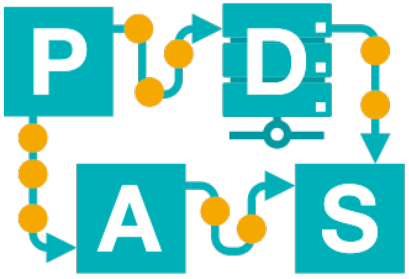
Introduction to Data Science (IDS) course

Decision Tree

Lecture 3 Instruction

IDS-I-L3



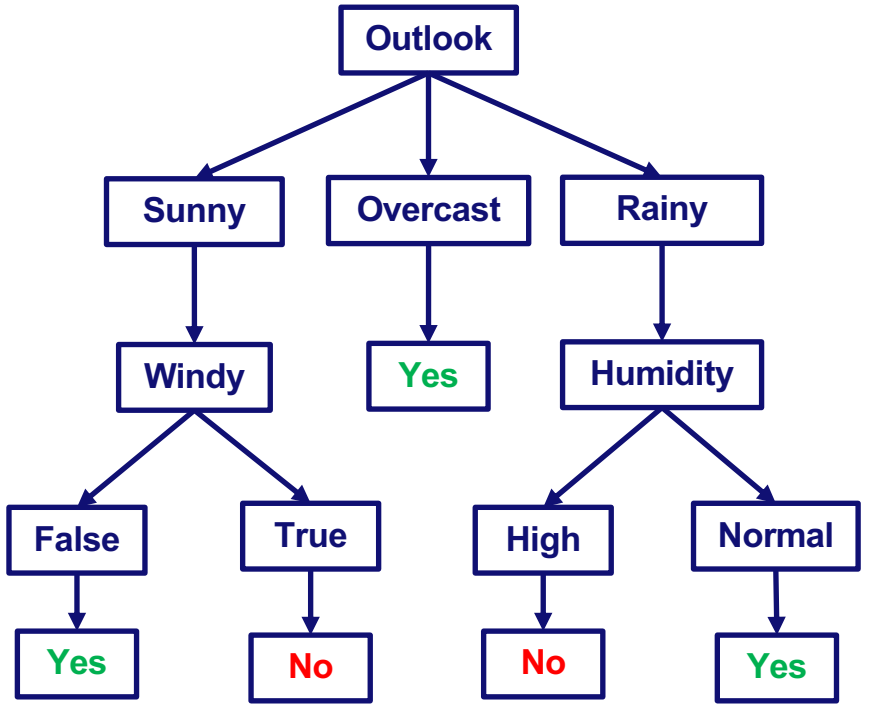
Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Q1. ID3 Complete example

Descriptive Features | Target Feature

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Chair of Process and Data Science

Q1. ID3 Complete example

1. Calculate entropy of the target feature.

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Play Golf	
Yes	No
9	5

$$\boldsymbol{Entropy\,(PlayGolf) = -(0.36log_2 0.36) - (0.64log_2 0.64) = 0.94}$$

Chair of Process and Data Science

Q1. ID3 Complete example

2. Entropy after splitting by “Outlook”.

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

$$E(3,2) = -(0.6log_20.6) - (0.4log_20.4) = 0.97$$

$$E(4,0) = -(1\ log_21) - (0log_20) = 0$$

$$E(2,3) = E(3,2) = 0.97$$

$$Entropy\ (PlayGolf)$$
$$= P(Sunny)\times E(3,2) + P(Overcast)\times E(4,0) + P(Rainy)\times E(2,3)$$
$$= \left(\frac{5}{14}\right)\times 0.97 + \left(\frac{4}{14}\right)\times 0 + \left(\frac{5}{14}\right)\times 0.97 = 0.69$$

$$Information\ Gain = 0.94 - 0.69 = 0.25$$

Chair of Process and Data Science

Q1. ID3 Complete example

3. Calculate information gain after splitting by each descriptive feature.

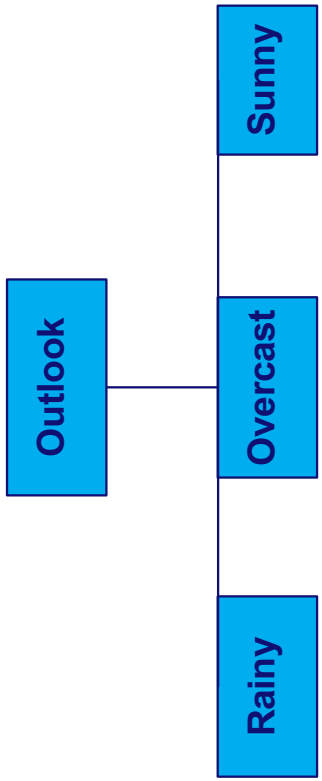
		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.25			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.02			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.15			

