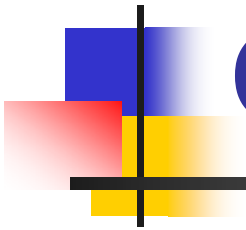
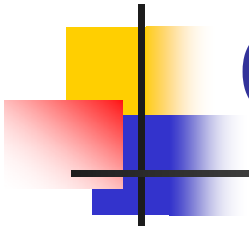


COMP1942

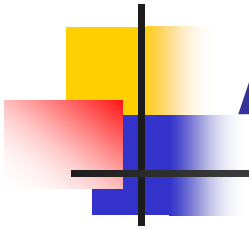
Classification: More Concept

Prepared by Raymond Wong
Presented by Raymond Wong
raywong@cse



Classification Concept

- Other Five Measurements
  - Accuracy
  - Precision
  - Recall
  - f-measure
  - Specificity
- False Positive/False Negative
- Two Phases



Accuracy

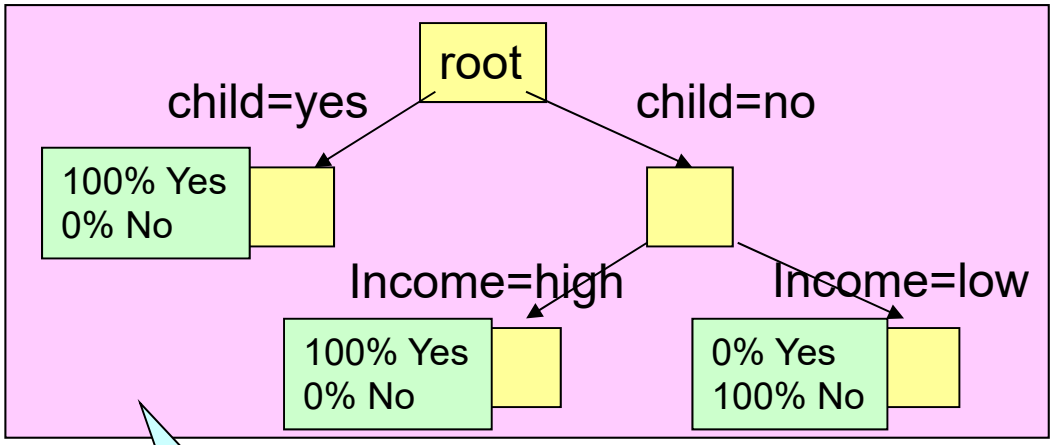
- Given
  - a classification model (e.g., decision tree)
  - an arbitrary dataset where its target attribute is known
- the **accuracy** of a classification model is defined to be the proportion of the values in the target attribute correctly predicted.

Accuracy

e.g.1

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no

Training set

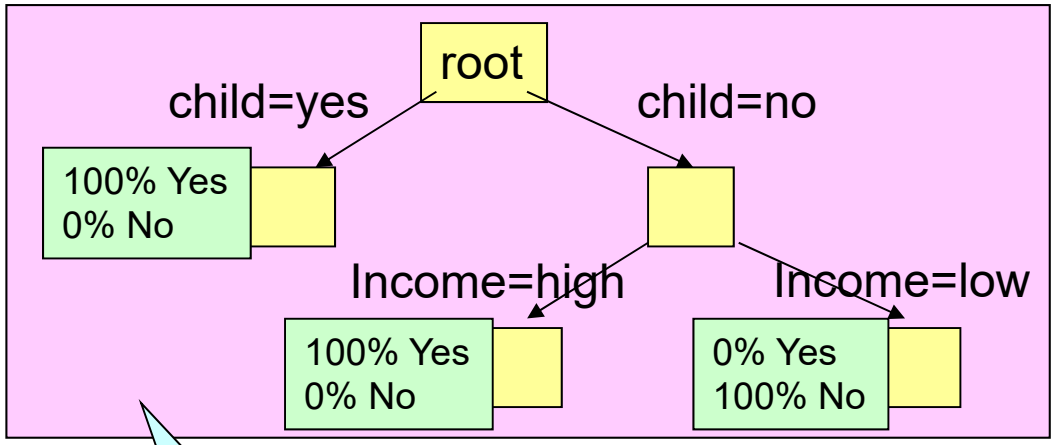


Decision tree

Accuracy

e.g.1

Race	Income	Child	Insuranc	Predicted
black	high	no	yes	yes
white	high	yes	yes	yes
white	low	yes	yes	yes
white	low	yes	yes	yes
black	low	no	no	no
black	low	no	no	no
black	low	no	no	no
white	low	no	no	no



Accuracy = 8/8*100 = 100%

Training set

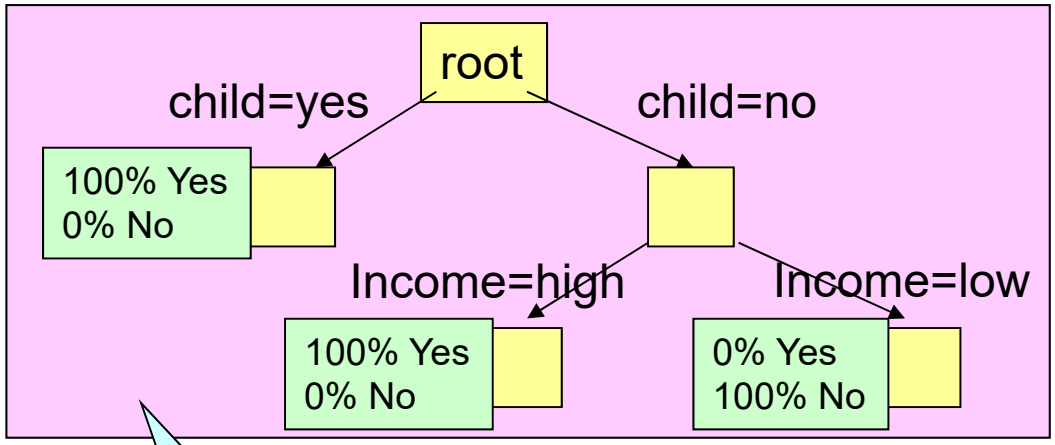
Decision tree

Accuracy

e.g.1

Race	Income	Child	Insurance
white	high	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	high	no	no

