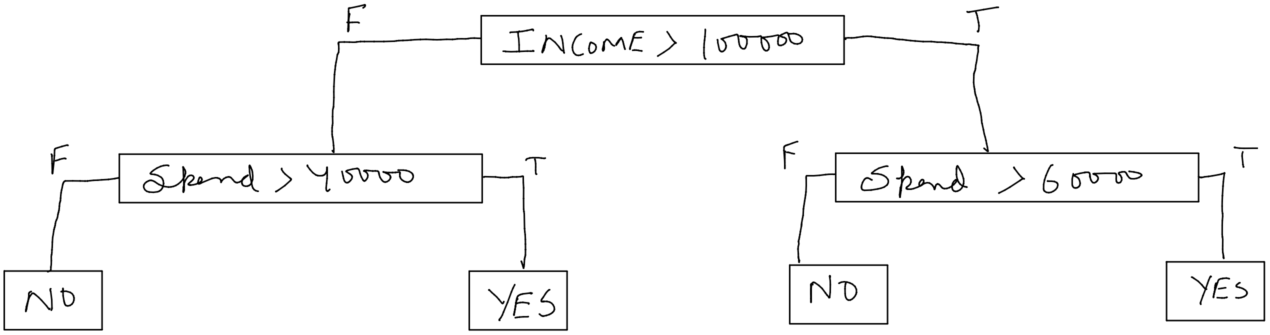
Decision Trees

2 January 2024 08:13 AM

- Decision Trees recursively partition the dataset into multiple segments.
- Each internal node corresponds to a test or a condition.
- Each branch is a result of the test.
- Each leaf node assigns a category.

Ex 1

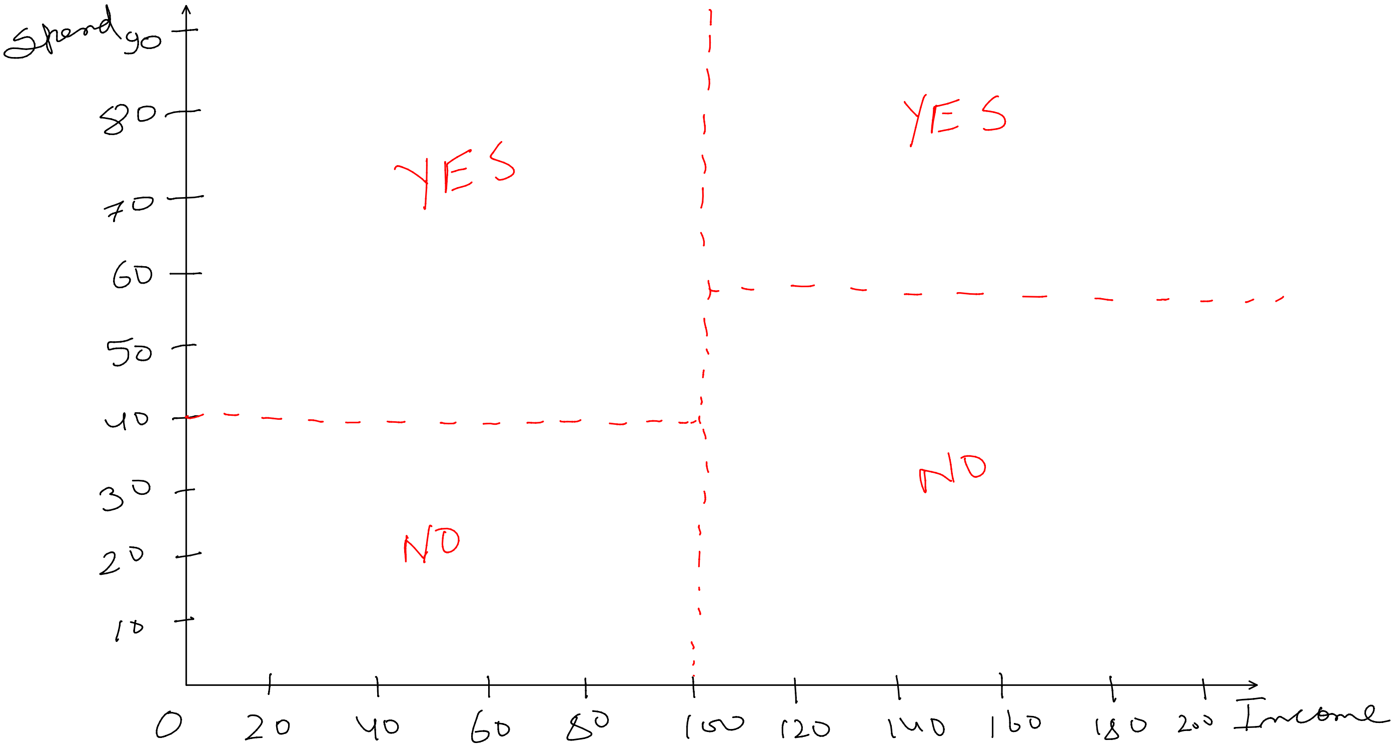
Income	Spend	Loan Default
–	–	Yes
–	–	NO
–	–	NO



