

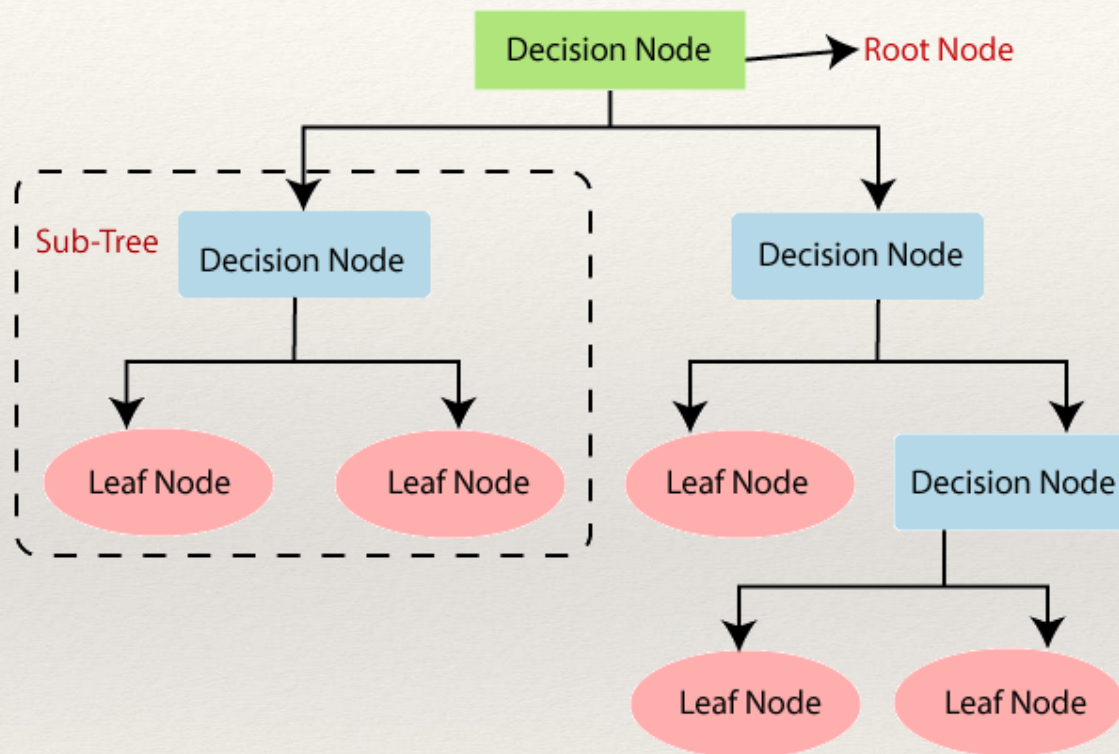
Machine learning

Entropy, Information Gain, Decision Trees & Accuracy

Exercise III

פיתוח:
ד"ר יהונתן שלר
משה פרידמן

עצי החלטה – משמעות וקריאת עץ החלטה



עץ החלטה: סדרת השאלות שמביאה אותנו להחלטה

צומת שורש: השאלה הראשונה בעץ ההחלטה

צמתי ביניים: שאלות המשך

צמתי עלים: ההחלטה המתקבלת

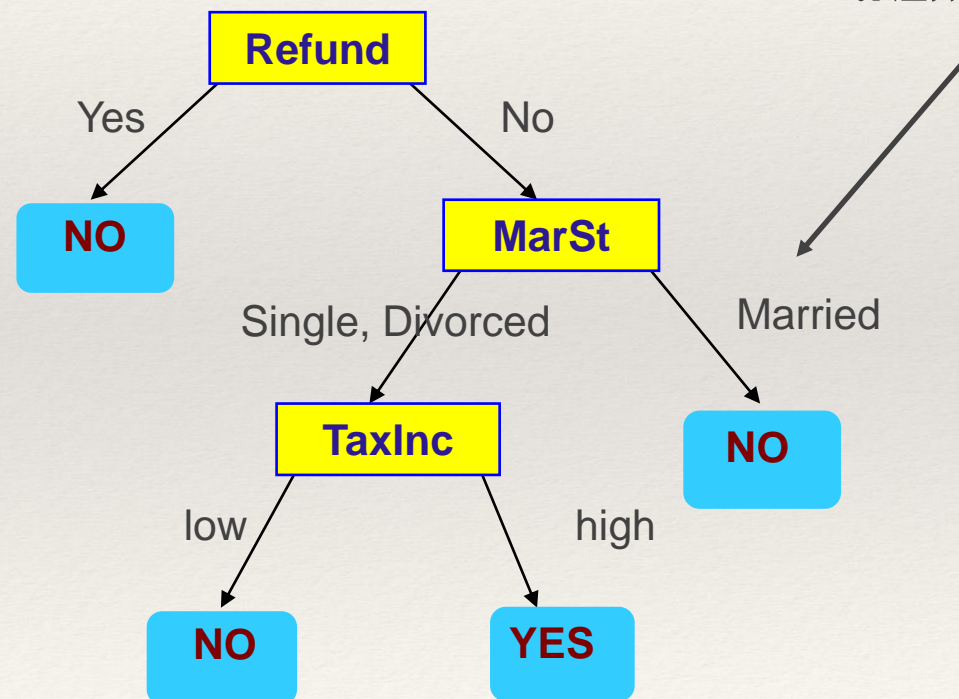
עצי החלטה – משמעות וקריאת עץ החלטה – תרגיל 1

categorical
categorical
categorical
class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	high	No
2	No	Married	high	No
3	No	Single	low	No
4	Yes	Married	high	No
5	No	Divorced	high	Yes
6	No	Married	low	No
7	Yes	Divorced	high	No
8	No	Single	high	Yes
9	No	Married	low	No
10	No	Single	high	Yes