An arbitrary dataset

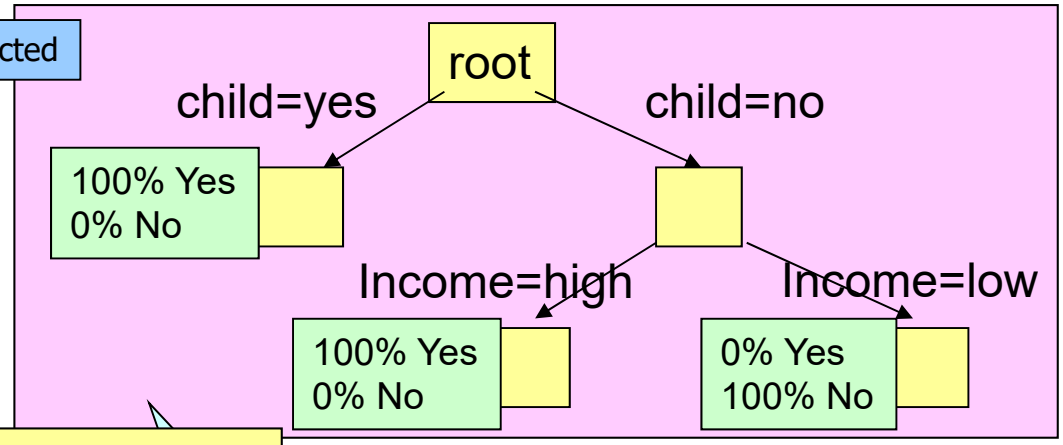


Decision tree

Accuracy

e.g.1

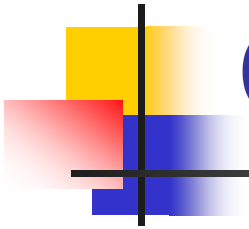
Race	Income	Child	Insurance	Predicted
white	high	yes	yes	yes
white	low	yes	yes	yes
black	low	no	no	no
black	low	no	no	no
black	low	no	no	no
white	high	no	no	yes



Accuracy = 5/6 * 100 = 83.3%

An arbitrary dataset

Decision tree



Classification Concept

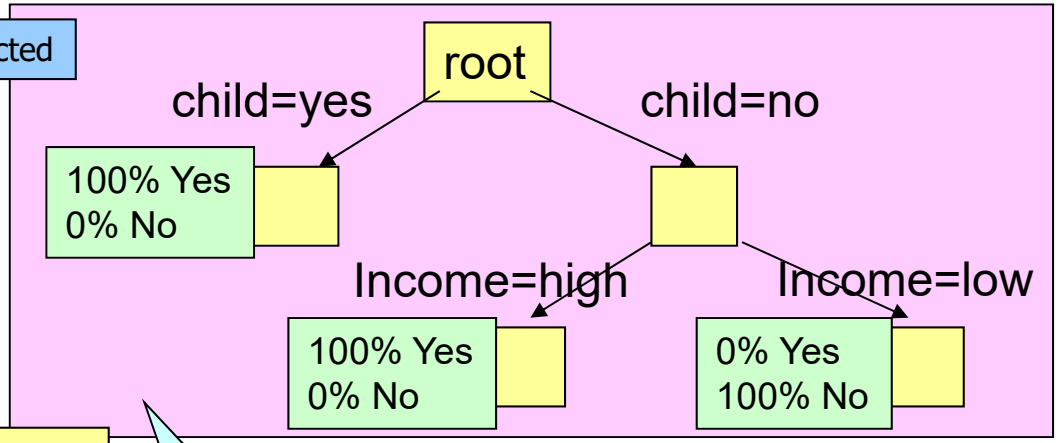
- Other Five Measurements
  - Accuracy
  - Precision
  - Recall
  - f-measure
  - Specificity
- False Positive/False Negative
- Two Phases

Precision

- Precision

$$= \frac{\text{Total no. of values in the target attribute correctly predicted as “Yes”}}{\text{Total no. of values in the target attribute predicted as “Yes”}}$$

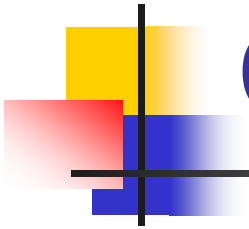
Race	Income	Child	Insurance	Predicted
white	high	yes	yes	yes
white	low	yes	yes	yes
black	low	no	no	no
black	low	no	no	no
black	low	no	no	no
white	high	no	no	yes



Precision = 2/3 = 0.67

An arbitrary dataset

Decision tree



Classification Concept

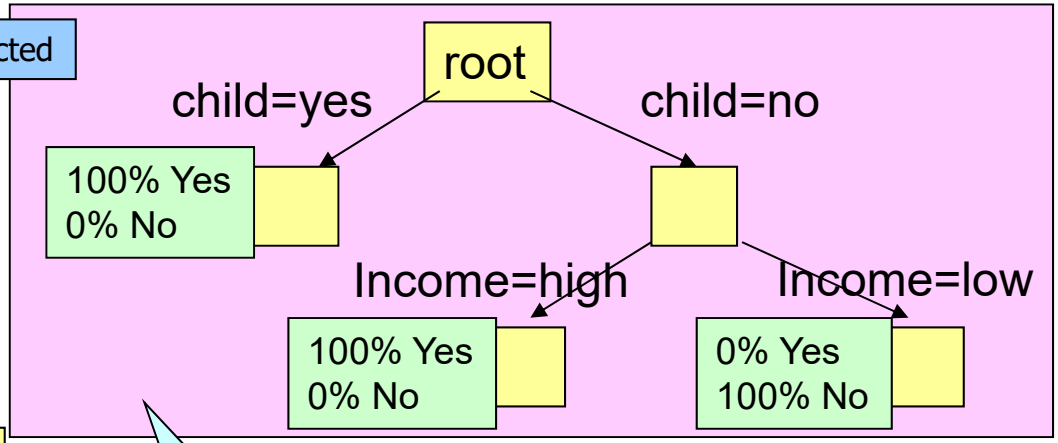
- Other Five Measurements
  - Accuracy
  - Precision
  - Recall
  - f-measure
  - Specificity
- False Positive/False Negative
- Two Phases

Recall

- Recall

$$= \frac{\text{Total no. of values in the target attribute correctly predicted as “Yes”}}{\text{Total no. of actual values in the target attribute equal to “Yes”}}$$

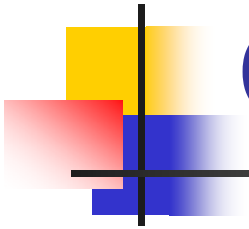
Race	Income	Child	Insurance	Predicted
white	high	yes	yes	yes
white	low	yes	yes	yes
black	low	no	no	no
black	low	no	no	no
black	low	no	no	no
white	high	no	no	yes



Recall = 2/2 = 1.0

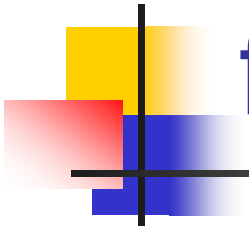
An arbitrary dataset

Decision tree



Classification Concept

- Other Five Measurements
  - Accuracy
  - Precision
  - Recall
  - f-measure
  - Specificity
- False Positive/False Negative
- Two Phases



