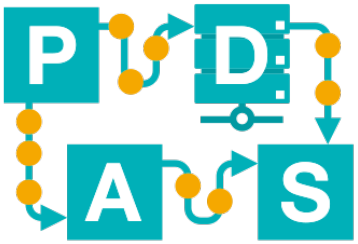
Responsible Data Science

Lecture 19 and 20 Instruction

IDS-L19-L20



Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Q1. Discrimination

Consider the following potentially discriminatory (PD) and the base rules with the mentioned confidence values.
What range for α causes the PD rule to be α-discriminatory?

Base Rule	$B \Rightarrow C$	**Confidence: 0.25**
PD Rule	$A, B \Rightarrow C$	**Confidence: 0.55**

Q1. Discrimination (Solution)

Consider the following potentially discriminatory (PD) and the base rules with the mentioned confidence values.
What range for α causes the PD rule to be α-discriminatory?

Base Rule	$B \Rightarrow C$	**Confidence: 0.25**
PD Rule	$A, B \Rightarrow C$	**Confidence: 0.55**

$$elift = \frac{confidence(A, B \Rightarrow C)}{confidence(B \Rightarrow C)} \qquad elift = \frac{0.55}{0.25} = 2.2$$

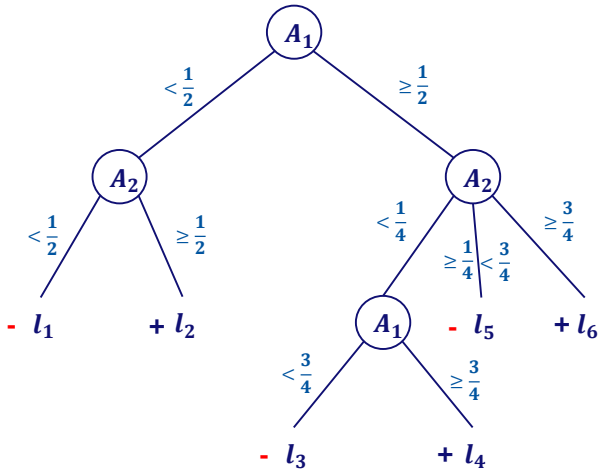
If $\alpha \leq 2.2$, then the PD rule is α-discriminatory.

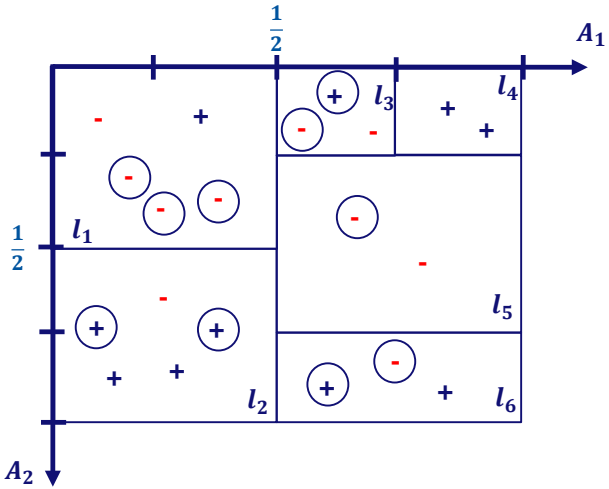
Q2. Discrimination (Your Turn)

Consider the following potentially discriminatory (PD) and the base rules with the mentioned support values.
What range for α causes the PD rule to be α-discriminatory?

Base Rule	$B \Rightarrow C$	$Support(\{B, C\}): 30$	$Support(\{B\}): 100$
PD Rule	$A, B \Rightarrow C$	$Support(\{A, B, C\}): 20$	$Support(\{A, B\}): 40$