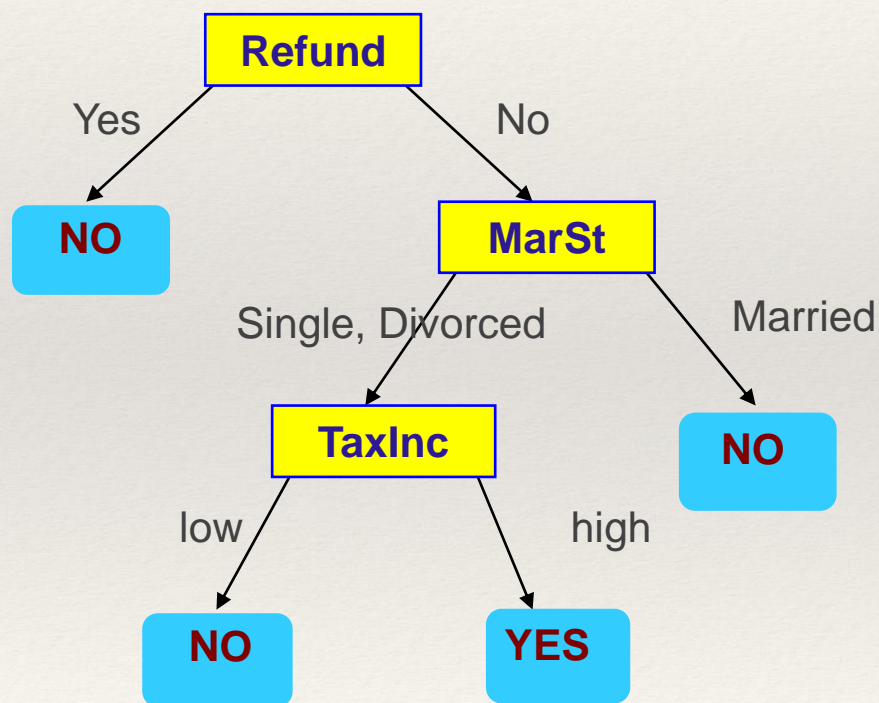
לאחר שלב האימון, של ה- trainset הנתון (מימין),

התקבל העץ הבא:



עצי החלטה – משמעות וקריאת עץ החלטה – תרגיל 1

מהו הסיווג של דוגמת ה-test-
הבאה?



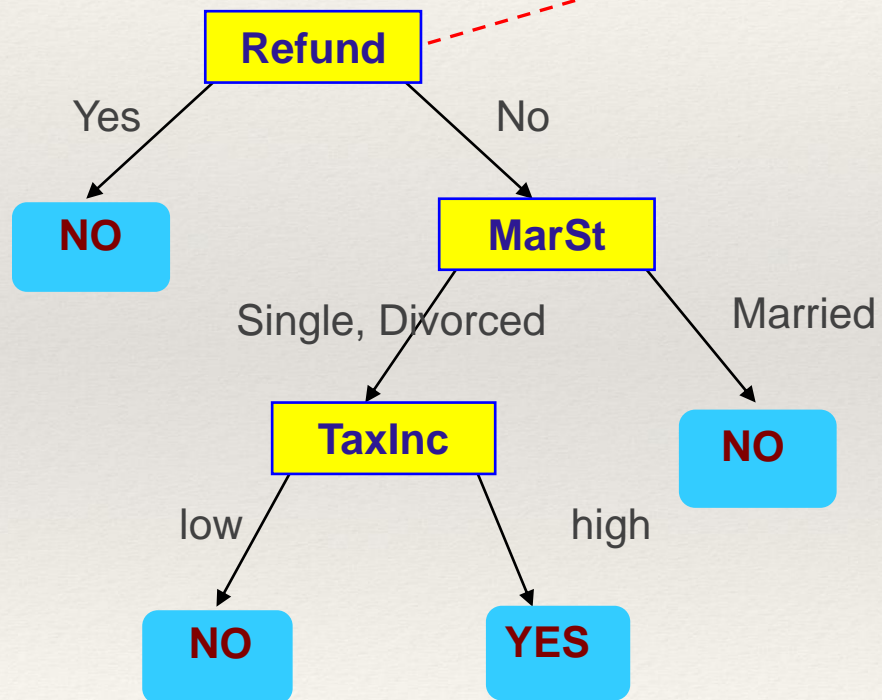
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	low	?

