Sr. No.	Module	Detailed Contents	Hrs
1	**Business Intelligence-**	Introduction and overview of BI-Effective and timely decisions, Data Information and knowledge, BI Architecture, Ethics and BI. BI Applications- Balanced score card, Fraud detection, Telecommunication Industry, Banking and finance, Market segmentation.	**06**
2	**Prediction methods and models for BI**	Data preparation, Prediction methods-Mathematical method, Distance methods, Logic method, heuristic method-local optimization technique, stochastic hill climber, evaluation of models	**06**
3	**BI using Data Warehousing**	Introduction to DW, DW architecture, ETL Process, Top-down and bottom-up approaches, characteristics and benefits of data mart, Difference between OLAP and OLTP. Dimensional analysis- Define cubes. Drill- down and roll- up – slice and dice or rotation, OLAP models- ROLAP and MOLAP. Define Schemas- Star, snowflake and fact constellations.	**08**
4	**Data Mining and Preprocessing**	Data mining- definition and functionalities, KDD Process, Data Cleaning: - Missing values, Noisy data, data integration and transformations. Data Reduction: - Data cube aggregation, dimensionality reduction- data compression, Numerosity reduction- discretization and concept hierarchy.	**06**
5	**Associations and Correlation**	Association rule mining:-support and confidence and frequent item sets, market basket analysis, Apriori algorithm, Incremental ARM, Associative classification- Rule Mining.	**06**
6	**Classification and Prediction**	Introduction, Classification methods:-Decision Tree- ID3, CART, Bayesian classification- Baye'stheorem(Naïve Bayesian classification),Linear and nonlinear regression.	**08**
7	**Clustering**	Introduction, categorization of Major, Clustering Methods:- partitioning methods- K-Means. Hierarchical- Agglomerative and divisive methods, Model- based- Expectation and Maximization.	**08**
8	**Web mining and Text**	Text data analysis and Information retrieval, text retrieval methods, dimensionality reduction for text	**04**

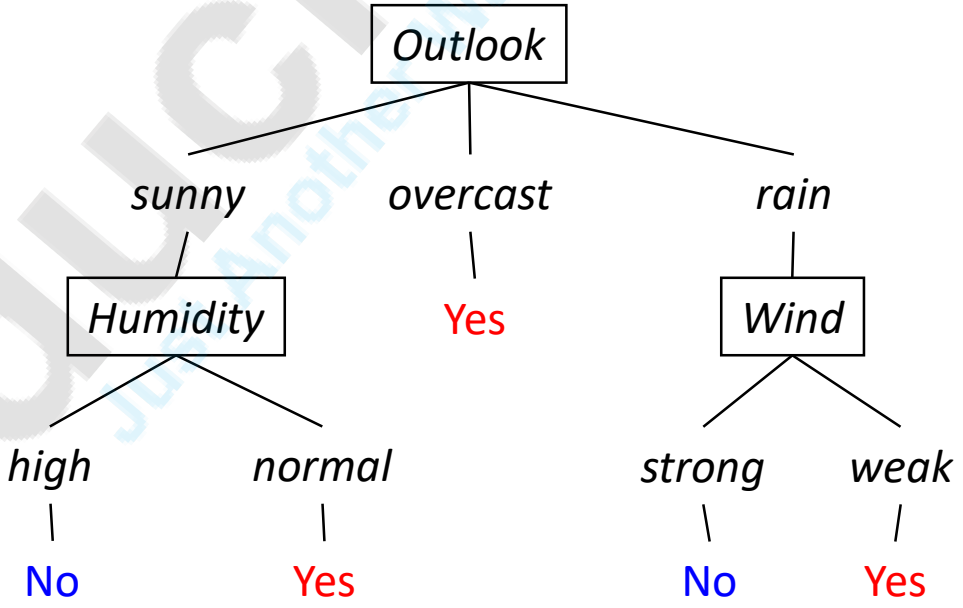
Decision trees

- Purpose is to expose the structural information contained in the data
- Classification trees
  - Response variable (class variable) is nominal or categorical
  - E.g. drug A or B
  - ID3 (induction decision tree version 3)
  - C4.5
- Regression trees
  - Response variable (class variable) is continuous
  - E.g. income
  - CART (classification and regression trees)

Decision Trees

- Decision tree to represent learned target functions
  - Each internal node tests an attribute
  - Each branch corresponds to attribute value
  - Each leaf node assigns a classification