Q3. Discrimination





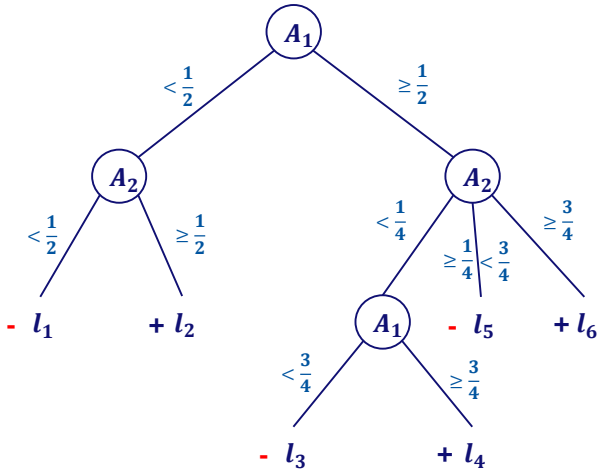
1. **Classify the regions based on their majority label.**
2. **Compute the accuracy and also the discrimination of the classifier w.r.t. discriminatory attribute (B).**
3. **If we want to relabel l_1, what would be the new label? and how this relabeling would affect the accuracy and discrimination?**

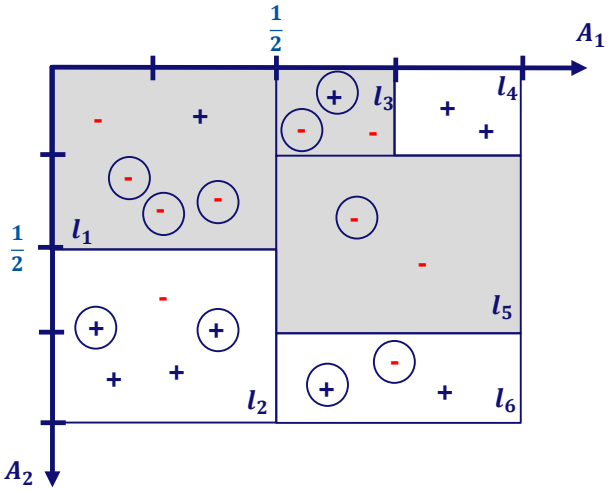
Note that encircled examples are discriminatory (have B=1).

Chair of Process and Data Science

Q3. Discrimination (Solution)

1. Classify the regions based on their majority label.

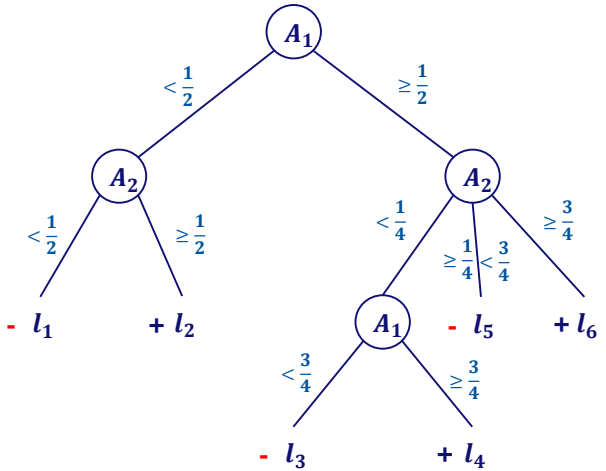


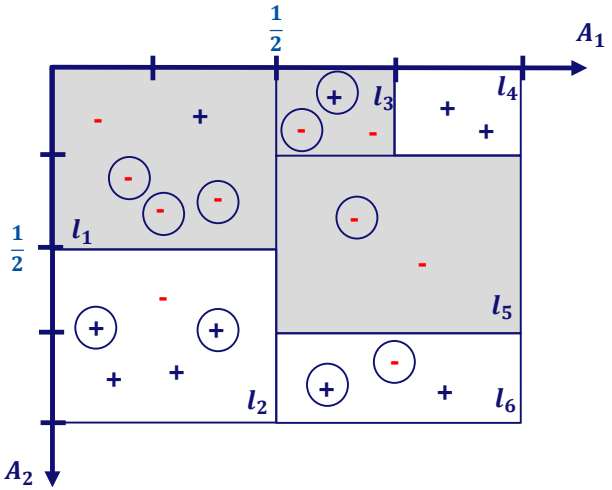


Chair of Process and Data Science

Q3. Discrimination (Solution)

