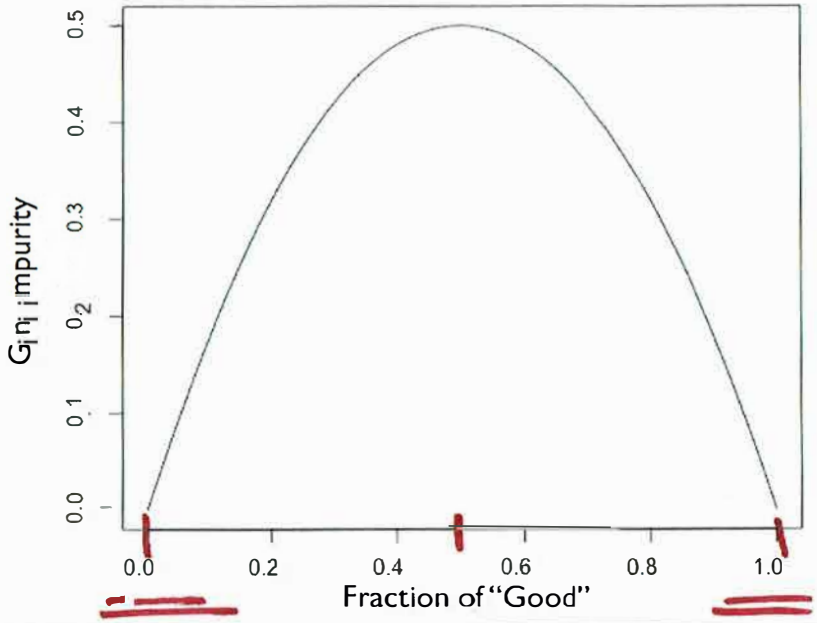


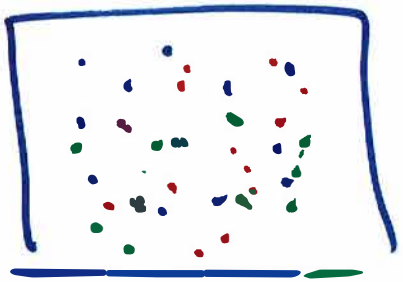
Gini impurity

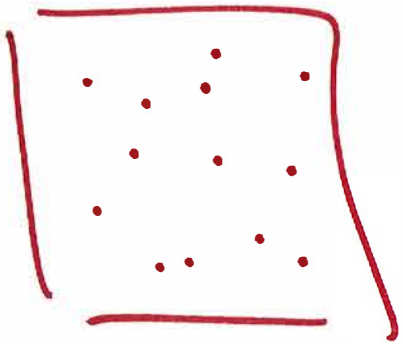
- Used by the CART
- Is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.
- Can be computed by summing the probability of an item with label i being chosen (p_i), times the probability of a mistake ($1 - p_i$) in categorizing that item.
- Simplifying gives, the Gini impurity of a set:

$$1 - \sum_{i=1}^{J} p_i^2$$









greatlearning
Learning for Life

P_1 → ① ⇒ $P_1(1-P_1)$ ← $P_1P_2 + P_1P_3$

P_2 → ② ⇒ $P_2(1-P_2)$ ← $P_2P_3 + P_2P_1$

P_3 → ③ ⇒ $P_3(1-P_3)$ ← $P_3P_1 + P_3P_1$

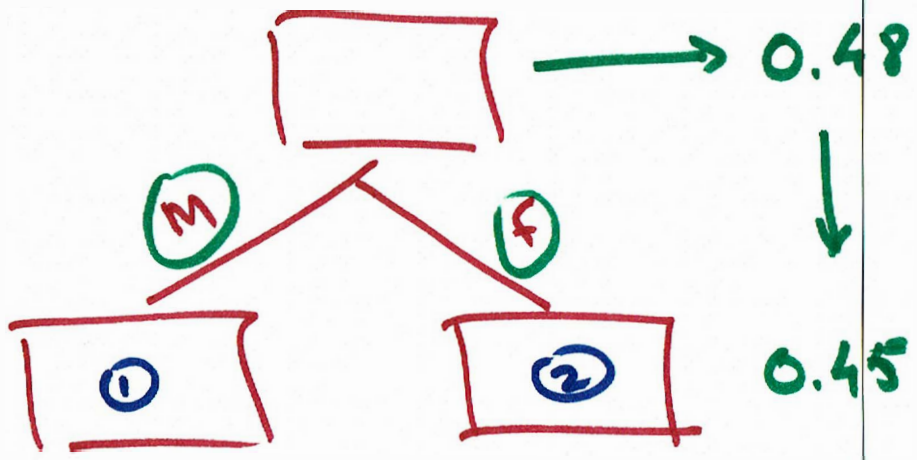
$$\Sigma P_i(1-P_i)$$

$$\Sigma P_i - \Sigma P_i^2 \Rightarrow 1 - \Sigma P_i^2$$

CART: An Example

greatlearning
Learning for Life

Cust_ID	Gender	Occupation	Age	Target
1	M	Sal	22	1
2	M	Sal	22	0
3	M	Self-Emp	23	1
4	M	Self-Emp	23	0
5	M	Self-Emp	24	1
6	M	Self-Emp	24	0
7	F	Sal	25	1
8	F	Sal	25	0
9	F	Sal	26	0
10	F	Self-Emp	26	0



