DECISION TREE

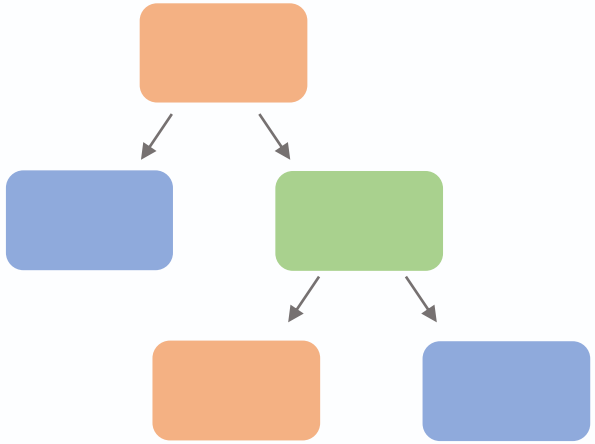
GINI Index

March 2023

DECISION TREE

INDUCE A DECISION TREE

DECISION TREE MODEL (TRAINED)



Restaurant Example

"Are they willing to wait for a table in a restaurant?"

- Label:
  - Will Wait: Yes/No
- All Features:
  - Alternatives
  - Bar
  - Friday
  - Hunger
  - Patrons
  - Price
  - Rain
  - Reservation
  - Type
  - Estimated Time

Decision Tree: Example

"Are they willing to wait for a table in a restaurant?"

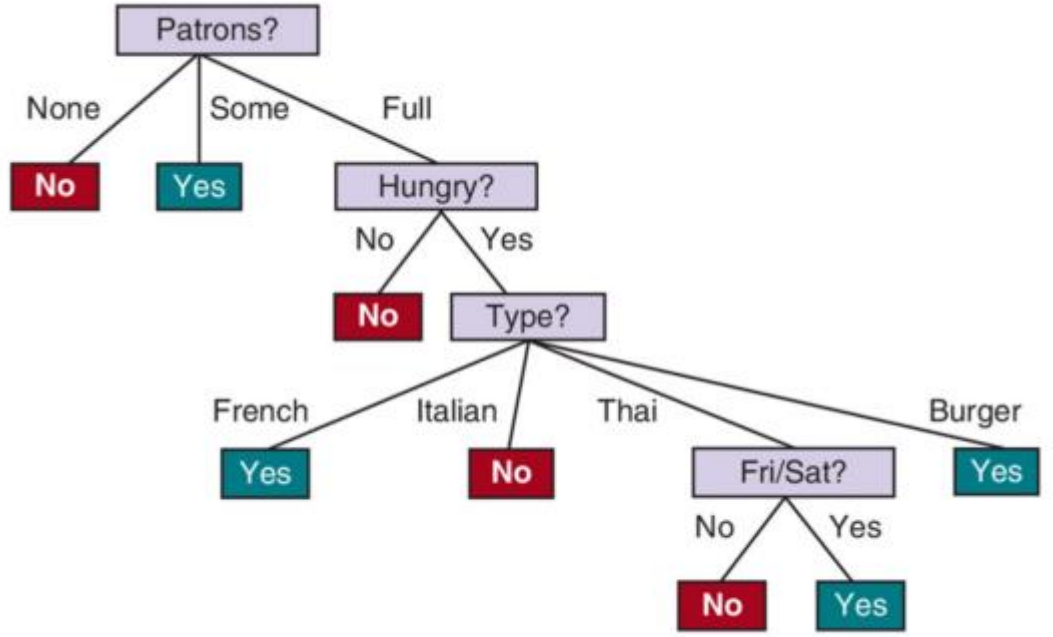
Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

Features and labels for the restaurant dataset

Restaurant Example

"Are they willing to wait for a table in a restaurant?"

- Label:
  - Will Wait: Yes/No
- Features Used For the tree:
  - Patrons (None/Some/Full)
  - Hunger (Yes/No)
  - Type (French/Italian/Thai/Burger)
  - Friday (Yes/No)



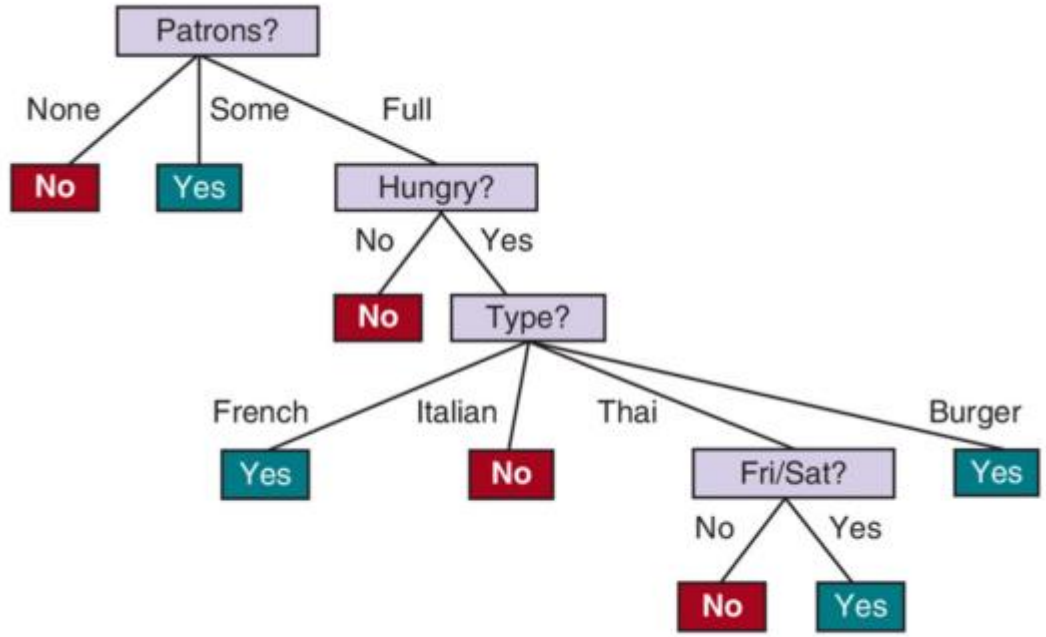
The decision tree induced from the 12-example training set.

Decision Tree: Example

"Are they willing to wait for a table in a restaurant?"

Making a decision tree:

1. How do we branch and determine terminal node?
2. How do we choose which features to use for branching?

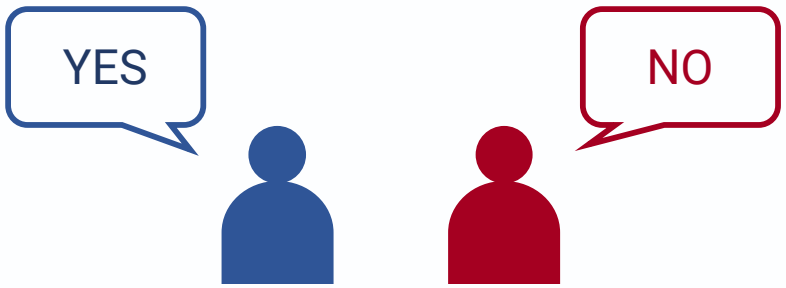


The decision tree induced from the 12-example training set.

Legend (for the slide set)

Each row in our feature table represents a person

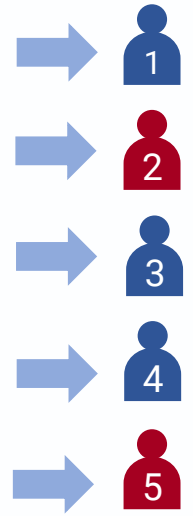
- Blue: Yes (they waited)
- Red: No (they did not wait)