2. Compute the accuracy and also the discrimination of the classifier w.r.t. discriminatory attribute (B).





Class	-	+	
Pred.	- / +	- / +	
$B = 1$	U_1/U_2	V_1/V_2	b
$B = 0$	W_1/W_2	X_1/X_2	$\bar{b}$
	N_1/N_2	P_1/P_2	1

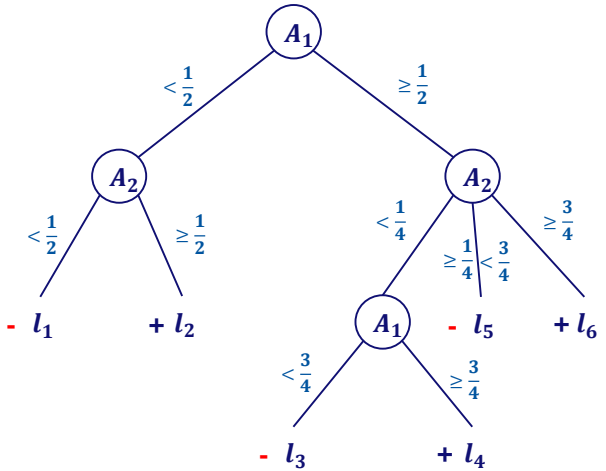
Class	-	+	
Pred.	- / +	- / +	
$B = 1$	$\frac{5}{20}/\frac{1}{20}$	$\frac{1}{20}/\frac{3}{20}$	$\frac{10}{20}$
$B = 0$	$\frac{3}{20}/\frac{1}{20}$	$\frac{1}{20}/\frac{5}{20}$	$\frac{10}{20}$
	$\frac{8}{20}/\frac{2}{20}$	$\frac{2}{20}/\frac{8}{20}$	1

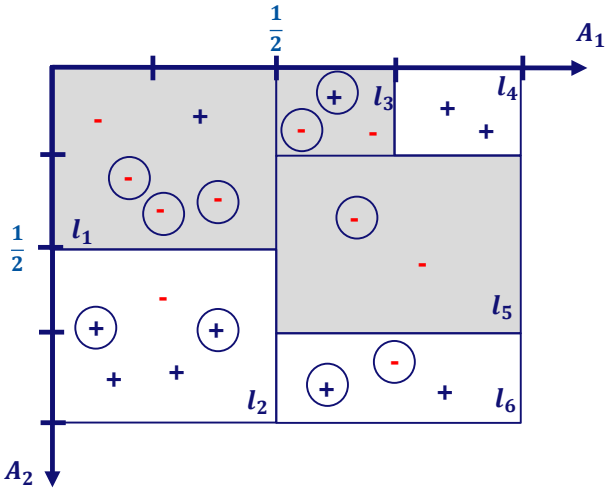
$$acc_T = N_1 + P_2 = \frac{8}{20} + \frac{8}{20} = 0.8$$

$$disc_T = \frac{W_2 + X_2}{\bar{b}} - \frac{U_2 + V_2}{b} = \frac{\frac{1}{20} + \frac{5}{20}}{\frac{1}{2}} - \frac{\frac{1}{20} + \frac{3}{20}}{\frac{1}{2}} = 0.2$$

Q3. Discrimination (Solution)

3. If we want to relabel l_1, what would be the new label? and how this relabeling would affect the accuracy and discrimination?





