Data :

	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	NO
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	NO
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	NO
9	Red	SUV	Imported	Yes
10	Red	Sports	Imported	NO

Split level 1

* Color:

- Color = Red:
  - + 5 counts
  - + P(Yes | Red) = 3/5 = 0,6
  - + P(No | Red) = 2/5 = 0,4
  - + Gini(Red) = $1 - (0,6)^2 - (0,4)^2$
    $= 0,48$

- Color = Yellow
  - + 5 counts
  - + P(Yes | Yellow) = 2/5 = 0,4
  - + P(No | Yellow) = 3/5 = 0,6
  - + Gini(Yellow) = $1 - (0,4)^2 - (0,6)^2$
    $= 0,48$

$$\text{WeightedGini(Color)} = \frac{5}{10} \cdot 0,48 + \frac{5}{10} \cdot 0,48$$
$$= 0,48$$

* Type:

- Type = Sport:
  + Counts = 6
  + $P(\text{Yes} \mid \text{Sport}) = 4/6 = 2/3$
  + $P(\text{No} \mid \text{Sport}) = 2/6 = 1/3$
  + $\text{Gini}(\text{Sport}) = 1 - (\frac{2}{3})^2 - (\frac{1}{3})^2$
    $\approx 0,44$

- Type = SUV:
  + Counts = 4
  + $P(\text{Yes} \mid \text{SUV}) = 1/4 = 0,25$
  + $P(\text{No} \mid \text{SUV}) = 3/4 = 0,75$
  + $\text{Gini}(\text{SUV}) = 1 - 0,25^2 - 0,75^2$
    $= 0,375$

$$\text{Weighted Gini (Type)} = \frac{6}{10} \cdot 0,44 + \frac{4}{10} \cdot 0,375$$
$$= 0,414$$

* Origin:

- Origin = Domestic:
  + Counts = 5
  + P(Yes | Domestic) = 2/5 = 0,4
  + P(No | Domestic) = 3/5 = 0,6
  + Gini (Domestic) = $1 - 0,4^2 - 0,6^2$
    $= 0,48$

- Origin = Imported:
  + Counts = 5
  + P(Yes | Imported) = 3/5 = 0,6
  + P(No | Imported) = 2/5 = 0,4
  + Gini (Imported) = $1 - 0,6^2 - 0,4^2$
    $= 0,48$

weighted Gini (Origin)

$= \frac{5}{10} \cdot 0,48 + \frac{5}{10} \cdot 0,48$

$= 0,48$

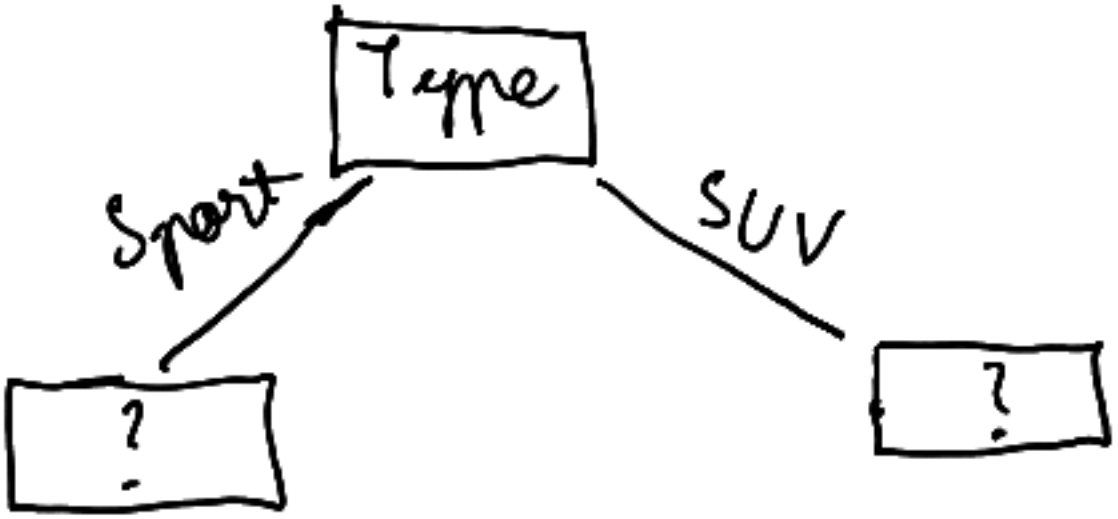
- Weighted Gini (color) = 0,48
  - (Type) = 0,414
  - (origin) = 0,48

⇒ the impurities of "Type" is lowest

⇒ choose "Type" as the 1st split

Split level 2

- has 2 values: Sport & SUV
- Each of these values will be a new branch.