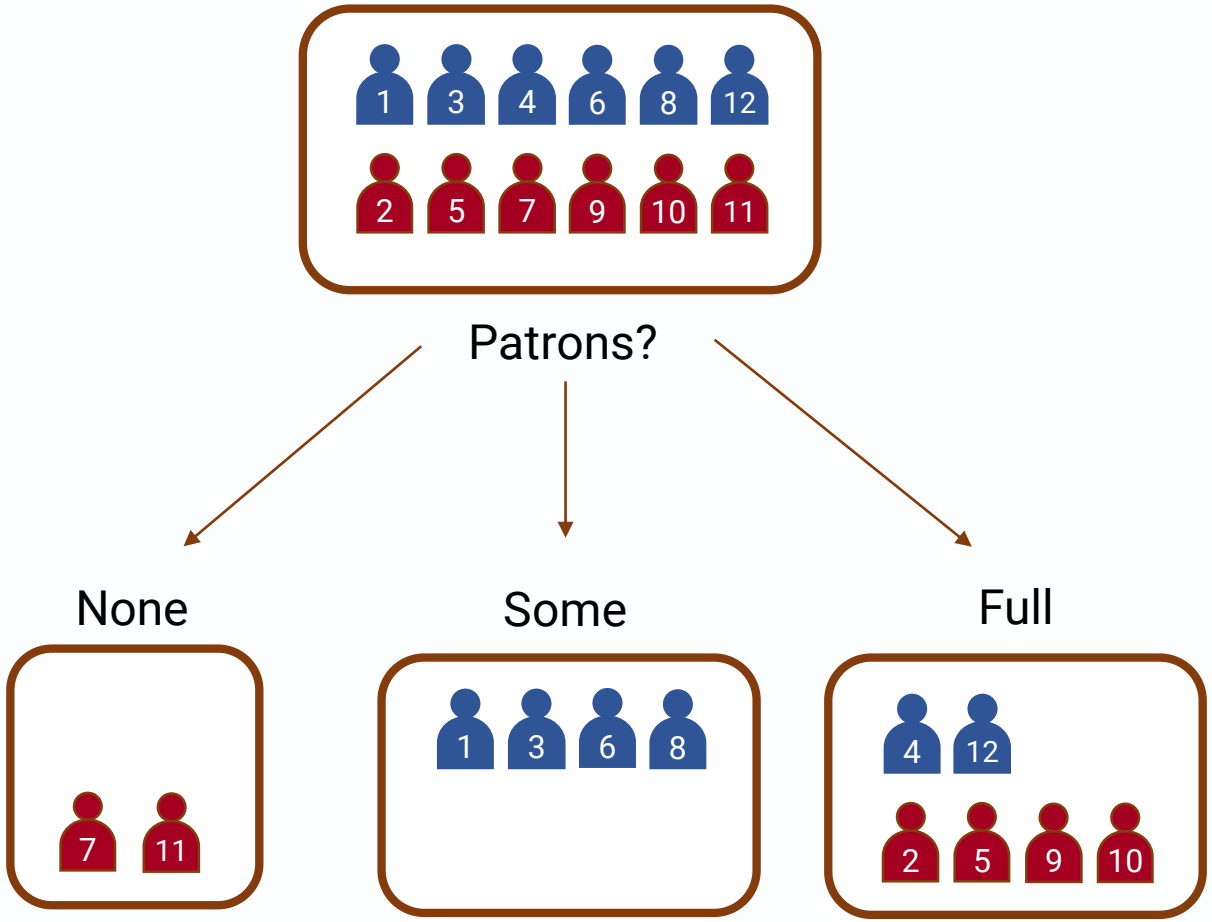
FEATURES

LABELS

Patrons	Hungry	Type	Friday	Will Wait
Some	Yes	French	No	Yes
Full	Yes	Thai	No	No
Some	No	Burger	No	Yes
Full	Yes	Thai	Yes	Yes
Full	No	French	Yes	No
. . .				. . .



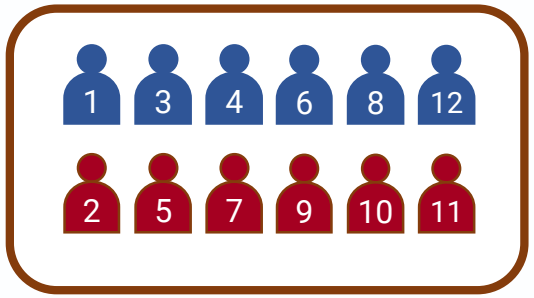
Group and Split Overview



"Persons" are grouped then split according to a feature

1. Start with everyone in one group
2. Select a feature (based on gini index)
3. Then split into multiple groups depending on their "answer" to the feature

Group and Split Overview



Patrons?

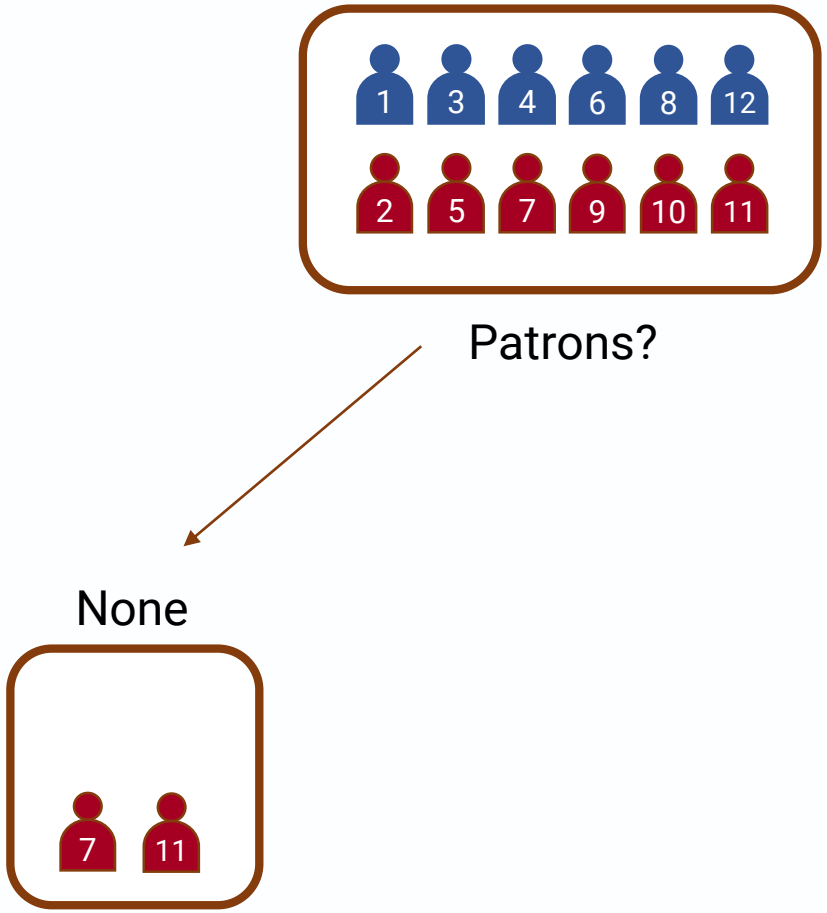
Feature: Patrons

(Are there people in the restaurant?)

Possible answers/categories:

- None:
- Some
- Full

Group and Split Overview



Feature: Patrons

(Are there people in the restaurant?)

Possible answers/categories:

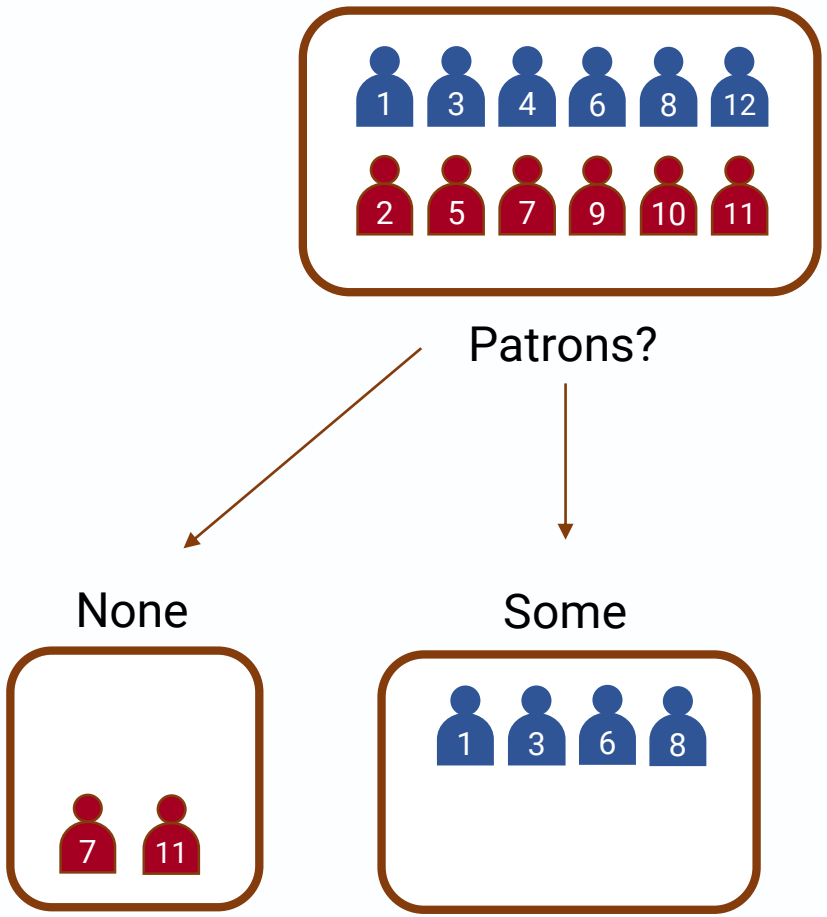
- None:

  Yes: 0 No: 2

- Some

- Full

Group and Split Overview



Feature: Patrons

(Are there people in the restaurant?)

Possible answers/categories:

- None:

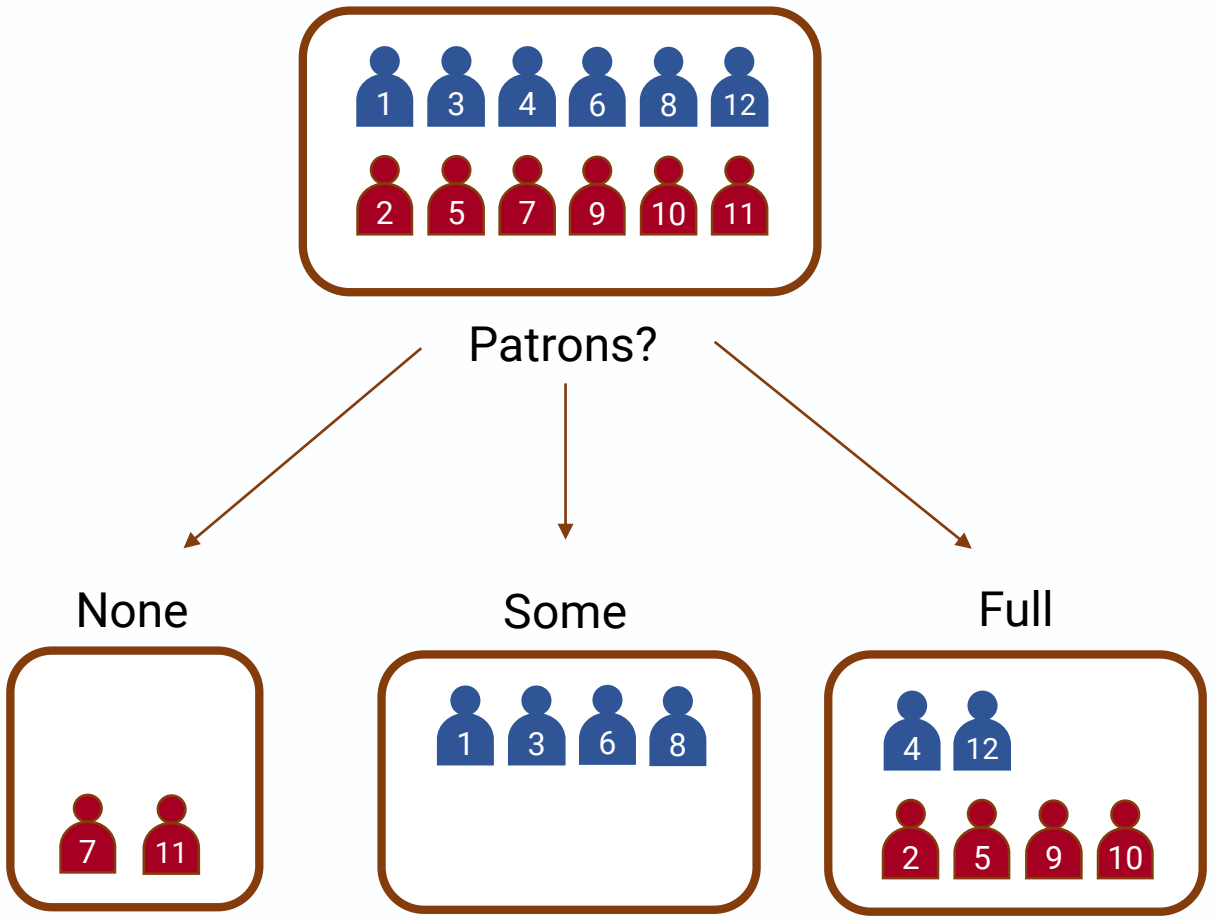
  Yes: 0 No: 2

- Some

  Yes: 4 No: 0

- Full

Group and Split Overview



Feature: Patrons

(Are there people in the restaurant?)

Possible answers/categories:

- None:

  Yes: 0 No: 2

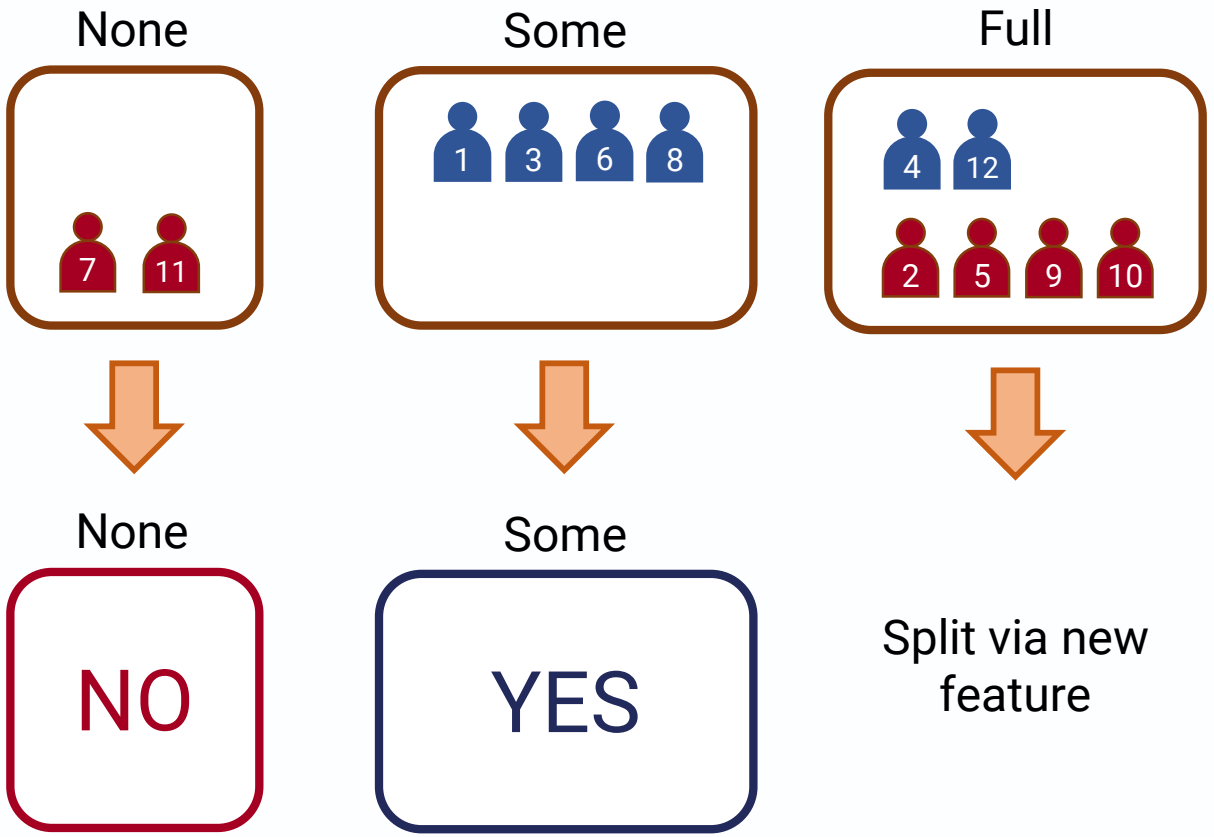
- Some

  Yes: 4 No: 0

- Full

  Yes: 2 No: 4

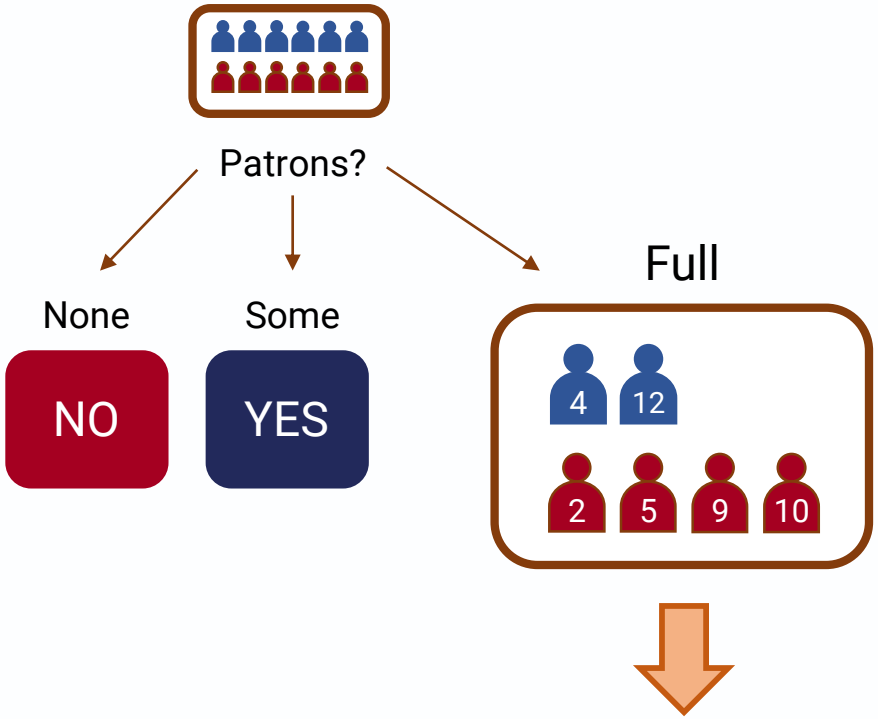
Group and Split Overview



What to do with new groups?

- Leaf (terminal node)
  - All instances in the group have the same label
  - e.g. all yes or all no
- Branch
  - Instances have mixed labels
  - e.g. mix of yes and no
  - Considered as a new group to split