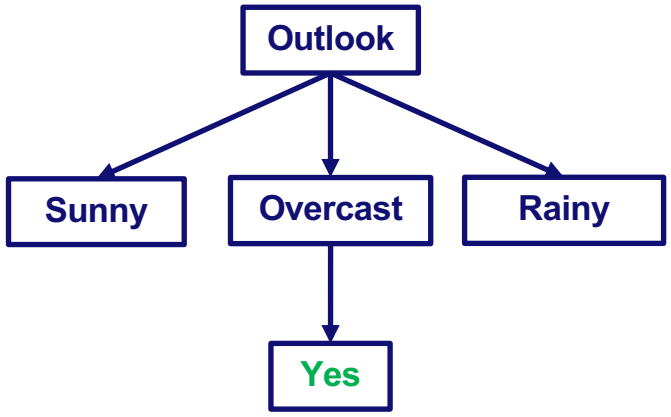
		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.04			

Q1. ID3 Complete example



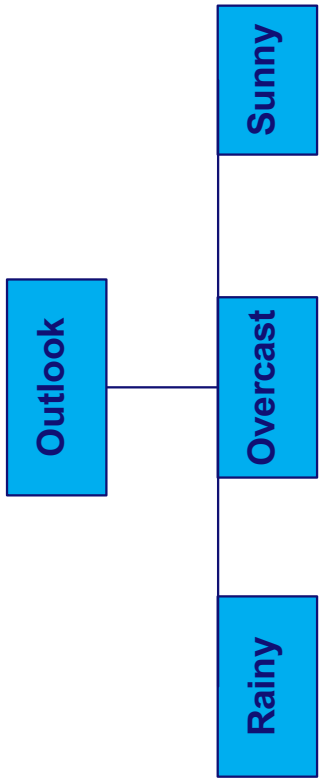
Outlook	Temp	Humidity	Windy	Play Golf
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Sunny	Mild	Normal	False	Yes
Sunny	Mild	High	True	No
Overcast	Hot	High	False	Yes
Overcast	Cool	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Rainy	Mild	Normal	True	Yes