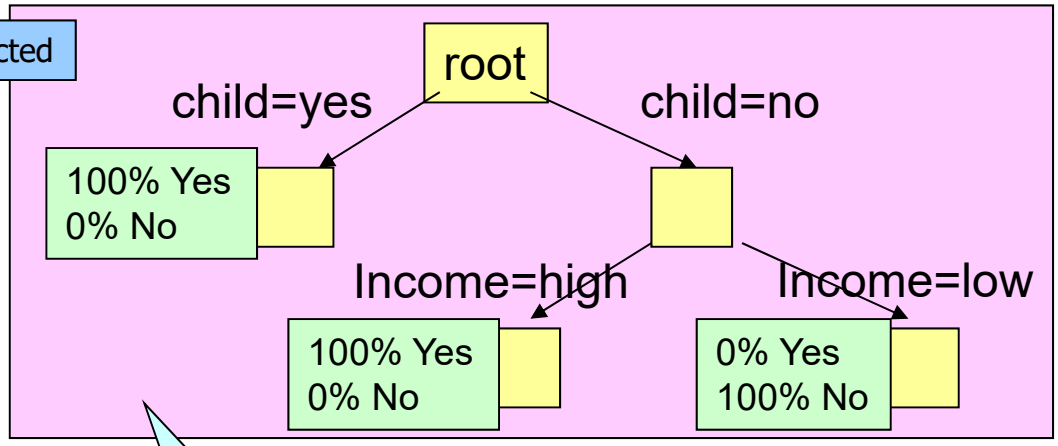
f-measure

- f-measure is also called “f1-score”.
- f-measure is a measurement by considering precision and recall together

- $$\text{f-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

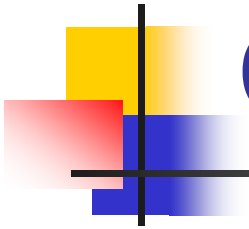
f-measure

Race	Income	Child	Insurance	Predicted
white	high	yes	yes	yes
white	low	yes	yes	yes
black	low	no	no	no
black	low	no	no	no
black	low	no	no	no
white	high	no	no	yes



$$\text{f-measure} = \frac{2 \times 0.67 \times 1.0}{0.67 + 1.0} = 0.80$$

Decision tree



Classification Concept

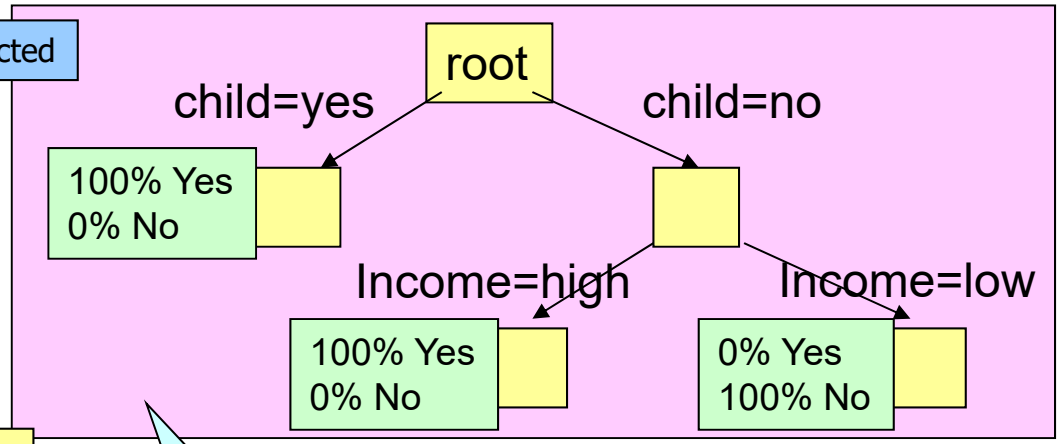
- Other Five Measurements
  - Accuracy
  - Precision
  - Recall
  - f-measure
  - Specificity
- False Positive/False Negative
- Two Phases

Specificity

- Specificity (similar to Recall (reverse way))

$$= \frac{\text{Total no. of values in the target attribute correctly predicted as “No”}}{\text{Total no. of actual values in the target attribute equal to “No”}}$$

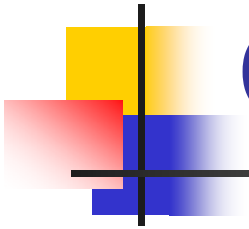
Race	Income	Child	Insurance	Predicted
white	high	yes	yes	yes
white	low	yes	yes	yes
black	low	no	no	no
black	low	no	no	no
black	low	no	no	no
white	high	no	no	yes



Recall = 3/4 = 0.75

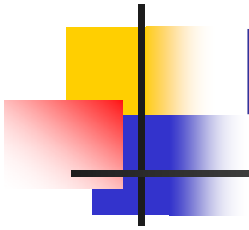
An arbitrary dataset

Decision tree



Classification Concept

- Other Five Measurements
  - Accuracy
  - Precision
  - Recall
  - f-measure
  - Specificity
- False Positive/False Negative
- Two Phases



False Positive/False Negative

- **True Positive**
  - The value is predicted as “Yes” and the actual value is “Yes”
- **False Positive**
  - The value is predicted as “Yes” but the actual value is “No”
- **True Negative**
  - The value is predicted as “No” and the actual value is “No”
- **False Negative**
  - The value is predicted as “No” but the actual value is “Yes”

False Positive/False Negative

- E.g.1

Race	Income	Child	Insurance	Predicted
white	high	yes	yes	yes
white	low	yes	yes	yes
black	low	yes	yes	no
black	low	yes	yes	no
black	low	no	no	no
white	high	no	no	yes

No. of true positives = 2

No. of false positives = 1

No. of true negatives = 1

No. of false negatives = 2

False Positive/False Negative

- E.g.2

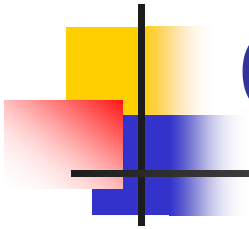
Race	Income	Child	Insurance	Predicted
white	high	yes	yes	yes
white	low	yes	yes	no
black	low	yes	yes	no
black	low	yes	yes	no
black	low	no	no	yes
white	high	no	no	yes

No. of true positives = ?

No. of false positives = ?

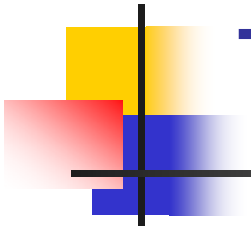
No. of true negatives = ?

No. of false negatives = ?