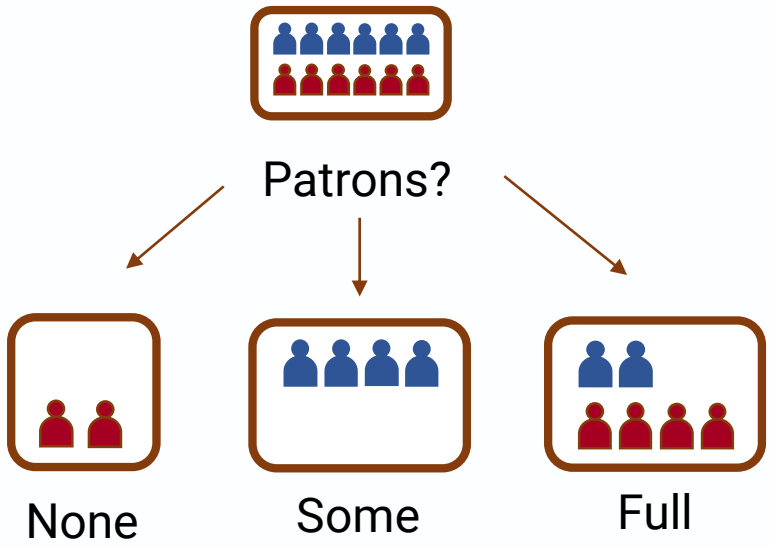
Choosing the Feature



Split via new feature…

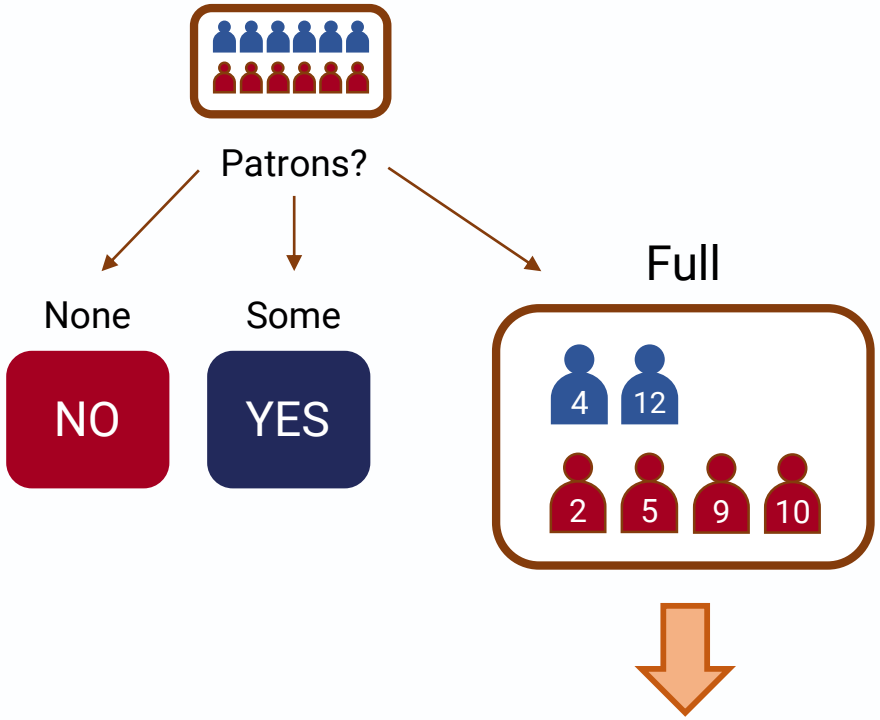
… But which feature?

Choosing the Feature



Why did we use the Patrons feature for branching?

Choosing the Feature



Split via new feature…

But which feature?

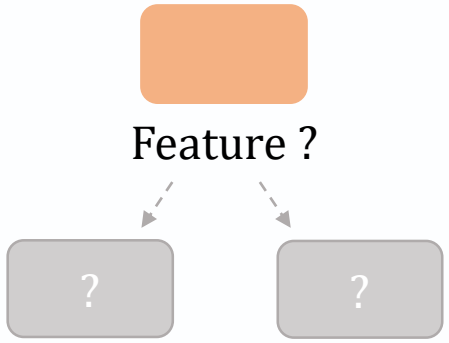
How do we choose which feature to use?

GINI INDEX!

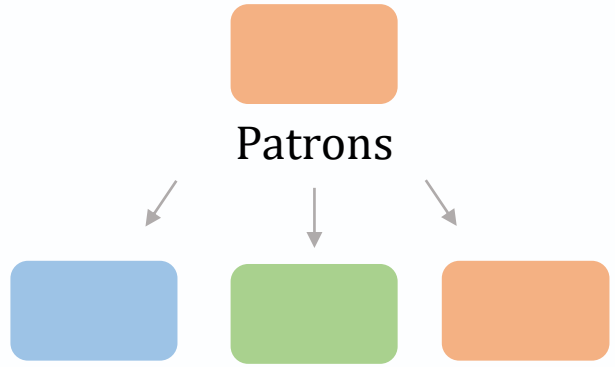
1. For each feature, compute Gini index for each of its categories
2. Compute for the weighted average of the feature's Gini indices
3. Select the feature with smallest (weighted average) Gini index

DECISION TREE

BRANCHING



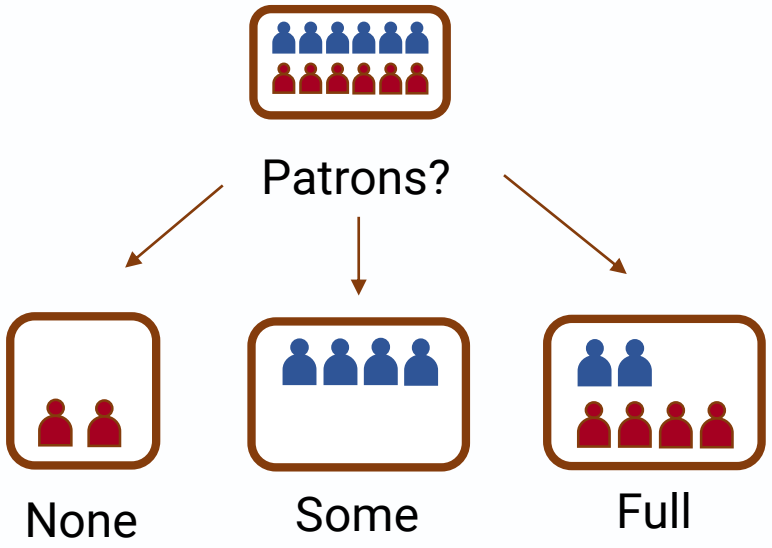
Feature	GINI Score
Patrons	0.28
Hungry	0.37
Type	0.50
Friday	0.49



FIND FEATURE WITH THE SMALLEST GINI SCORE

SPLIT/BRANCH ACCORDING TO THAT FEATURE

GINI Index



How do we choose which feature to use?

1. For each feature, compute Gini index for each of its categories
2. Compute for the weighted average of the feature's Gini indices
3. Select the feature with smallest (weighted average) Gini index

GINI Index

Measure of Impurity: GINI

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

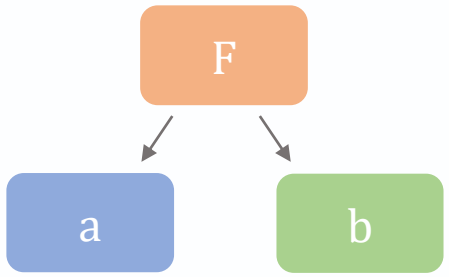
t : node (e.g. the category like none/some/full for Patrons)

j : class (e.g. the label like Yes/No for Will Wait)

$p(j|t)$: relative frequency of the class in the group

DECISION TREE

GINI INDEX FOR A FEATURE



PICK A FEATURE AND SPLIT

COMPUTE GINI INDEX FOR EACH NODE

$$GINI(a)\left(\frac{\#\ in\ a}{total\ \#}\right) + GINI(b)\left(\frac{\#\ in\ b}{total\ \#}\right)$$

GET THE WEIGHTED AVERAGE

Gini Index

In the Restaurant Example…

"Patrons" Feature

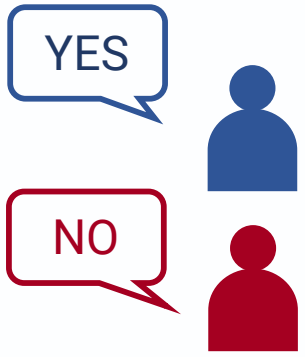
- 3 Categories (None/Some/Full)
- 2 Labels for Will Wait (Yes/No)

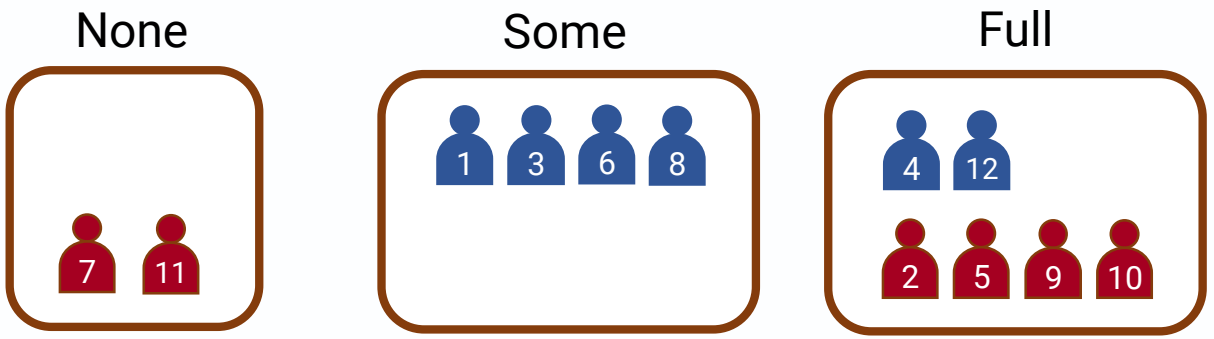
3 GINI indices:

GINI (None)

GINI (Some)

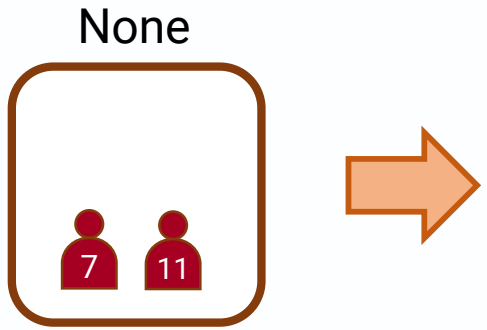
GINI (Full)





