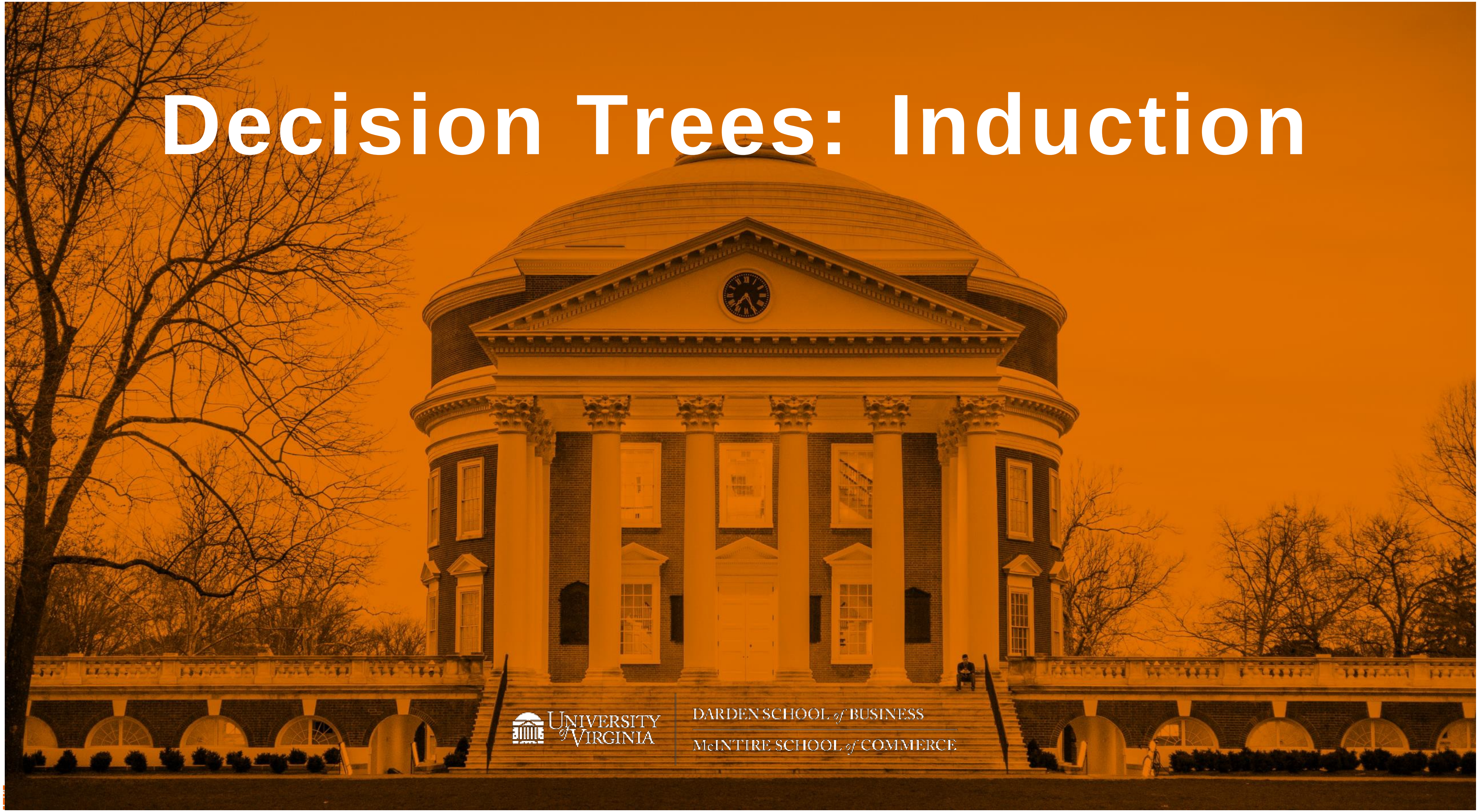
# Decision Trees: Induction



UNIVERSITY of VIRGINIA

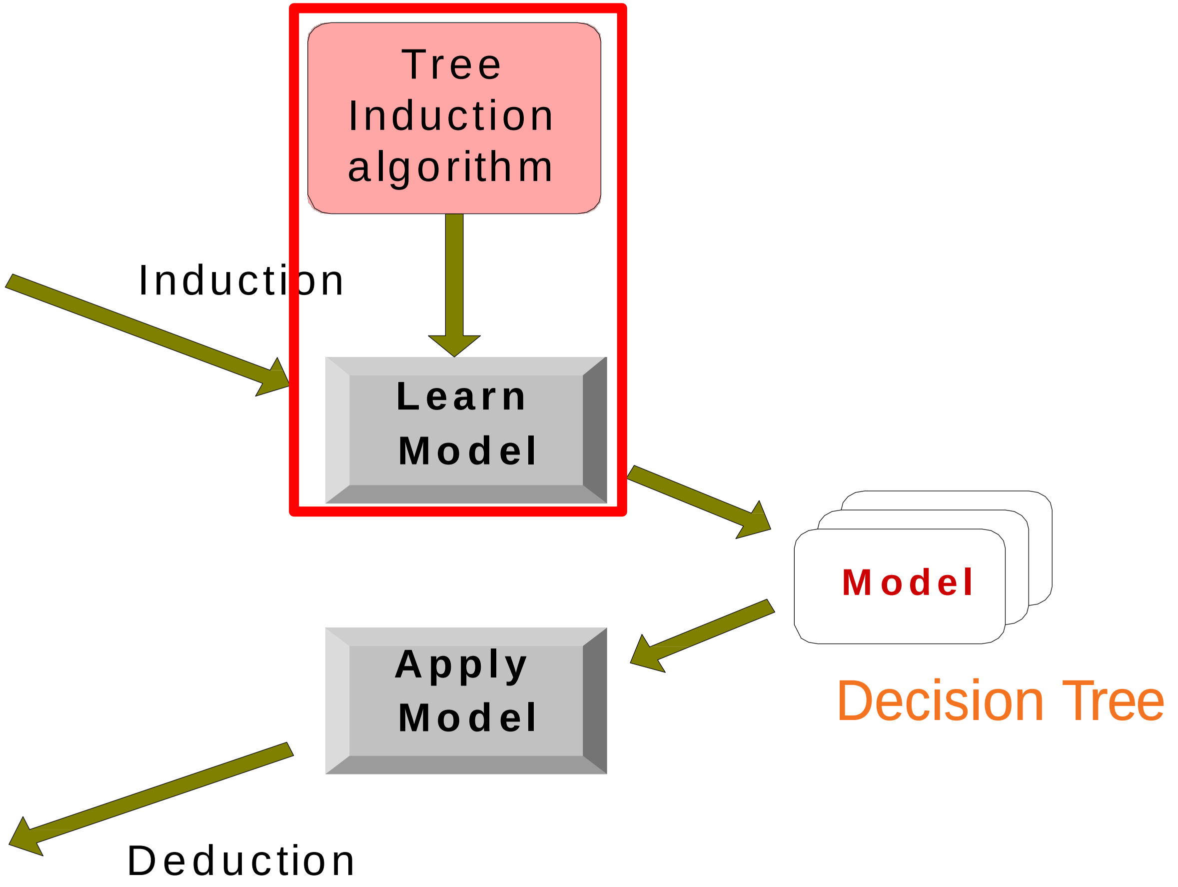
DARDEN SCHOOL of BUSINESS

McINTIRE SCHOOL of COMMERCE

# TREE INDUCTION ALGORITHMS

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 1 | Yes | Large | 125K | No |
| 2 | No | Medium | 100K | No |
| 3 | No | Small | 70K | No |
| 4 | Yes | Medium | 120K | No |
| 5 | No | Large | 95K | Yes |
| 6 | No | Medium | 60K | No |
| 7 | Yes | Large | 220K | No |
| 8 | No | Small | 85K | Yes |
| 9 | No | Medium | 75K | No |
| 10 | No | Small | 90K | Yes |

Training Set




| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 11 | No | Small | 55K | ? |
| 12 | Yes | Medium | 80K | ? |
| 13 | Yes | Large | 110K | ? |
| 14 | No | Small | 95K | ? |
| 15 | No | Large | 67K | ? |

