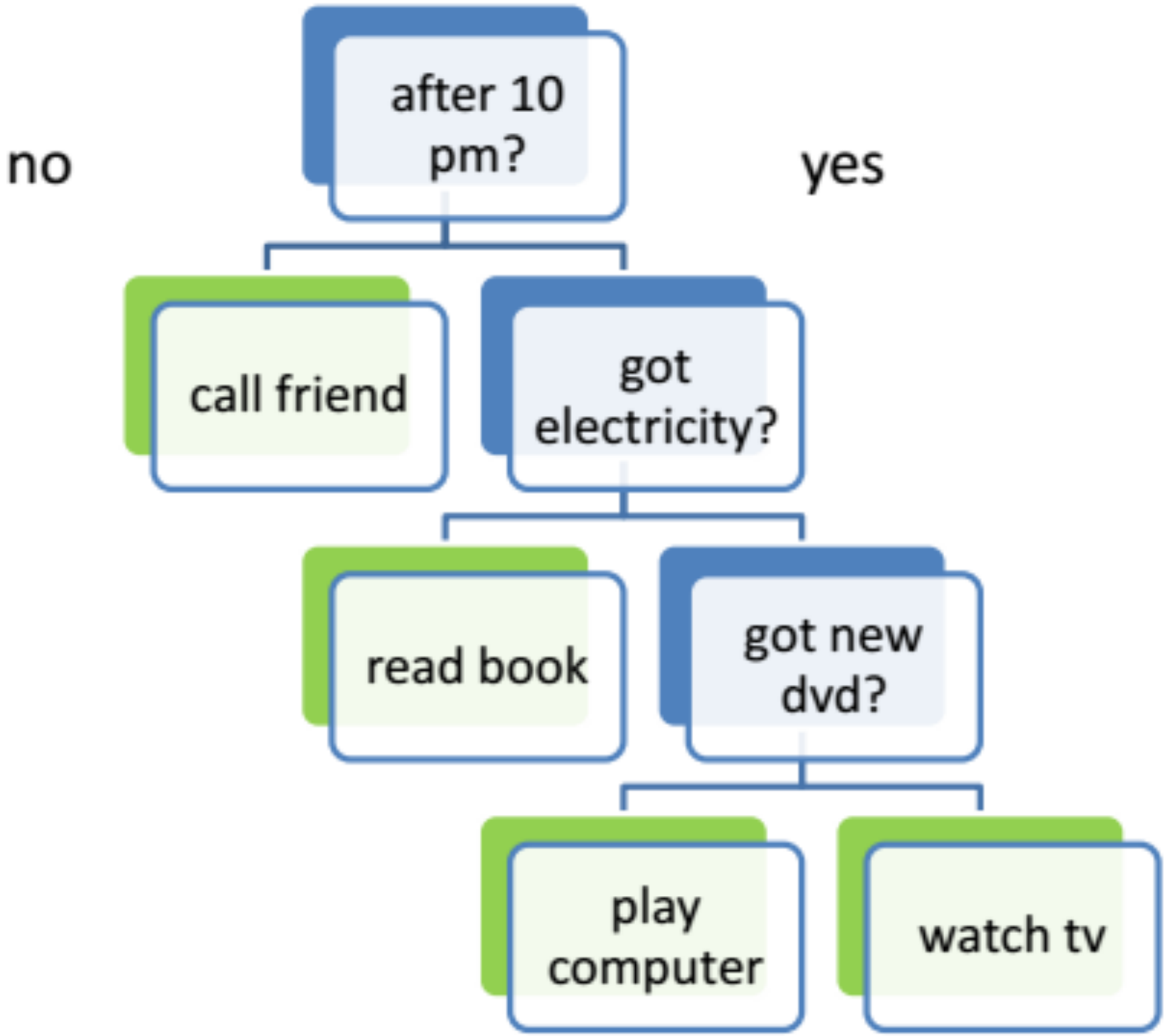
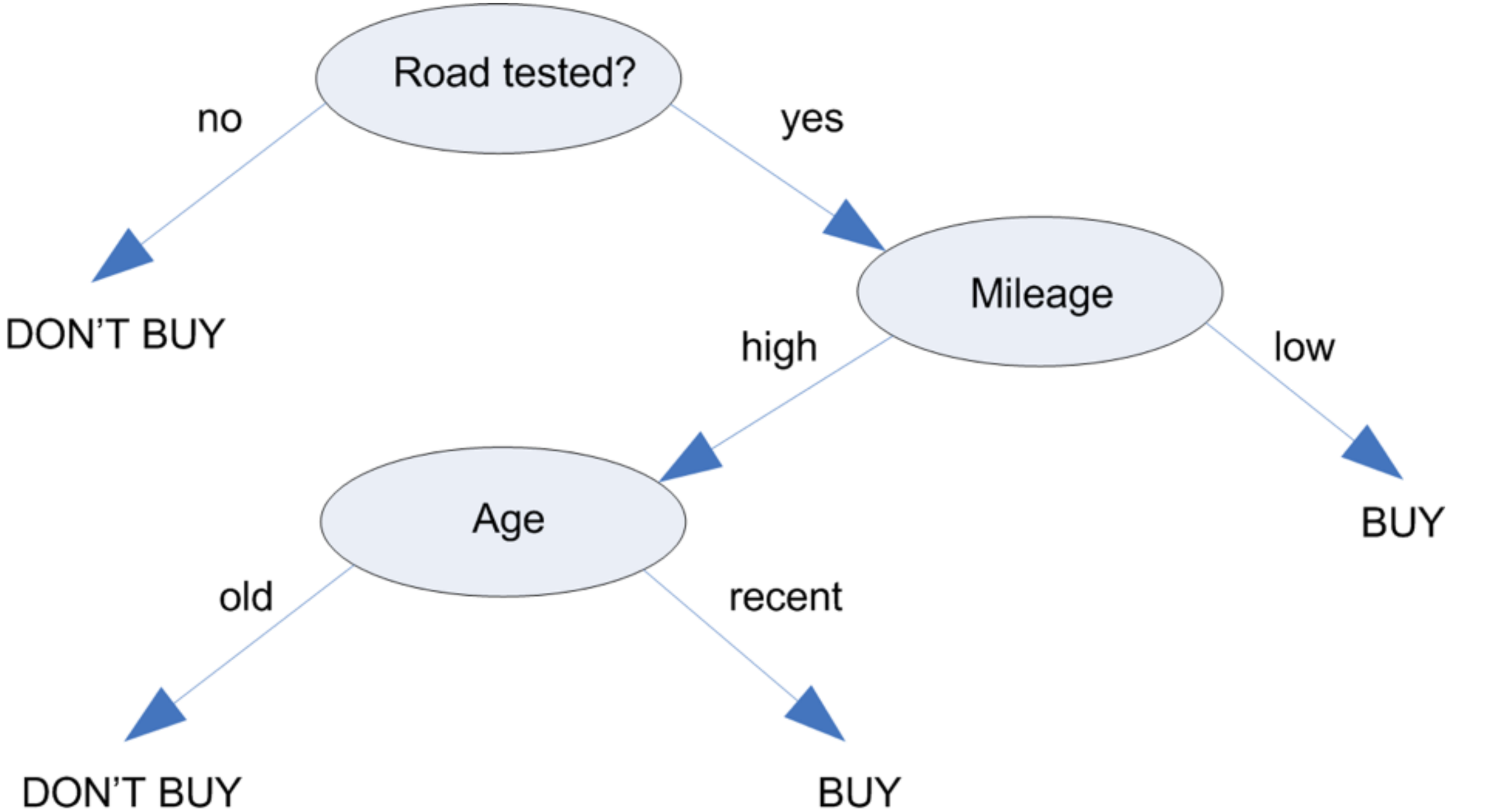
GENERAL ASSEMBLY