4. Split data based on the feature which has the maximum gain, and repeat steps 1-3 for each part.

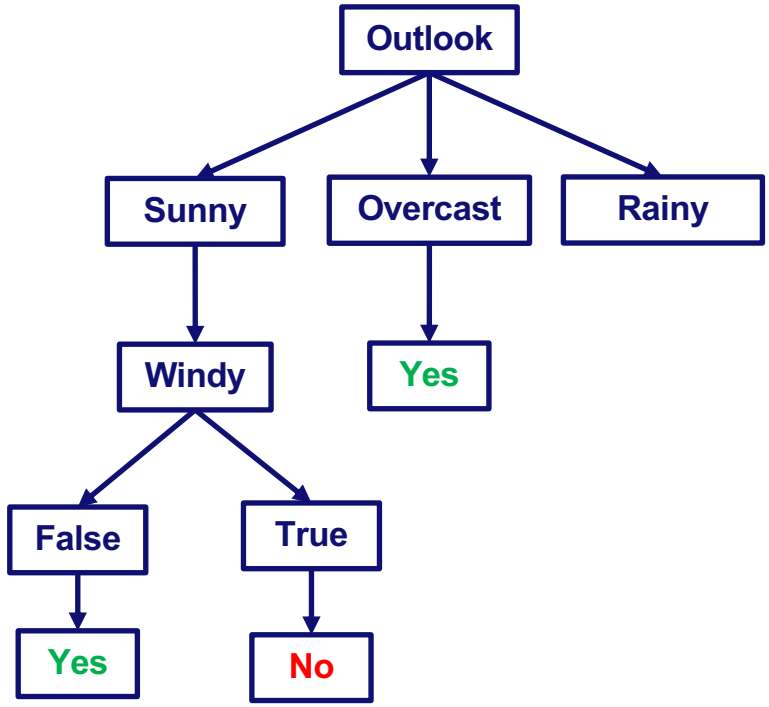


Chair of Process and Data Science

Q1. ID3 Complete example

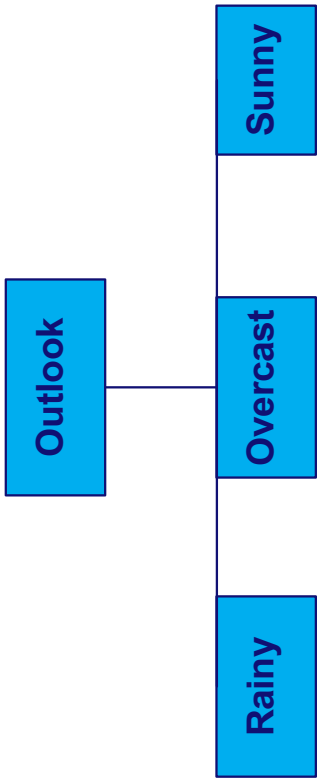


Outlook	Temp	Humidity	Windy	Play Golf
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Sunny	Mild	Normal	False	Yes
Sunny	Mild	High	True	No
Overcast	Hot	High	False	Yes
Overcast	Cool	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Rainy	Mild	Normal	True	Yes

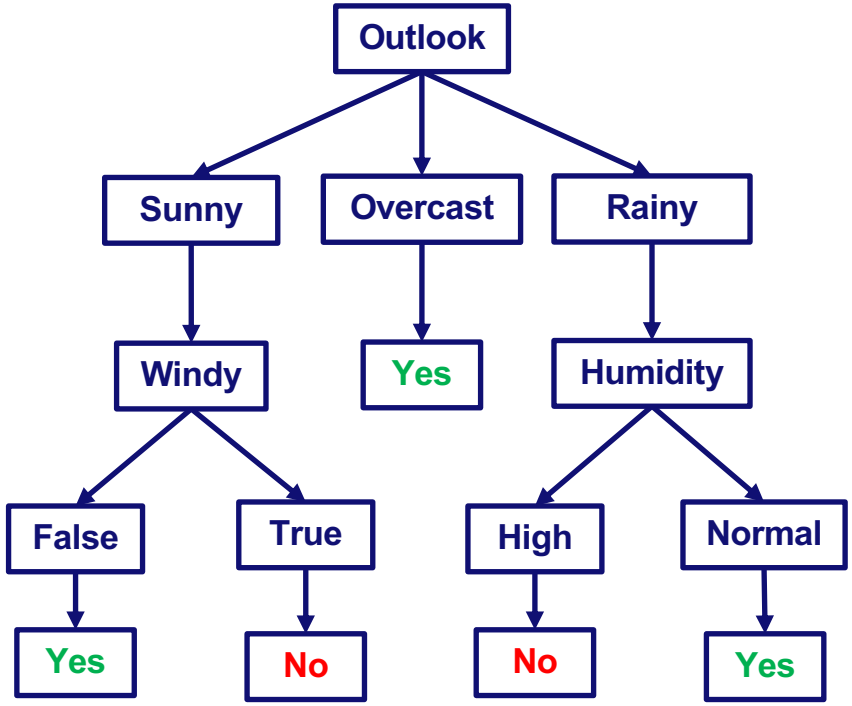


Chair of Process and Data Science

Q1. ID3 Complete example



Outlook	Temp	Humidity	Windy	Play Golf
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Sunny	Mild	Normal	False	Yes
Sunny	Mild	High	True	No
Overcast	Hot	High	False	Yes
Overcast	Cool	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Rainy	Mild	Normal	True	Yes



Chair of Process and Data Science

Q2. Your Turn

- **Suppose that the following data is about accepting or rejecting job applications based on "Experience", "Degree", and type of the job ("Job") that applicants applied for it. What is the decision tree for the following data set based on entropy?**

Experience	Degree	Job	Class
Exp >10	HS	Board	No
5< Exp <10	Uni	Board	Yes
Exp >10	HS	Board	No
5< Exp <10	HS	Hcare	Yes
Exp < 5	HS	Hcare	Yes
Exp < 5	HS	Board	No
Exp < 5	None	Edu	No
Exp >10	None	Hcare	No
Exp < 5	Uni	Edu	Yes
Exp >10	Uni	Board	Yes

Q2. Solution

1. Calculate entropy of the target feature.

Experience	Degree	Job	Class
Exp >10	HS	Board	No
5< Exp <10	Uni	Board	Yes
Exp >10	HS	Board	No
5< Exp <10	HS	Hcare	Yes
Exp < 5	HS	Hcare	Yes
Exp < 5	HS	Board	No
Exp < 5	None	Edu	No
Exp >10	None	Hcare	No
Exp < 5	Uni	Edu	Yes
Exp >10	Uni	Board	Yes