Classification Concept

- Other Five Measurements
  - Accuracy
  - Precision
  - Recall
  - f-measure
  - Specificity
- False Positive/False Negative
- Two Phases



Two Phases

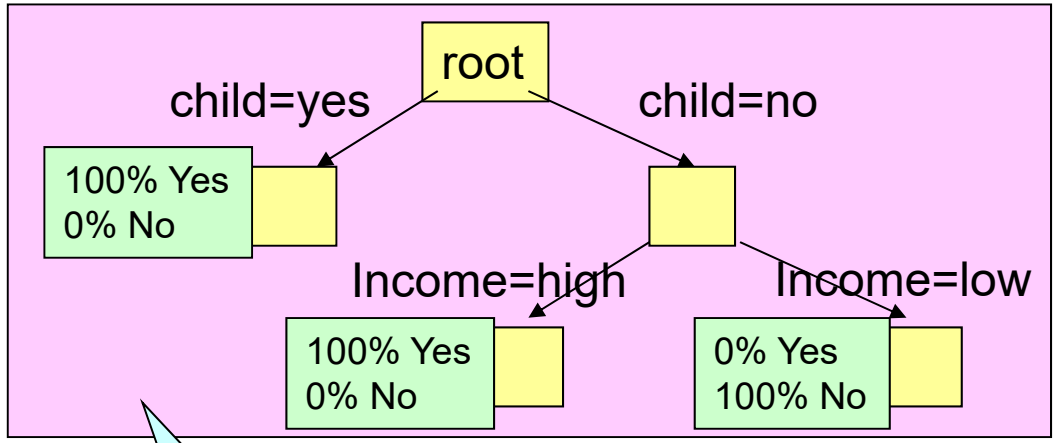
- Two Phases
  - Training Phase
    - Training Set
    - Validation Set
    - Test Set
  - Prediction Phase
    - New Set

Training Phase

- What we learnt is the following.

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no

Training set



Decision tree

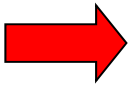
Training Phas

Original dataset

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
black	low	no	no
black	low	no	no
white	high	yes	yes
black	low	no	no
…	…	…	…
white	low	no	no

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no

Training set

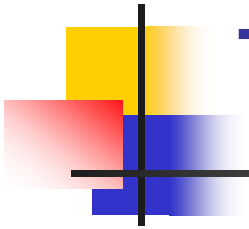


Race	Income	Child	Insurance
white	high	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no

Validation set

Race	Income	Child	Insurance
white	low	yes	yes
black	low	no	no
black	low	no	no

Test set



Two Phases

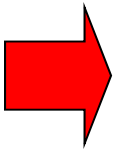
- Two Phases
  - Training Phase
    - Training Set
    - Validation Set
    - Test Set
  - Prediction Phase
    - New Set