Gini Index

In the Restaurant Example…



Yes: 0

No: 2

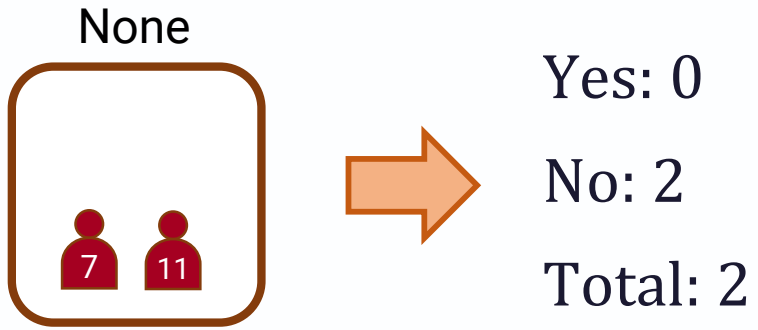
Total: 2

$$GINI(None) = 1 - \left(\frac{total\ \#\ "Yes"}{total\ under\ None}\right)^2 - \left(\frac{total\ \#\ "No"}{total\ under\ None}\right)^2$$

$$GINI(None) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2$$

Gini Index

In the Restaurant Example…

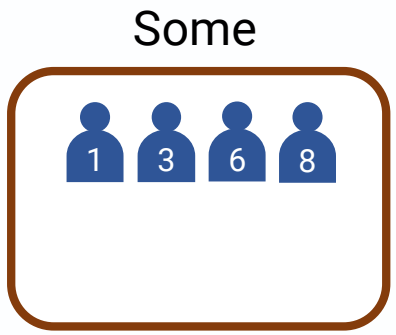


$$GINI(None) = 1 - \left(\frac{total\ \#\ "Yes"}{total\ under\ None}\right)^2 - \left(\frac{total\ \#\ "No"}{total\ under\ None}\right)^2$$

$$GINI(None) = 1 - 0 - 1 = 0$$

Gini Index

In the Restaurant Example…



Yes: 4

No: 0

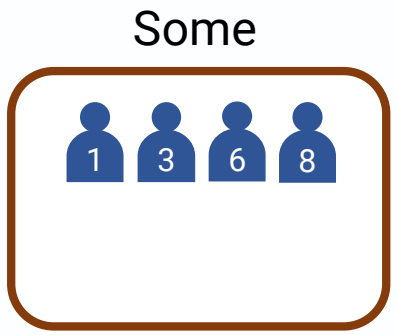
Total: 4

$$GINI(Some) = 1 - \left(\frac{total\ \#\ "Yes"}{total\ under\ Some}\right)^2 - \left(\frac{total\ \#\ "No"}{total\ under\ Some}\right)^2$$

$$GINI(Some) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2$$

Gini Index

In the Restaurant Example…



Yes: 4

No: 0

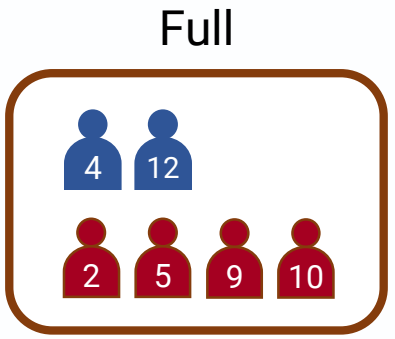
Total: 4

$$GINI(Some) = 1 - \left(\frac{total\ \#\ "Yes"}{total\ under\ Some}\right)^2 - \left(\frac{total\ \#\ "No"}{total\ under\ Some}\right)^2$$

$$GINI(Some) = 1 - 1 - 0 = 0$$

Gini Index

In the Restaurant Example…



Yes: 2

No: 4

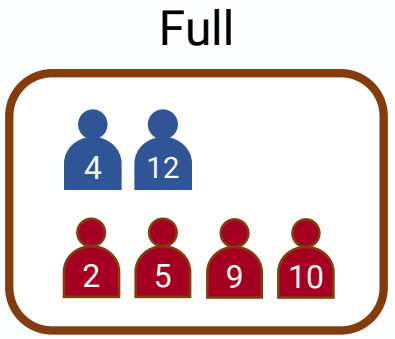
Total: 6

$$GINI(Full) = 1 - \left(\frac{total\ \#\ "Yes"}{total\ under\ Full}\right)^2 - \left(\frac{total\ \#\ "No"}{total\ under\ Full}\right)^2$$

$$GINI(Full) = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2$$

Gini Index

In the Restaurant Example…



Yes: 2

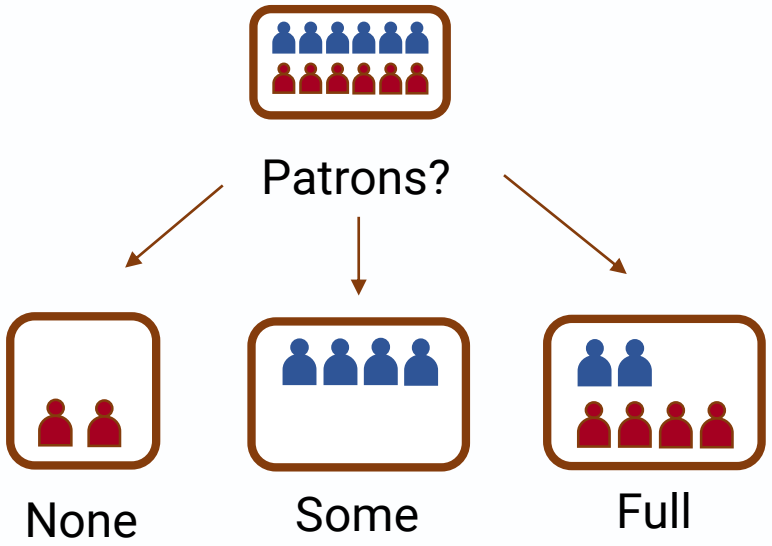
No: 4

Total: 6

$$GINI(Full) = 1 - \left(\frac{total\ \#\ "Yes"}{total\ under\ Full}\right)^2 - \left(\frac{total\ \#\ "No"}{total\ under\ Full}\right)^2$$

$$GINI(Full) = 1 - 0.111 - 0.444 = 0.56$$

GINI Index



$$\text{GINI (None)} = 0$$

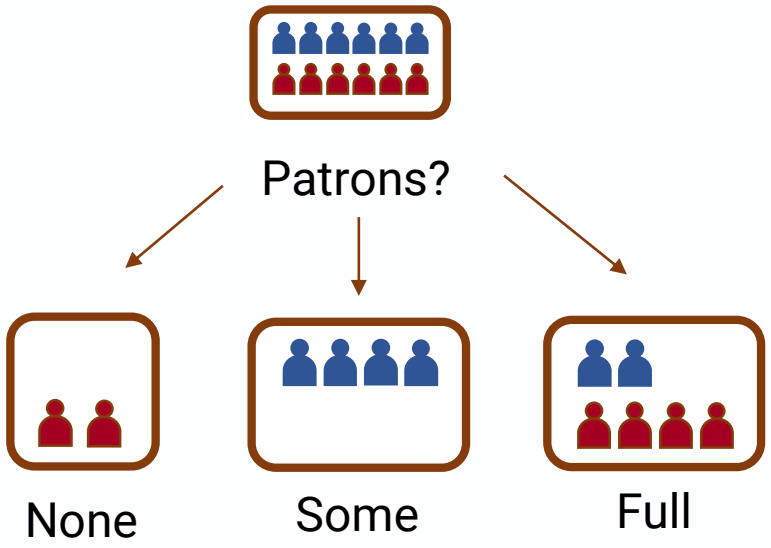
$$\text{GINI (Some)} = 0$$

$$\text{GINI (Full)} = 0.55$$

How do we choose which feature to use?

1. For each feature, compute Gini index for each of its categories
2. Compute for the weighted average of the feature's Gini indices
3. Select the feature with smallest (weighted average) Gini index

GINI Index



$\text{GINI(None)}\frac{2}{12}$ $\text{GINI(Some)}\frac{4}{12}$ $\text{GINI(Full)}\frac{6}{12}$

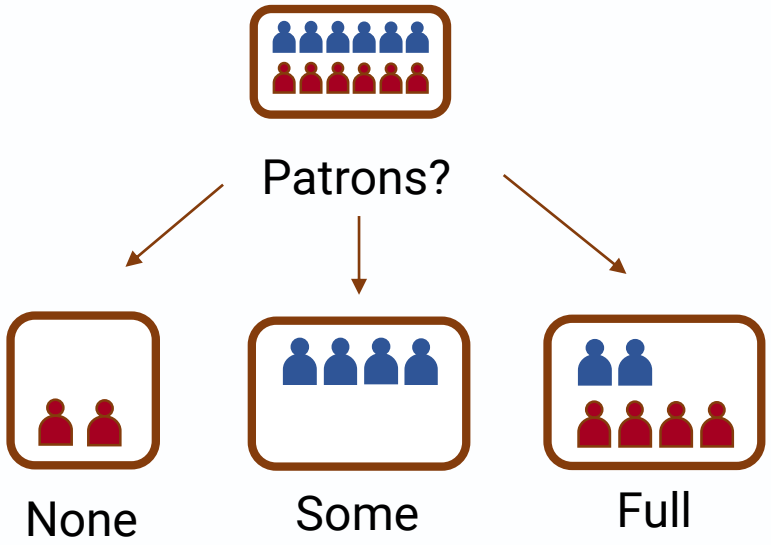
Computing for Weighted Average:

- For each category, multiply GINI(category) by:

$$\frac{total\ \#\ in < category >}{total\ \#\ in\ all\ categories}$$

- Then sum all values to get the weighted average
- Also called $GINI_{split}$

GINI Index



Total # in None: 2

Total # in Some: 4

Total # in Full: 6

Total # in all: 12

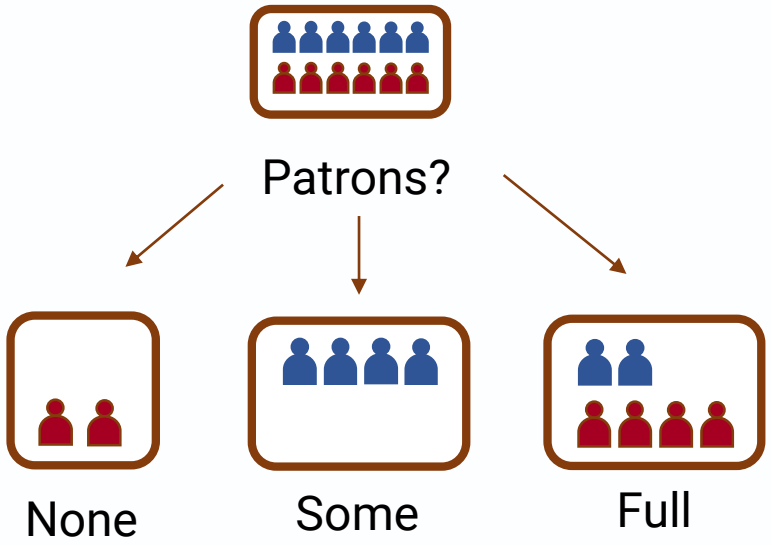
GINI (None) = 0

GINI (Some) = 0

GINI (Full) = 0.56

$$GINI_{split} = GINI(None)\frac{total\ \#\ in\ None}{total\ \#\ in\ all} + GINI(Some)\frac{total\ \#\ in\ Some}{total\ \#\ in\ all} + GINI(Full)\frac{total\ \#\ in\ Full}{total\ \#\ in\ all}$$