- Can be represented by logical formulas



Representation in decision trees

- Example of representing rule in DT's:

  if outlook = sunny AND humidity = normal

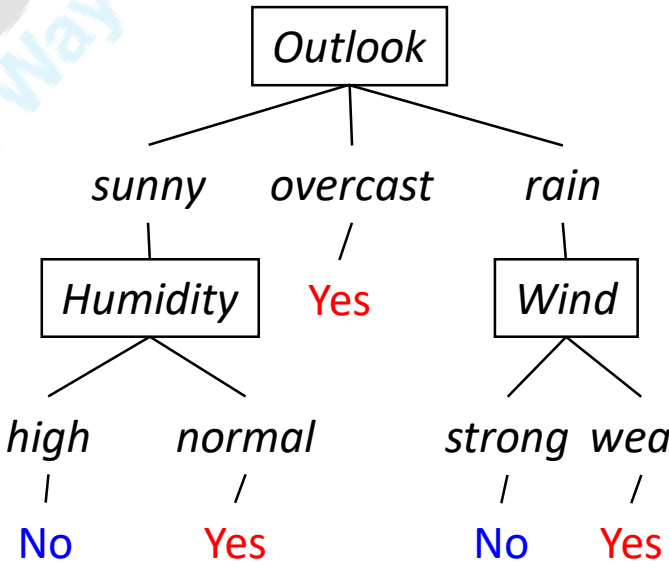
  OR

  if outlook = overcast

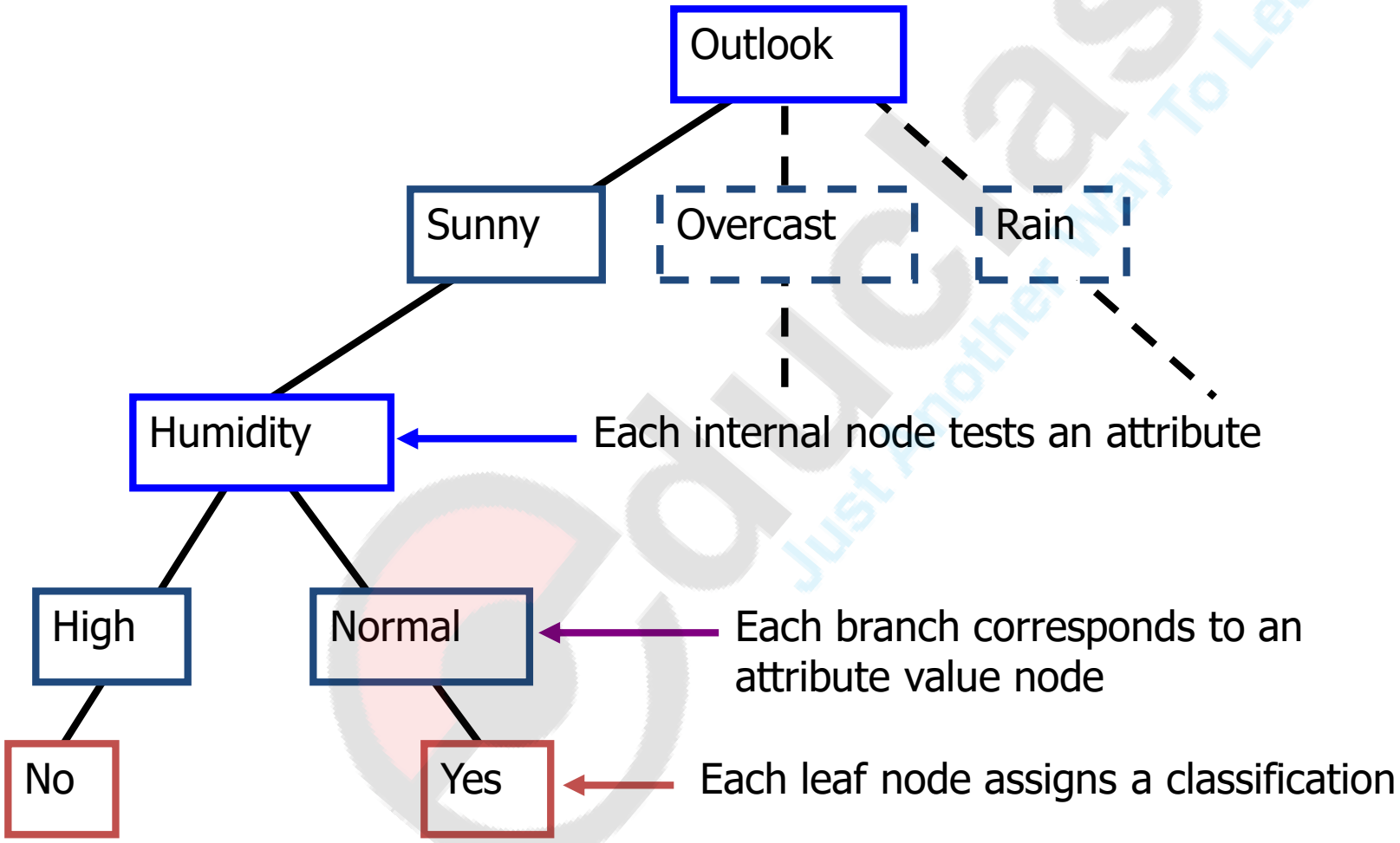
  OR

  if outlook = rain AND wind = weak

  then playtennis

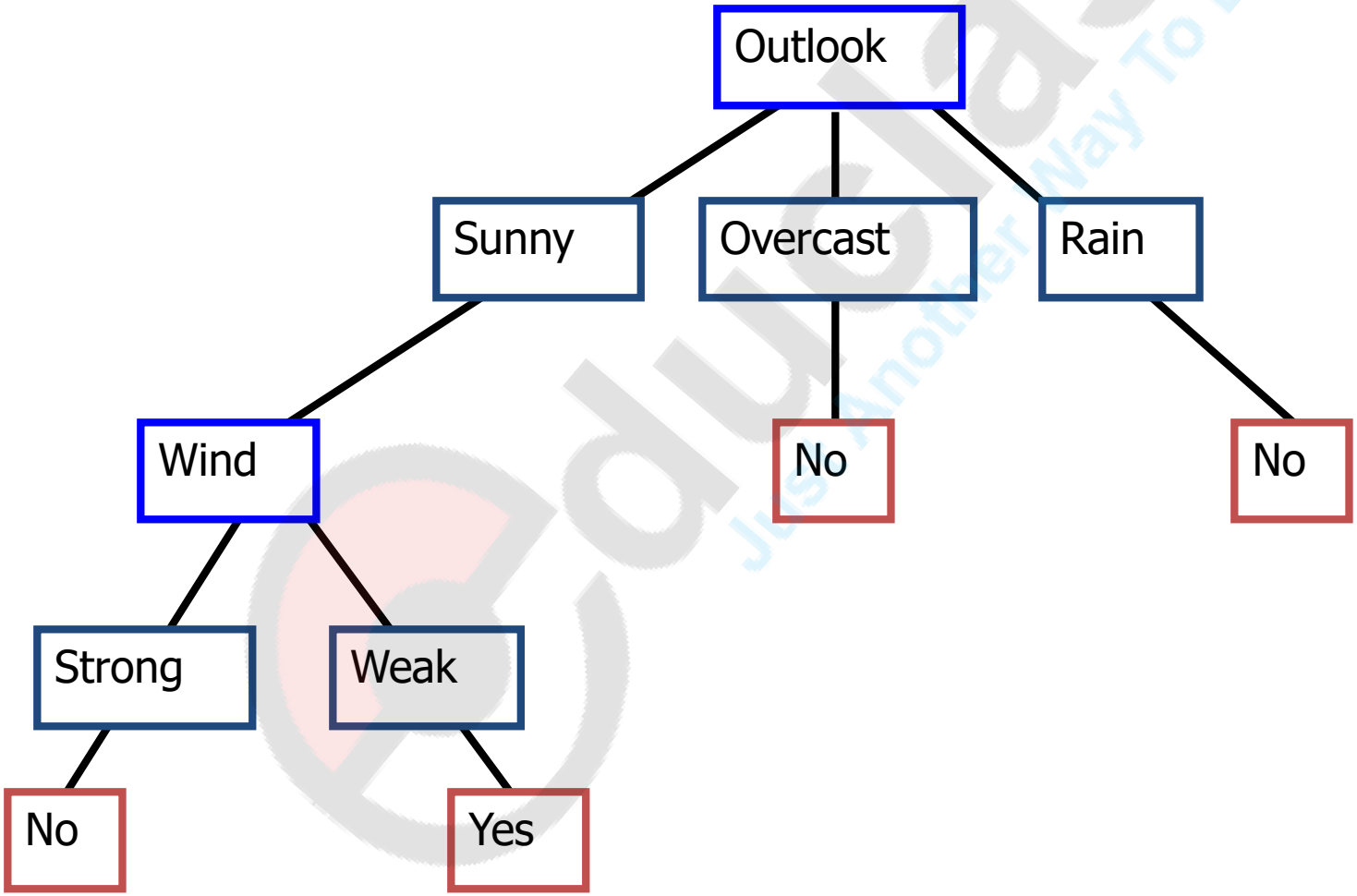


Decision Tree for PlayTennis



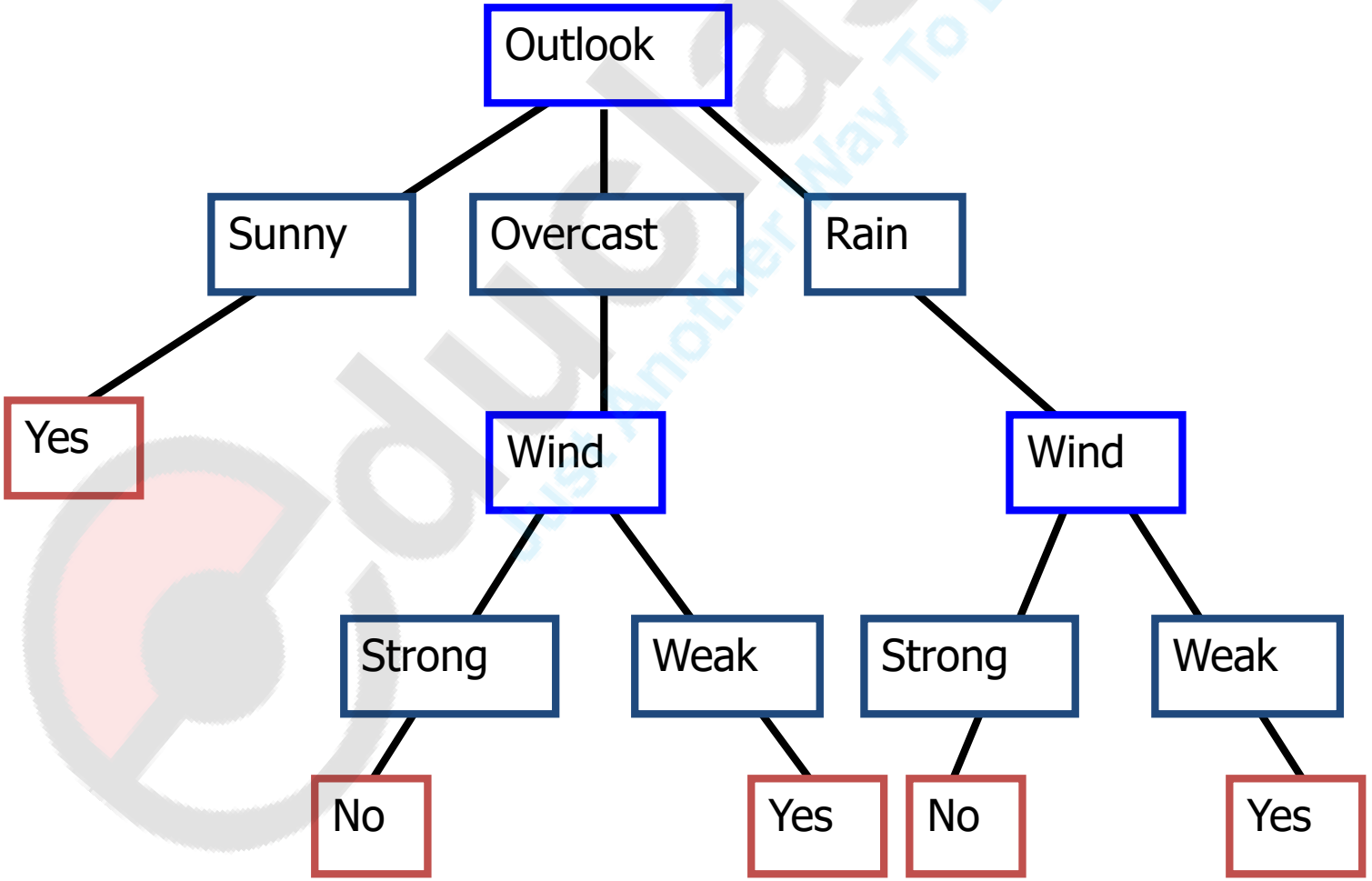
Decision Tree

Outlook=Sunny $\wedge$ Wind=Weak



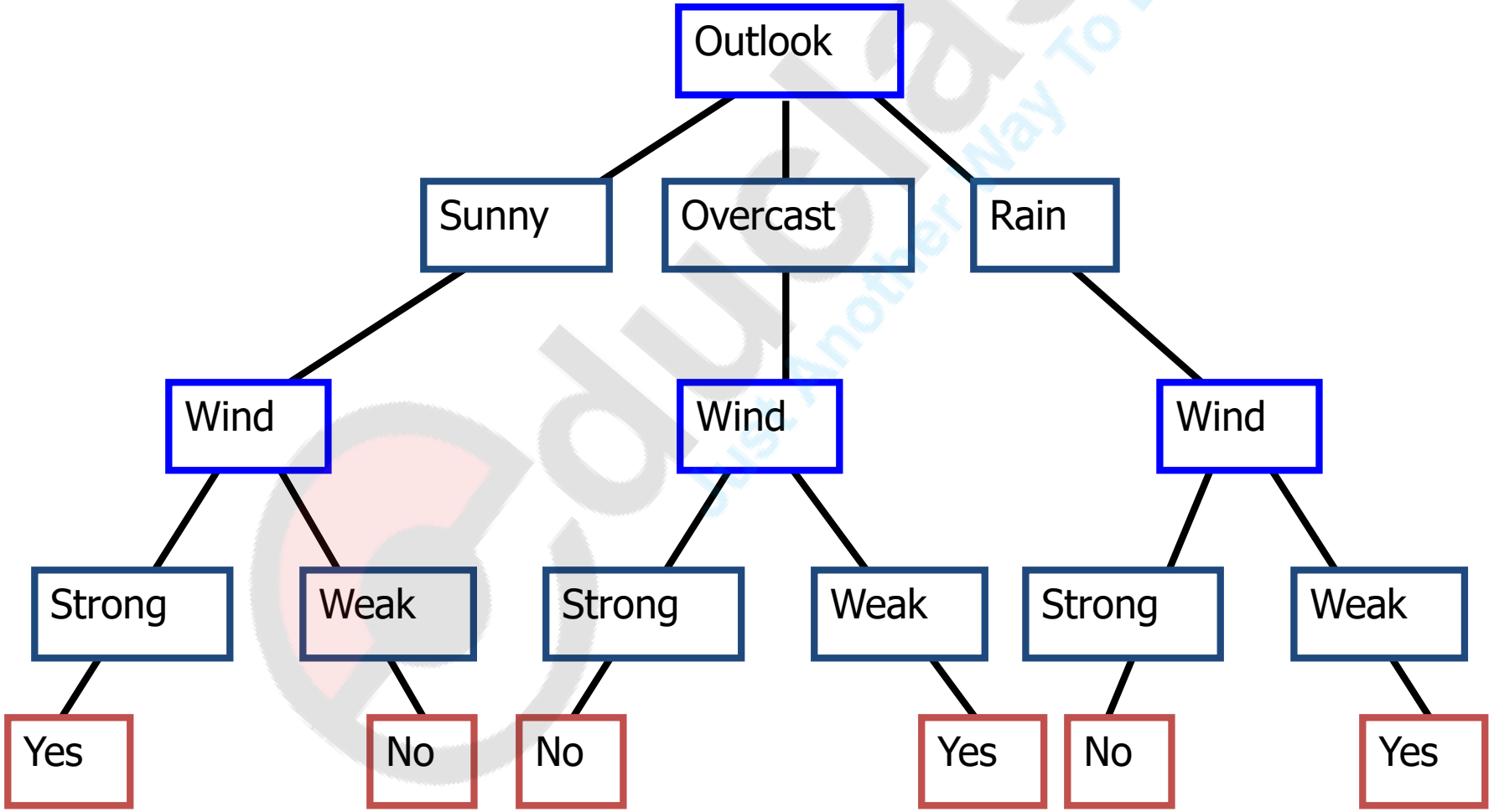
Decision Tree

Outlook=Sunny $\vee$ Wind=Weak



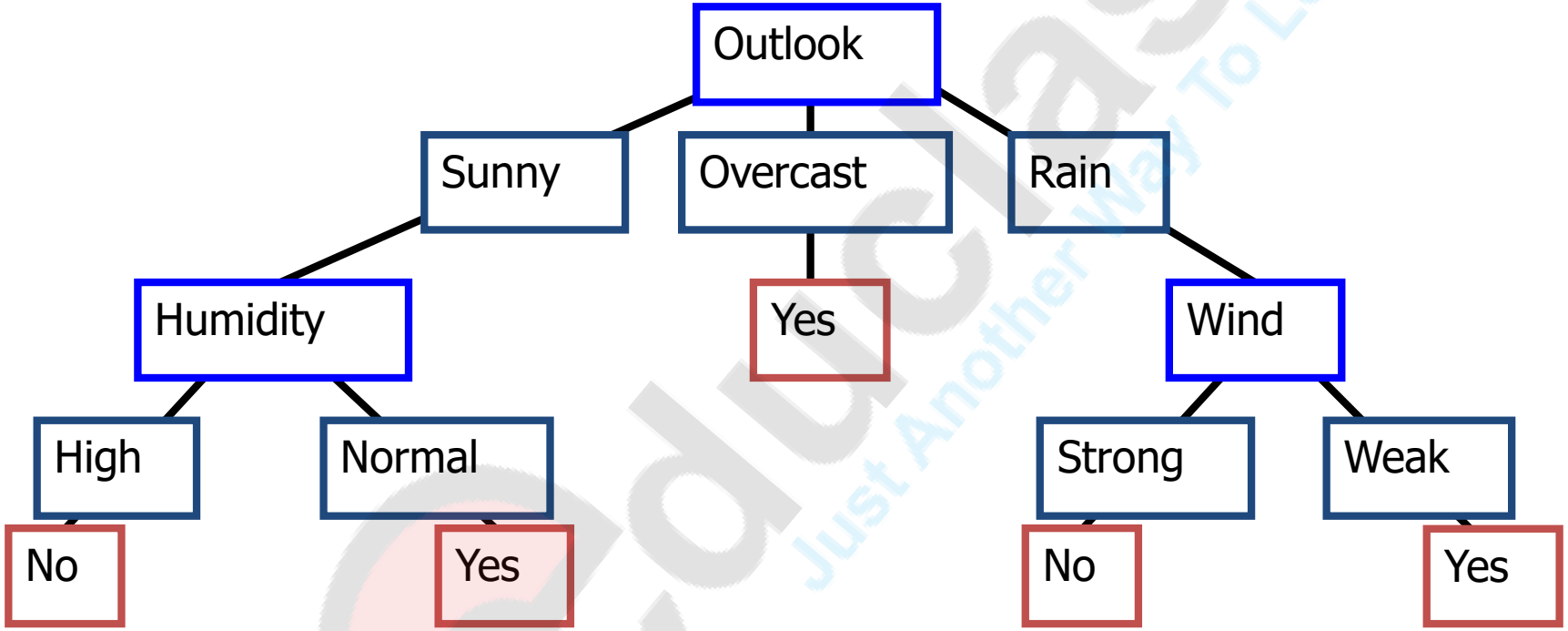