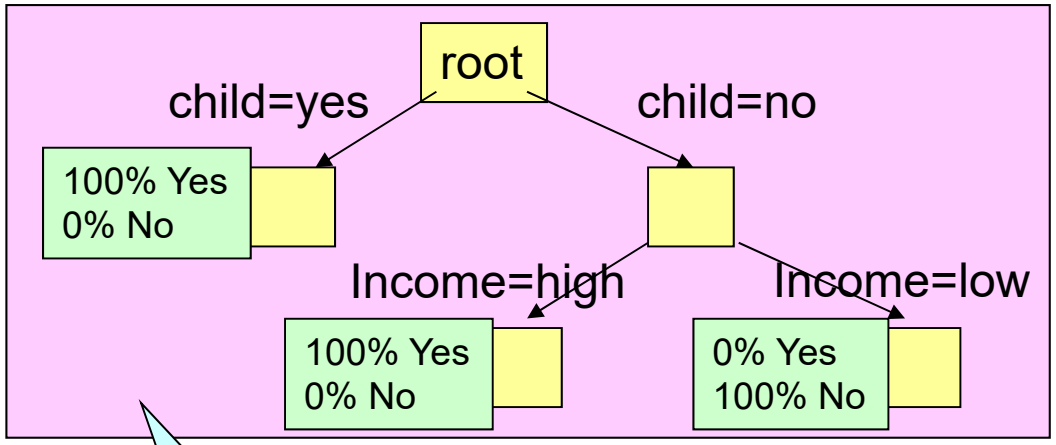
Training Set

- Used to train or build a model

e.g.1

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no





Accuracy = 100%

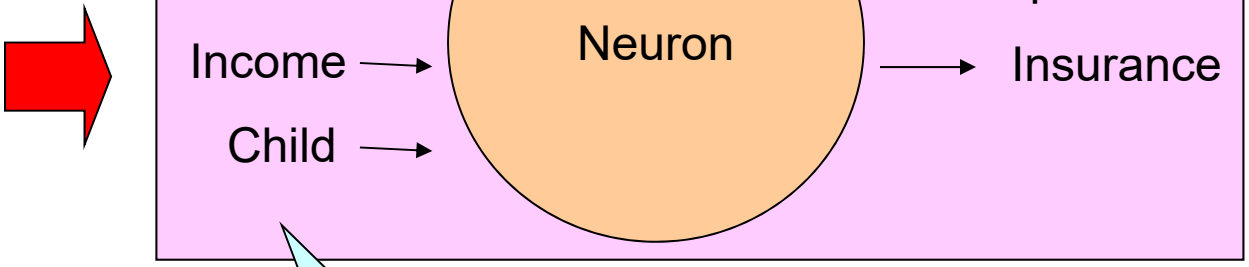
Training set

Decision tree

Training Set

e.g.2

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no



Accuracy = 100%

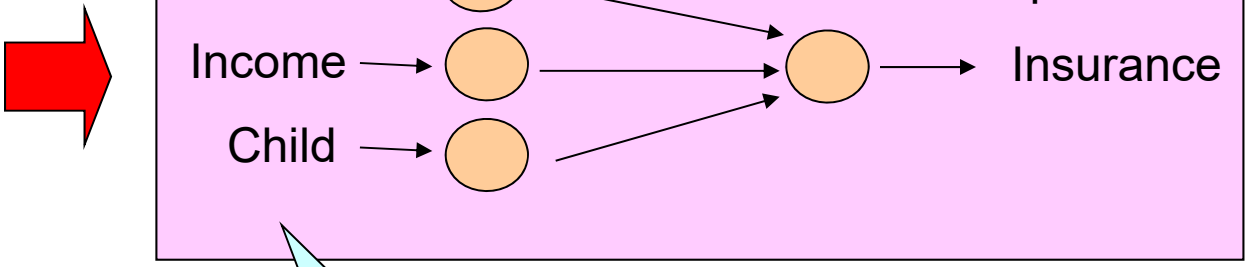
Training set

Neural Network

Training Set

e.g.3

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no



Accuracy = 100%

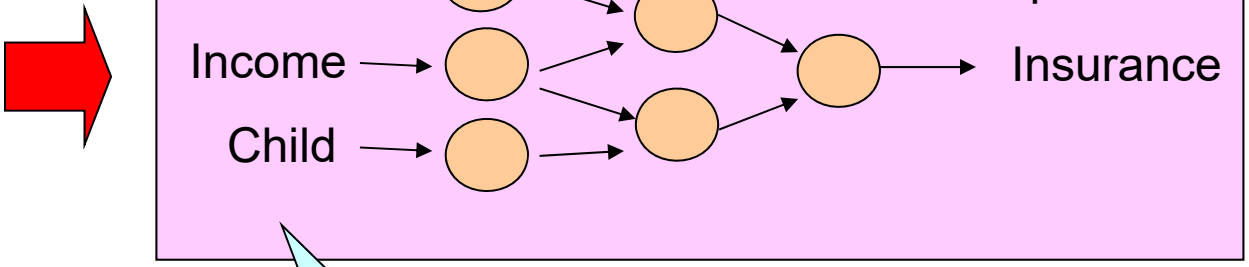
Training set

Neural Network

Training Set

e.g.4

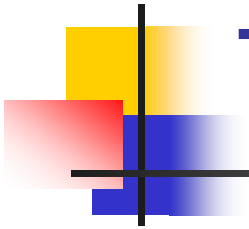
Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no



Accuracy = 100%

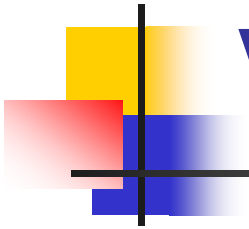
Training set

Neural Network



Two Phases

- Two Phases
  - Training Phase
    - Training Set
    - Validation Set
    - Test Set
  - Prediction Phase
    - New Set



Validation Set

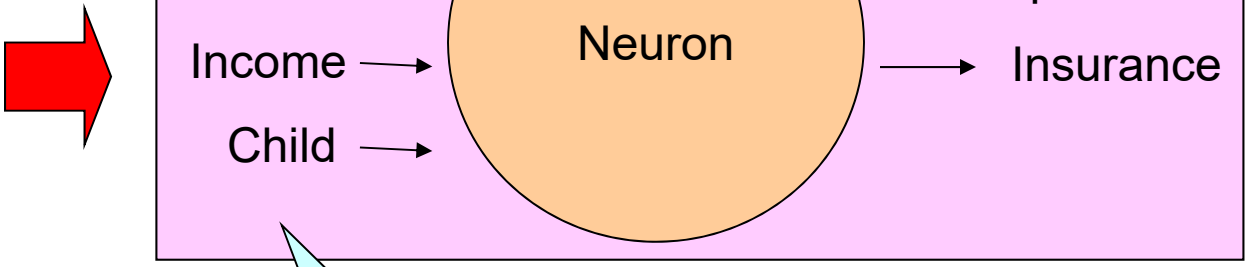
- We obtain a model from the training set.

- Validation set
  - Used to fine-tune the model
  - Different models have different ways of fine-tuning

Validation Set – Neural Network

e.g.1

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no



Accuracy = 100%

Training set

Neural Network

Validation Set – Neural Network

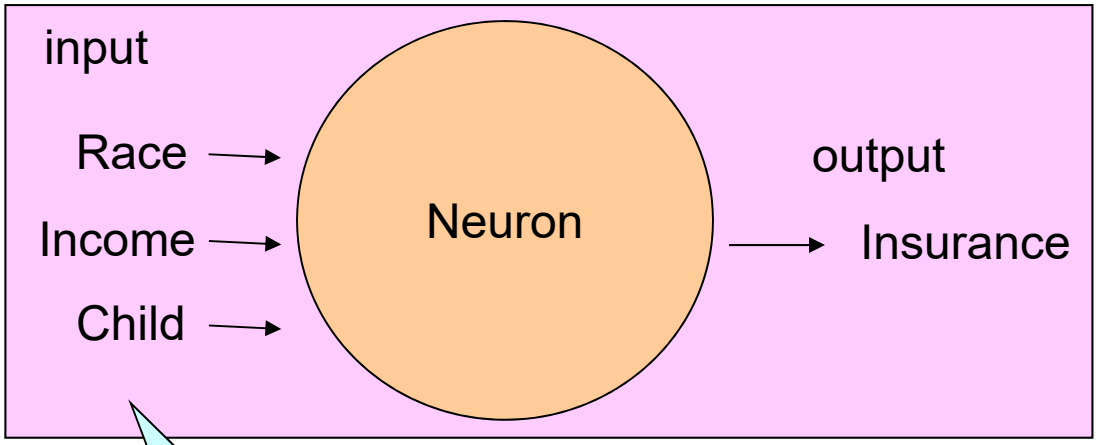
e.g.1

Race	Income	Child	Insurance
white	high	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no