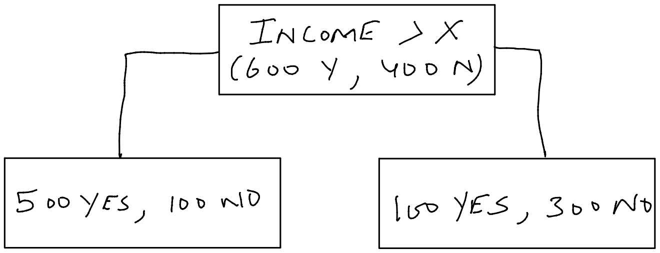
i) Income : 95000, Spend : 50000 YES

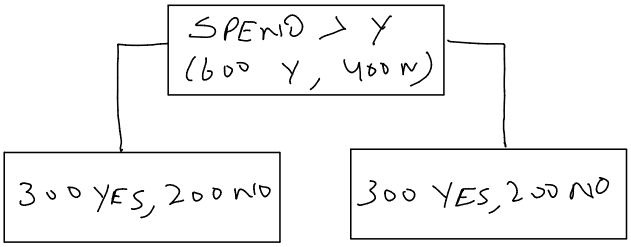
ii) Income : 120000, Spend : 50000 NO

iii) Income : 100000, Spend : 45000 YES



- The algorithm works by recursively partitioning the dataset into multiple segments.
- At each step the algorithm selects the most predictive feature to split the data.
- The most predictive feature is the one which gives maximum separation between the classes.





- At each step it is important to determine the best test condition, which feature and value are most significant.
  - Criteria for classification : Entropy or Gini Index
  - Criteria for Regression : MSE (RSS)

Decision Tree algorithm for classification

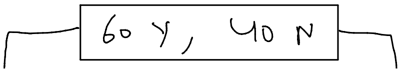
- Choose attributes and values from dataset.
- Calculate Entropy before and after split.
- Choose an attribute and value (Test Condition) which give us maximum reduction in entropy.
  - Purer Split
  - Maximum Separation
- Perform the split based on that test condition.
- Repeat the steps recursively.

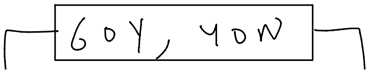
Entropy

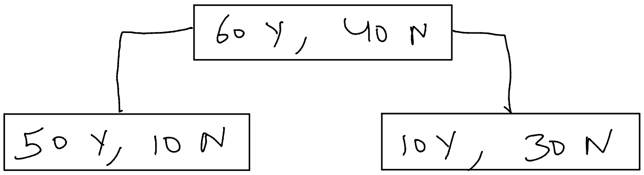
- Entropy is measure of purity / impurity of a node
- Lower the entropy, purity is more
- Higher the entropy, purity is less

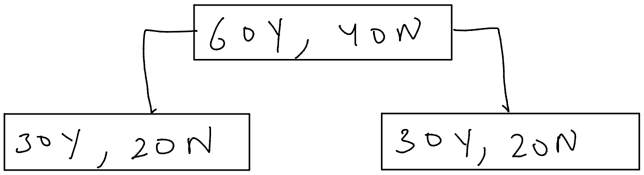
$$\boxed{\text{ENTROPY} = -\sum p_i \log_2 (p_i)}$$