עצי החלטה – משמעות וקריאת עץ החלטה – תרגיל 1

Test Data

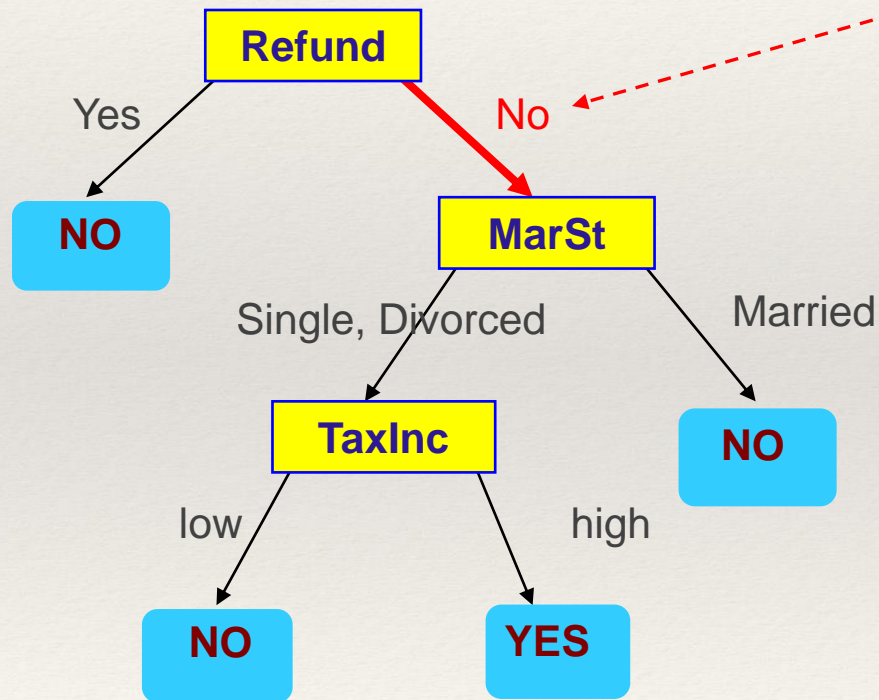
Refund	Marital Status	Taxable Income	Cheat
No	Married	low	?



עצי החלטה – משמעות וקריאת עץ החלטה – תרגיל 1

Test Data

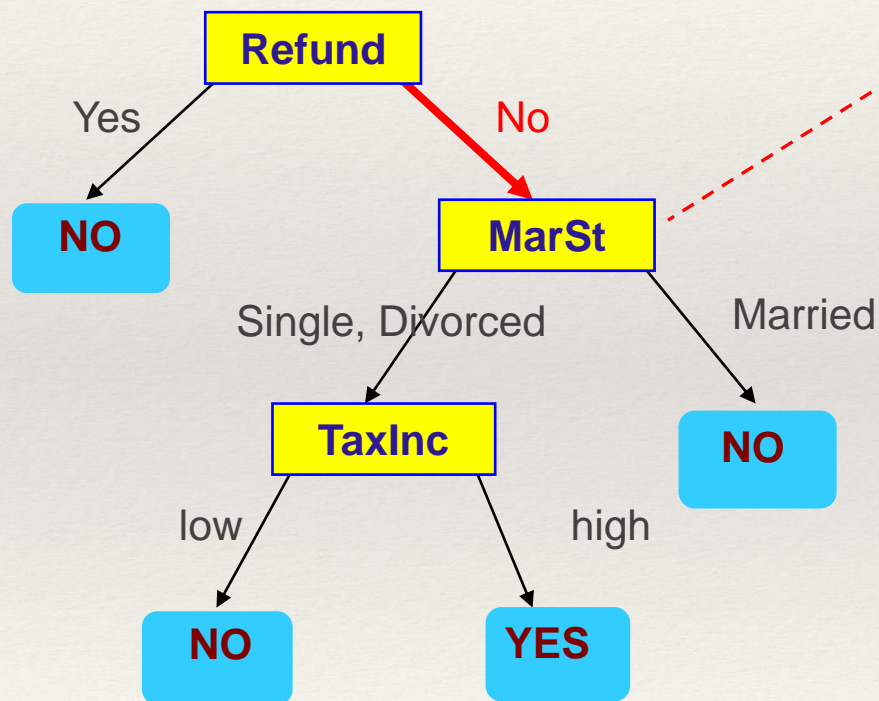
Refund	Marital Status	Taxable Income	Cheat
No	Married	low	?