Class	-	+	
$B=1$	u	v	b
$B=0$	w	x	$\bar{b}$
	n	p	a

Class	-	+	
$B=1$	3/20	0	3/20
$B=0$	1/20	1/20	2/20
	4/20	1/20	5/20

$$n > p$$

New label would be +

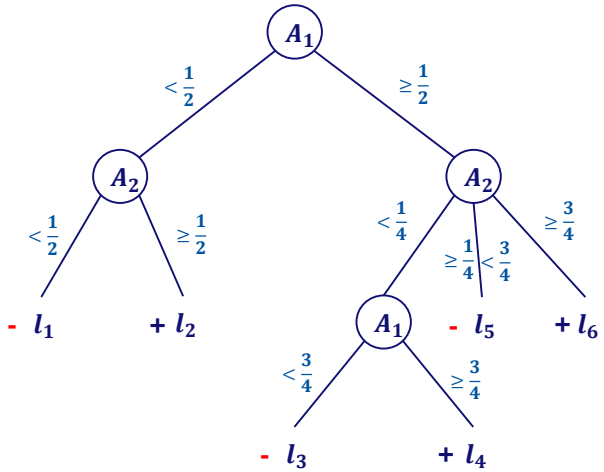
$$\Delta acc_l = p - n = -3/20$$

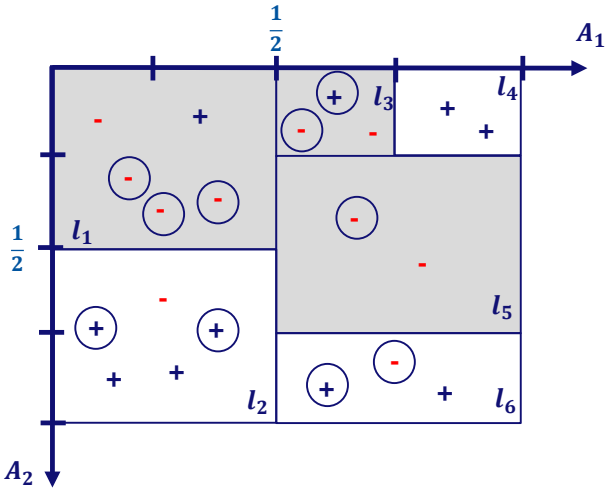
$$\Delta disc_l = -\frac{u+v}{b} + \frac{w+x}{\bar{b}} = -\frac{\frac{3}{20}}{\frac{1}{2}} + \frac{\frac{2}{20}}{\frac{1}{2}} = -0.1$$

Chair of Process and Data Science

Q4. Discrimination (Your Turn)

If we want to relabel l_6, what would be the new label? and how this relabeling would affect the accuracy and discrimination?

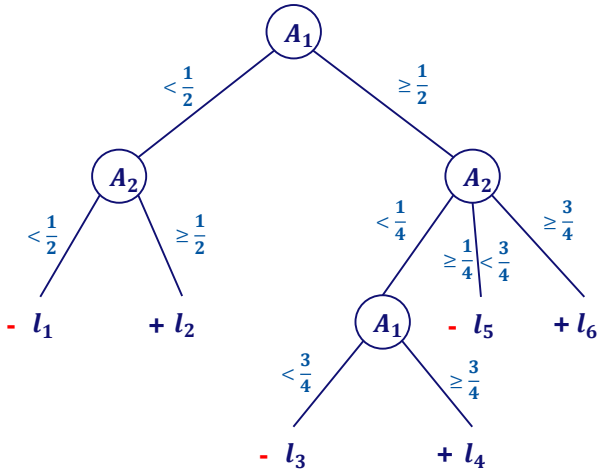


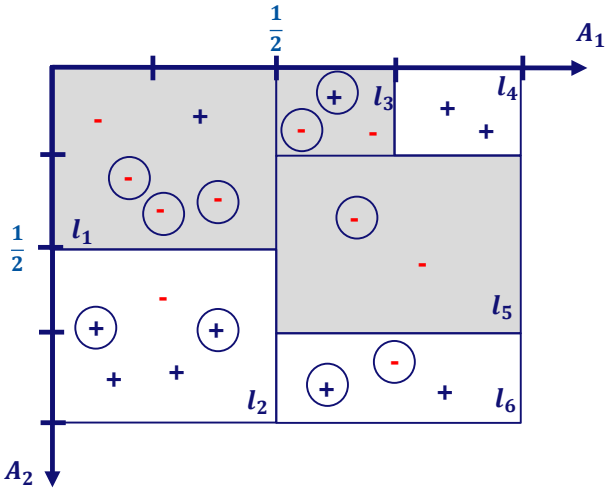


Chair of Process and Data Science

Q5. Discrimination (Your Turn)

If we want to relabel l_4, what would be the new label? and how this relabeling would affect the accuracy and discrimination?