* Sport branch: 6 counts

Sport → Color

- Color = Red + Yellow

- Sport. Red:
  + Counts = 4
  + P(Yes | Sport. Red) = 3/4 = 0,75
  + P(No | Sport. Red) = 1/4 = 0,25
  + Gini (Sport → Red) = $1 - 0{,}75^2 - 0{,}25^2$
    $= 0{,}375$

- Sport. Yellow:
  + Counts = 2
  + P(Yes | Sport. Yellow) = 1/2 = 0,5
  + P(No | Sport. Yellow) = 1/2 = 0,5
  + Gini (Sport → Yellow) = $1 - 0{,}5^2 - 0{,}5^2$
    $= 0{,}5$

Weighted Gini (Sport → Color) = $\frac{4}{6} \cdot 0{,}375 + \frac{2}{6} * 0{,}5$
$\simeq 0{,}417$

Sport → Origin

- Origin = Domestic + Imported
- Sport. Domestic:
  + Counts = 4
  + P(Yes | Sport. Domestic) = 2/4 = 0,5
  + P(No | Sport. Domestic) = 2/4 = 0,5
  + Gini (Sport → Domestic) = $1 - 0,5^2 - 0,5^2$
    = 0,5
- Sport. Imported
  + Counts = 2
  + P(Yes | Sport. Imported) = 2/2 = 1
  + P(No | Sport. Imported) = 0/2 = 0
  + Gini (Sport → Domestic) = $1 - 1^2 - 0^2$
    = 0

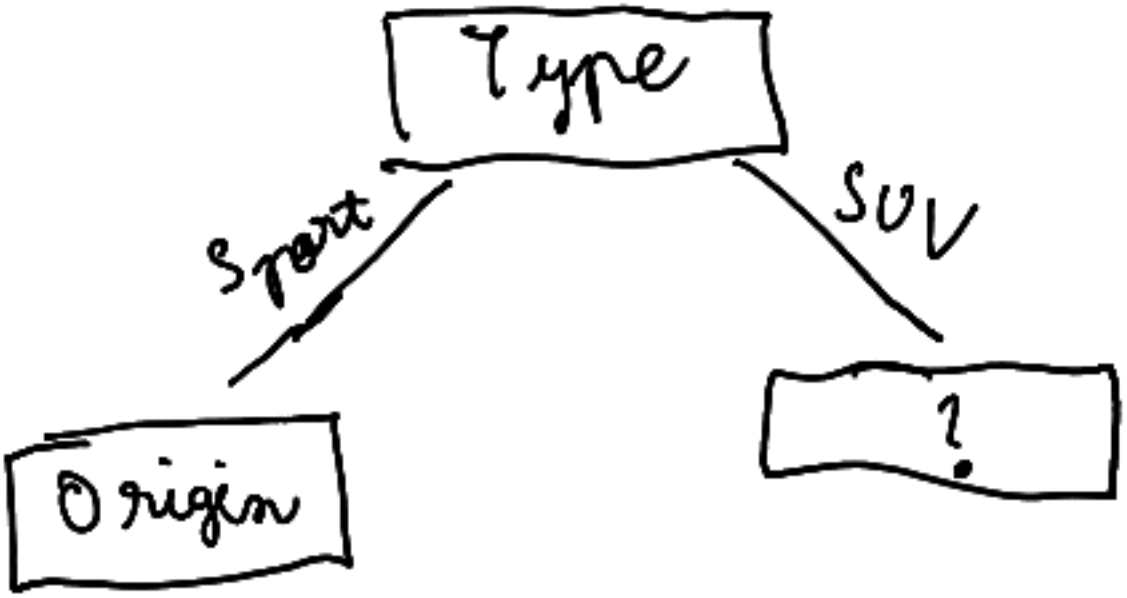
Weighted Gini (Sport → Origin) = $\frac{4}{6} \cdot 0,5 + \frac{2}{6} \cdot 0$
= 0,33

Weighted Gini (Sport → Color) = 0,417

(Sport → Origin) = 0,33

⇒ "Origin" has lower impurity

⇒ choose "Origin" as next node on "Sport" branch



* SUV branch: 4 counts.