עצי החלטה – משמעות וקריאת עץ החלטה – תרגיל 1

Test Data

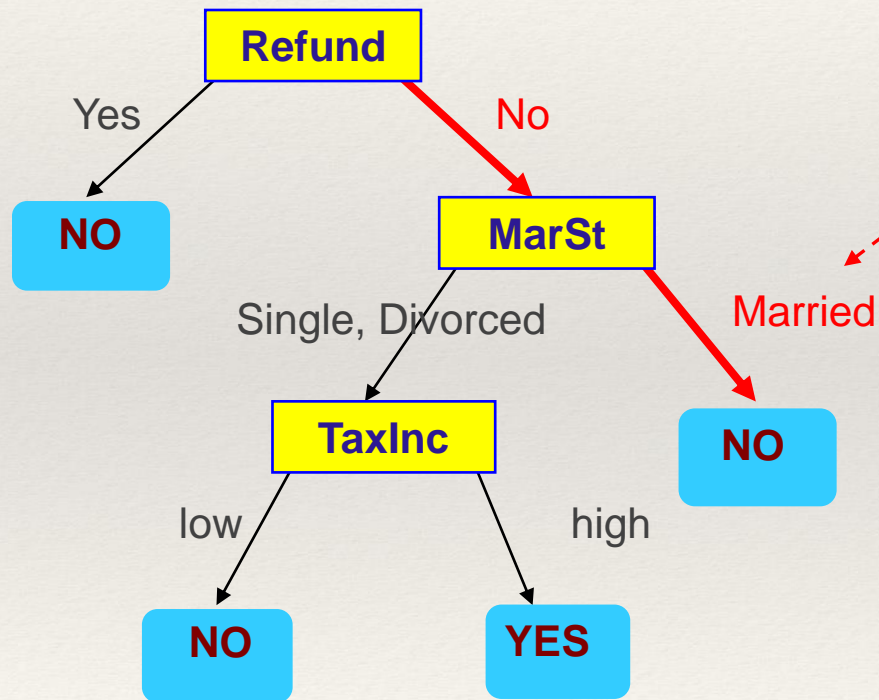
Refund	Marital Status	Taxable Income	Cheat
No	Married	low	?



עצי החלטה – משמעות וקריאת עץ החלטה – תרגיל 1

Test Data

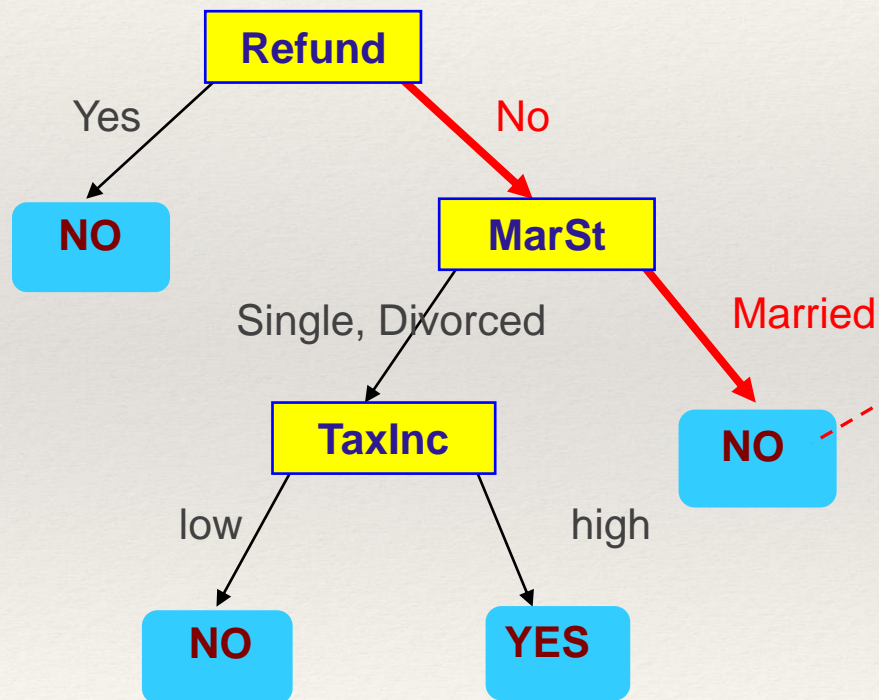
Refund	Marital Status	Taxable Income	Cheat
No	Married	low	?



עצי החלטה – משמעות וקריאת עץ החלטה – תרגיל 1

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	low	?



Assign Cheat to “No”

עץ החלטה: כיצד נבחר את התכונה הבאה

העיקרון הבסיסי

❖ ננסה לייצר את "המסלול" הקצר ביותר

❖ הרעיון: בכל רמה בעץ ננסה לשאול את השאלה שתשפר לנו בצורה הטובה ביותר את רמת הוודאות בחיזוי

נתרגל 2 פונקציות לבחירת מאפיין לצומת:

- אנטרופיה
- Information gain

אנטרופיה - entropy

❖ נתבונן במקרה הכללי בו נתונות לנו הסתברויות:

$$P(X=\alpha_1) = p_1, P(X=\alpha_2) = p_2, \dots, P(X=\alpha_n) = p_n$$

נגדיר $H(X)$ כאנטרופיה של X (Entropy)

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \dots - p_n \log_2 p_n$$

$$= -\sum_{j=1}^n p_j \log_2 p_j$$

אנטרופיה גבוהה, משמעותה, שההתפלגות יותר דומה לאחידה.