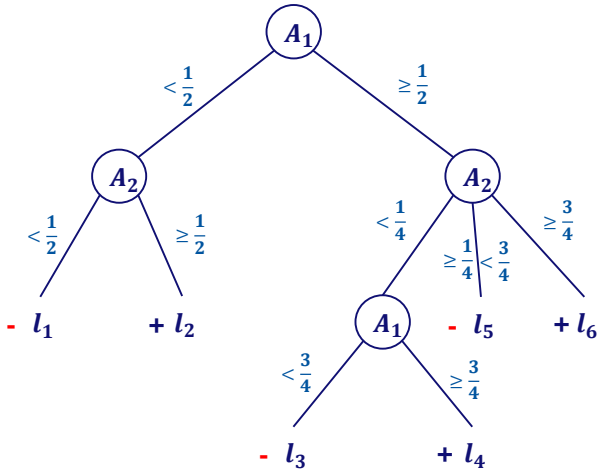


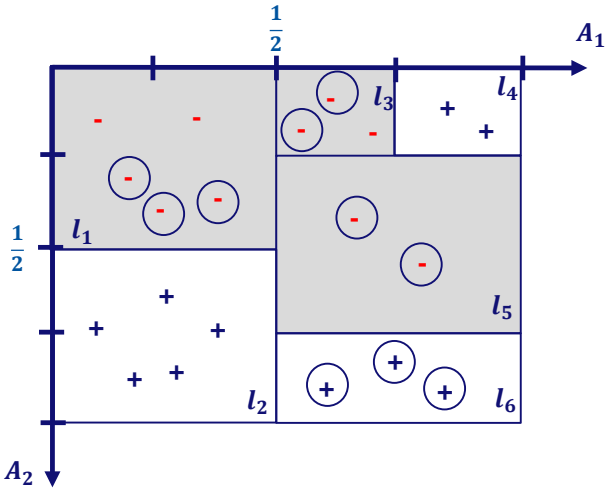


Chair of Process and Data Science

Q6. Discrimination (Your Turn)

In the following DT classifier, relabeling which leaf leads to the maximum reduction on the discrimination?

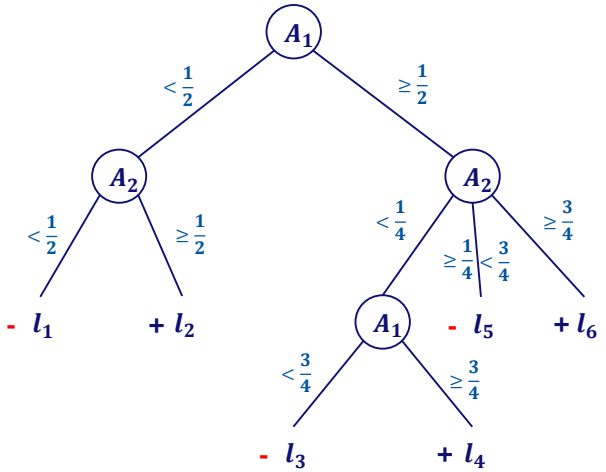


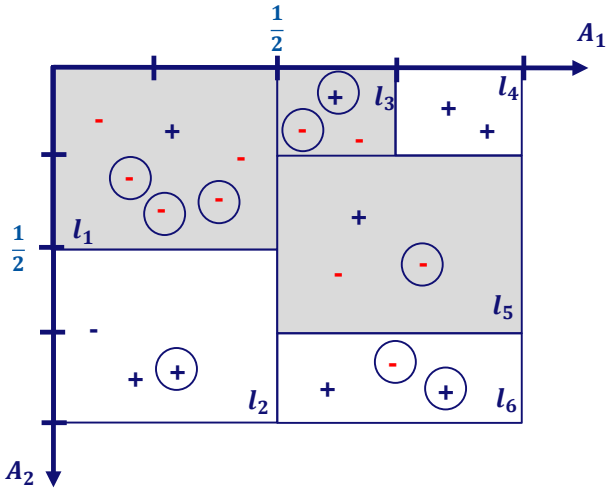


Chair of Process and Data Science

Q7. Discrimination (Your Turn)

In the following DT classifier, relabeling which leaf has the maximum effect on the accuracy?