SUV → Color

- Color = Red + Yellow

- SUV.Red
  + Counts = 1
  + P(Yes | SUV.Red) = 0/1 = 0
  + P(No | SUV.Red) = 1/1 = 1
  + Gini (SUV → Red) = $1 - 0^2 - 1^2$
    = 0

- SUV.Yellow
  + Counts = 3
  + P(Yes | SUV.Yellow) = 1/3
  + P(No | SUV.Yellow) = 2/3
  + Gini (SUV → Yellow) = $1 - (\frac{1}{3})^2 - (\frac{2}{3})^2$
    = 0,44

Weighted Gini(SUV → Color) = $\frac{1}{4} \cdot 0 + \frac{3}{4} \cdot 0,44$
= 0,33

SUV → Origin

- Origin = Domestic + Imported
- SUV . Domestic:
  + Counts = 1
  + P(Yes | SUV . Domestic) = 0/1 = 0
  + P(No | SUV . Domestic) = 1/1 = 1
  + Gini (SUV → Domestic) = $1 - 0^2 - 1^2$ = 0
- SUV . Imported:
  + Counts = 3
  + P(Yes | SUV . Imported) = 1/3
  + P(No | SUV . Imported) = 2/3
  + Gini (SUV → Imported) = $1 - \frac{1}{9} - \frac{4}{9}$ = 0,44

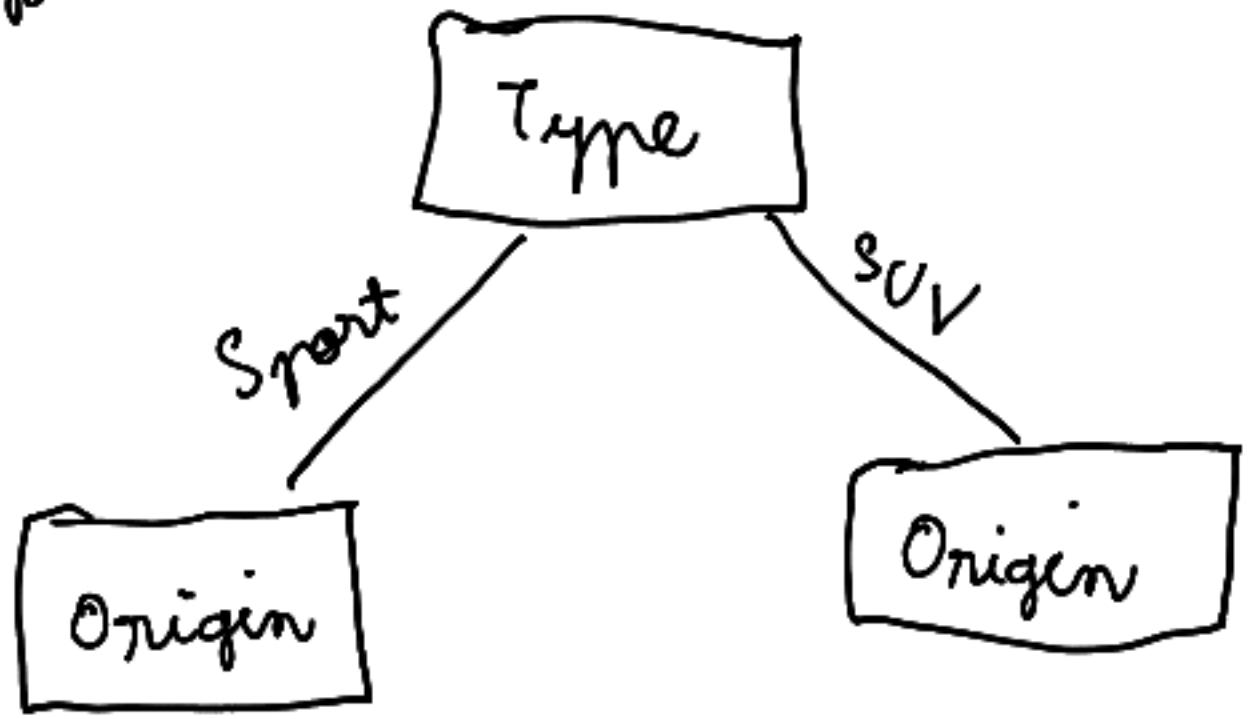
Weighted Gini (SUV → Origin)