אנטרופיה נמוכה ואנטרופיה גבוהה

❖ אנטרופיה גבוהה

❖ התפלגות דומה לאחידה

❖ קשה לחזות

❖ אי וודאות

❖ רמת אי סדר גבוהה

❖ אנטרופיה נמוכה

❖ התפלגות מגוונת (בעלת צורה של גבעות ועמקים)

❖ יותר קל לחזות

❖ רמת וודאות גבוהה

❖ רמת אי סדר נמוכה

דוגמה 2א - חישוב אנטרופיה

נתונה הטבלה הבאה:

תרגיל: חשבו את האנטרופיות של X ו- Y

X "גיל"	Y "עובד בקורונה"
צעיר	"כן"
צעיר	"כן"
בינוני	"כן"
צעיר	"לא"
בינוני	"כן"
זקן	"לא"
צעיר	"לא"
זקן	"לא"

$$P(Y = \text{כן}) = 0.5$$

$$P(Y = \text{לא}) = 0.5$$

$$H(Y) = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = -0.5 \times (-1) - 0.5 \times (-1) = 1$$

$$P(X = \text{צעיר}) = 0.5$$

$$P(X = \text{בינוני}) = 0.25$$

$$P(X = \text{זקן}) = 0.25$$

$$H(X) = -0.5 \log_2 0.5 - 0.25 \log_2 0.25 - 0.25 \log_2 0.25 = -0.5 \times (-1) - 0.25 \times (-2) - 0.25 \times (-2) = 1.5$$

$$H(Y) = 1, \quad H(X) = 1.5$$

אנטרופיה מותנית

האנטרופיה מותנית $H(Y|X)$ הינה ממוצע משוקלל של האנטרופיות ה"ספציפיות" של Y

$$H(Y | X) = \sum_j P(X = \alpha_j) H(Y | X = \alpha_j)$$

דוגמה 2ב - אנטרופיה מותנית מסוימת – Specific Conditional Entropy

נגדיר "אנטרופיה מותנית מסוימת"

(Specific Conditional Entropy)

$$H(Y|X=\alpha)$$

כאנטרופיה של Y בין כל אותם הרשומות שבהן
 X מקבל את הערך α

חשבו $H(Y|X = \text{צעיר})$

X "גיל"	Y "עובד בקורונה"
צעיר	"כן"
צעיר	"כן"
בינוני	"כן"
צעיר	"לא"
בינוני	"כן"
זקן	"לא"
צעיר	"לא"
זקן	"לא"

$$P(Y = \text{כן} | X = \text{צעיר}) = 0.5$$

$$P(Y = \text{לא} | X = \text{צעיר}) = 0.5$$

$$H(Y|X = \text{צעיר}) = -0.5 \log 0.5 - 0.5 \log 0.5 = 1$$

דוגמה 2 - אנטרופיה מותנית מסוימת – Specific Conditional Entropy

חשבו את שאר האנטרופיות המותנות המסוימות

X	Y
"גיל"	"עובד בקורונה"
צעיר	"כן"
צעיר	"כן"
בינוני	"כן"
צעיר	"לא"
בינוני	"כן"
זקן	"לא"
צעיר	"לא"
זקן	"לא"

$$\begin{aligned}P(Y = \text{כן} \mid X = \text{בינוני}) &= 1 \\P(Y = \text{לא} \mid X = \text{בינוני}) &= 0 \\H(Y \mid X = \text{בינוני}) &= -1\log 1 = 0\end{aligned}$$

$$\begin{aligned}P(Y = \text{כן} \mid X = \text{זקן}) &= 0 \\P(Y = \text{לא} \mid X = \text{זקן}) &= 1 \\H(Y \mid X = \text{זקן}) &= -1\log 1 = 0\end{aligned}$$

$$\begin{aligned}P(Y = \text{כן} \mid X = \text{צעיר}) &= 0.5 \\P(Y = \text{לא} \mid X = \text{צעיר}) &= 0.5 \\H(Y \mid X = \text{צעיר}) &= -0.5\log 0.5 - 0.5\log 0.5 = 1\end{aligned}$$

$$H(Y \mid X = \text{צעיר}) = 1$$

$$H(Y \mid X = \text{בינוני}) = 0$$

$$H(Y \mid X = \text{זקן}) = 0$$

דוגמה 2 – אנטרופיה מותנית – Conditional Entropy

X	Y
"גיל"	"עובד בקורונה"
צעיר	"כן"
צעיר	"כן"
בינוני	"כן"
צעיר	"לא"
בינוני	"כן"
זקן	"לא"
צעיר	"לא"
זקן	"לא"

האנטרופיה המותנית $H(Y|X)$ הינה ממוצע משוקלל של האנטרופיות ה"ספציפיות" של Y

$$H(Y | X) = \sum_j P(X = \alpha_j) H(Y | X = \alpha_j)$$

$$H(Y|X=\text{צעיר}) = 1 \quad P(X=\text{צעיר}) = 0.5$$

$$H(Y|X=\text{בינוני}) = 0 \quad P(X=\text{בינוני}) = 0.25$$

$$H(Y|X=\text{זקן}) = 0 \quad P(X=\text{זקן}) = 0.25$$

חשבו את
 $H(Y|X)$

$$H(Y|X) = 0.5 \times 1 + 0.25 \times 0 + 0.25 \times 0 = 0.5$$

Information Gain

Gain($Y|X$) הינה ההפחתה הצפויה באנטרופיה של Y בגלל מיון עפ"י תכונה X ❖

$$Gain(Y | X) = H(Y) - H(Y | X)$$

תרגיל 3 – חשבו את $IG(\text{wealth} | \text{relation})$

wealth values: poor rich			
relation	Husband	10870	8846
	Not_in_family	11307	1276
	Other_relative	1454	52
	Own_child	7470	111
	Unmarried	4816	309
	Wife	1238	1093