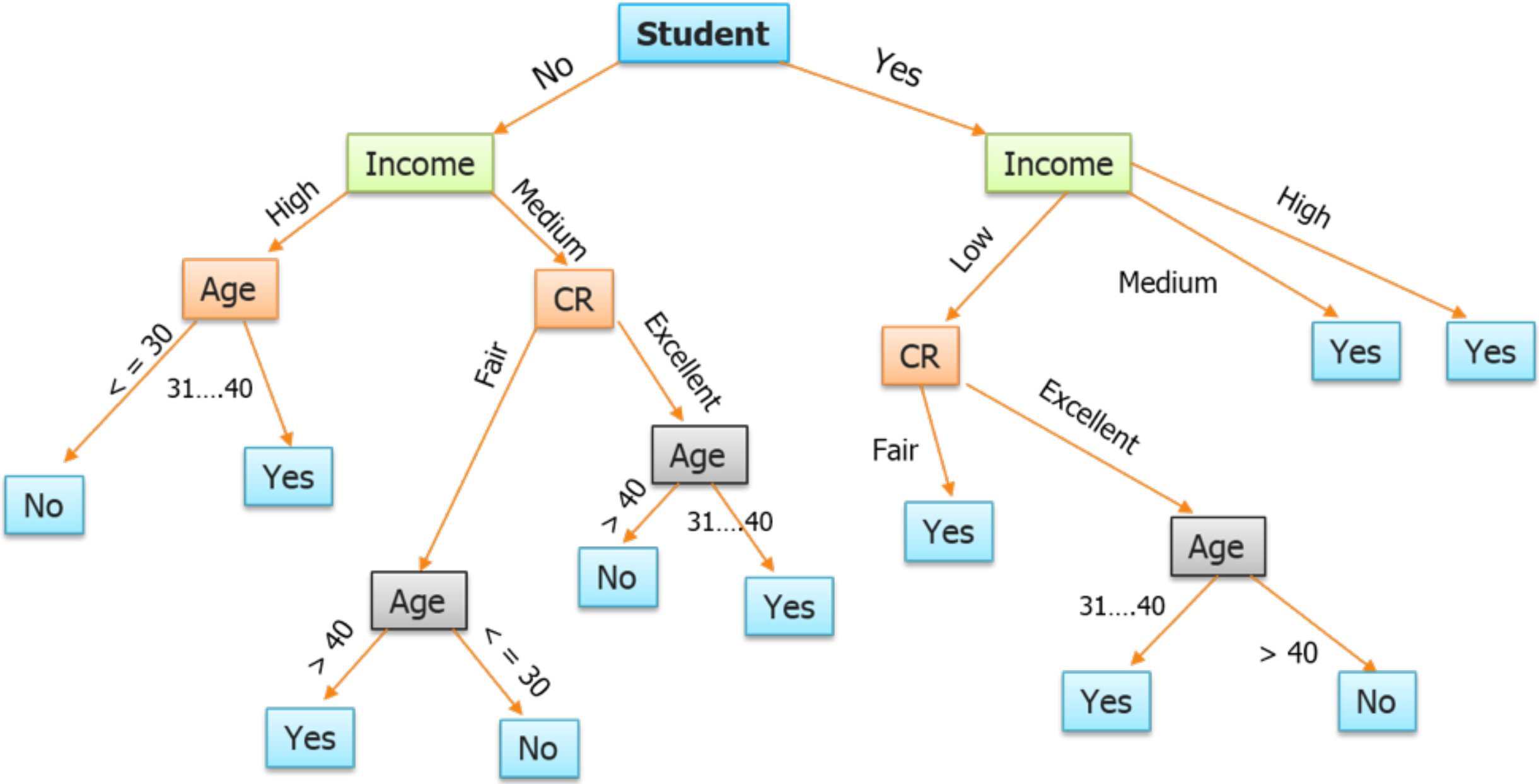
DECISION TREES

DSI NYC 1
July 11 2016


no
after 10 pm?
yes
call friend
got electricity?
read book
got new dvd?
play computer
watch tv


Road tested?
no
yes
DON’T BUY
Mileage
high
low
Age
BUY
old
recent
DON’T BUY
BUY

Which students bought a computer?



OBJECTIVE:

Understand when and how to use a decision tree model.

AGENDA:

Go over how a decision tree works.
Build a simple decision tree by hand.
Review the advantages and disadvantages of a single-tree model.
Look over case studies of decision trees in action

DECISION TREES

NON-PARAMETRIC[1] HIERARCHICAL[2] CLASSIFICATION AND REGRESSION[3] TECHNIQUE

DECISION TREES

NON-PARAMETRIC[1] HIERARCHICAL[2] CLASSIFICATION AND REGRESSION[3] TECHNIQUE

1 - not defined in terms of parameters. No assumptions of normality within features.

DECISION TREES

NON-PARAMETRIC[1] HIERARCHICAL[2] CLASSIFICATION AND REGRESSION[3] TECHNIQUE

1 - not defined in terms of parameters. No assumptions of normality within features.

2 - the model will lead to a class label when applied to any record. Think of it like a recipe.

DECISION TREES

NON-PARAMETRIC[1] HIERARCHICAL[2] CLASSIFICATION AND REGRESSION[3] TECHNIQUE

1 - not defined in terms of parameters. No assumptions of normality within features.

2 - the model will lead to a class label when applied to any record. Think of it like a recipe.

3 - can be used for classification or regression

DECISION TREES *… simplified*

PREDICTIONS

Based on a series of if-then rules

FITTING A MODEL

Create if-then rules one at a time to cover each case as well as you can

DECISION TREES *... simplified*

PREDICTIONS

Based on a series of if-then rules

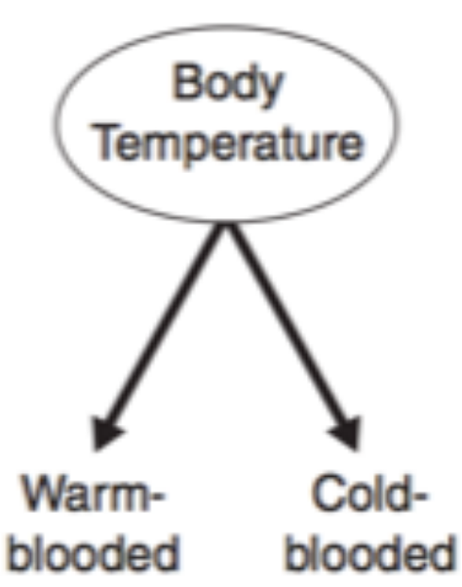
FITTING A MODEL

Create if-then rules one at a time to cover each case as well as you can

> *get as much information as you can from each rule*

TYPES OF SPLITS

- For distinct features:



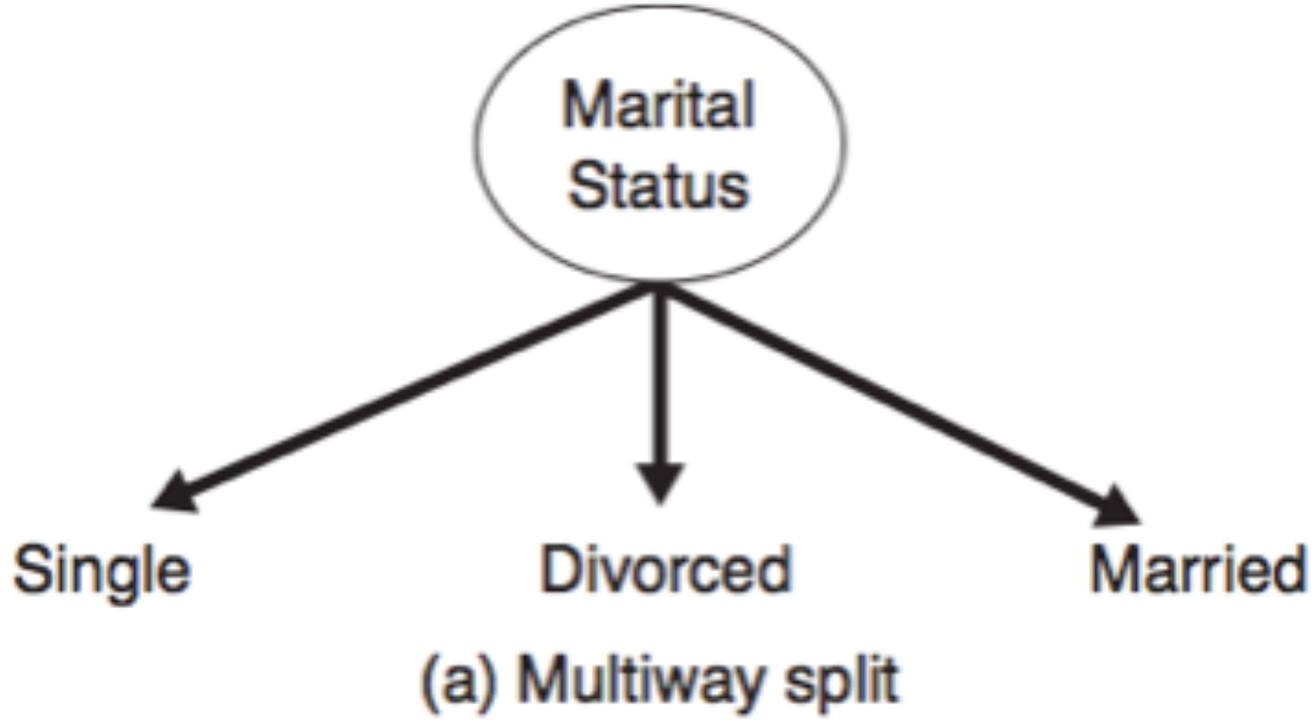


Figure 4.8. Test condition for binary attributes.