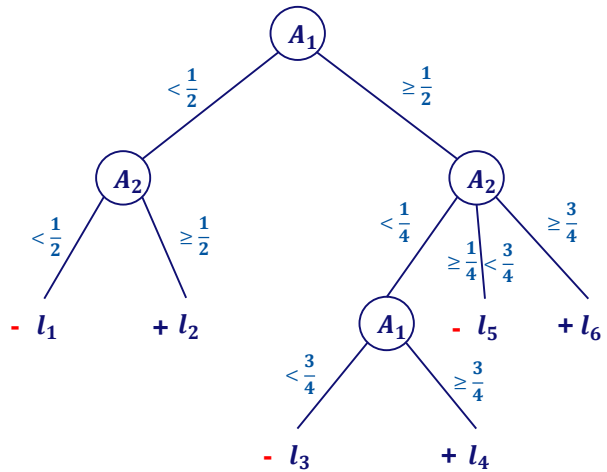


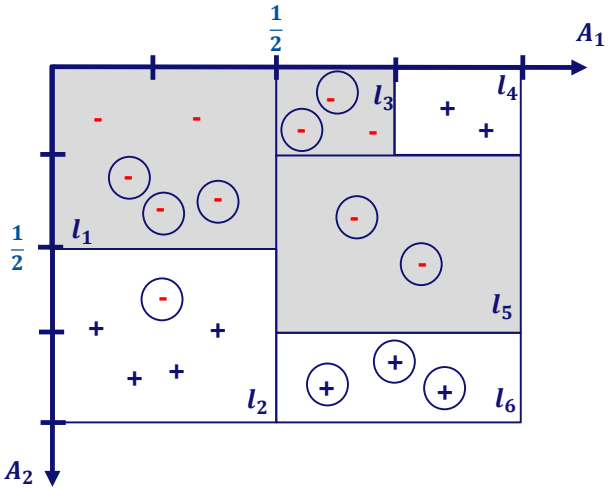


Chair of Process and Data Science

Q8. Discrimination (Homework)

In the following DT classifier, relabeling which leaf leads to the maximum reduction of the discrimination, and minimum reduction of the accuracy (the best leaf for relabeling)?





Chair of Process and Data Science

Q9. Discrimination (Homework)

What is the first node of the decision tree for the following table of data with respect to accuracy and fairness? (use $IGC - IGS$)

Sex	Exp	Degree	Job	Class
F	Exp >10	HS	Board	-
M	5< Exp <10	Uni	Board	+
M	Exp >10	HS	Board	-
M	5< Exp <10	HS	Hcare	+
M	Exp < 5	HS	Hcare	+
F	Exp < 5	HS	Board	-
M	Exp < 5	None	Edu	-
F	Exp >10	None	Hcare	-
M	Exp < 5	Uni	Edu	+
M	Exp >10	Uni	Board	+

$$IGC := H_{Class}(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} H_{Class}(D_i) \qquad IGS := H_B(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} H_B(D_i)$$

Q10. Confidentiality

Suppose that we have the following table of information about people and what they bought from an online grocery shop.

- **Using suppression and generalization anonymize this table such that it has 2-anonymity and distinct 2-diversity.**

Explicit identifier	Quasi-identifiers				Sensitive
Name	**Age**	**Gender**	**State of domicile**	**Religion**	**Product**
Ramsha	22	Female	Tamil Nadu	Hindu	Pea
Yadu	24	Female	Kerala	Hindu	Bean
Salima	25	Female	Tamil Nadu	Muslim	Peanut
Sunny	25	Male	Karnataka	Buddhist	Pea
Joan	24	Female	Kerala	Muslim	Bean
Bahuksana	23	Male	Karnataka	Buddhist	Lentil
Rambha	19	Male	Kerala	Christian	Peanut
Kishor	24	Male	Karnataka	Buddhist	Lentil
Johnson	17	Male	Kerala	Christian	Peanut
John	19	Male	Kerala	Christian	Pea