Decision Tree for XOR

Outlook=Sunny XOR Wind=Weak



Decision Tree

- decision trees represent disjunctions of conjunctions



(Outlook=Sunny ∧ Humidity=Normal)
∨ (Outlook=Overcast)
∨ (Outlook=Rain ∧ Wind=Weak)

ID3

- Recursive procedure to construct a decision tree from data.
- ID3 algorithm assumes that a good decision tree is the simplest decision tree
  - We should always accept the simplest answer that correctly fits our data
  - The smallest decision tree that correctly classifies all given examples
- The simplest decision tree that covers all examples should be the least likely to include unnecessary constraints

Important values

- **Information gain**

$$I(p,n) = \frac{-p}{p+n}\log_2\left[\frac{p}{p+n}\right] - \frac{n}{p+n}\log_2\left[\frac{n}{p+n}\right]$$

- **Entropy** of the attribute

$$E(A) = \sum \frac{p_i+n_i}{p+n}\left[\, I(p,n) \,\right]$$

- **Gain**(A) $= I(p,n) - E(A)$

p= yes=9

n=no =5

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D3	Overcast	Hot	High	Weak	*Yes*
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D7	Overcast	Cool	Normal	Strong	*Yes*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D10	Rain	Mild	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

p= yes=9

n=no =5

- Information gain (IG) *of the table*

$$I(p,n) = \frac{-p}{p+n}\log_2\left[\frac{p}{p+n}\right] - \frac{n}{p+n}\log_2\left[\frac{n}{p+n}\right]$$

$$I(9,5) = \frac{-9}{14}\log_2\left[\frac{9}{14}\right] - \frac{5}{14}\log_2\left[\frac{5}{14}\right]$$

$$I(9,5) = \frac{-9}{14}\log_2\left[0.642\right] - \frac{5}{14}\log_2\left[0.357\right]$$

$$I(9,5) = \frac{-9}{14}\frac{\log(0.642)}{\log 2} - \frac{5}{14}\frac{\log(0.357)}{\log 2}$$

$$= [-9(-0.639)]/14 - [5(-1.485)]/14 = 0.941$$

- P= yes=9
- N=no =5

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D3	Overcast	Hot	High	Weak	*Yes*
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D7	Overcast	Cool	Normal	Strong	*Yes*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D10	Rain	Mild	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

$$I(p,n) = \frac{-p}{p+n}\log_2\left[\frac{p}{p+n}\right] - \frac{n}{p+n}\log_2\left[\frac{n}{p+n}\right]$$

p= yes=9

n=no =5

- **Entropy** of the attribute

$$E(A) = \sum \frac{p_i+n_i}{p+n}\left[I(p_i,n_i) \right]$$

Outlook	Tennis?
Sunny	*No*
Sunny	*No*
Overcast	*Yes*
Rain	*Yes*
Rain	*Yes*
Rain	*No*
Overcast	*Yes*
Sunny	*No*
Sunny	*Yes*
Rain	*Yes*
Sunny	*Yes*
Overcast	*Yes*
Overcast	*Yes*
Rain	*No*