TYPES OF SPLITS

- For continuous features:

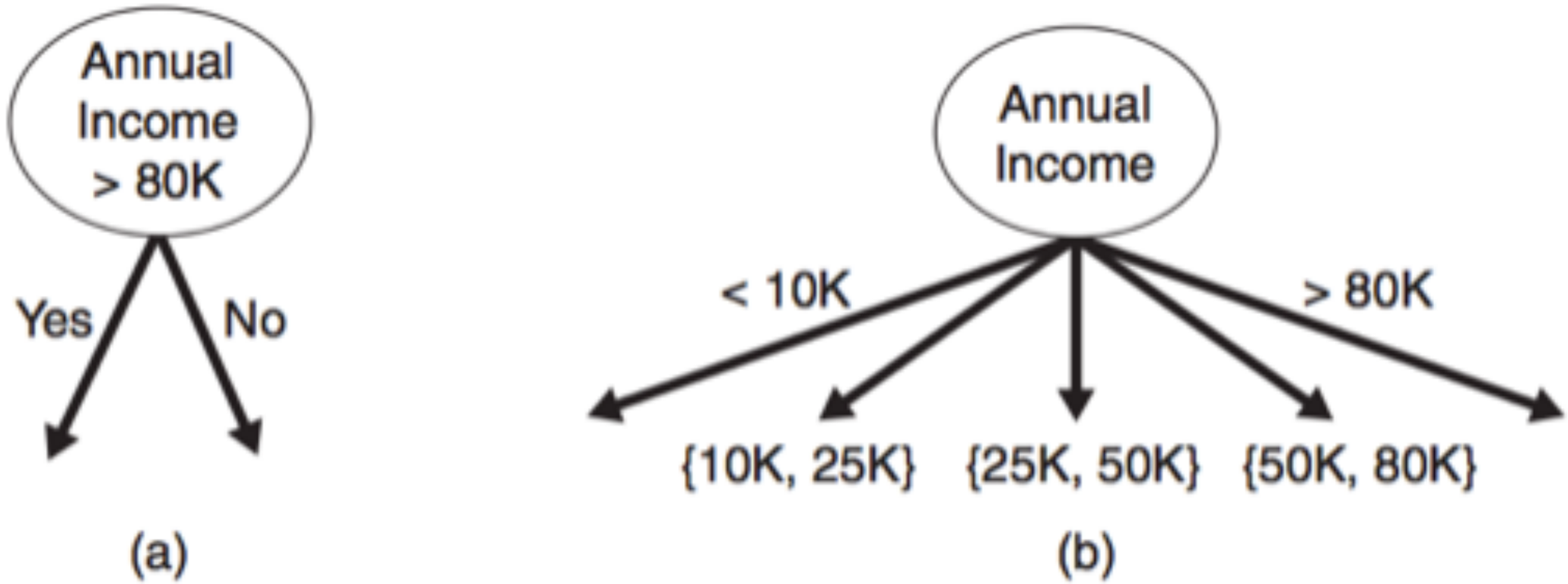
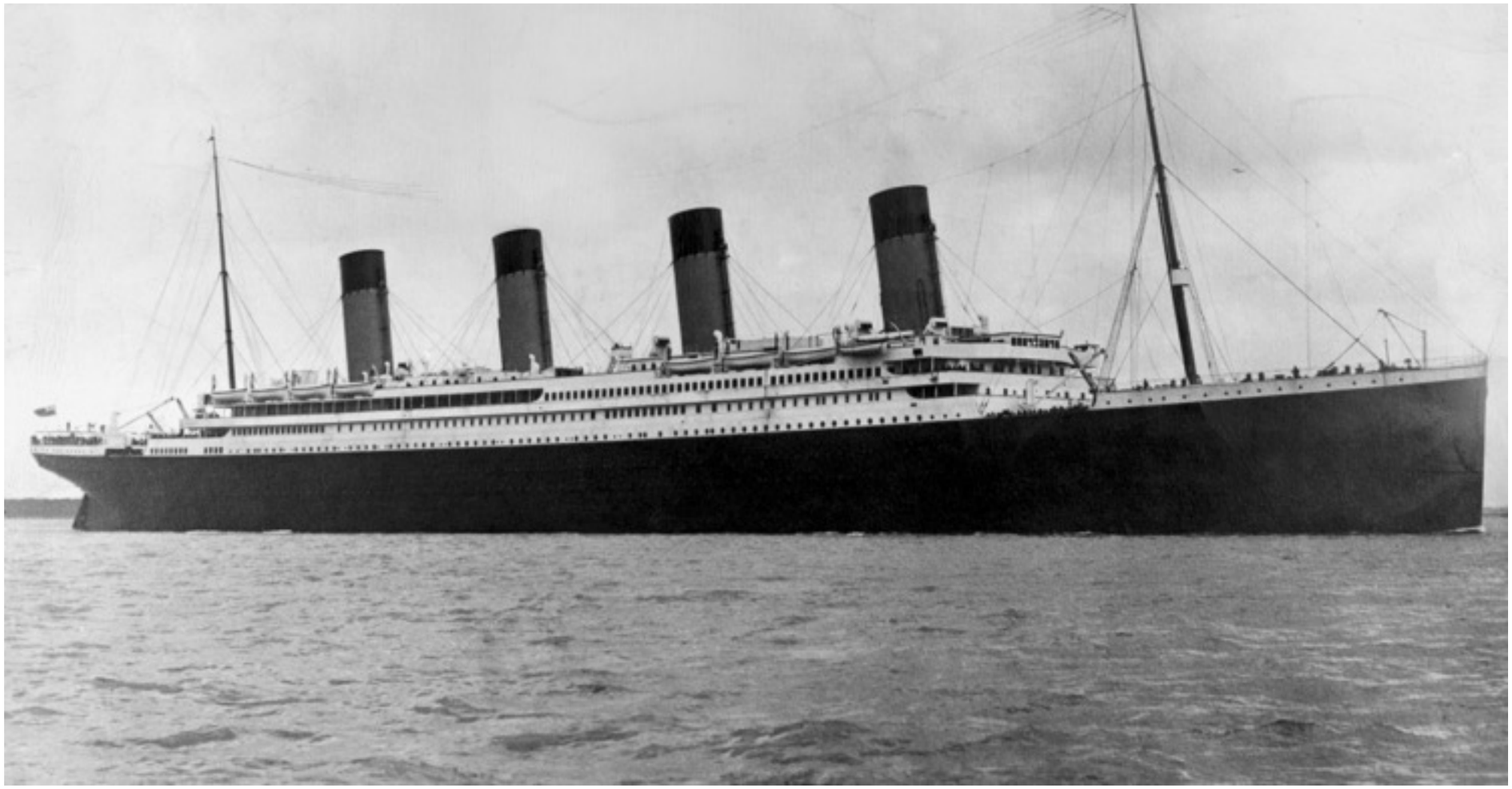


Figure 4.11. Test condition for continuous attributes.