= $\frac{1}{4} \cdot 0 + \frac{3}{4} \cdot 0{,}44$ = 0,11

Weighted Gini (SUV → Color) = 0,33

——————— (SUV → Origin) = 0,11

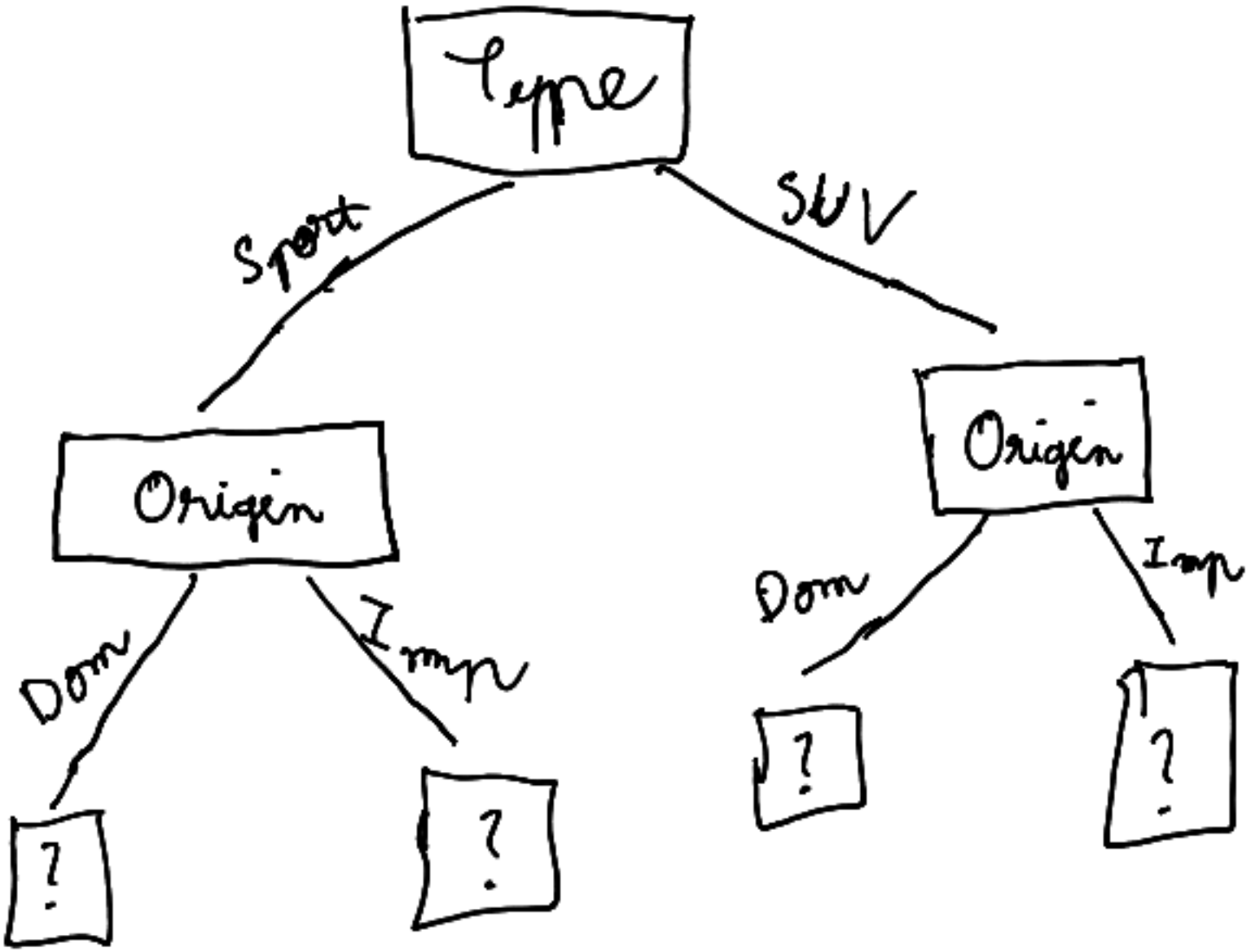
⇒ choose "Origin" as next node for "SUV" branch