Root node : $P_1 = 0.4$ $P_2 = 0.6$

$GI = 1 - (0.4)^2 - (0.6)^2$

$= 0.48$

1: $P_1 = 0.5$, $P_2 = 0.5$

$1 - 0.5^2 - 0.5^2$

$= 0.5$

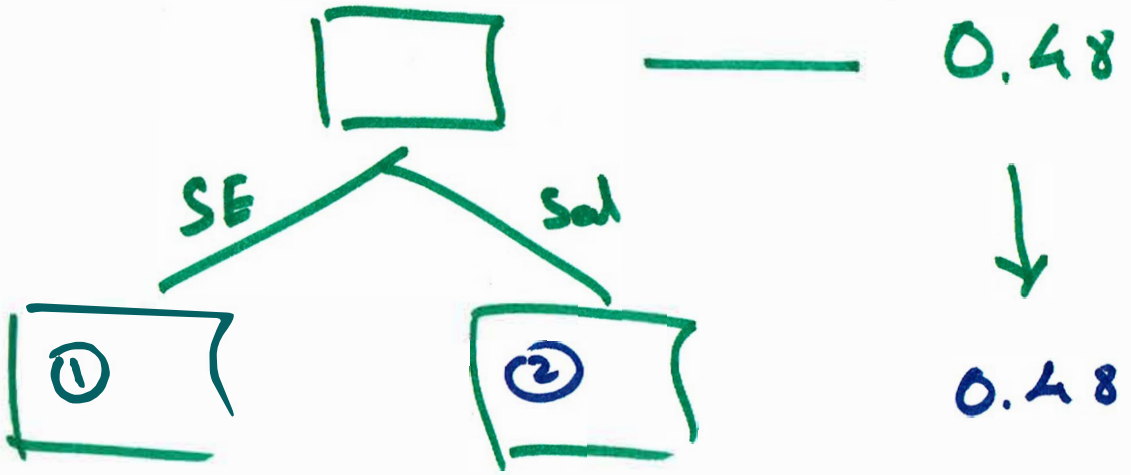
2: $P_1 = 0.25$, $P_2 = 0.75$

$1 - 0.25^2 - 0.75^2$

$= 0.375$

$GI = \frac{6}{10}(0.5) + \frac{4}{10}(0.375)$

$= 0.45$



① G.I = $1 - 0.4^2 - 0.6^2$

= 0.48

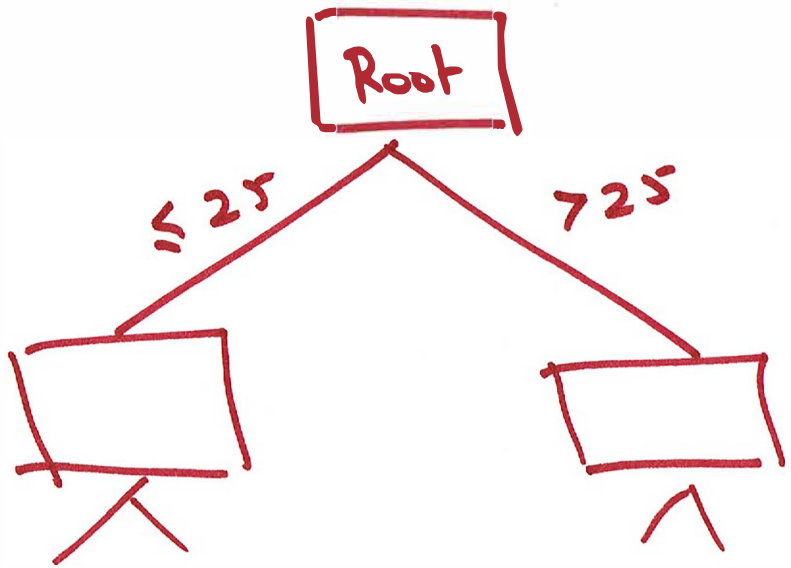
② G.I = $1 - 0.4^2 - 0.6^2$

= 0.48

$$G.I = \frac{5}{10}(0.48) + \frac{5}{10}(0.48) = 0.48$$

	left	right	Gini Split
≤22, >22	0.5	0.47	0.48
≤23, >23	0.5	0.44	0.47
≤24, >24	0.5	0.39	0.45
≤25, >25	0.5	0	0.40