Class	
No	Yes
5	5

$$Entropy\,(Class) = -(0.5log_2 0.5) - (0.5log_2 0.5) = -0.5(-1) - 0.5(-1) = 1$$

Chair of Process and Data Science

Q2. Solution

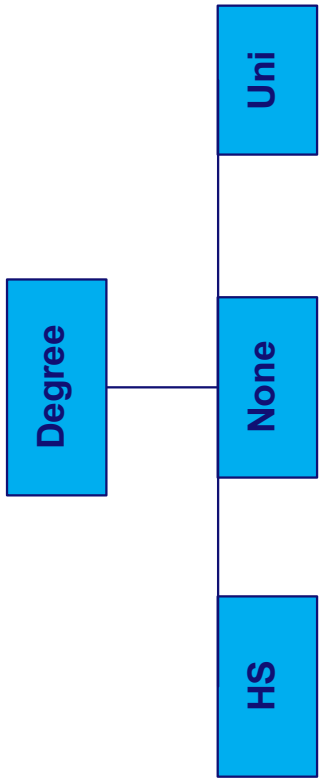
2. Calculate information gain after splitting by each descriptive feature.

		Play Golf	
		No	Yes
Experience	Exp>10	3	1
	5<Exp<10	0	2
	Exp<5	2	2
Gain = 0.27			

		Play Golf	
		No	Yes
Degree	HS	3	2
	Uni	0	3
	None	2	0
Gain = 0.52			

		Play Golf	
		No	Yes
Job	Board	3	2
	Hcare	1	2
	Edu	1	1
Gain = 0.05			

Q2. Solution

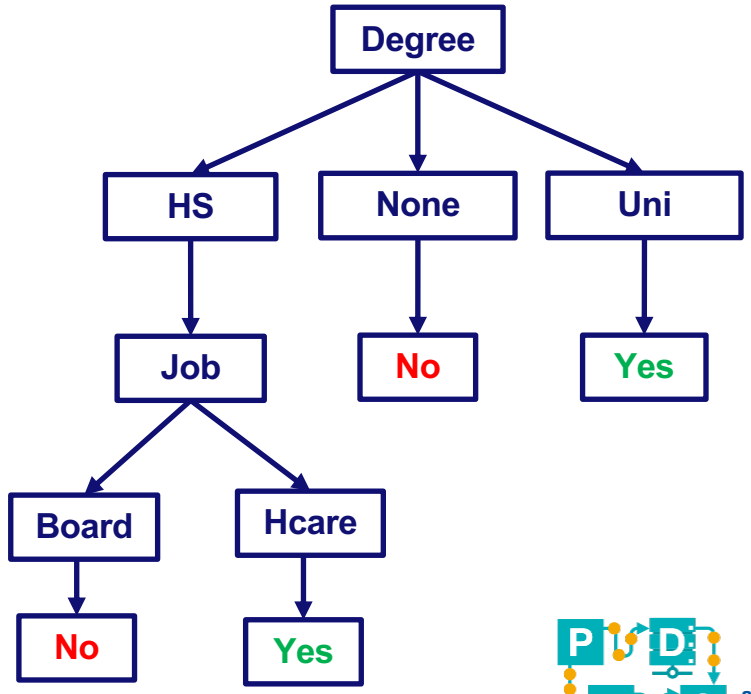


5< Exp <10	Uni	Board	Yes
Exp < 5	Uni	Edu	Yes
Exp >10	Uni	Board	Yes

Exp < 5	None	Edu	No
Exp >10	None	Hcare	No

Exp >10	HS	Board	No
Exp >10	HS	Board	No
5< Exp <10	HS	Hcare	Yes
Exp < 5	HS	Hcare	Yes
Exp < 5	HS	Board	No

3. Split data based on the feature which has the maximum gain, and repeat the same steps for each part.



Chair of Process and Data Science

Q3. Numerical Descriptive Features

- **What are possible categories for “Experience” feature?**

Experience	Degree	Job	Class
12	HS	Board	No
20	Uni	Board	Yes
15	HS	Board	No
10	HS	Hcare	Yes
3	HS	Hcare	No
7	HS	Board	Yes
5	None	Edu	No
2	None	Hcare	No
6	Uni	Edu	Yes
11	Uni	Board	Yes