נתונה הטבלה הבאה:

חשבו את $IG(\text{wealth} | \text{relation})$

תרגיל 3 – חשבו את $IG(wealth | relation)$ נחשב את ההסתברויות ...

wealth values: poor rich

relation	Husband	10870	8846		$H(wealth relation = Husband) = 0.992385$
	Not_in_family	11307	1276		$H(wealth relation = Not_in_family) = 0.473439$
	Other_relative	1454	52		$H(wealth relation = Other_relative) = 0.216617$
	Own_child	7470	111		$H(wealth relation = Own_child) = 0.110192$
	Unmarried	4816	309		$H(wealth relation = Unmarried) = 0.328606$
	Wife	1238	1093		$H(wealth relation = Wife) = 0.997207$
				$H(wealth) = 0.793844$	
				$H(wealth relation) = 0.628421$	
				$IG(wealth relation) = 0.165423$	

ראשית נחשב את ההסתברויות ...

$$total\ poor = 10870 + 11307 + 1454 + 7470 + 4816 + 1238 = 37,155$$

$$total\ rich = 8846 + 1276 + 52 + 111 + 309 + 1093 = 11,687$$

$$total = 11,687 + 37,155 = 48,842$$

$$p(wealth = poor) = \frac{37,155}{48,842} = 0.761 \quad p(wealth = rich) = \frac{11,687}{48,842} = 0.239$$

$$p(relation = husband) = \frac{10870 + 8846}{48842} = 0.403$$

$$p(relation = not_in_family) = \frac{11307 + 1276}{48842} = 0.257$$

$$p(relation = other_relative) = \frac{1454 + 52}{48842} = 0.030$$

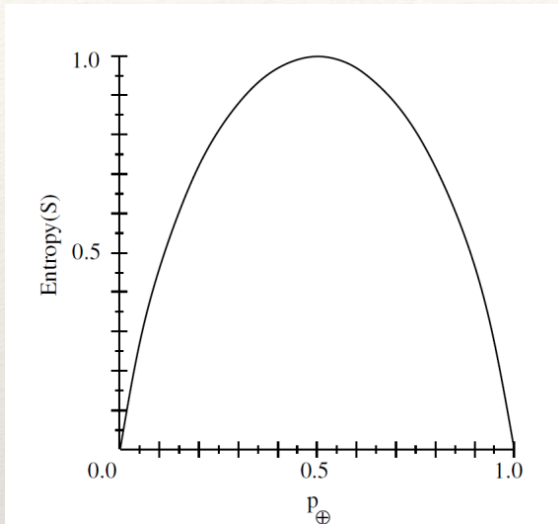
$$p(relation = own_child) = \frac{7470 + 111}{48842} = 0.155$$

$$p(relation = unmarried) = \frac{4816 + 309}{48842} = 0.104$$

$$p(relation = wife) = \frac{1238 + 1093}{48842} = 0.047$$

תרגיל 3 – חשבו את $IG(\text{wealth} | \text{relation})$ – כעת נחשב $H(\text{wealth})$

תזכורת - אנטרופיה של המחלקה (עבור 2 מחלקות אפשריות)



❖ S הינה קבוצה של דוגמאות

❖ P^+ הינו החלק היחסי של הדוגמאות החיוביות
בקבוצה

❖ P^- הינו החלק היחסי של הדוגמאות השליליות
בקבוצה

❖ האנטרופיה של S :

$$H(S) = -p^+ \log_2 p^+ - p^- \log_2 p^-$$

תרגיל 3 – חשבו את $IG(\text{wealth} | \text{relation})$

כעת נחשב $H(\text{wealth})$

wealth values: poor rich			
relation	Husband	10870	8846
	Not_in_family	11307	1276
	Other_relative	1454	52
	Own_child	7470	111
	Unmarried	4816	309
	Wife	1238	1093

$$p(\text{wealth} = \text{poor}) = 0.761$$

$$p(\text{wealth} = \text{rich}) = 0.239$$

$$\begin{aligned} H(\text{wealth}) &= -\frac{37,155}{48,842} \cdot \log_2 \left(\frac{37,155}{48,842} \right) - \frac{11,687}{48,842} \cdot \log_2 \left(\frac{11,687}{48,842} \right) = \\ &= -0.761 \cdot \log_2(0.761) - 0.239 \cdot \log_2(0.239) = \\ &= -0.761 \cdot -0.394 - 0.239 \cdot -2.063 = 0.3 + 0.494 = 0.794 \end{aligned}$$

תרגיל 3 – חשבו את $I_G(\text{wealth} | \text{relation})$ – כעת נחשב $H(\text{wealth} | \text{relation})$
תזכורת - אנטרופיה מותנית