Jain → 0.08

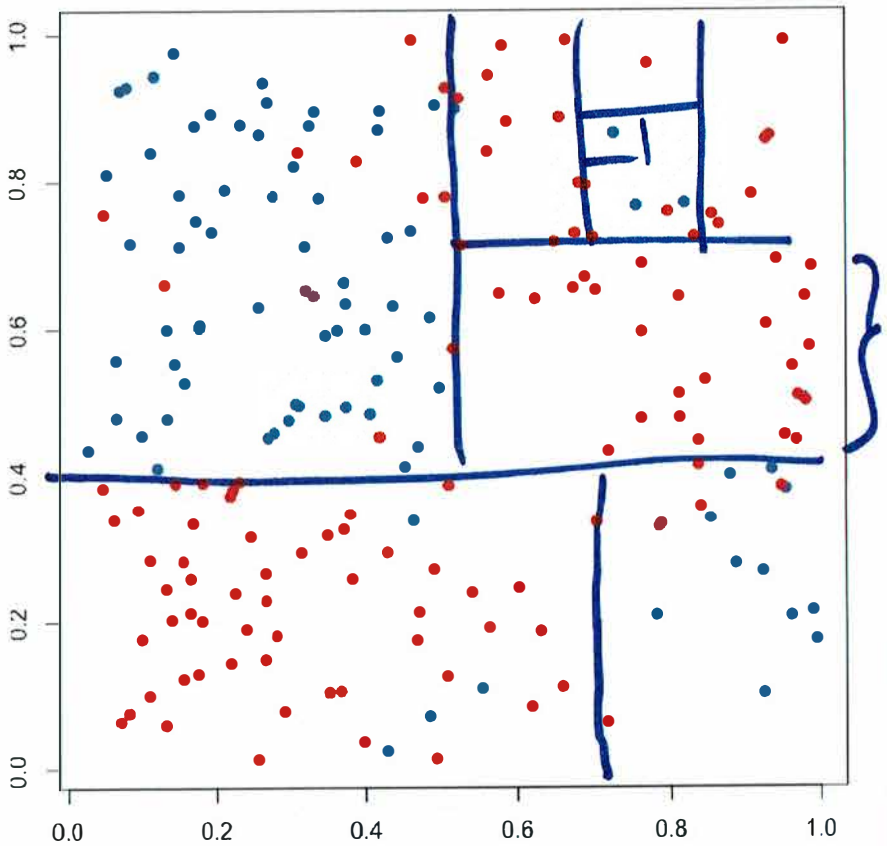


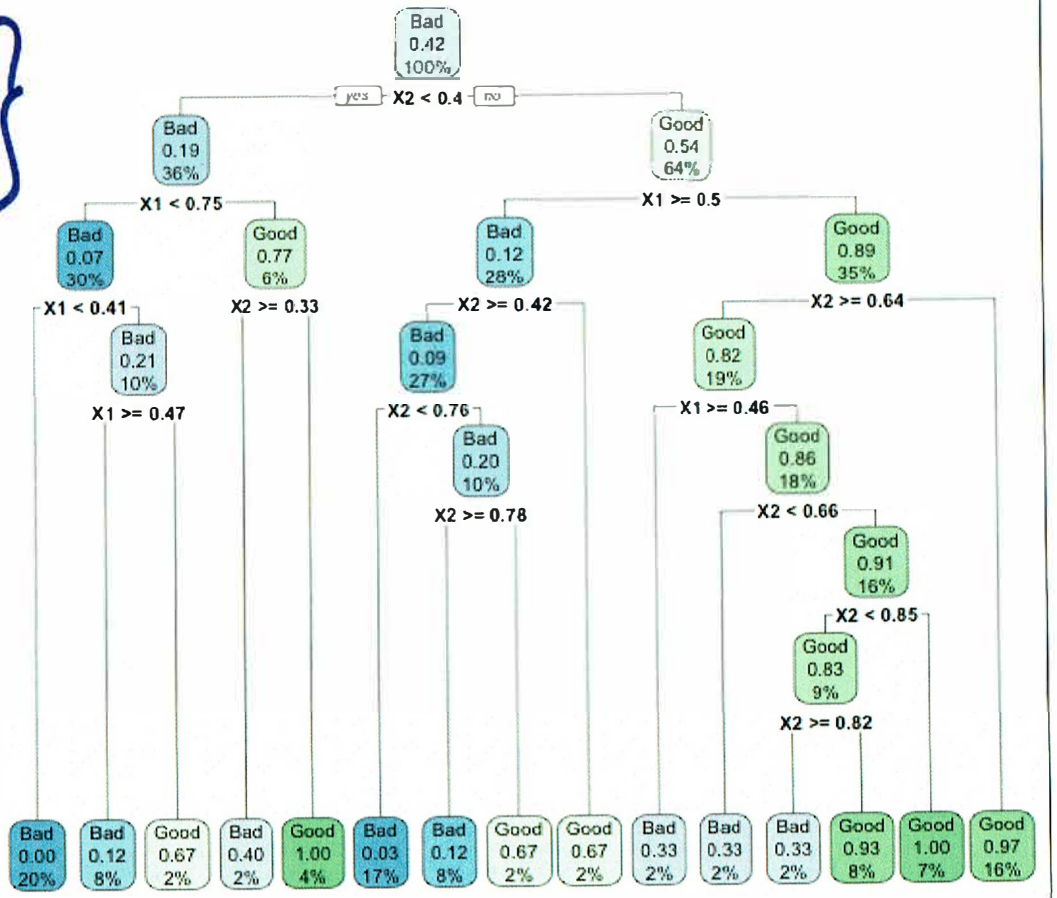
0.48

↓

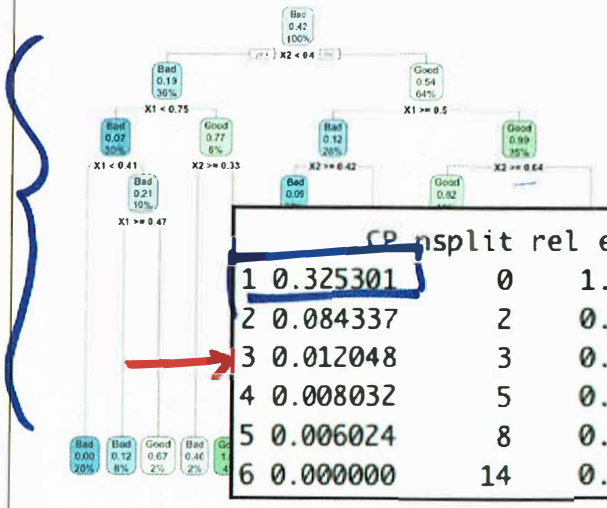
0.40

Overfitting in Decision Trees





Pruning



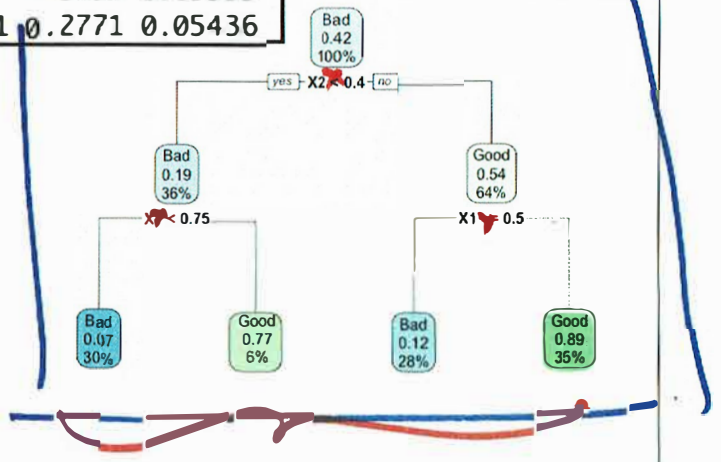
	CP	nsplit	rel error	xerror	xstd
1	0.325301	0	1.0000	1.0000	0.08395
2	0.084337	2	0.3494	0.3494	0.05999
3	0.012048	3	0.2651	0.2771	0.05436