FINDING THE BEST SPLIT

- Maximize ‘purity’ or ‘information gain’ at each split
- There are algorithms that do this work for you.
  - One example: Hunt’s algorithm. A greedy, recursive solution.
    - Greedy: stores lots of possible outcomes in memory
    - Recursive: splits tasks into subtasks and solves each the same way

PREDICTING SURVIVORS ON THE TITANIC



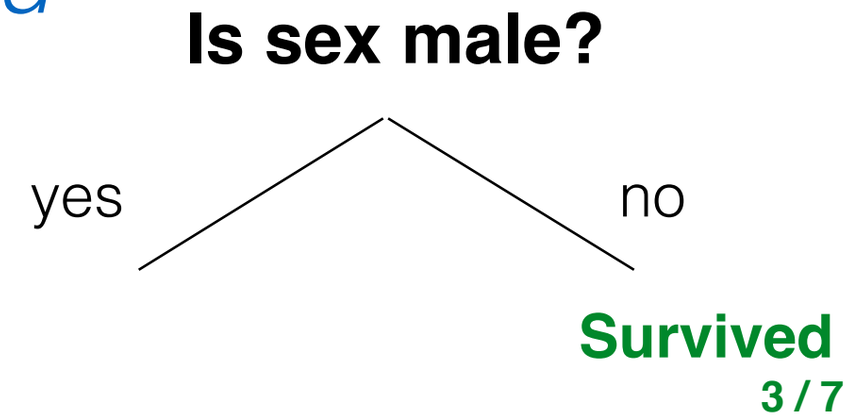
PREDICTING SURVIVORS ON THE TITANIC

PREDICTING SURVIVORS ON THE TITANIC

#	Sex	Age	# of family members	Ticket price	Survived?
1	Male	48	0	$4	No
2	Male	23	0	$4	No
3	Female	34	2	$59	Yes
4	Female	19	1	$59	Yes
5	Male	7	0	$128	Yes
6	Male	68	2	$4	No
7	Female	24	3	$4	Yes