GINI Index



Total # in None: 2

Total # in Some: 4

Total # in Full: 6

Total # in all: 12

GINI (None) = 0

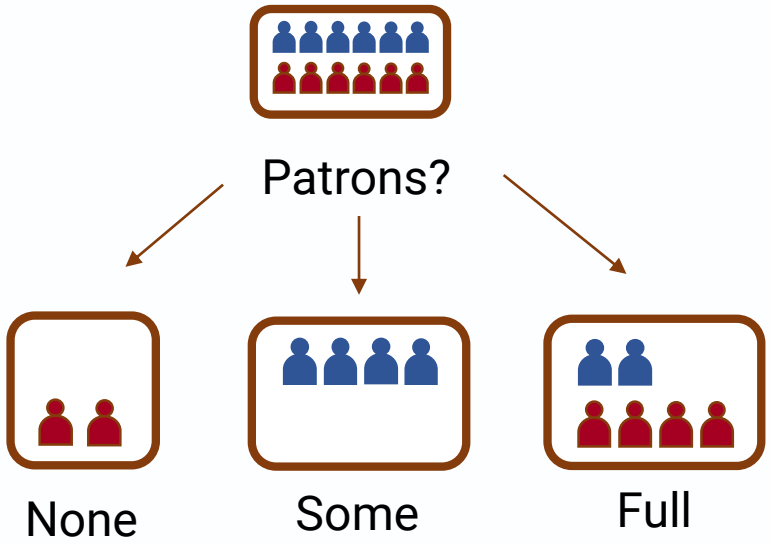
GINI (Some) = 0

GINI (Full) = 0.56

$$GINI_{split} = GINI(None)\frac{total\ \#\ in\ None}{total\ \#\ in\ all} + GINI(Some)\frac{total\ \#\ in\ Some}{total\ \#\ in\ all} + GINI(Full)\frac{total\ \#\ in\ Full}{total\ \#\ in\ all}$$

$$GINI_{split} = GINI(None)\left(\frac{2}{12}\right) + GINI(Some)\left(\frac{4}{12}\right) + GINI(Full)\left(\frac{6}{12}\right)$$

GINI Index



Total # in None: 2

Total # in Some: 4

Total # in Full: 6

Total # in all: 12

GINI (None) = 0

GINI (Some) = 0

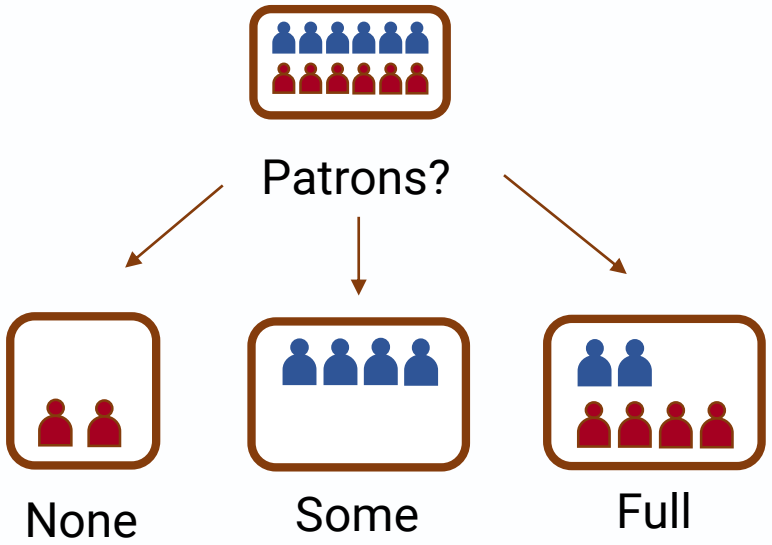
GINI (Full) = 0.56

$$GINI_{split} = GINI(None)\frac{total\ \#\ in\ None}{total\ \#\ in\ all} + GINI(Some)\frac{total\ \#\ in\ Some}{total\ \#\ in\ all} + GINI(Full)\frac{total\ \#\ in\ Full}{total\ \#\ in\ all}$$

$$GINI_{split} = GINI(None)\left(\frac{2}{12}\right) + GINI(Some)\left(\frac{4}{12}\right) + GINI(Full)\left(\frac{6}{12}\right)$$

$$GINI_{split} = 0\left(\frac{2}{12}\right) + 0\left(\frac{4}{12}\right) + 0.56\left(\frac{6}{12}\right) = 0.28$$

GINI Index

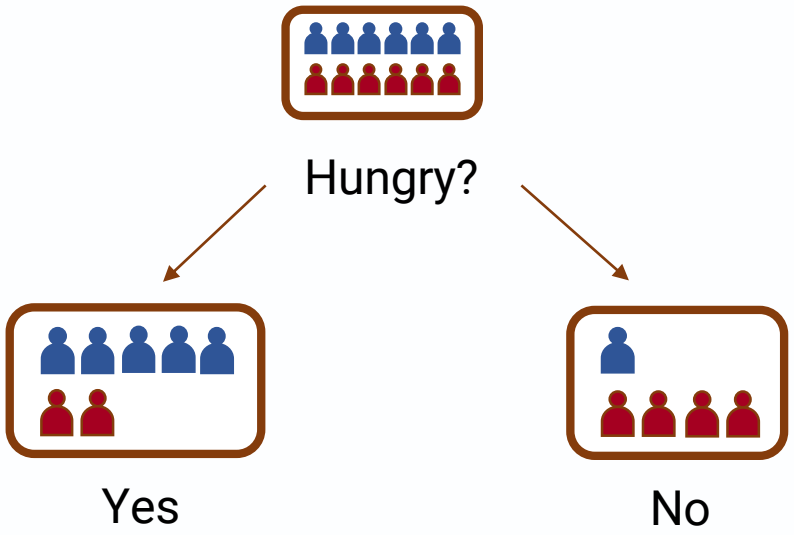


GINI for Patrons: 0.28

How do we choose which feature to use?

1. For each feature, compute Gini index for each of its categories
2. Compute for the weighted average of the feature's Gini indices
3. Select the feature with smallest (weighted average) Gini index

GINI Index



GINI for Patrons: 0.28

GINI for Hungry: _____

How do we choose which feature to use?

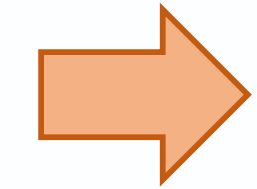
1. For each feature, compute Gini index for each of its categories
2. Compute for the weighted average of the feature's Gini indices
3. Select the feature with smallest (weighted average) Gini index

Gini Index

In the Restaurant Example…

"Hungry" Feature

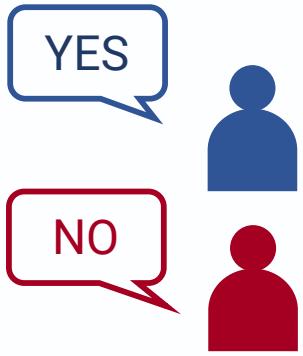
- 2 Categories for hungry (Yes/No)
- 2 Labels for Will Wait (Yes/No)



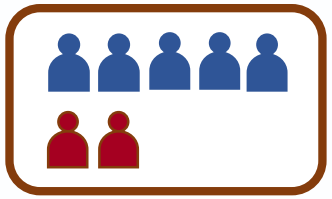
2 GINI indices:

GINI (Hungry)

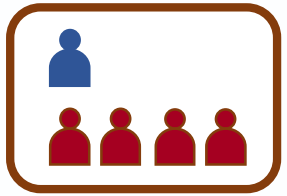
GINI (Not Hungry)



Yes (Hungry)



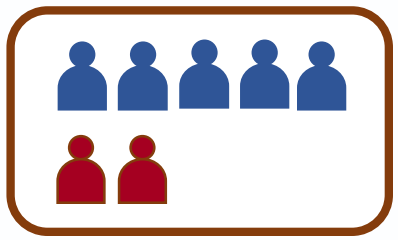
No (Not Hungry)



Gini Index

In the Restaurant Example…

Yes (Hungry)