Q10. Confidentiality (Solution)

2-anonymity

- **Data is k-anonymity if each equivalence class contains at least k records.**
- **Equivalence class is a set of records that have the same values for the quasi-identifiers.**

Name	Age	Gender	State of domicile	Religion	Product
*	20 < Age ≤ 25	Female	Tamil Nadu	*	Pea
*	20 < Age ≤ 25	Female	Kerala	*	Bean
*	20 < Age ≤ 25	Female	Tamil Nadu	*	Peanut
*	20 < Age ≤ 25	Male	Karnataka	*	Pea
*	20 < Age ≤ 25	Female	Kerala	*	Bean
*	20 < Age ≤ 25	Male	Karnataka	*	Lentil
*	Age ≤ 20	Male	Kerala	*	Peanut
*	20 < Age ≤ 25	Male	Karnataka	*	Lentil
*	Age ≤ 20	Male	Kerala	*	Peanut
*	Age ≤ 20	Male	Kerala	*	Pea

2-anonymity, distinct 2-diversity

- **Data is distinct l-diversity if there are at least l distinct values for the sensitive attribute in each equivalence class.**

Name	Age	Gender	State of domicile	Religion	Product
*	20 < Age ≤ 25	Female	*	Hindu	Pea
*	20 < Age ≤ 25	Female	*	Hindu	Bean
*	20 < Age ≤ 25	Female	*	Muslim	Peanut
*	20 < Age ≤ 25	Male	*	Buddhist	Pea
*	20 < Age ≤ 25	Female	*	Muslim	Bean
*	20 < Age ≤ 25	Male	*	Buddhist	Lentil
*	Age ≤ 20	Male	*	Christian	Peanut
*	20 < Age ≤ 25	Male	*	Buddhist	Lentil
*	Age ≤ 20	Male	*	Christian	Peanut
*	Age ≤ 20	Male	*	Christian	Pea

Q11. Confidentiality (Your Turn)

What is the maximum l value for entropy l-diversity in the following table which has 2-anonimity?

Explicit identifier | Quasi-identifiers | Sensitive

Name	Age	Gender	State of domicile	Religion	Product
*	20 < Age ≤ 25	Female	*	Hindu	Pea
*	20 < Age ≤ 25	Female	*	Hindu	Bean
*	20 < Age ≤ 25	Female	*	Muslim	Peanut
*	20 < Age ≤ 25	Male	*	Buddhist	Pea
*	20 < Age ≤ 25	Female	*	Muslim	Bean
*	20 < Age ≤ 25	Male	*	Buddhist	Lentil
*	Age ≤ 20	Male	*	Christian	Peanut
*	20 < Age ≤ 25	Male	*	Buddhist	Lentil
*	Age ≤ 20	Male	*	Christian	Peanut
*	Age ≤ 20	Male	*	Christian	Pea

A table is said to have entropy l-diversity if for every equivalence class E, $Entropy(E) \geq log(l)$.

Q12. Confidentiality (Your Turn)

- **Assume the following list as the list of frequency of sensitive values in an equivalence class.**
  - **Does the corresponding equivalence class have recursive (1,2)-diversity?**
  - **Does the corresponding equivalence class have recursive (2,3)-diversity?**

- $Frequency\ list\ =\ (\boldsymbol{r_1 = 500, r_2 = 400, r_3 = 200, r_4 = 50, r_5 = 20})$

$$\boldsymbol{r_1 < c(r_l + r_{l+1} + \cdots + r_m)}$$

Q13. Confidentiality (Your Turn)

- **Consider "Age" and "Gender" as the quasi-identifiers:**
  - **Anatomize the following table with the minimum number of groups in order to have 2 distinct sensitive values in each group.**
  - **What is the response for the following query in the intermediate generalized table and in the anatomized tables?** $Count(Age = 40, Disease = 'Flu')$

Age	Gender	Disease
30	Female	Hepatitis
31	Female	Hepatitis
32	Female	HIV
35	Male	Hepatitis
38	Male	HIV
36	Male	HIV
42	Female	Flu
40	Female	Flu
43	Female	Heart
45	Female	Heart