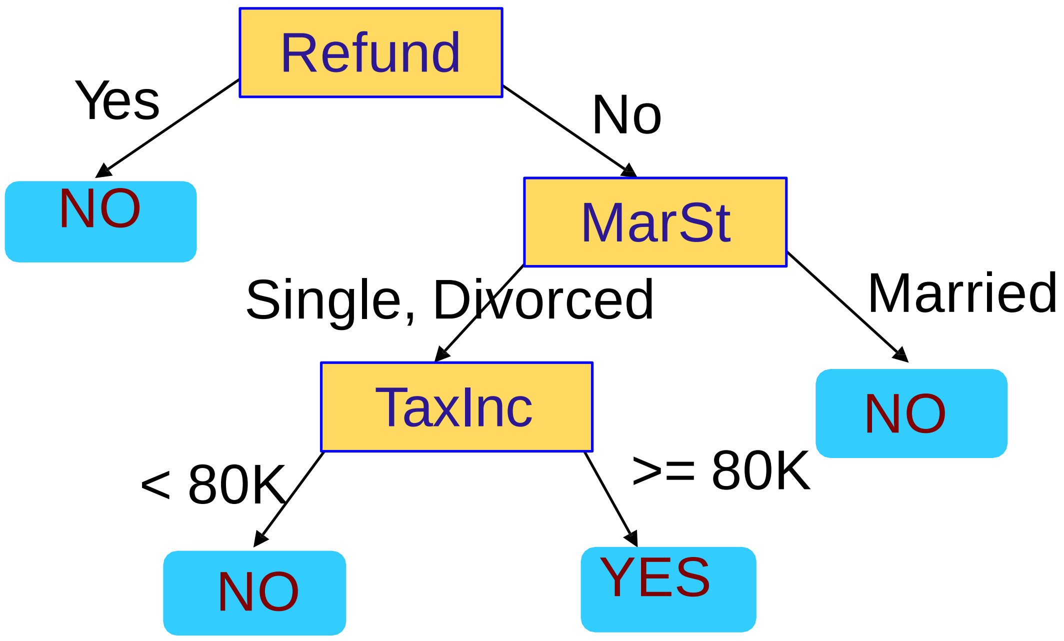
Test Set

# OUR PREVIOUS EXAMPLE…

class

| *Tid* | Refund | Marital Status | Taxable Income | Fraud |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | NO |
| 2 | No | Married | 100K | NO |
| 3 | No | Single | 70K | NO |
| 4 | Yes | Married | 120K | NO |
| 5 | No | Divorced | 95K | YES |
| 6 | No | Married | 60K | NO |
| 7 | Yes | Divorced | 220K | NO |
| 8 | No | Single | 85K | YES |
| 9 | No | Married | 75K | NO |
| 10 | No | Single | 90K | YES |