Yes: 5

No: 2

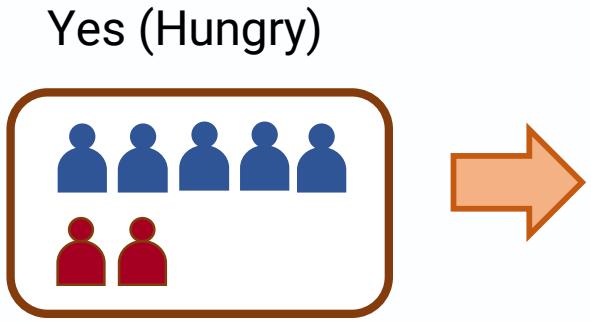
Total: 7

$$GINI(Hungry) = 1 - \left(\frac{total\ \#\ "Yes"}{total\ under\ Hungry}\right)^2 - \left(\frac{total\ \#\ "No"}{total\ under\ Hungry}\right)^2$$

$$GINI(Hungry) = 1 - \left(\frac{5}{7}\right)^2 - \left(\frac{2}{7}\right)^2$$

Gini Index

In the Restaurant Example…



Yes: 5

No: 2

Total: 7

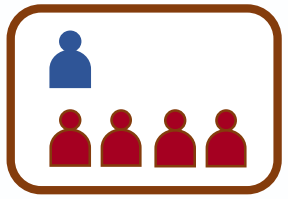
$$GINI(Hungry) = 1 - \left(\frac{total\ \#\ "Yes"}{total\ under\ Hungry}\right)^2 - \left(\frac{total\ \#\ "No"}{total\ under\ Hungry}\right)^2$$

$$GINI(Hungry) = 1 - 0.51 - 0.08 = 0.41$$

Gini Index

In the Restaurant Example…

No (Not Hungry)



Yes: 1

No: 4

Total: 5

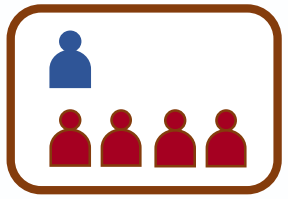
$$GINI(Not\ Hungry) = 1 - \left(\frac{total\ \#\ "Yes"}{total\ under\ Not\ Hungry}\right)^2 - \left(\frac{total\ \#\ "No"}{total\ under\ Not\ Hungry}\right)^2$$

$$GINI(Not\ Hungry) = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2$$

Gini Index

In the Restaurant Example…

No (Not Hungry)



Yes: 1

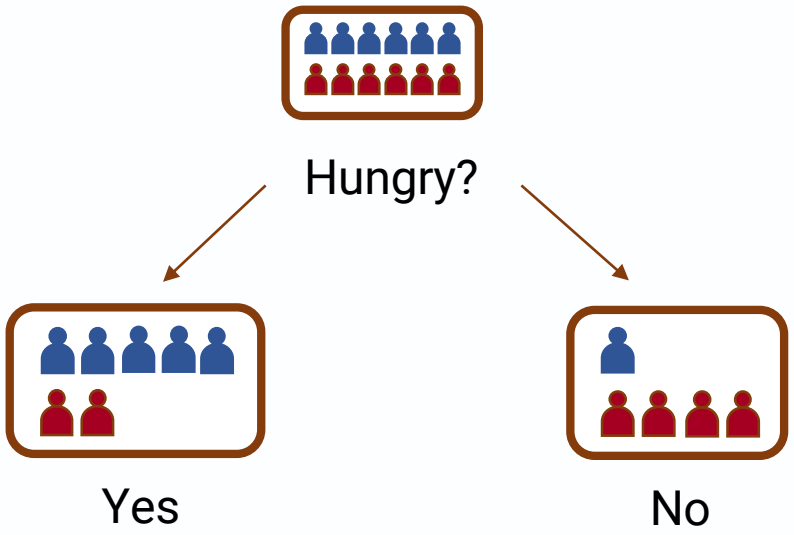
No: 4

Total: 5

$$GINI(Not\ Hungry) = 1 - \left(\frac{total\ \#\ "Yes"}{total\ under\ Not\ Hungry}\right)^2 - \left(\frac{total\ \#\ "No"}{total\ under\ Not\ Hungry}\right)^2$$

$$GINI(Not\ Hungry) = 1 - 0.04 - 0.64 = 0.32$$

GINI Index



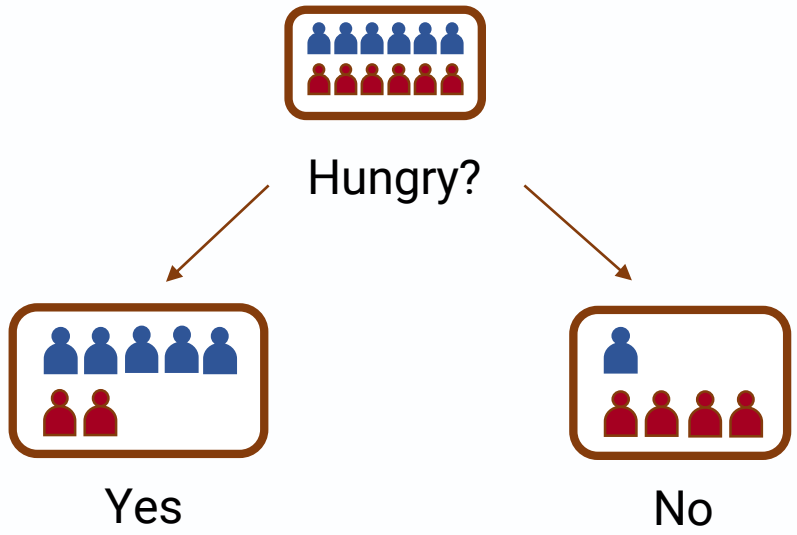
$$\text{GINI}(\text{Hungry}) = 0.41$$

$$\text{GINI}(\text{Not Hungry}) = 0.32$$

How do we choose which feature to use?

1. For each feature, compute Gini index for each of its categories
2. Compute for the weighted average of the feature's Gini indices
3. Select the feature with smallest (weighted average) Gini index

GINI Index



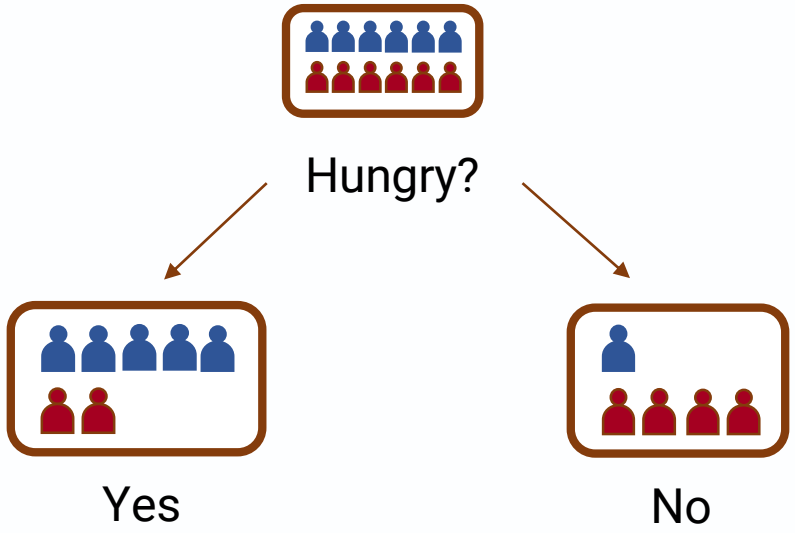
Hungry: $\frac{7}{12}$ Not Hungry: $\frac{5}{12}$

Weighted Average

Sum of the following value for all categories:

$$\text{GINI(category)}\left(\frac{total\ \#\ in\langle category\rangle}{total\ \#\ in\ all\ categories}\right)$$

GINI Index



Total # in Hungry: 7

Total # in Not Hungry: 5

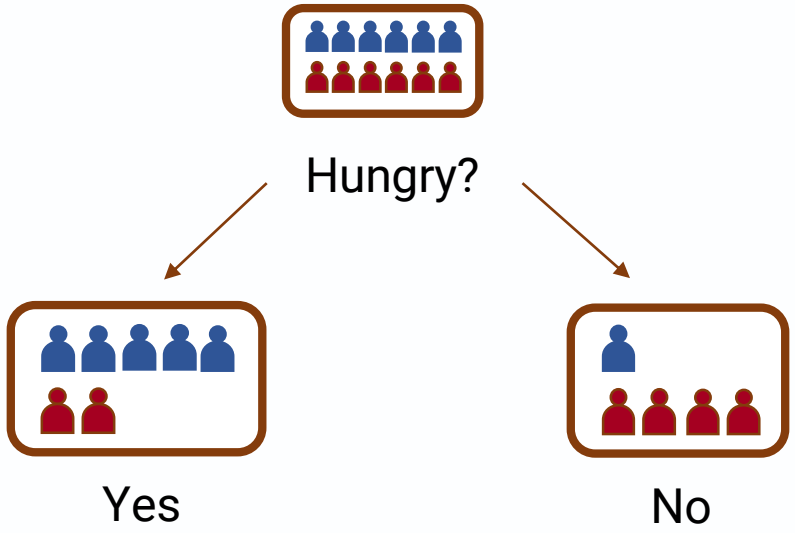
Total # in all: 12

GINI (Hungry) = 0.41

GINI (Not Hungry) = 0.32

$$GINI_{split} = GINI(Hungry)\frac{total\ \#\ in\ Hungry}{total\ \#\ in\ all} + GINI(Not\ Hungry)\frac{total\ \#\ in\ Not\ Hungry}{total\ \#\ in\ all}$$