Q3. Solution

- **Sort the data based on the numerical feature and select borders based on the transitions in the target feature.**

Experience	Degree	Job	Class
12	HS	Board	No
20	Uni	Board	Yes
15	HS	Board	No
10	HS	Hcare	Yes
3	HS	Hcare	No
7	HS	Board	Yes
5	None	Edu	No
2	None	Hcare	No
6	Uni	Edu	Yes
11	Uni	Board	Yes

Sort based on "Experience"

	Experience	Degree	Job	Class
Experience < 6	2	None	Hcare	No
	3	HS	Hcare	No
	5	None	Edu	No
6 <= Experience < 12	6	Uni	Edu	Yes
	7	HS	Board	Yes
	10	HS	Hcare	Yes
	11	Uni	Board	Yes
12 <= Experience < 20	12	HS	Board	No
	15	HS	Board	No
Experience >= 20	20	Uni	Board	Yes

Chair of Process and Data Science

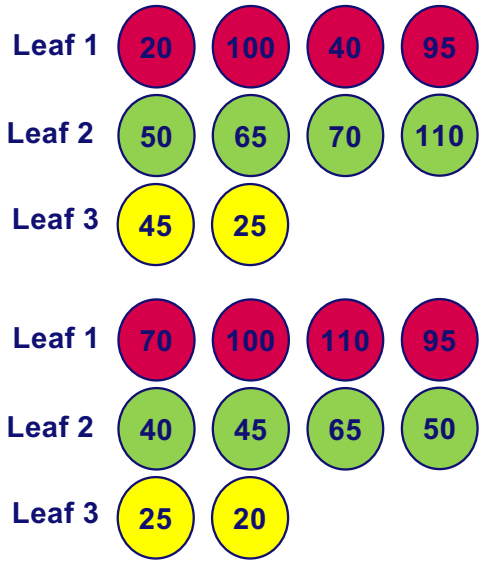
Q4. Numerical Target Feature

- **Suppose that we have the following leaves after splitting the data set. Which classification is better and why?**

$$var(a) = \frac{\sum_{i=1}^{n} (a_i - \bar{a})^2}{n-1}$$

Descriptive Features | Target Features

Experience	Degree	Job	Salary (K)
*	*	*	20
*	*	*	40
*	*	*	50
*	*	*	65
*	*	*	70
*	*	*	25
*	*	*	95
*	*	*	100
*	*	*	110
*	*	*	45

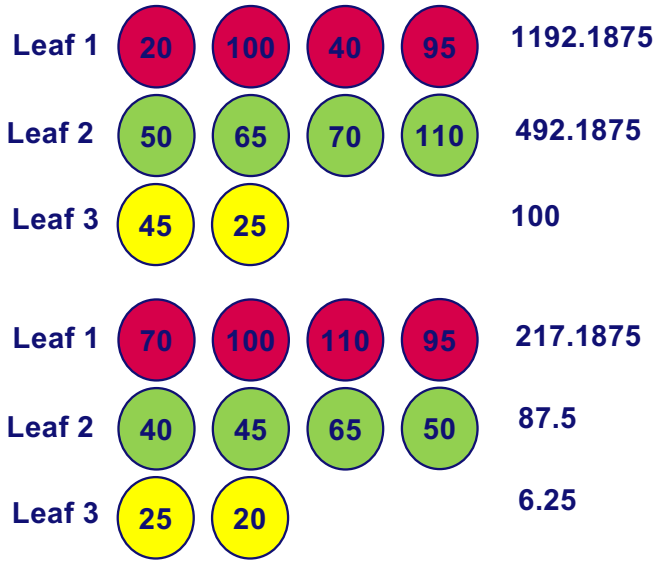


Chair of Process and Data Science

Q4. Solution

- **Suppose that we have following leaves after splitting the data set. Which classification is better and why?**

Descriptive Features			Target Features
Experience	Degree	Job	Salary (K)
*	*	*	20
*	*	*	40
*	*	*	50
*	*	*	65
*	*	*	70
*	*	*	25
*	*	*	95
*	*	*	100
*	*	*	110
*	*	*	45



Since the variances are better, this classification is better

Chair of Process and Data Science

Q5. Homework

- We would like to predict the sex of a person based on two binary attributes: leg-cover (pants or skirts) and facial-hair (some or none). We have a data set of 1000 individuals, half male and half female. 50% of females wear skirt, and no male wears skirt. 75% of males and 25% of females have facial hair.
- Which attribute should be used as the root of the decision tree based on Entropy?