Outlook	p	n	I(p,n)
Sunny	2	3	0.970
Overcast	4	0	0
Rain	3	2	0.970

I(2,3)= **0.970**

I(4,0)= **0** , if p or n is zero

I(3,2)= **0.970** , same for 3,2 or 2,3

E(outlook)= **[**(2+3)/(9+5)**]**(0.970) **+** 0 **+** **[**(3+2)/14**]**(0.970)

= 0.692

Gain (outlook)= IG - **E**(outlook)

= 0.940 – 0.692

= 0.248

E(humidity)= ?

E(wind)= ?

E(temp)= ?

Gain (humidity)= ?

Gain(wind)= ?

Gain (temp)= ?

$$I(p,n) = \frac{-p}{p+n}\log_2\left[\frac{p}{p+n}\right] - \frac{n}{p+n}\log_2\left[\frac{n}{p+n}\right]$$

p= yes=9
n=no =5

- **Entropy** $E(A) = \sum \frac{p_i+n_i}{p+n}\left[I(p_i,n_i) \right]$

  of the attribute

Humidity	Tennis ?
High	*No*
High	*No*
High	*Yes*
High	*Yes*
Normal	*Yes*
Normal	*No*
Normal	*Yes*
High	*No*
Normal	*Yes*
Normal	*Yes*
Normal	*Yes*
High	*Yes*
Normal	*Yes*
High	*No*

High → p=3 n=4 I(3,4)= ?
Normal → p=6 n=1 I(6,1)= ?

E(humidity)= ?

Gain (humidity)= IG - **E**(humidity)
= 0.151

gain(wind)= 0.045
gain(temp)= 0.029

Gain (outlook)= IG - **E**(outlook)

= 0.940 – 0.692

= 0.248

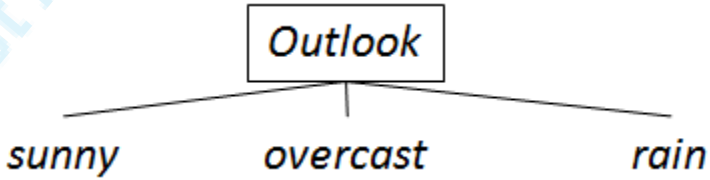
Maximum gain is of outlook therefore tree starts from outlook ie. root

Gain (humidity)= IG - **E**(humidity)

= 0.151

gain(wind)= 0.045

gain(temp)= 0.029



T

p= yes=9

n=no =5

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D3	Overcast	Hot	High	Weak	*Yes*
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D7	Overcast	Cool	Normal	Strong	*Yes*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D10	Rain	Mild	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

T_{sunny}

p= yes=2

n=no =3

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*

Entropy $E(A) = \sum \frac{p_i+n_i}{p+n} [I(p_i,n_i)]$

TEMP	p	n	I(p,n)
Hot	0	2	
Mild	1	1	
cold	1	0	

humidity	p	n	I(p,n)
High	0	3	
normal	2	0	

wind	p	n	I(p,n)
Weak	1	2	
strong	1	1	

E(temp)=

E(humidity)=

E(wind)=

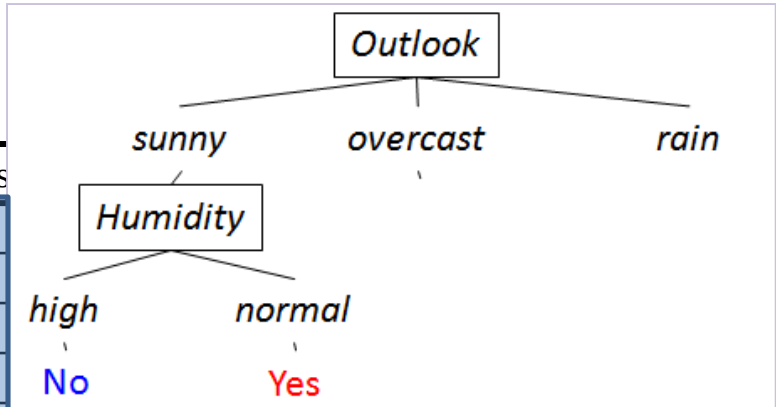
Gain(temp)= IG_{sunny} -E(temp)=0.571 Gain(humidity)= 0.971