4	0.008032	5	0.2410	0.3133	0.05730
5	0.006024	8	0.2169	0.3133	0.05730
6	0.000000	14	0.1807	0.4217	0.06474

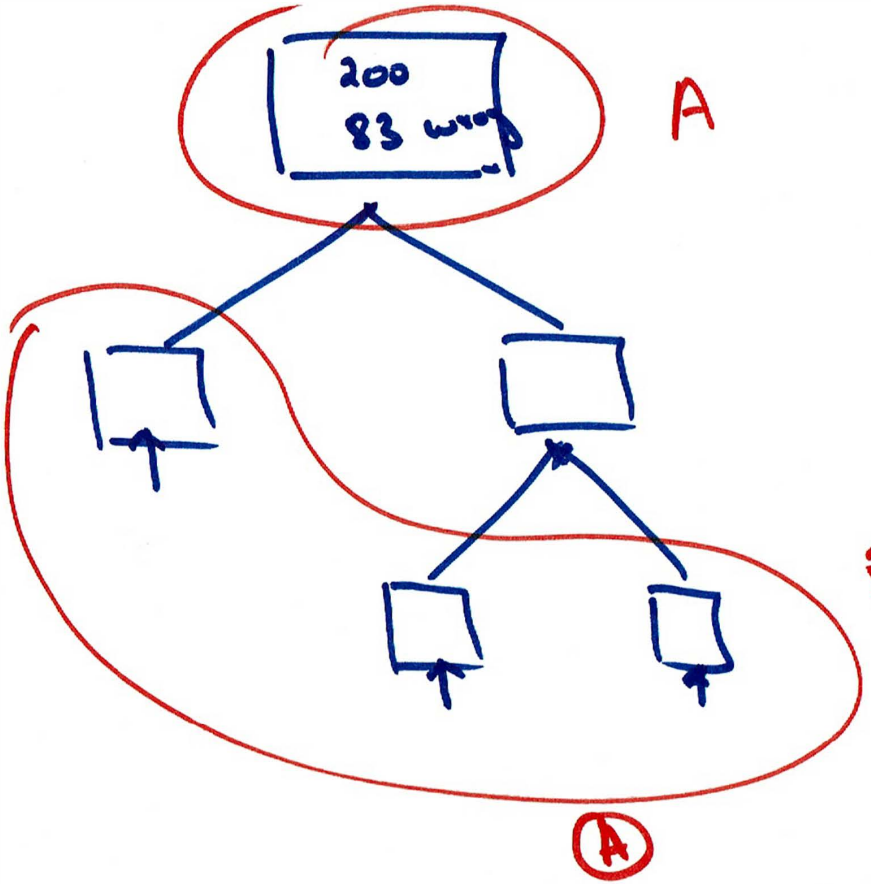
A

B

Prune with $\alpha = 0.15$

	CP	nsplit	rel error	xerror	xstd
1	0.325301	0	1.0000	1.0000	0.08395
2	0.084337	2	0.3494	0.3494	0.05999
3	0.012048	3	0.2651	0.2771	0.05436





A → 83

B → 29

(A) (B)