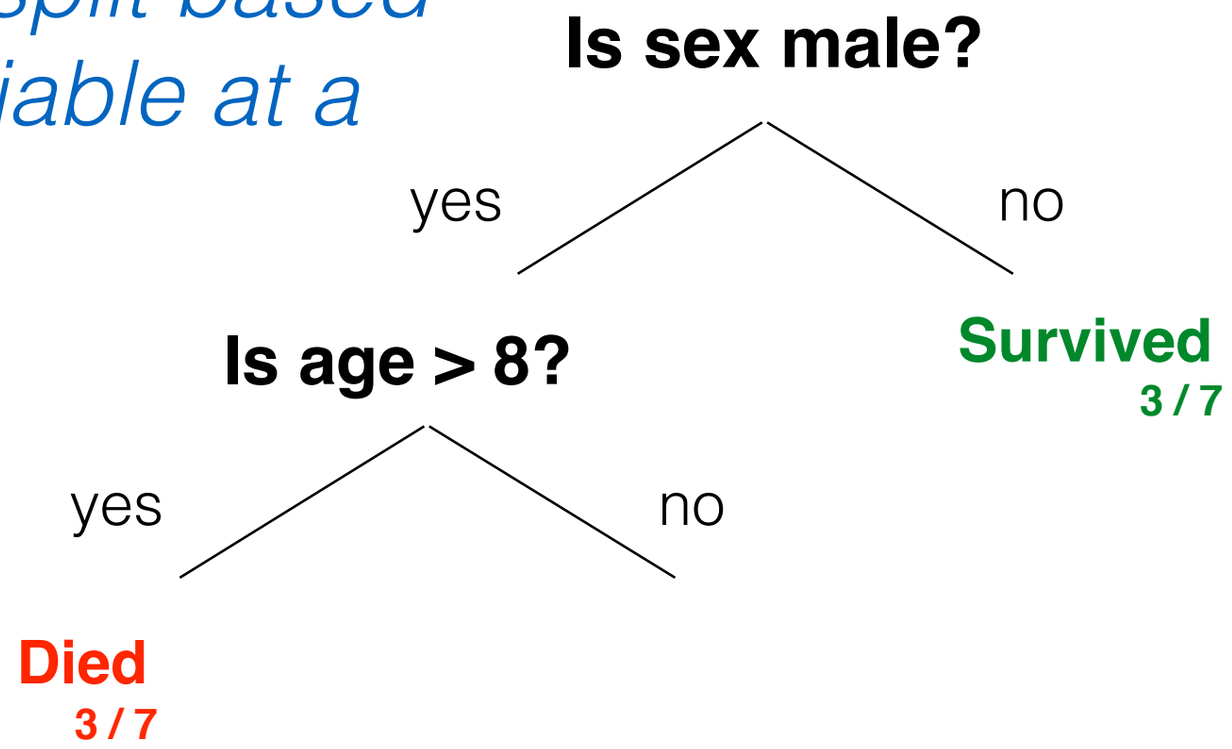
PREDICTING SURVIVORS ON THE TITANIC

make a split based on 1 variable at a time



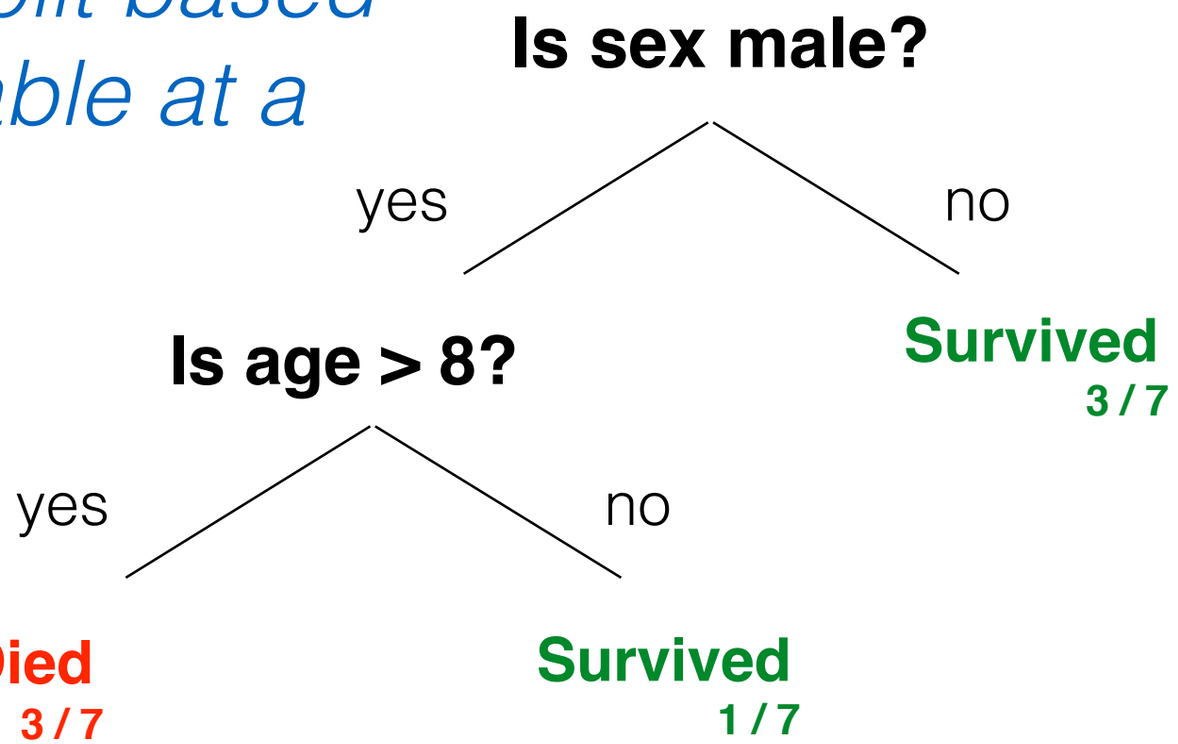
PREDICTING SURVIVORS ON THE TITANIC

make a split based on 1 variable at a time



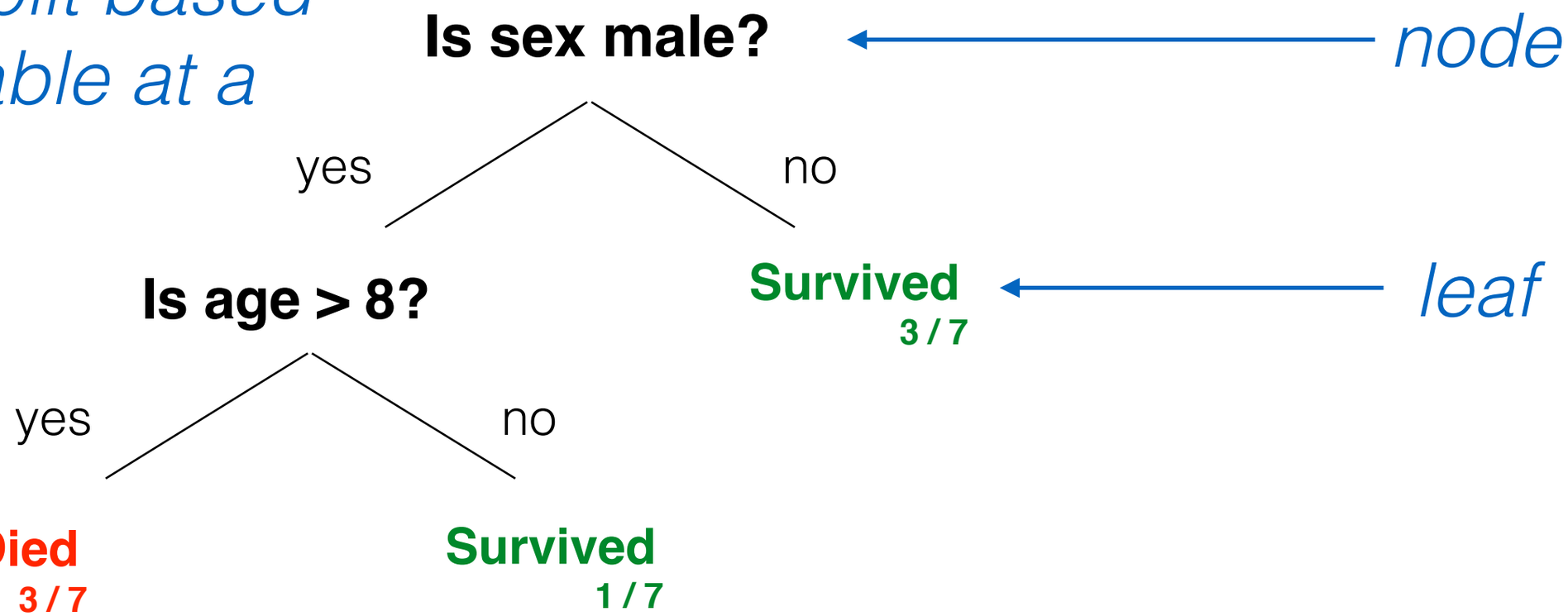
PREDICTING SURVIVORS ON THE TITANIC

make a split based on 1 variable at a time



PREDICTING SURVIVORS ON THE TITANIC

make a split based on 1 variable at a time



SHOULD I PLAY TENNIS?

SHOULD I PLAY TENNIS?

Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcast	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No

SHOULD I PLAY TENNIS?

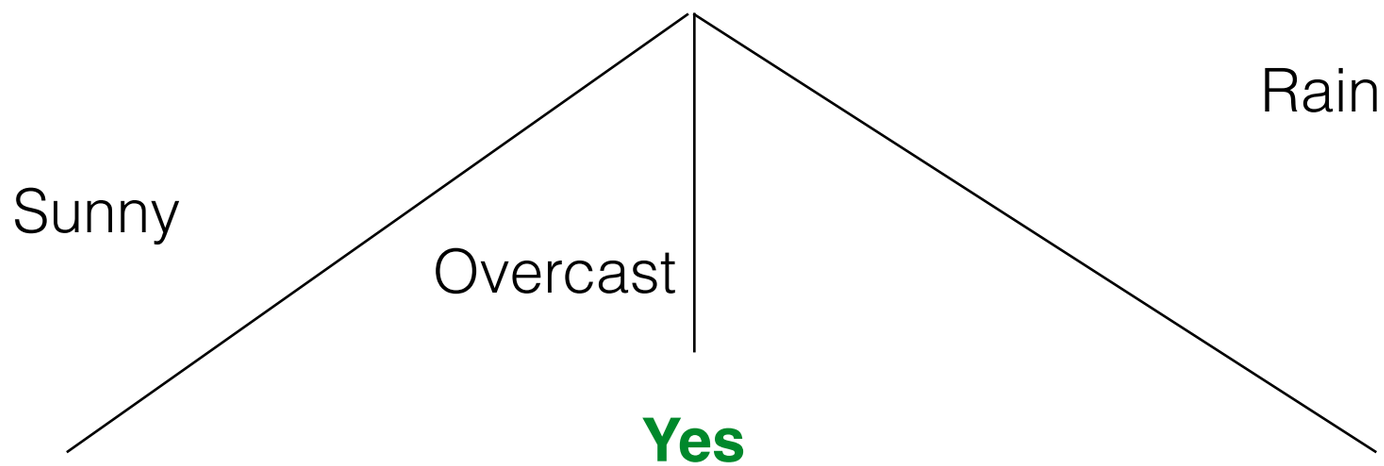
SHOULD I PLAY TENNIS?

activity:
build your own tree

SHOULD I PLAY TENNIS?

activity:
build your own tree

What's the outlook?



SHOULD I PLAY TENNIS?

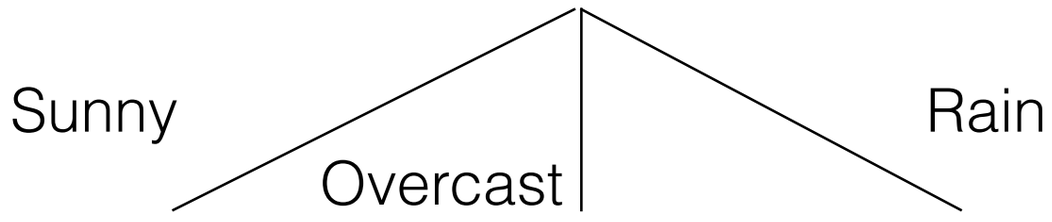
Day	Outlook	Temp	Humidity	Wind	Play Tennis?
15	Overcast	Cool	High	Weak	
16	Rain	Mild	High	Weak	
17	Overcast	Cool	Low	Strong	
18	Overcast	Hot	Low	Weak	
19	Sunny	Mild	High	Weak	
20	Sunny	Cool	Low	Weak	
21	Rain	Cool	Low	Weak	
22	Sunny	Mild	Low	Strong	
23	Rain	Mild	High	Strong	

SHOULD I PLAY TENNIS?

Day	Outlook	Temp	Humidity	Wind	Play Tennis?
15	Overcast	Cool	High	Weak	No
16	Rain	Mild	High	Weak	Yes
17	Overcast	Cool	Low	Strong	Yes
18	Overcast	Hot	Low	Weak	Yes
19	Sunny	Mild	High	Weak	No
20	Sunny	Cool	Low	Weak	Yes
21	Rain	Cool	Low	Weak	No
22	Sunny	Mild	Low	Strong	Yes
23	Rain	Mild	High	Strong	No

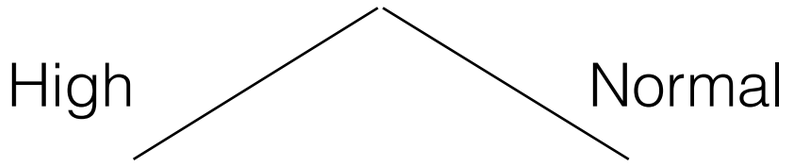
SHOULD I PLAY TENNIS?

What's the outlook?



Yes 4/14

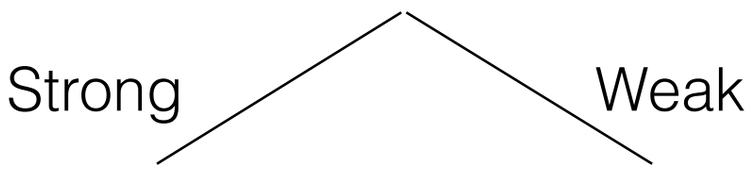
How's the humidity?



No 3/14

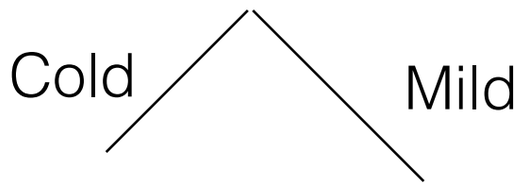
Yes 2/14

How strong is the wind?



No 2/14

How cold is it?



No

Yes

SHOULD I PLAY TENNIS?

problem:
overfitting your tree to the train set

What's the outlook?

Sunny | Overcast | Rain

Yes 4/14

How's the humidity?

High | Normal

No 3/14 | **Yes** 2/14

How strong is the wind?