Gain(wind)= 0.020

We find that gain for humidity is max

$IG(T_{sunny}) =$

Day	Outlook	Temp	Humidity	Wind	Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes



T

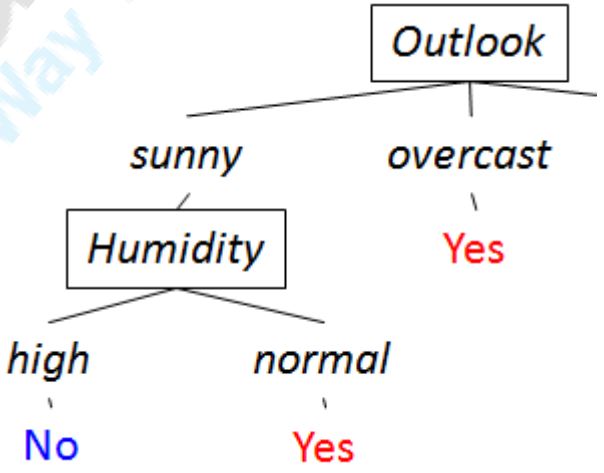
Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D3	Overcast	Hot	High	Weak	*Yes*
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D7	Overcast	Cool	Normal	Strong	*Yes*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D10	Rain	Mild	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

p= yes=9

n=no =5

All “yes”

Thus leaf node found



T overcast

Day	Outlook	Temp	Humidity	Wind	Tennis?
D3	Overcast	Hot	High	Weak	*Yes*
D7	Overcast	Cool	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*

p= yes=4

n=no =0

T

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D3	Overcast	Hot	High	Weak	*Yes*
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D7	Overcast	Cool	Normal	Strong	*Yes*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D10	Rain	Mild	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

p= yes=9

n=no =5

As humidity is already considered we will exclude it.

T_{Rain}

Day	Outlook	Temp	Humidity	Wind	Tennis?
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D10	Rain	Mild	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

p= yes=3

n=no =2

Entropy $E(A) = \sum \frac{p_i+n_i}{p+n} [I(p_i,n_i)]$

TEMP	p	n	I(p,n)
Hot	0	0	0
Mild	2	1	
cold	1	1	

wind	p	n	I(p,n)
Weak	3	0	0
strong	0	2	0

E(temp)=0.950978

E(wind)= 0

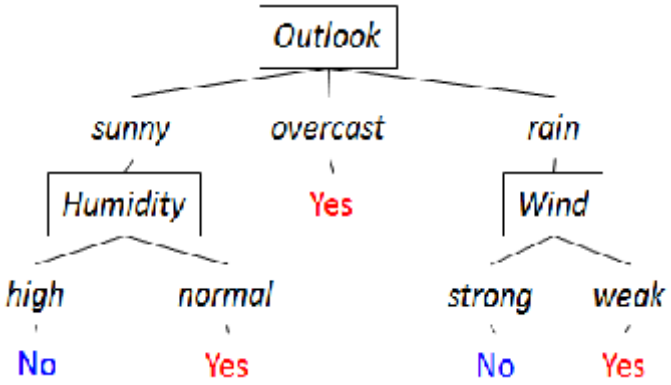
Gain(temp)= IG_{sunny} -E(temp)= 0.019973

Gain(wind)= IG_{sunny} -E(temp)=0.970951

Wind is the next branch

T Rain

Day	Outlook	Temp	Humidity	Wind	Tennis?
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D10	Rain	Mild	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



Example 1:

	age	competition	Type	Profit
1	old	Yes	software	down
2	old	No	Software	Down
3	old	No	Hardware	Down
4	Mid	Yes	Software	Down
5	Mid	Yes	Hardware	Down
6	Mid	No	Hardware	Up
7	Mid	No	Software	Up
8	new	Yes	Software	Up
9	new	No	Hardware	Up
10	new	No	Software	Up

TABLE 4.1: Data for Height Classification

Name	Gender	Height	Output1
Kristina	F	1.6 m	Short
Jim	M	2 m	Tall
Maggie	F	1.9 m	Medium
Martha	F	1.88 m	Medium
Stephanie	F	1.7 m	Short
Bob	M	1.85 m	Medium
Kathy	F	1.6 m	Short
Dave	M	1.7 m	Short
Worth	M	2.2 m	Tall
Steven	M	2.1 m	Tall
Debbie	F	1.8 m	Medium
Todd	M	1.95 m	Medium
Kim	F	1.9 m	Medium
Amy	F	1.8 m	Medium
Wynette	F	1.75 m	Medium