Split level 3

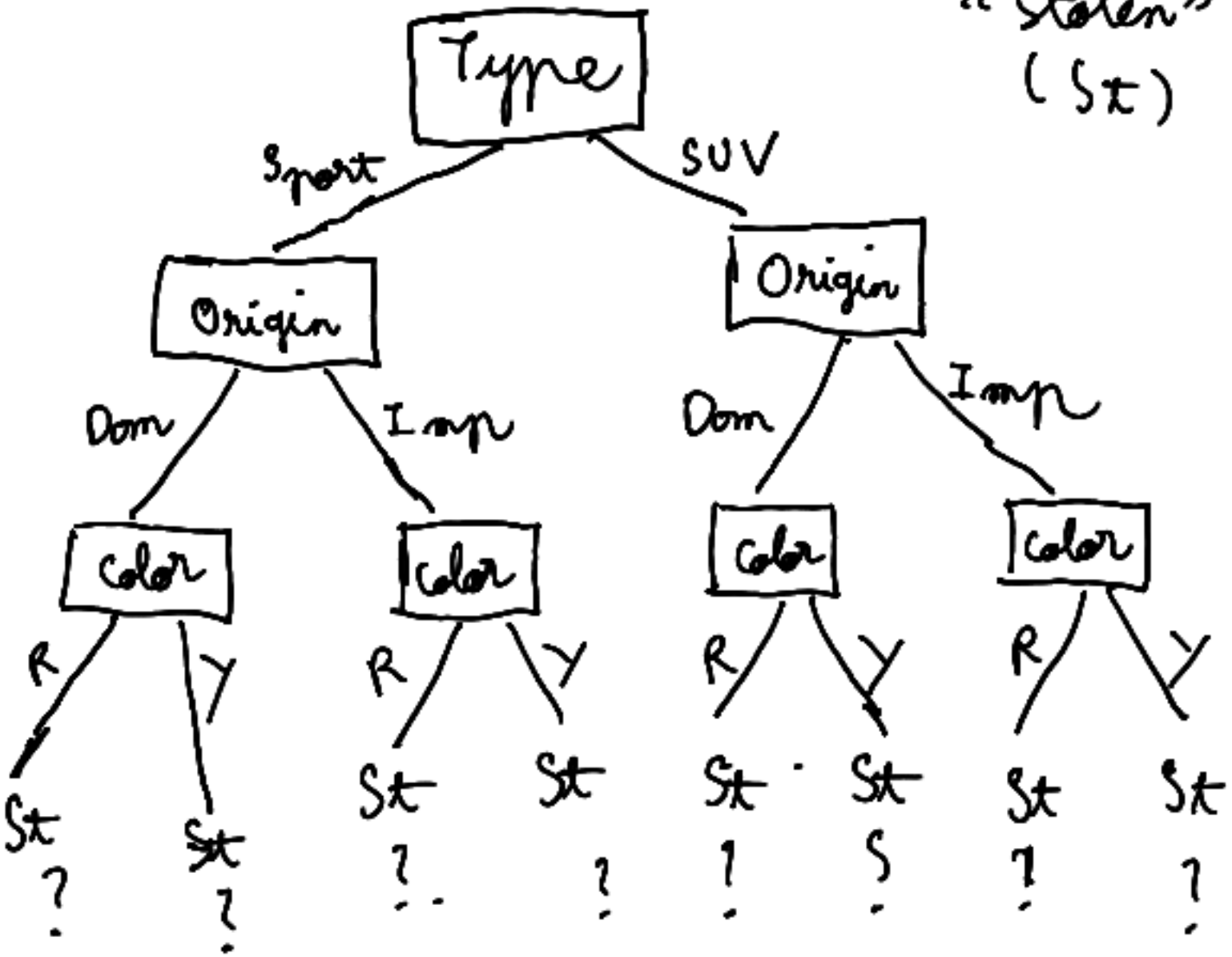
- Origin has 2 values : Domestic and Imported

→ 2 branches



- "Color" is the only feature left, so it's the final node (leaf node)

- Color ⟶ 2 branches: Red + Yellow
- These 2 branches lead to outcome "Stolen" (St)



- $P(Yes \mid Sport. Dom. Red) = 2/3$
  $P(No \mid Sport. Dom. Red) = 1/3$

  $\Rightarrow$ "Stolen" = Yes

- $P(Yes \mid Sport. Dom. Yellow) = 0/1 = 0$
  $P(No \mid Sport. Dom. Yellow) = 1/1 = 1$
  $\Rightarrow$ "stolen" = No

- $P(Yes \mid Sport. Imp. Red) = 1/1 = 1$
  $P(No \mid Sport - Imp. Red) = 0/1 = 0$
  $\Rightarrow$ "Stolen" = Yes

- $P(Yes \mid Sport. Imp. Yellow) = 1/1 = 1$
  $P(No \mid Sport. Imp. Yellow) = 0/1 = 0$
  $\Rightarrow$ "stolen" = Yes