Training Data Matrix



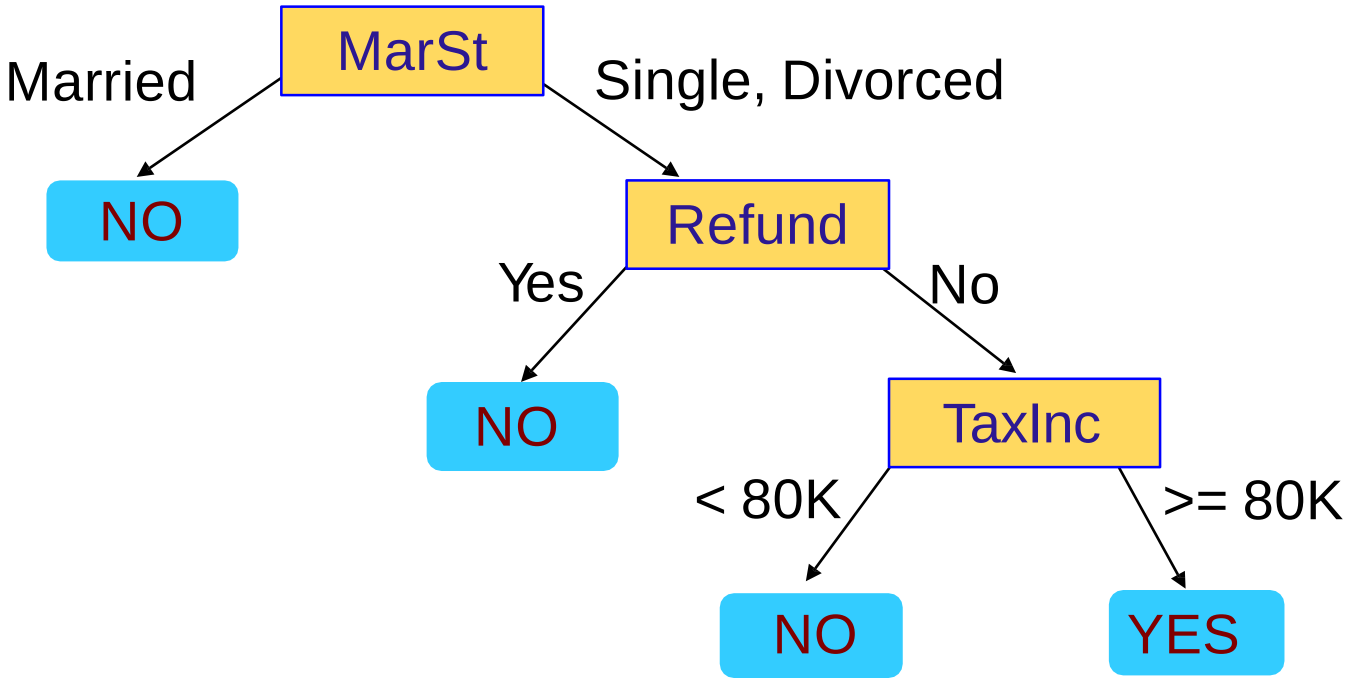

Decision Tree Model

# A DIFFERENT MODEL!

class

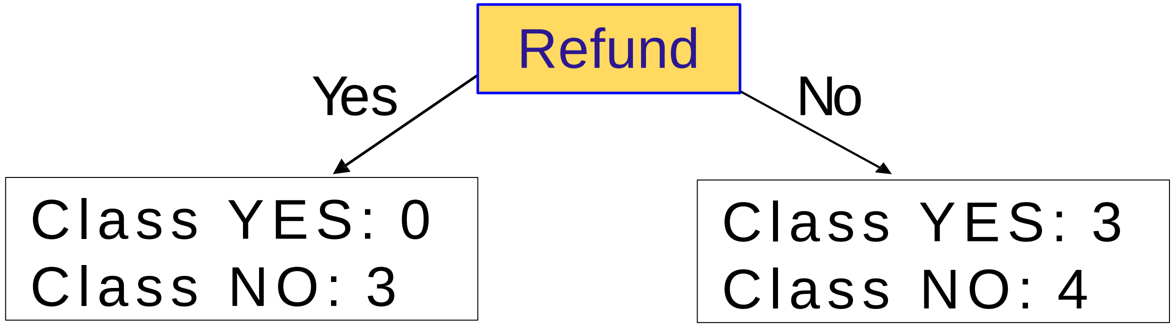
| *Tid* | Refund | Marital Status | Taxable Income | Fraud |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | NO |
| 2 | No | Married | 100K | NO |
| 3 | No | Single | 70K | NO |
| 4 | Yes | Married | 120K | NO |
| 5 | No | Divorced | 95K | YES |
| 6 | No | Married | 60K | NO |
| 7 | Yes | Divorced | 220K | NO |
| 8 | No | Single | 85K | YES |
| 9 | No | Married | 75K | NO |
| 10 | No | Single | 90K | YES |

Training Data Matrix




Decision Tree Model

# ATTRIBUTE SELECTION




A pure node (all of the cases have the same class label)

A node with impurity (cases are more evenly distributed across classes)

# ATTRIBUTE SELECTION

Marital Status

Married

Class YES: 0
Class NO: 4

Single, Divorced

Class YES: 3
Class NO: 3

A pure node (all of the cases have the same class label)

A node with impurity (more evenly distributed across classes)

# ATTRIBUTE SELECTION

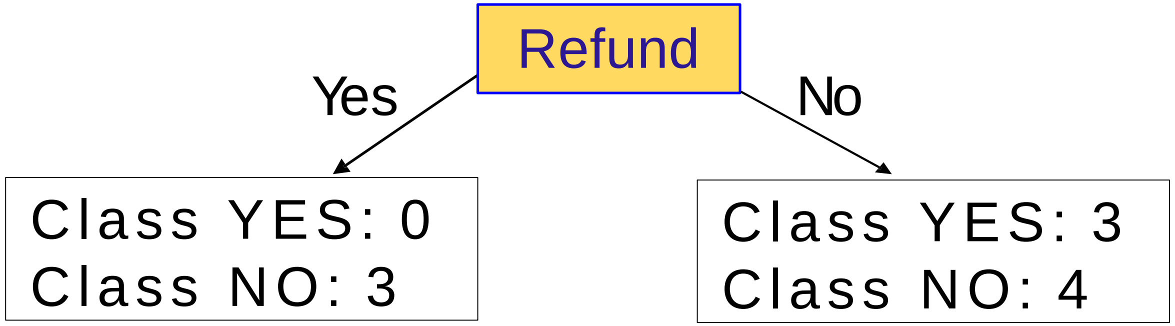
Entropy at a given node $\boldsymbol{t}$:

$$Entropy = -\sum_{i=0}^{c-1} p_i(t)\, log_2\, p_i(t)$$

Where $\boldsymbol{p_i(t)}$ is the frequency of class $\boldsymbol{i}$ at node $\boldsymbol{t}$, and $\boldsymbol{c}$ is the total number of classes

- Maximum entropy when records are equally distributed among all classes, implying the least beneficial situation for classification
- Minimum entropy when all records belong to one class, implying most beneficial situation for classification

# ATTRIBUTE SELECTION



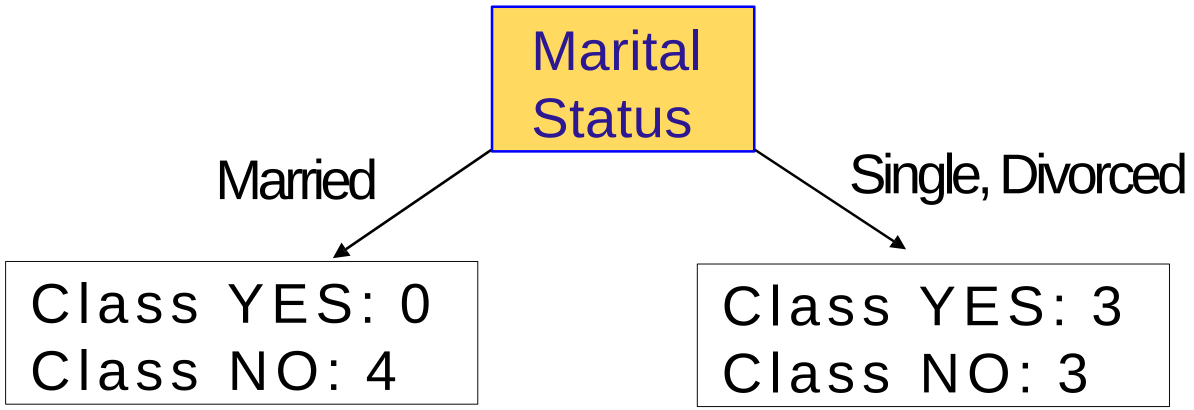

P(Class Yes) = 0/3 = 0
P(Class No) = 3/3 = 1

$$\text{Entropy} = -\,0 \log_2 0 - 1 \log_2 1 = -\,0 - 0 = 0$$

P(Class Yes) = 3/7 = 0.43
P(Class No) = 4/7 = 0.57

$$\text{Entropy} = -\,(0.43) \log_2 (0.43) - (0.57) \log_2 (0.57)$$
$$= -\,(0.43) \times (-1.22) - (0.57) \times (-0.81)$$
$$= 0.52 + 0.46 = 0.98$$

# ATTRIBUTE SELECTION




P(Class Yes) = 0/4 = 0
P(Class No) = 4/4 = 1

Entropy $= -\,0 \log_2 0 - 1 \log_2 1 = -\,0 - 0 = 0$

P(Class Yes) = 3/6 = 0.5
P(Class No) = 3/6 = 0.5

Entropy $= -\,(0.5) \log_2 (0.5) - (0.5) \log_2 (0.5)$
$= -\,(0.5) \times (-1) - (0.5) \times (-1)$
$= 0.5 + 0.5 = 1$

# ATTRIBUTE SELECTION

Information Gain:

$$Gain_{split} = Entropy(p) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$$

Parent node, $p$ is split into $k$ nodes (children)
$n_i$ is number of records in child node $i$

- Choose the split that achieves most reduction (maximizes GAIN)
- Used in the ID3 and C4.5 decision tree algorithms