אנטרופיה מותנית $H(Y|X)$ הינה ממוצע משוקלל
של האנטרופיות המותנות "ספציפיות" של Y

$$H(Y | X) = \sum_j P(X = \alpha_j) H(Y | X = \alpha_j)$$

תרגיל 3 – חשבו את $H(\text{wealth} | \text{relation})$ – נחשב $H(\text{wealth} | \text{relation})$

נחשב קודם את האנטרופיות המותנות המסוימות

wealth values: poor rich

relation	Husband	10870	8846	<div><div style="width: 55.1%;"></div></div>	$H(\text{wealth} \text{relation} = \text{Husband}) = 0.992385$
	Not_in_family	11307	1276	<div><div style="width: 9.0%;"></div></div>	$H(\text{wealth} \text{relation} = \text{Not_in_family}) = 0.473439$
	Other_relative	1454	52	<div><div style="width: 1.0%;"></div></div>	$H(\text{wealth} \text{relation} = \text{Other_relative}) = 0.216617$
	Own_child	7470	111	<div><div style="width: 0.9%;"></div></div>	$H(\text{wealth} \text{relation} = \text{Own_child}) = 0.110192$
	Unmarried	4816	309	<div><div style="width: 0.3%;"></div></div>	$H(\text{wealth} \text{relation} = \text{Unmarried}) = 0.328606$
	Wife	1238	1093	<div><div style="width: 53.1%;"></div></div>	$H(\text{wealth} \text{relation} = \text{Wife}) = 0.997207$
$H(\text{wealth}) = 0.793844$ $H(\text{wealth} \text{relation}) = 0.628421$					
$IG(\text{wealth} \text{relation}) = 0.165423$					

$H(\text{wealth} | \text{relation} = \text{husband})$

$$H(Y|X) = \sum_j P(X = \alpha_j) H(Y|X = \alpha_j)$$

Relation	
Poor	Rich
10,870	8,846
total	10,870+8,846 = 19,716
$P(\text{Poor} \text{relation} = \text{husband}) = 10,870/19,716 = 0.551$	$P(\text{Rich} \text{relation} = \text{husband}) = 8,846/19,716 = 0.448$
$\log_{10}(0.551) = -0.258$	$\log_{10}(0.448) = -0.348$
$\log_{10}(2) = 0.301$	$\log_{10}(2) = 0.301$
$\log_2(0.551) = -0.258/0.301 = -0.857$	$\log_2(0.448) = -0.348/0.301 = -1.156$
$H(\text{wealth} \text{relation} = \text{husband}) = -0.551 \cdot -0.857 - 0.448 \cdot -1.156 = 0.99$	

$$\log_a(x) = \frac{\log_b(x)}{\log_b(a)}$$

$$\log_2(x) = \frac{\log_{10}(x)}{\log_{10}(2)}$$

תרגיל 3 – חשבו את $I_G(\text{wealth} | \text{relation})$ – נחשב $H(\text{wealth} | \text{relation})$ כעת, נחשב את שאר האנטרופיות המותנות המסוימות







number of examples	Poor	Rich	total
relation=husband	8846	10870	19716
relation=not_in_family	1276	11307	12583
relation=other_relative	52	1454	1506
relation=own_child	111	7470	7581
relation=unmarried	309	4816	5125
relation=wife	1093	238	1331

conditional probabilities	pr(wealth=poor ...)	pr(wealth=rich ...)
relation=husband	0.44867113	0.55132887
relation=not_in_family	0.10140666	0.89859334
relation=other_relative	0.034528552	0.965471448
relation=own_child	0.014641868	0.985358132
relation=unmarried	0.060292683	0.939707317
relation=wife	0.821187077	0.178812923

specific conditional entropy	$H(\text{wealth} \dots)$
$H(\text{wealth} \text{relation}=\text{husband})$	0.99238459
$H(\text{wealth} \text{relation}=\text{not_in_family})$	0.47343881
$H(\text{wealth} \text{relation}=\text{other_relative})$	0.21661703
$H(\text{wealth} \text{relation}=\text{own_child})$	0.11019232
$H(\text{wealth} \text{relation}=\text{unmarried})$	0.32860567
$H(\text{wealth} \text{relation}=\text{wife})$	0.67747326

תרגיל 3 – חשבו את $H(\text{wealth} | \text{relation})$ – נחשב $H(\text{wealth} | \text{relation})$ כעת, נחשב האנטרופיה המותנת $H(\text{wealth} | \text{relation})$