Strong | Weak

No 2/14

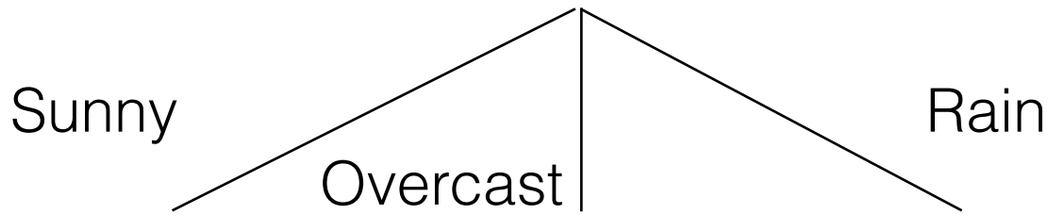
How cold is it?

Cold | Mild

No | **Yes**

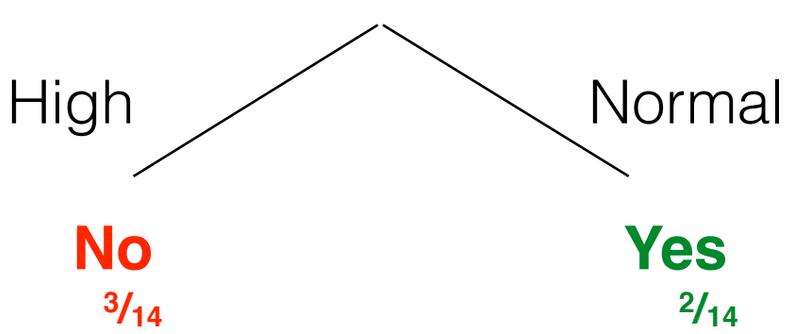
SHOULD I PLAY TENNIS?

What's the outlook?



Yes
4/14

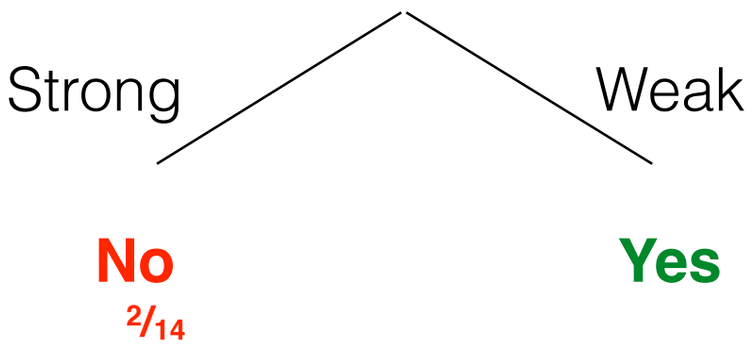
How's the humidity?



No
3/14

Yes
2/14

How strong is the wind?



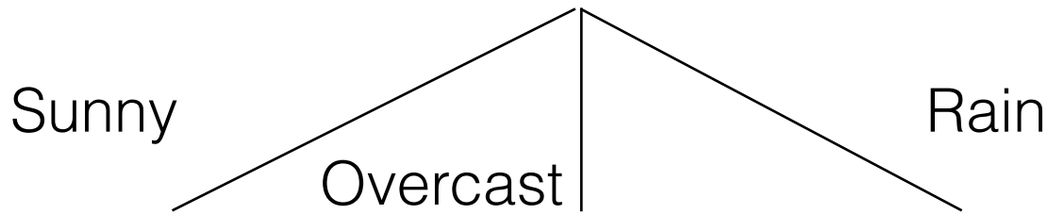
No
2/14

Yes

SHOULD I PLAY TENNIS?

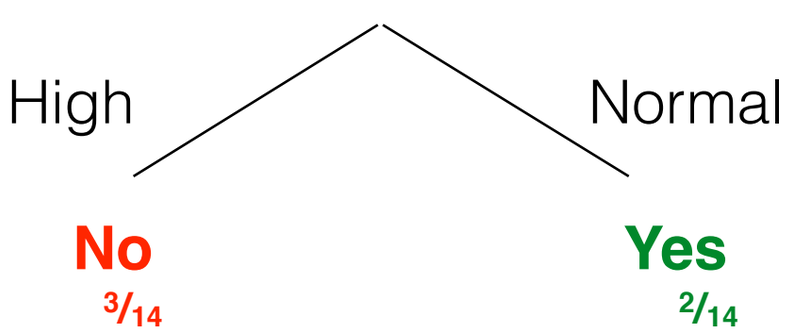
one solution: pruning

What's the outlook?



Yes
4/14

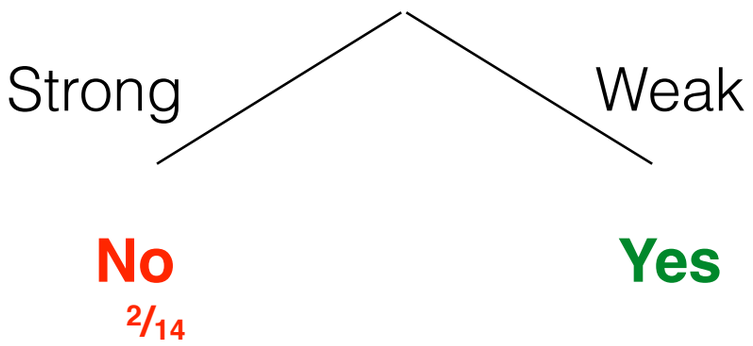
How's the humidity?



No
3/14

Yes
2/14

How strong is the wind?



No
2/14

Yes

SHOULD I PLAY TENNIS?

one solution: pruning

(limit the depth of the tree)

What's the outlook?

Sunny | Overcast | Rain

Yes 4/14

How's the humidity?

High | Normal

No 3/14 | **Yes** 2/14

How strong is the wind?

Strong | Weak

No 2/14 | **Yes**

EXERCISE 20

KEY OBJECTIVE(S)

Brainstorm the advantages and disadvantages of trees

AGENDA

10:00 min

1. Read r2d3's visualization of a decision tree
2. With your table, come up with a list of advantages and disadvantages of decision trees

DELIVERABLE

A list of at least 3 advantages and 3 disadvantages

RESOURCES

www.r2d3.us

ADVANTAGES:

- The decisions are easy to understand and interpret.
- The weight and importance of each feature becomes clear.
- Both numerical and categorical features can be used naturally.
- Trees are a natural multi-class classifier.

ADVANTAGES:

- The decisions are easy to understand and interpret.
- The weight and importance of each feature becomes clear.
- Both numerical and categorical features can be used naturally.
- Trees are a natural multi-class classifier.

DISADVANTAGES:

- Can overfit to training data with complex trees.
- Small changes in input data can result in totally different trees.
- Can make mistakes with unbalanced classes.
- No confidence intervals (regression).
- Requires large datasets to build robust rules.

CASE STUDIES

BigML is a service that offers machine learning solutions to its customers

https://bigml.com/gallery/models

7 min

1. Read through a few decision tree examples on the Big ML website.
2. Take notes on one case. Be prepared to present to the class on the problem, dataset, nodes and splits. Be ready to walk through at least one branch.

5 min

3. Share out with the class

```
from sklearn.tree import DecisionTreeClassifier
```

```
from sklearn.tree import DecisionTreeRegressor
```

http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=None, splitter='best')
```

http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=None, splitter='best')
```

FEATURE IMPORTANCE

Decision trees don't give your features coefficient's.

But you can get their relative importance to the model. Normalized by how many rows are accurately sorted through that feature.

Another solution to overfitting

Another solution to overfitting

LOTS OF TREES!

Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcast	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No

Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcast	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No

Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcast	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No

train

Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcast	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No

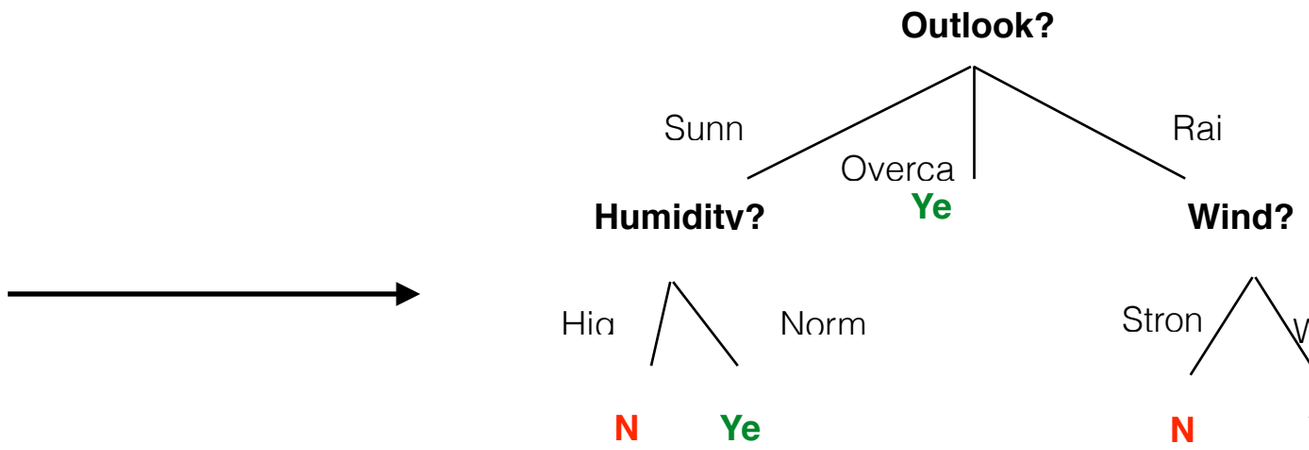
train

Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcast	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No

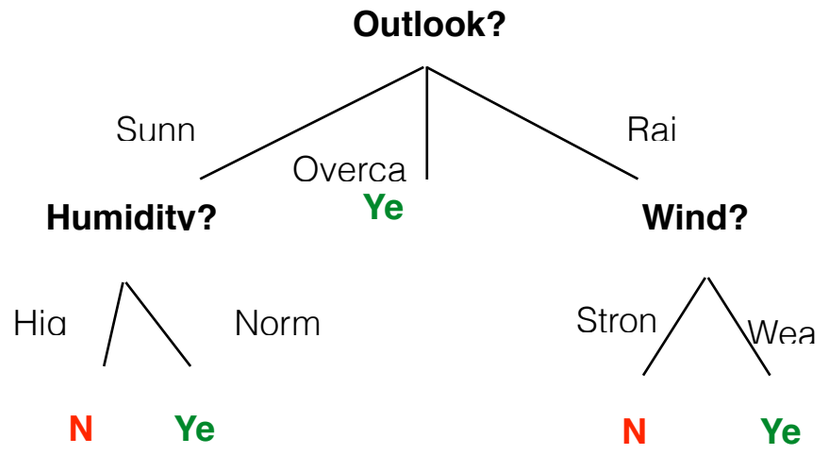
train

test

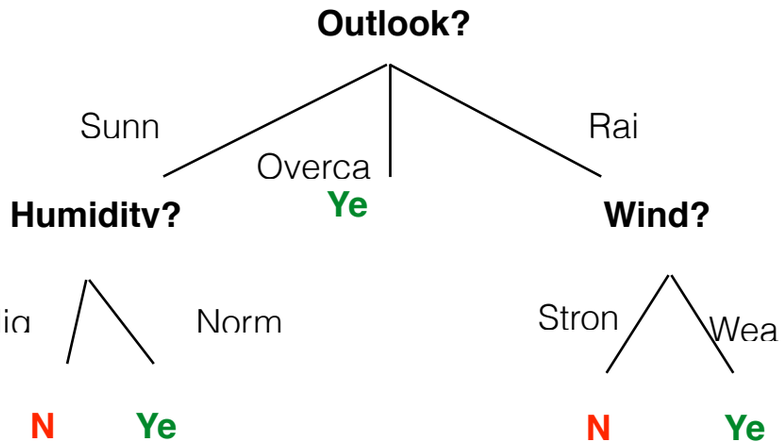
Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcast	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No



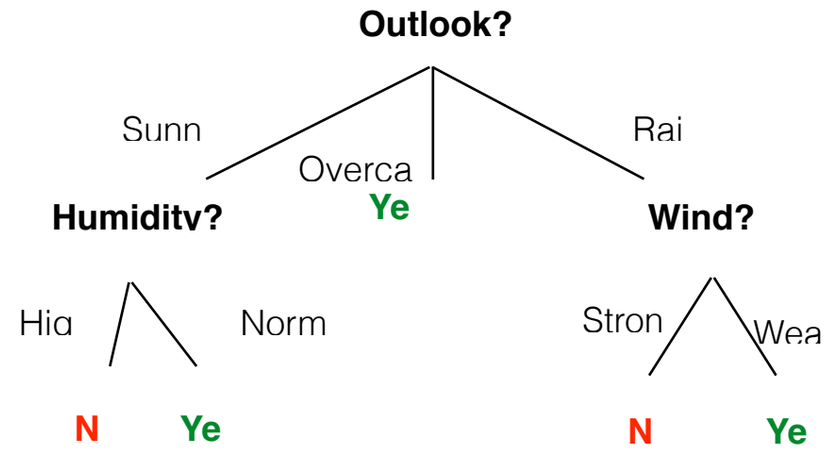
Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcast	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No



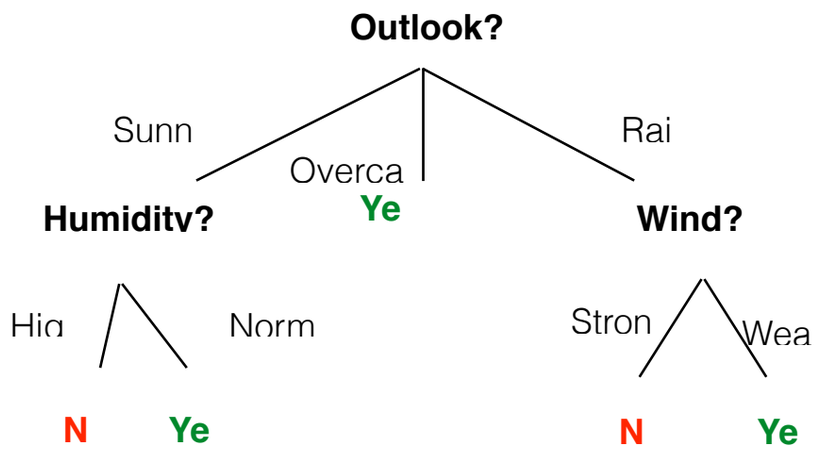
Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcast	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No



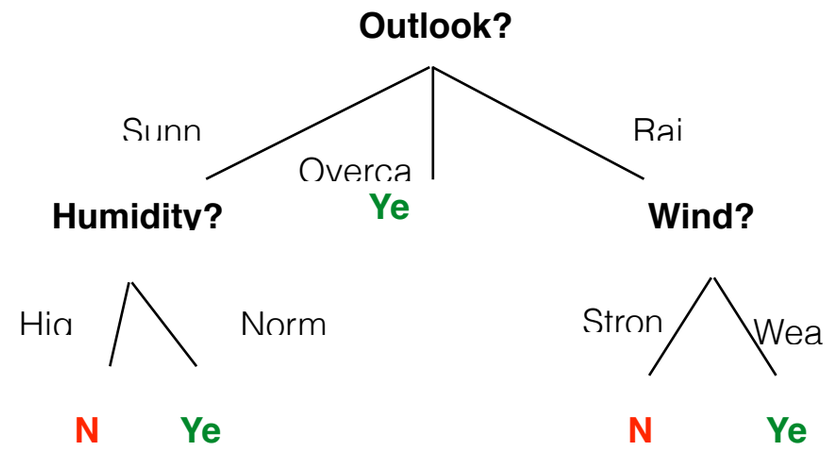
Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcast	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No



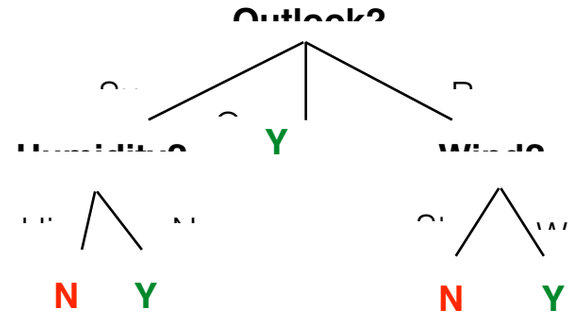
Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcast	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No



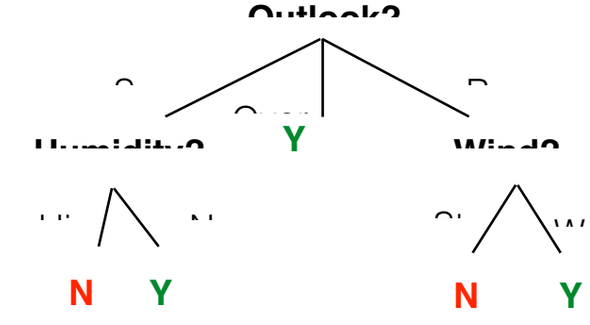
Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcast	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No