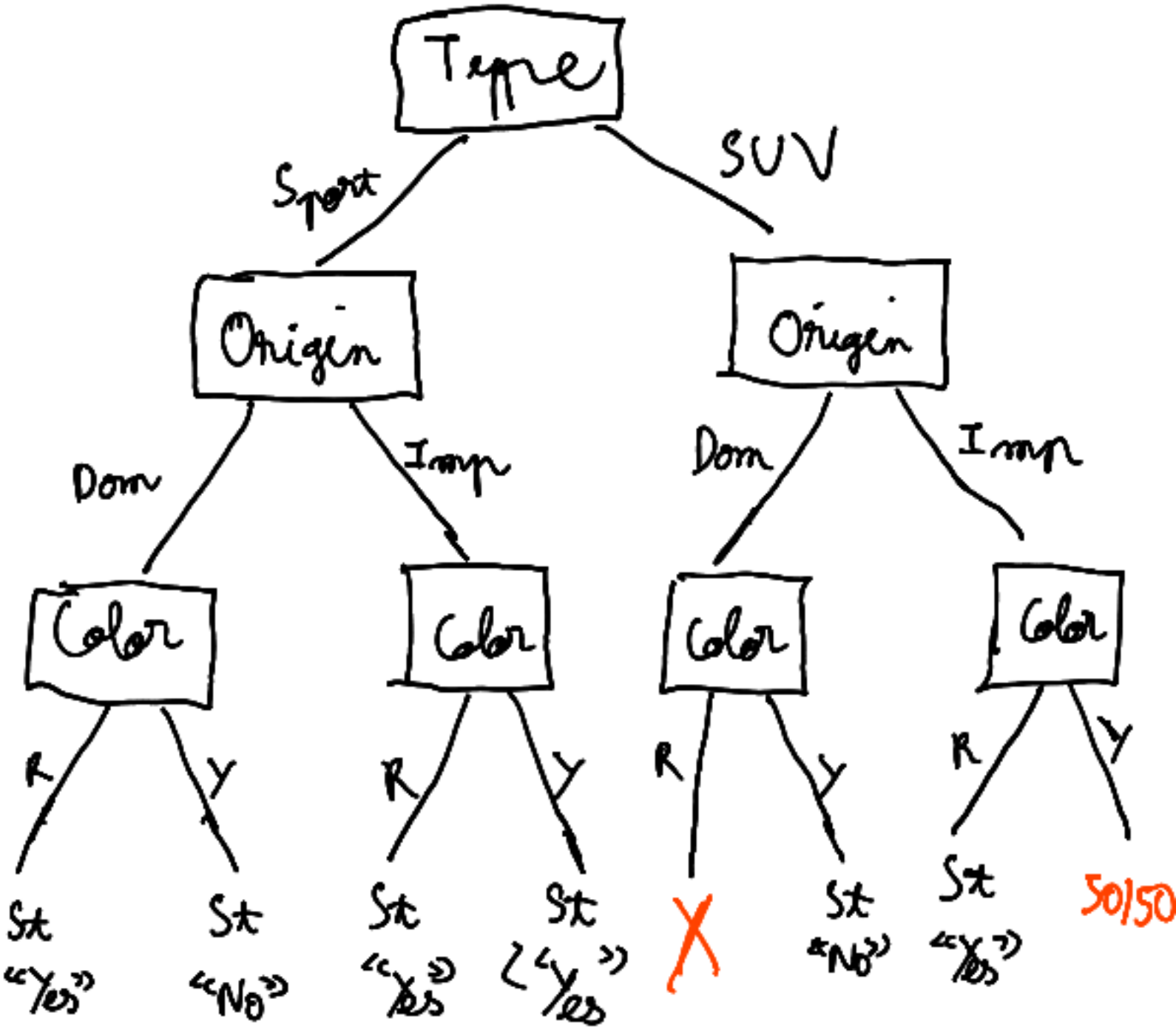
- $P(Yes \mid SUV.Dom.Red) = X$
  $P(No \mid SUV.Dom.Red) = X$
  } No data
  No predict

- $P(Yes \mid SUV.Dom.Yellow) = 0/1 = 0$
  $P(No \mid SUV.Dom.Yellow) = 1/1 = 1$
  $\Rightarrow$ "Stolen" = No

- $P(Yes \mid SUV.Imp.Red) = 0/1 = 0$
  $P(No \mid SUV.Imp.Red) = 1/1 = 1$
  $\Rightarrow$ "Stolen" = Yes

- $P(Yes \mid SUV.Imp.Yellow) = 1/2$
  $P(No \mid SUV.Imp.Yellow) = 1/2$
  $\Rightarrow$ "Stolen" = ... unsure ....