DECISION TREE MODELS: INDUCTION

# ATTRIBUTE SELECTION

| Tid | Refund | Marital Status | Taxable Income | Fraud |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | NO |
| 2 | No | Married | 100K | NO |
| 3 | No | Single | 70K | NO |
| 4 | Yes | Married | 120K | NO |
| 5 | No | Divorced | 95K | YES |
| 6 | No | Married | 60K | NO |
| 7 | Yes | Divorced | 220K | NO |
| 8 | No | Single | 85K | YES |
| 9 | No | Married | 75K | NO |
| 10 | No | Single | 90K | YES |

Parent Node: Class YES: 3
Class NO: 7

$/10) \log_2 (3/10) - (7/10) \log_2 (7/10)$

$- (0.7) \times (-0.51) = 0.52 + 0.36 = 0.88$

Refund

Yes: Class YES: 0
Class NO: 3

No: Class YES: 3
Class NO: 4

$/10) \times (0) - (7/10) \times (0.98) =$ 0.19

Entropy at Parent Node

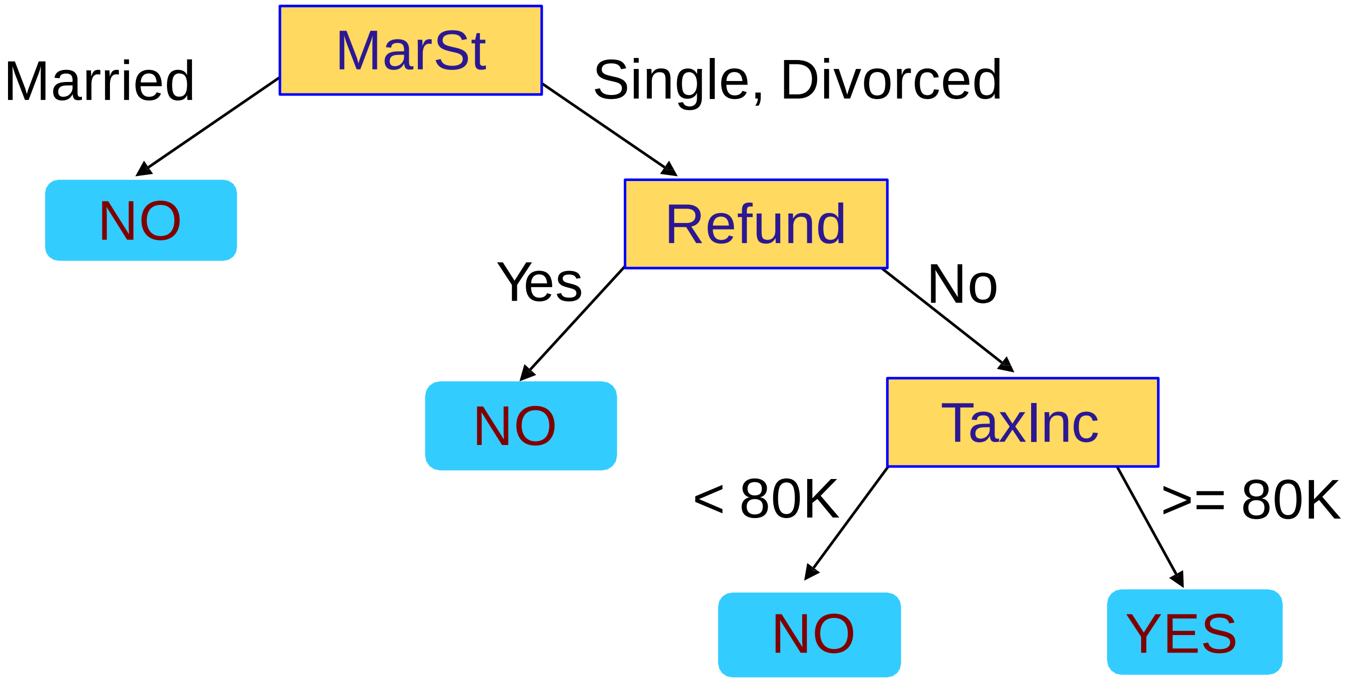
Entropy at Child Nodes

Marital Status

Married: Class YES: 0
Class NO: 4

Single, Divorced: Class YES: 3
Class NO: 3

$Gain_{split} = 0.88 - (4/10) \times (0) - (6/10) \times (1) =$ 0.28


MarSt
Married
Single, Divorced
NO
Refund
Yes
No
NO
TaxInc
< 80K
>= 80K
NO
YES