Accuracy = 80%

Validation set

Structure 1

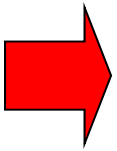


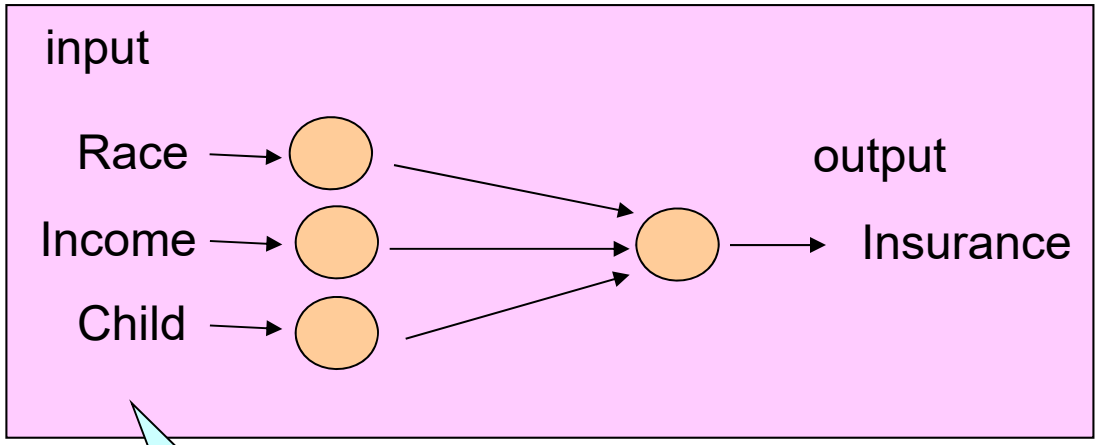
Neural Network

Validation Set – Neural Network

e.g.2

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no





Accuracy = 100%

Training set

Neural Network

Validation Set – Neural Network

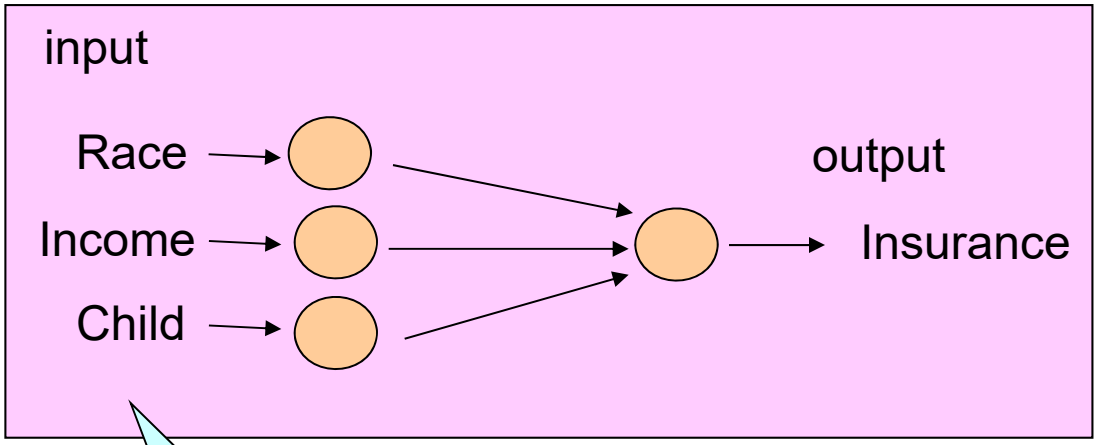
e.g.2

Race	Income	Child	Insurance
white	high	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no

Accuracy = 95%

Validation set

Structure 2

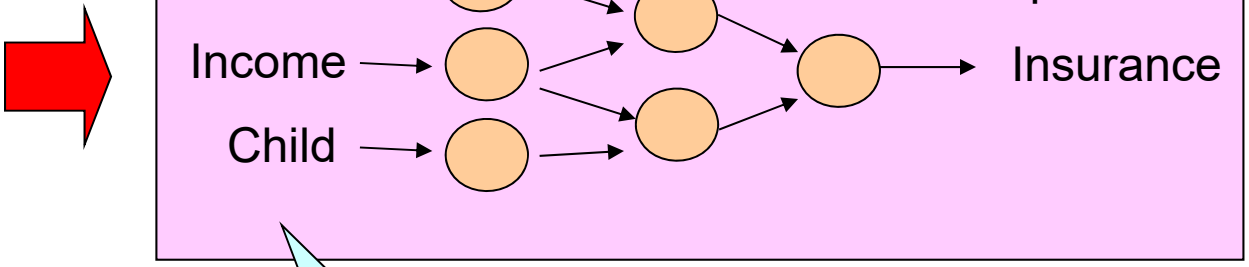


Neural Network

Validation Set – Neural Network

e.g.3

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no



Accuracy = 100%

Training set

Neural Network

Validation Set – Neural Network

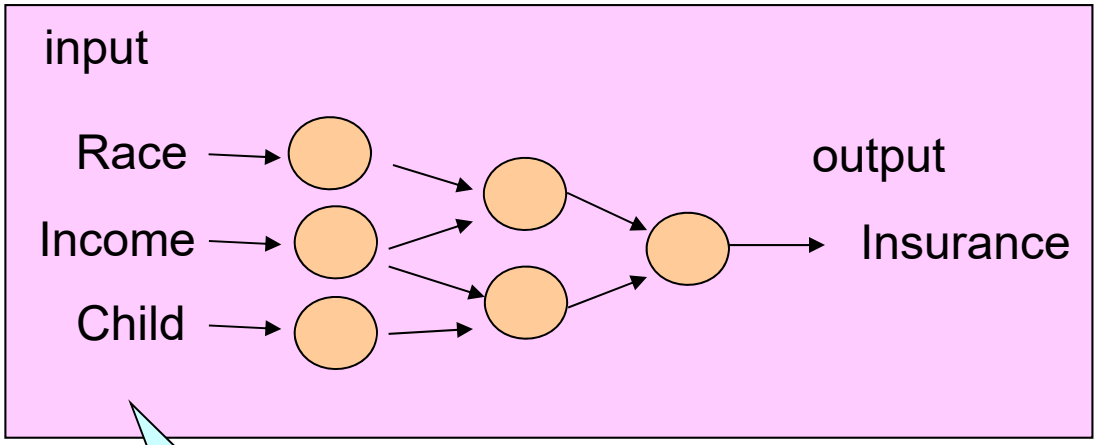
e.g.3

Race	Income	Child	Insurance
white	high	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no

Accuracy = 70%

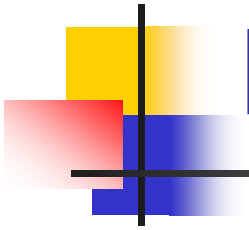
Validation set

Structure 3



Neural Network

Validation Set – Neural Network

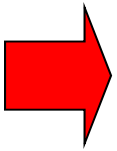


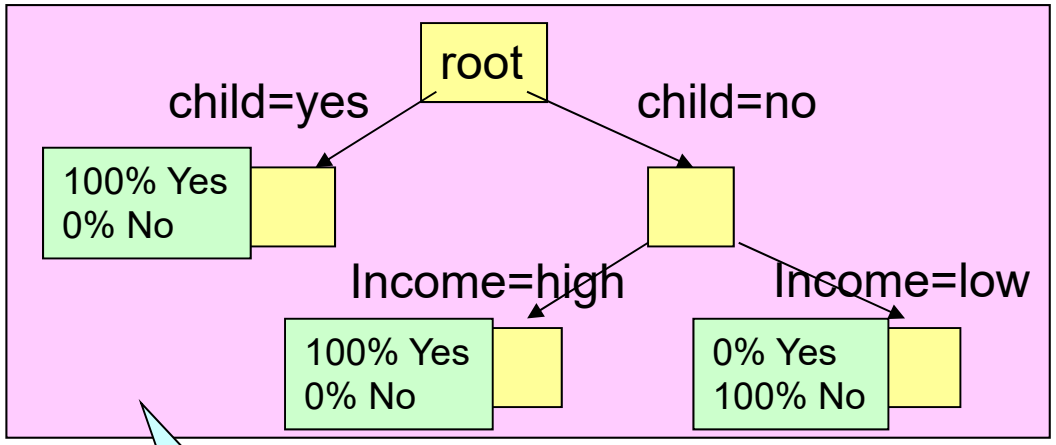
- Which structure is the best?
- Why?