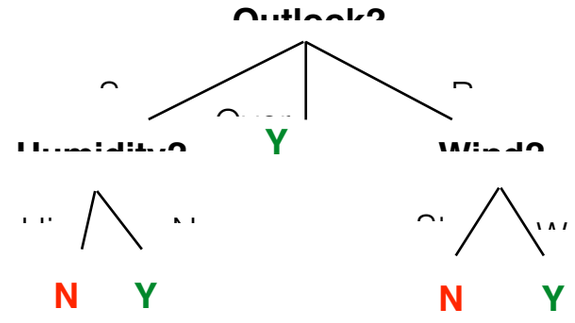
Da	Outlook	Temp	Humidit	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcas	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcas	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcas	Mild	High	Strong	Yes
13	Overcas	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No



Da	Outlook	Temp	Humidit	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcas	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcas	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcas	Mild	High	Strong	Yes
13	Overcas	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No

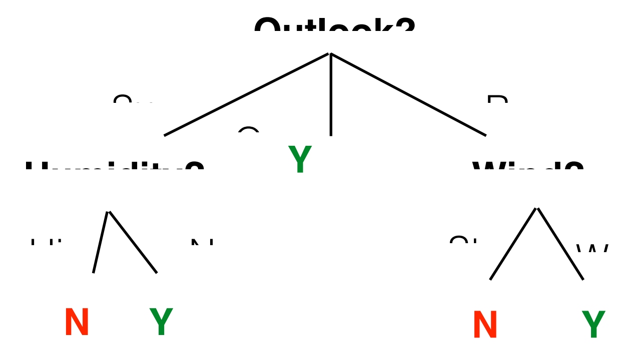


Da	Outlook	Temp	Humidit	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcas	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcas	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcas	Mild	High	Strong	Yes
13	Overcas	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No

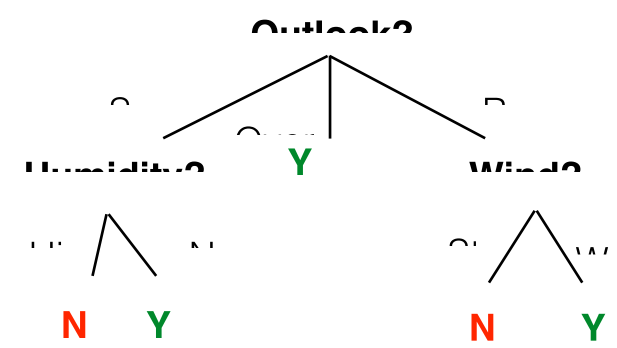


For classification, each tree gets one “vote”

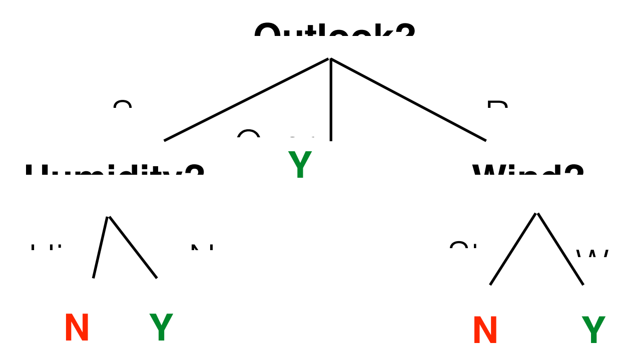
Da	Outlook	Temp	Humidit	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcas	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcas	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcas	Mild	High	Strong	Yes
13	Overcas	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No



Da	Outlook	Temp	Humidit	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcas	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcas	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcas	Mild	High	Strong	Yes
13	Overcas	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No

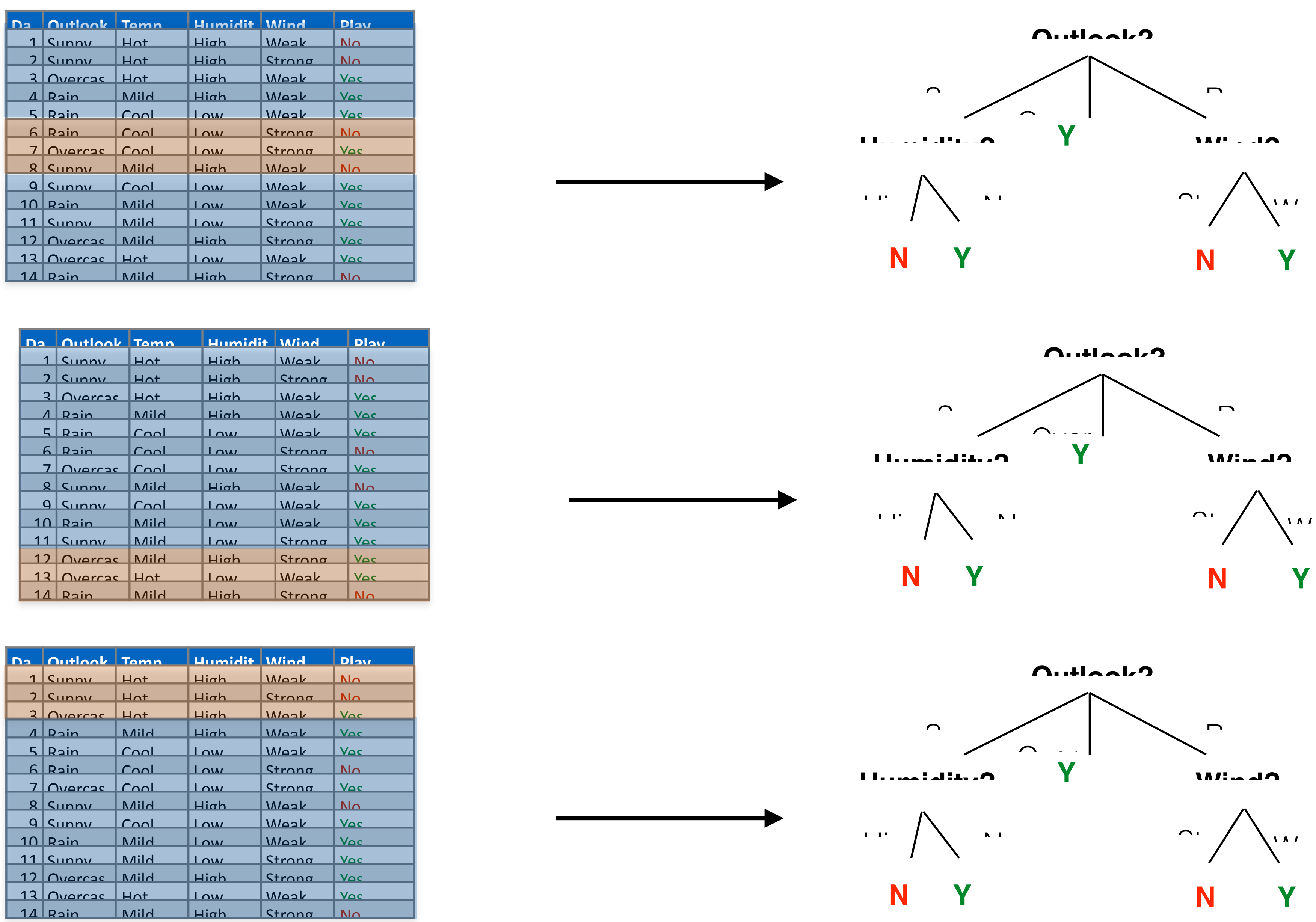


Da	Outlook	Temp	Humidit	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcas	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Low	Weak	Yes
6	Rain	Cool	Low	Strong	No
7	Overcas	Cool	Low	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Low	Weak	Yes
10	Rain	Mild	Low	Weak	Yes
11	Sunny	Mild	Low	Strong	Yes
12	Overcas	Mild	High	Strong	Yes
13	Overcas	Hot	Low	Weak	Yes
14	Rain	Mild	High	Strong	No



“Random Forest”

For classification, each tree gets one “vote”



```
from sklearn.tree import RandomForestClassifier
```

```
from sklearn.tree import RandomForestRegressor
```