Final tree:



- New data:

Color	Type	Origin
Yellow	Sport	Imported

Type = Sport → origin = Imp → color = Yellow → Stolen: "Yes"

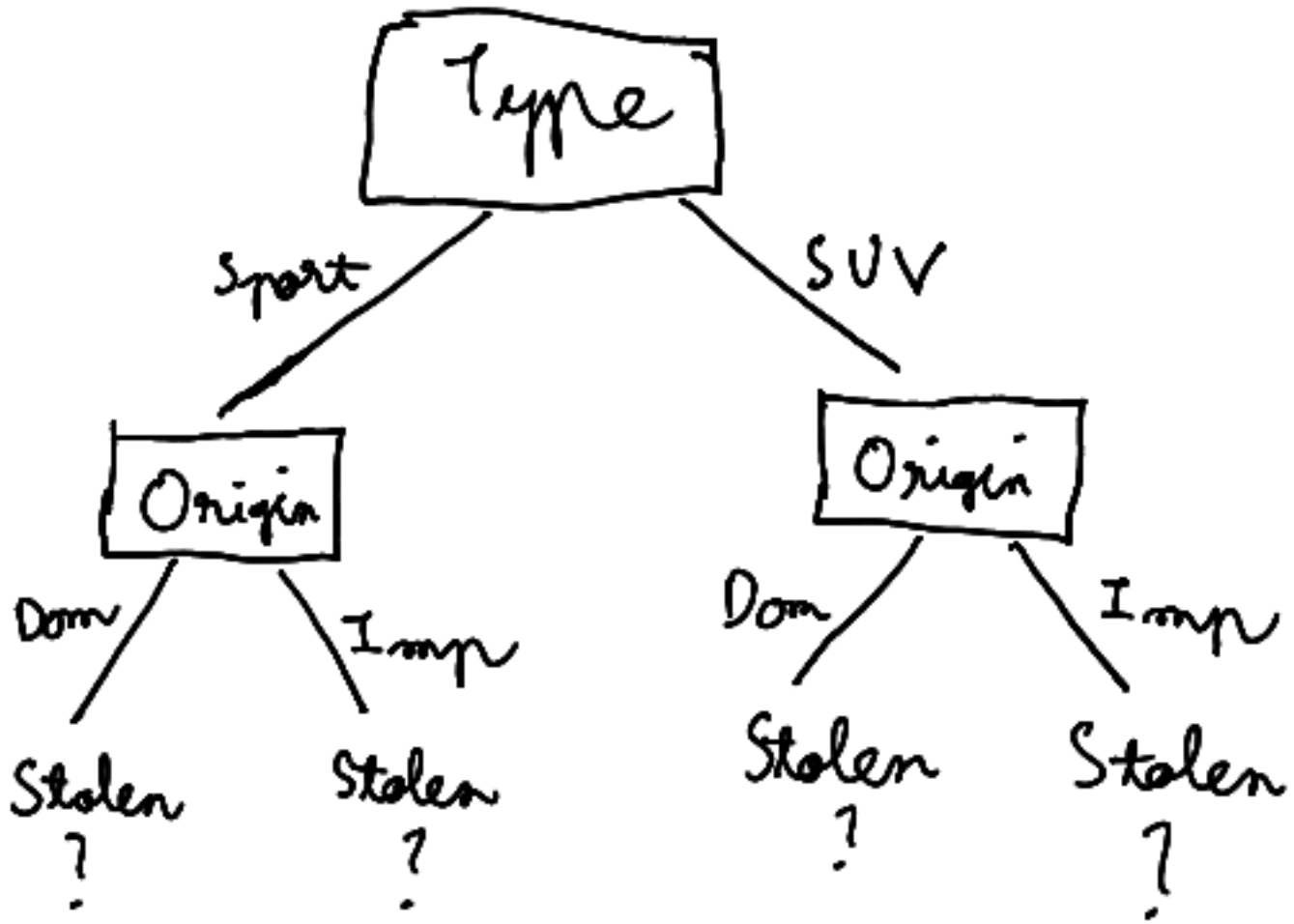
Simplify the tree

- If the tree is too detailed
  ↳ can cause "Overfitting"

(as we can see in the example above, the tree cannot predict for "SUV. Domestic. Red" case due to lack of data,
and cannot predict "SUV. Imported, Yellow" due to tied probabilities
$P(Yes) = P(No) = \frac{1}{2}$)

⇒ Must simplify the tree, by stop the splitting at a certain level.