Validation Set – Decision Tree

e.g.1

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no





Accuracy = 100%

Training set

Decision tree

Validation Set – Decision Tree

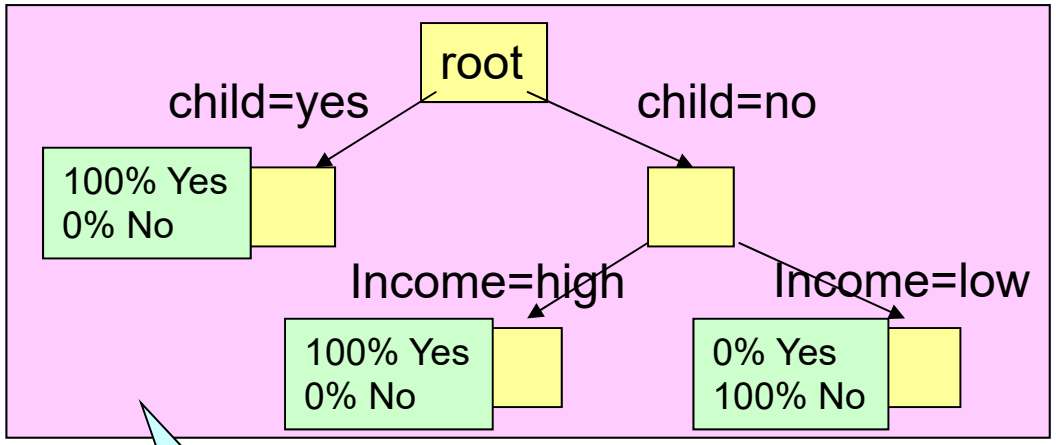
e.g.1

Race	Income	Child	Insurance
white	high	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no

Accuracy = 70%

Validation set

Original tree



Decision tree

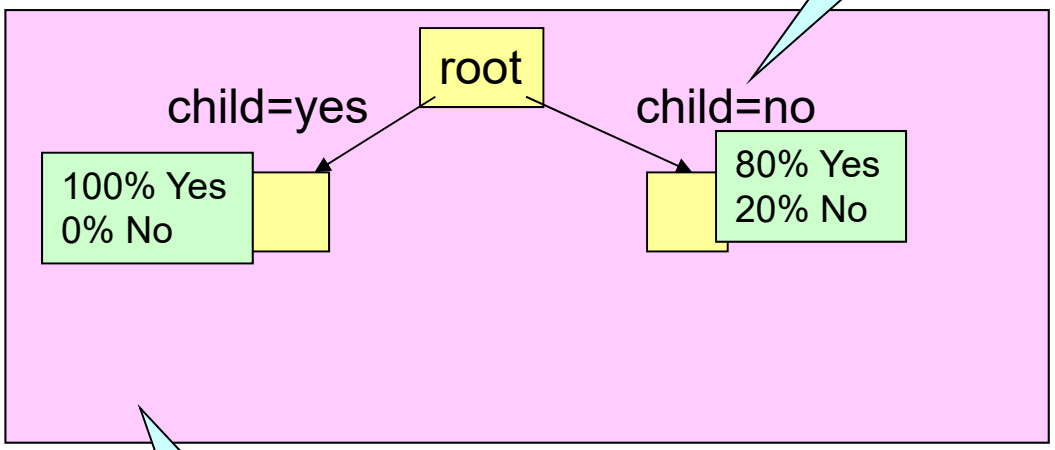
Validation Set – Decision Tree

An operation to remove the whole subtree

Pruning

e.g.1

Variation of the original tree

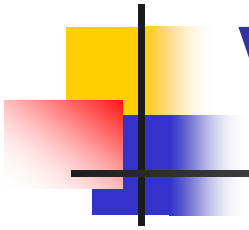


Race	Income	Child	Insurance
white	high	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no

Accuracy = 95%

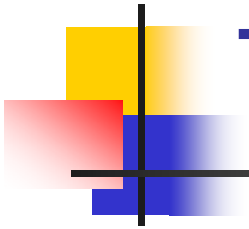
Validation set

Decision tree



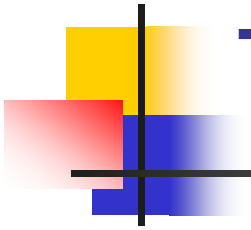
Validation Set – Decision Tree

- Which tree is the best?
- Why?



Two Phases

- Two Phases
  - Training Phase
    - Training Set
    - Validation Set
    - Test Set
  - Prediction Phase
    - New Set



Test Set

- The validation test is often used to fine-tune the model
- In such a case, when a model is finally chosen, its accuracy with the validation set is still an **optimistic** estimate of how it would perform with **unseen data**
- Thus, we use **test set** to **evaluate** the accuracy of the model on completely unseen data

Test set – Neural Network

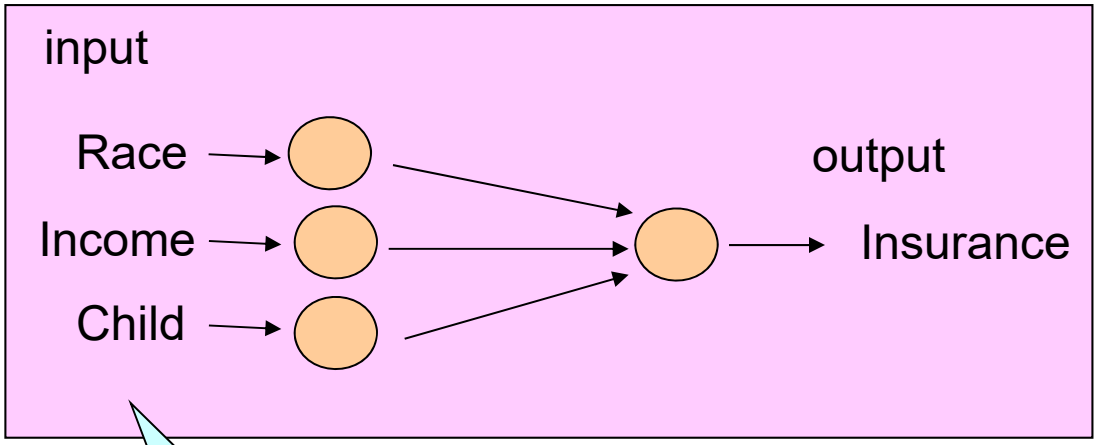
e.g.1

Race	Income	Child	Insurance
white	high	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no