Ex: 100, 60 YES, 40 NO









$$\text{Entropy} = -\left(p_Y \cdot \log_2(p_Y) + p_N \cdot \log_2(p_N)\right)$$

$$\text{Entropy}_B = -\left(\frac{6}{10} \cdot \log_2\left(\frac{6}{10}\right) + \frac{4}{10} \cdot \log_2\left(\frac{4}{10}\right)\right)$$

$$= 0.97$$

$\text{Entropy}_{\text{left}} = 0.65$

$\text{Entropy}_{\text{right}} = 0.81$

$$\text{Entropy}_A = \frac{60 \times 0.65 + 40 \times 0.81}{100}$$

$$= 0.71$$

$$\text{Entropy}_B - \text{Entropy}_A = 0.97 - 0.71$$
$$= 0.26$$

$\text{Entropy}_{\text{left}} = 0.97$

$\text{Entropy}_{\text{right}} = 0.97$

$$\text{Entropy}_A = \frac{50 \times 0.97 + 50 \times 0.97}{100}$$

$$= 0.97$$

$$\text{Entropy}_B - \text{Entropy}_A = 0.97 - 0.97$$
$$= 0$$

$$\boxed{\text{INFORMATION GAIN} = E_B - E_A}$$

———— X ———— X ———— X ————

* GINI INDEX :-

$$\boxed{GI = 1 - \Sigma p_i^2}$$

$$\boxed{GI = \Sigma p_i(1 - p_i)}$$

$$GI = P_1(1 - P_1) + P_2(1 - P_2)$$
$$= P_1 - P_1^2 + P_2 - P_2^2$$

$$GI = P_1(1-P_1) + P_2(1-P_2)$$

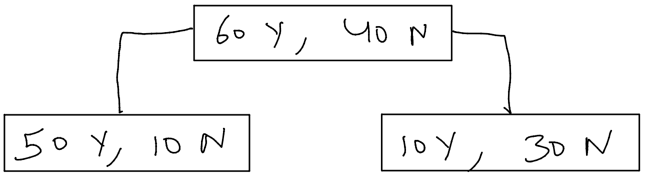
$$= P_1 - P_1^2 + P_2 - P_2^2$$

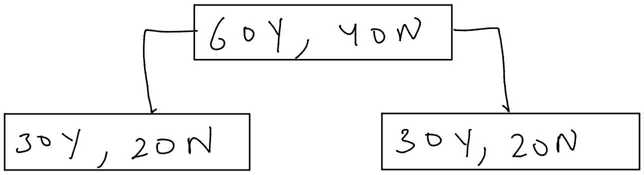
$$= P_1 + P_2 - P_1^2 - P_2^2$$

$$= 1 - (P_1^2 + P_2^2)$$

$$= 1 - \Sigma p_i^2$$

Ex: 100, 60 YES, 40 NO





$$GI = 1 - (P_Y^2 + P_N^2)$$

$$GI_B = 1 - \left(\left(\frac{60}{100}\right)^2 + \left(\frac{40}{100}\right)^2\right)$$

$$= 0.48$$

$$GI_{Left} = 0.27$$

$$GI_{Right} = 0.37$$

$$GI_A = \frac{0.27 \times 60 + 0.37 \times 40}{100}$$

$$= 0.31$$

$$GI_B - GI_A = 0.48 - 0.31$$
$$= 0.17$$

$$GI_{Left} = 0.48$$

$$GI_{Right} = 0.48$$

$$GI_A = \frac{0.48 \times 50 + 0.48 \times 50}{100}$$

$$= 0.48$$

$$GI_A - GI_B = 0.48 - 0.48$$
$$= 0$$

——— X ——— X ——— X ———

Decision Tree algorithm for Regression

- Choose attributes and values from dataset.
- Calculate MSE/RSS before and after split.

- Choose an attribute and value (Test Condition) which gives us maximum reduction in MSE.
- Perform the split based on that test condition.
- Repeat the steps recursively.