GINI Index



Total # in Hungry: 7

Total # in Not Hungry: 5

Total # in all: 12

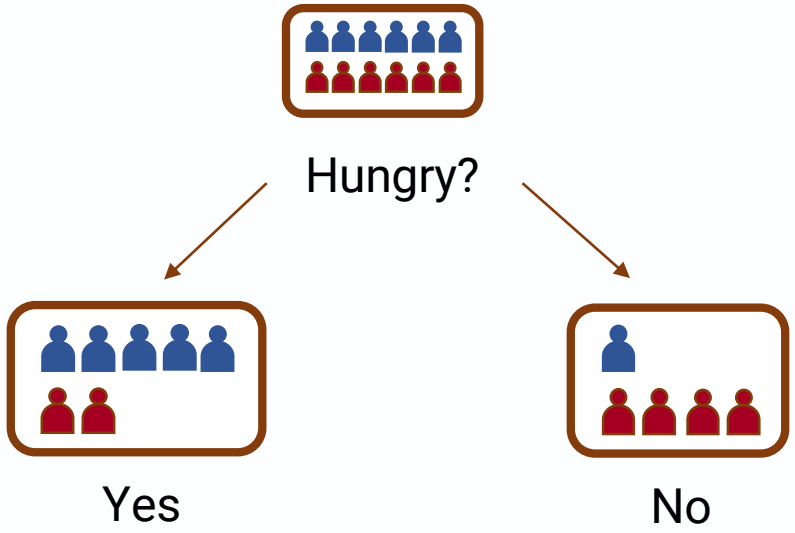
GINI (Hungry) = 0.41

GINI (Not Hungry) = 0.32

$$GINI_{split} = GINI(Hungry)\frac{total\ \#\ in\ Hungry}{total\ \#\ in\ all} + GINI(Not\ Hungry)\frac{total\ \#\ in\ Not\ Hungry}{total\ \#\ in\ all}$$

$$GINI_{split} = GINI(Hungry)\left(\frac{7}{12}\right) + GINI(Not\ Hungry)\left(\frac{5}{12}\right)$$

GINI Index



Total # in Hungry: 7

Total # in Not Hungry: 5

Total # in all: 12

GINI (Hungry) = 0.41

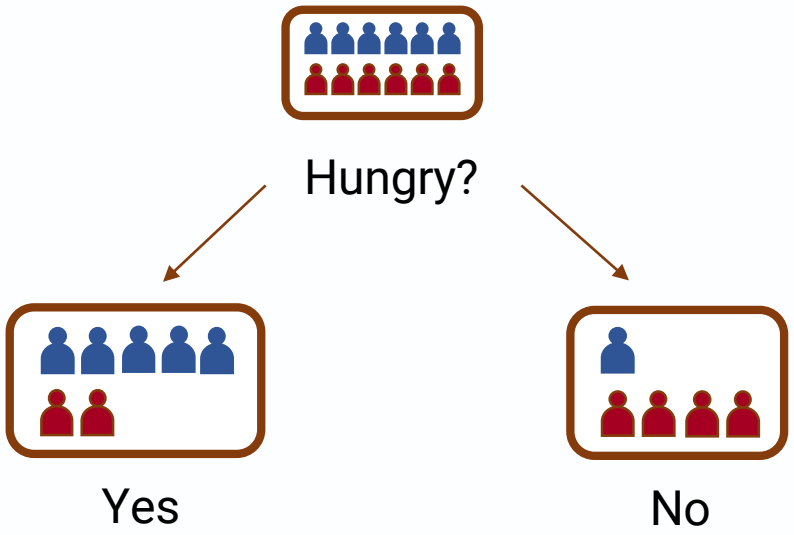
GINI (Not Hungry) = 0.32

$$GINI_{split} = GINI(Hungry)\frac{total\ \#\ in\ Hungry}{total\ \#\ in\ all} + GINI(Not\ Hungry)\frac{total\ \#\ in\ Not\ Hungry}{total\ \#\ in\ all}$$

$$GINI_{split} = GINI(Hungry)\left(\frac{7}{12}\right) + GINI(Not\ Hungry)\left(\frac{5}{12}\right)$$

$$GINI_{split} = 0.41\left(\frac{7}{12}\right) + 0.32\left(\frac{5}{12}\right) = 0.37$$

GINI Index



GINI for Patrons: 0.28

GINI for Hungry: 0.37

GINI for Type: 0.5

GINI for Fri/Sat: 0.49

How do we choose which feature to use?

1. For each feature, compute Gini index for each of its categories
2. Compute for the weighted average of the feature's Gini indices
3. Select the feature with smallest (weighted average) Gini index

Decision Tree: Example

Restaurant Example:

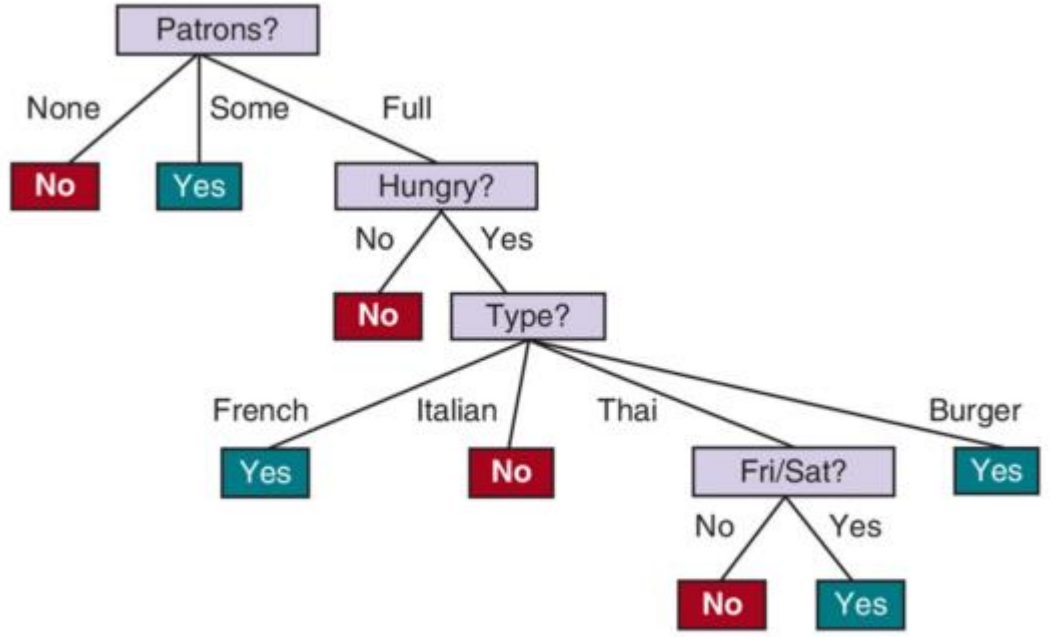
The feature Patrons is used as the root node because it has the smallest GINI index

GINI for Patrons: 0.28

GINI for Hungry: 0.37

GINI for Type: 0.5

GINI for Fri/Sat: 0.49

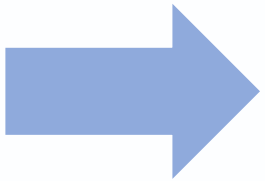


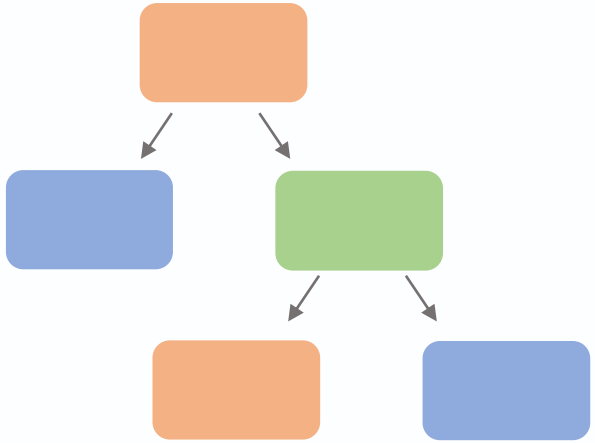
The decision tree induced from the 12-example training set.

Summary

INDUCE A DECISION TREE

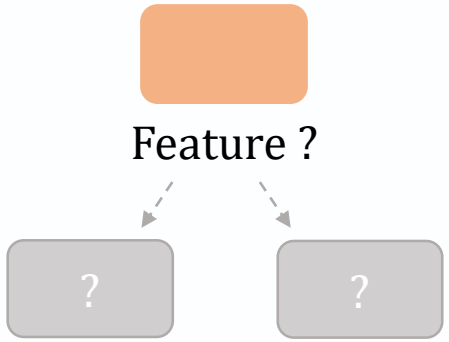
DECISION TREE MODEL (TRAINED)



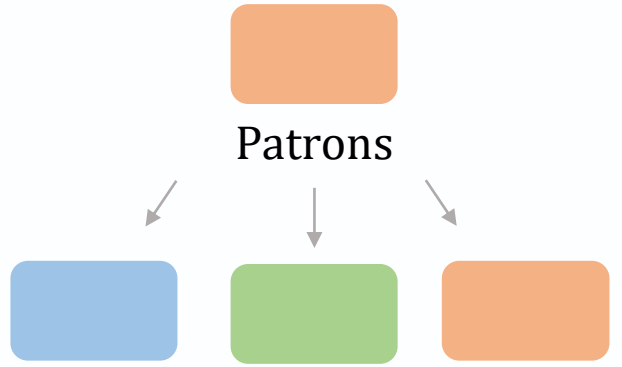


Summary

BRANCHING



Feature	GINI Score
Patrons	0.28
Hungry	0.37
Type	0.50
Friday	0.49

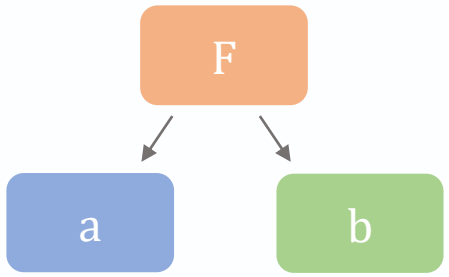


FIND FEATURE WITH THE SMALLEST GINI SCORE

SPLIT/BRANCH ACCORDING TO THAT FEATURE

Summary

GINI INDEX FOR A FEATURE



PICK A FEATURE AND SPLIT

COMPUTE GINI INDEX FOR EACH NODE

$$GINI(a)\left(\frac{\#\ in\ a}{total\ \#}\right) + GINI(b)\left(\frac{\#\ in\ b}{total\ \#}\right)$$

GET THE WEIGHTED AVERAGE