- In the above example, let's stop at "Origin"



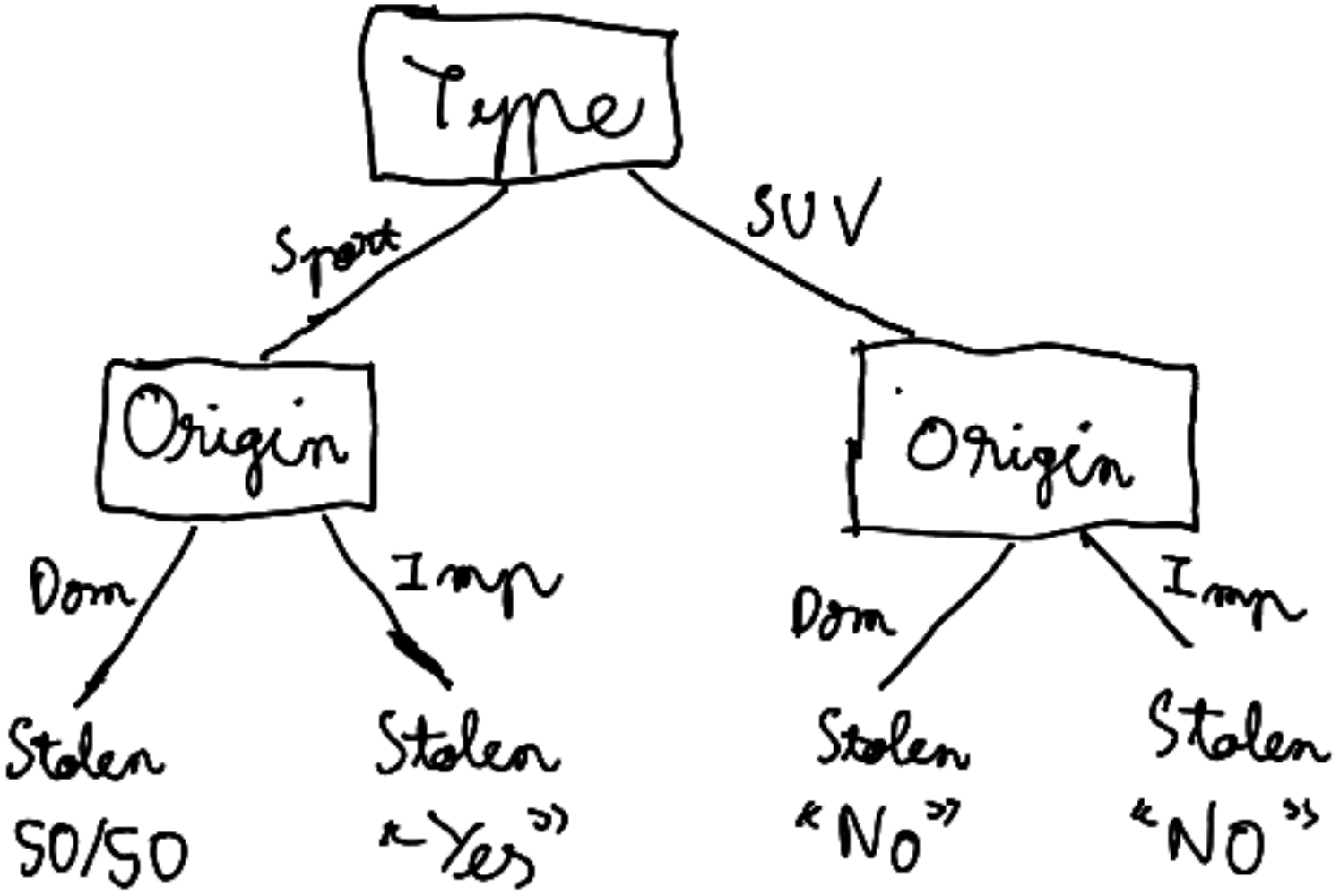
- $P(\text{Yes} \mid \text{Sport. Domestic}) = 2/4 = 1/2$
  $P(\text{No} \mid \text{Sport. Domestic}) = 2/4 = 1/2$
  $\Rightarrow \ldots$ Unsure.

- $P(\text{Yes} \mid \text{Sport. Imported}) = 2/2 = 1$
  $P(\text{No} \mid \text{Sport. Imported}) = 0/2 = 0$
  $\Rightarrow$ "Stolen" = Yes

- $P(\text{Yes} \mid \text{SUV. Domestic}) = 0/1 = 0$
  $P(\text{No} \mid \text{SUV. Domestic}) = 1/1 = 1$
  $\Rightarrow$ "Stolen" = No

- $P(\text{Yes} \mid \text{SUV. Imported}) = 1/3$
  $P(\text{No} \mid \text{SUV. Imported}) = 2/3$
  $\Rightarrow$ "Stolen" = No

Final Simplified Tree



- New data :

Color	Type	Origin
Yellow	Sport	Imported

check Type = Sport

↳ Checke Origin = Imp

↳ Stolen "Yes"