Validation set

Accuracy = 95%

Structure 2



Neural Network

Test set – Neural Network

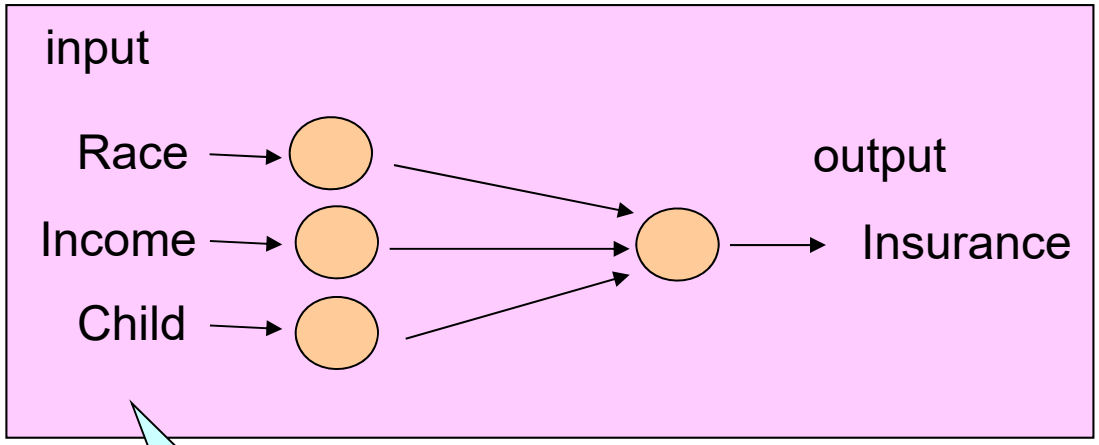
e.g.1

Race	Income	Child	Insurance
white	low	yes	yes
black	low	no	no
black	low	no	no

Test set

Accuracy = 90%

Structure 2



Neural Network

Test set – Decision Tree

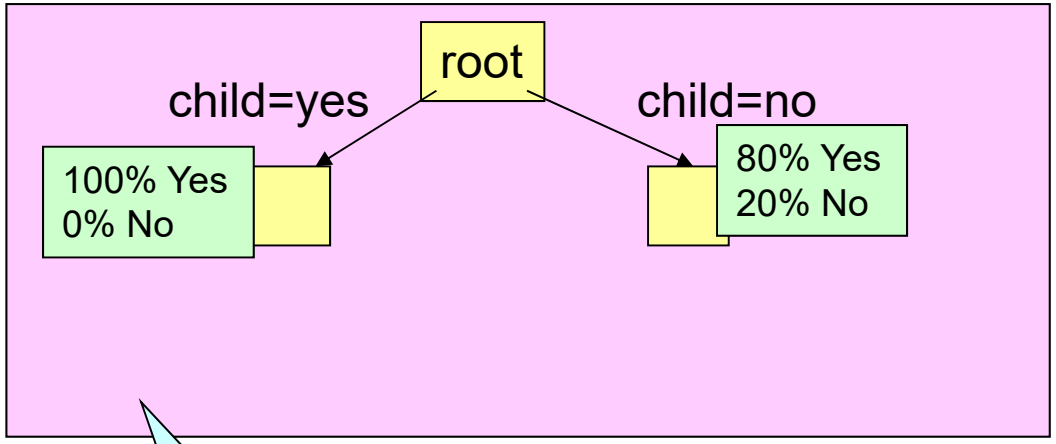
e.g.2

Race	Income	Child	Insurance
white	high	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no

Accuracy = 95%

Validation set

Variation of the original tree



Decision tree

Test set – Decision Tree

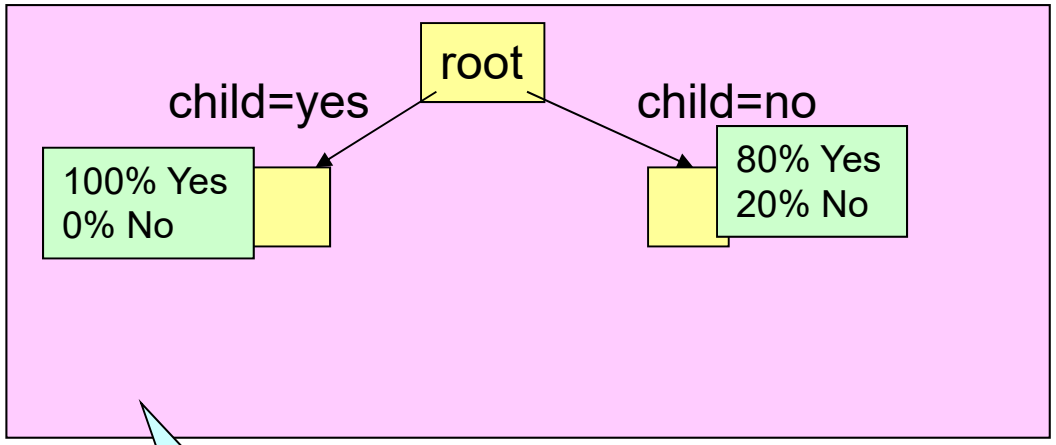
e.g.2

Race	Income	Child	Insurance
white	low	yes	yes
black	low	no	no
black	low	no	no

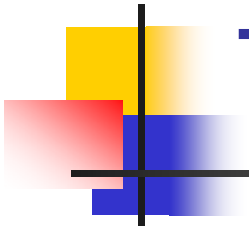
Test set

Accuracy = 90%

Variation of the original tree



Decision tree



Two Phases

- Two Phases
  - Training Phase
    - Training Set
    - Validation Set
    - Test Set
  - Prediction Phase
    - New Set

New Set

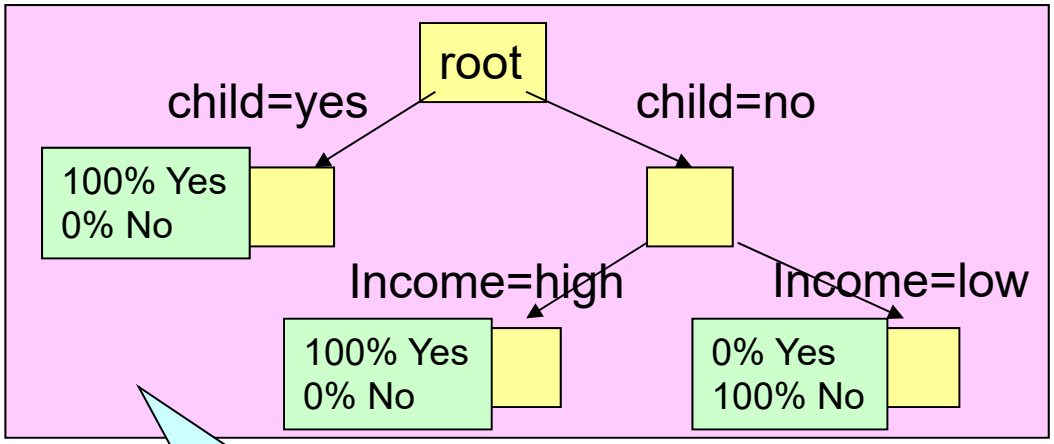
Suppose there is a person.

Race	Income	Child	Insurance
white	high	no	?

New set

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	low	no	no

Training set



Decision tree