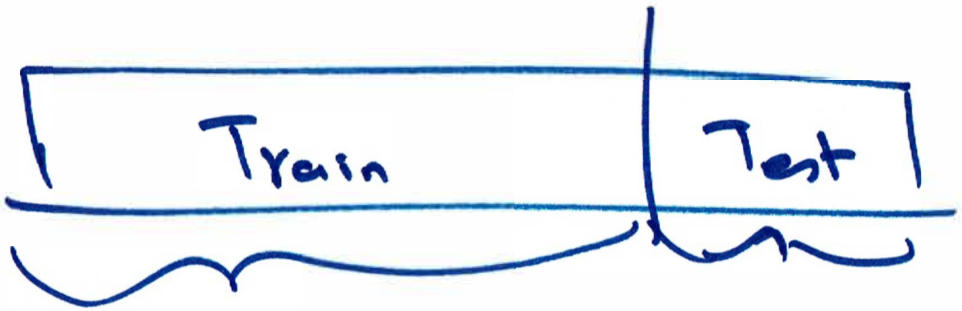
1 node → 3 nodes

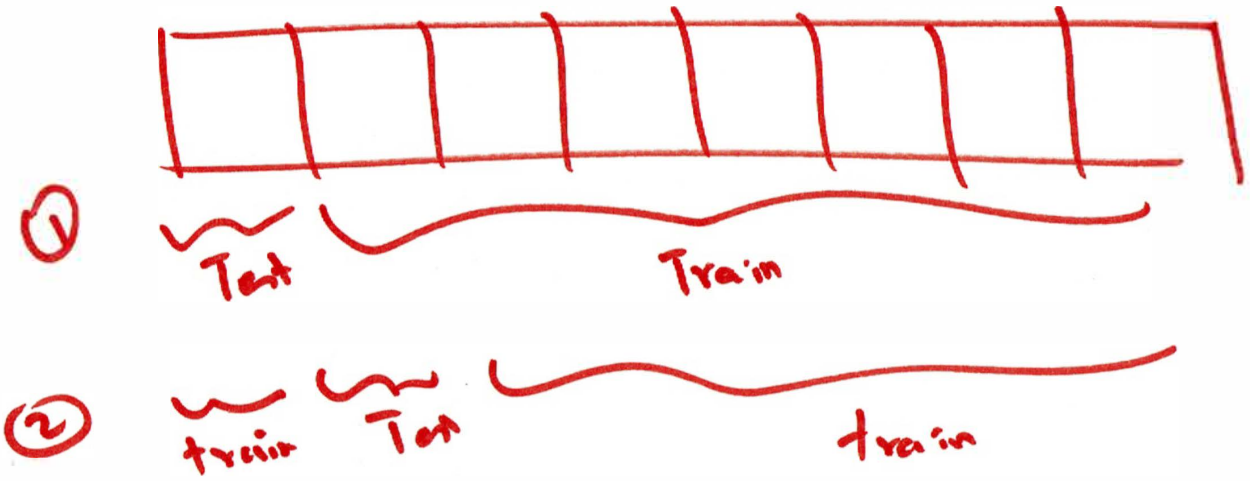
83 → 29

Rel. error dec = $\frac{54}{83} = 2\alpha$

$\alpha = 0.15$

(CP) $\Rightarrow \alpha = \frac{54}{83} \times \frac{1}{2} = 0.325$





Regression Trees

greatlearning
Learning for Life

	Price	Country	Reliability	Mileage	Type
Acura Integra 4	11950	Japan	Much better	NA	Small
Dodge Colt 4	6851	Japan	NA	NA	Small
Dodge Omni 4	6995	USA	Much worse	NA	Small
Eagle Summit 4	8895	USA	better	33	Small
Ford Escort 4	7402	USA	worse	33	Small
Ford Festiva 4	6319	Korea	better	37	Small
GEO Metro 3	6695	Japan	NA	NA	Small
GEO Prizm 4	10125	Japan/USA	Much better	NA	Small
Honda Civic 4	6635	Japan/USA	Much better	32	Small
Hyundai Excel 4	5899	Korea	worse	NA	Small
Mazda Protege 4	6599	Japan	Much better	32	Small