כבר חישבנו:

wealth values: poor rich			
relation Husband	10870	8846	 $H(\text{wealth} \text{relation} = \text{Husband}) = 0.992385$
Not_in_family	11307	1276	 $H(\text{wealth} \text{relation} = \text{Not_in_family}) = 0.473439$
Other_relative	1454	52	 $H(\text{wealth} \text{relation} = \text{Other_relative}) = 0.216617$
Own_child	7470	111	 $H(\text{wealth} \text{relation} = \text{Own_child}) = 0.110192$
Unmarried	4816	309	 $H(\text{wealth} \text{relation} = \text{Unmarried}) = 0.328606$
Wife	1238	1093	 $H(\text{wealth} \text{relation} = \text{Wife}) = 0.997207$
$H(\text{wealth}) = 0.793844$ $H(\text{wealth} \text{relation}) = 0.628421$			
$IG(\text{wealth} \text{relation}) = 0.165423$			

$$p(\text{relation} = \text{husband}) = 0.403$$

$$p(\text{relation} = \text{not_in_family}) = 0.257$$

$$p(\text{relation} = \text{other_relative}) = 0.03$$

$$p(\text{relation} = \text{own_child}) = 0.155$$

$$p(\text{relation} = \text{unmarried}) = 0.104$$

$$p(\text{relation} = \text{wife}) = 0.047$$

כעת נוכל לחשב את האנטרופיה המותנית $H(\text{wealth} | \text{relation})$

$$0.403 \times 0.99 + 0.257 \times 0.473 + 0.03 \times 0.216 + 0.155 \times 0.11 + 0.104 \times 0.328 + 0.047 \times 0.677 = 0.559$$

specific conditional entropy	$H(\text{wealth} \dots)$
$H(\text{wealth} \text{relation} = \text{husband})$	0.99238459
$H(\text{wealth} \text{relation} = \text{not_in_family})$	0.47343881
$H(\text{wealth} \text{relation} = \text{other_relative})$	0.21661703
$H(\text{wealth} \text{relation} = \text{own_child})$	0.11019232
$H(\text{wealth} \text{relation} = \text{unmarried})$	0.32860567
$H(\text{wealth} \text{relation} = \text{wife})$	0.67747326

תרגיל 3 – חשבו את $IG(\text{wealth} | \text{relation})$ – כעת נחשב $H(\text{wealth})$ תזכורת - Information Gain







Information Gain($Y | X$)

הינה ההפחתה הצפויה באנטרופיה של Y בגלל מיון עפ"י תכונה X

$$Gain(Y|X) = H(Y) - H(Y|X)$$

תרגיל 3 – חשבו את $IG(\text{wealth} | \text{relation})$ – נחשב $H(\text{wealth} | \text{relation})$ בעת, נחשב את $IG(\text{wealth} | \text{relation})$

wealth values: poor rich

relation	Husband	10870	8846		$H(\text{wealth} \text{relation} = \text{Husband}) = 0.992385$
	Not_in_family	11307	1276		$H(\text{wealth} \text{relation} = \text{Not_in_family}) = 0.473439$
	Other_relative	1454	52		$H(\text{wealth} \text{relation} = \text{Other_relative}) = 0.216617$
	Own_child	7470	111		$H(\text{wealth} \text{relation} = \text{Own_child}) = 0.110192$
	Unmarried	4816	309		$H(\text{wealth} \text{relation} = \text{Unmarried}) = 0.328606$
	Wife	1238	1093		$H(\text{wealth} \text{relation} = \text{Wife}) = 0.997207$
$H(\text{wealth}) = 0.793844$					
$H(\text{wealth} \text{relation}) = 0.628421$					
$IG(\text{wealth} \text{relation}) = 0.165423$					

$$H(\text{wealth}) = 0.794$$

$$H(\text{wealth} | \text{relation}) = 0.559$$

כבר חישבנו:

$$Gain(Y | X) = H(Y) - H(Y | X)$$

בעת נוכל לחשב את $IG(\text{wealth} | \text{relation})$

$$H(\text{wealth}) - H(\text{wealth} | \text{relation}) = 0.794 - 0.559 = 0.235$$

תרגיל 4 – חיזוי מוצלח

א. היכן תעדיפו לשחק? במשחק חזרתי של הטלת מטבע או הטלת קוביה?

האפשרות המועדפת
(מדוע?)



זריקת קוביה האנטרופיה

$$6 \cdot \left(-\frac{1}{6} \cdot \log_2 \frac{1}{6}\right) = 2.58$$

הטלת מטבע האנטרופיה

$$2 \cdot (-0.5 \cdot \log_2 0.5) = 1$$

ב. והיכן תעדיפו לשחק? במשחק חזרתי של חיזוי תוצאת סכום של שתי הטלות קוביה

או

בחירת מס' מבין 9 אפשרויות בהסתברות שווה

תרגיל 14 - סכום 2 הטלות או בחירה בין 9 אפשרויות

prob	sum
1/36	2
2/36	3
3/36	4
4/36	5
5/36	6
6/36	7
5/36	8
4/36	9
3/36	10
2/36	11
1/36	12

סכום 2 הטלות:

$$\begin{aligned} & -2 \times \frac{1}{36} \log_2 \frac{1}{36} - 2 \times \frac{2}{36} \log_2 \frac{2}{36} \\ & - 2 \times \frac{3}{36} \log_2 \frac{3}{36} - 2 \times \frac{4}{36} \log_2 \frac{4}{36} \\ & - 2 \times \frac{5}{36} \log_2 \frac{5}{36} - \frac{6}{36} \log_2 \frac{6}{36} = 3.2 \end{aligned}$$

בחירה בין 9 אפשרויות

$$9 \times \left(-\frac{1}{9} \log_2 \frac{1}{9} \right) = 3.17$$

האפשרות המועדפת
(מדוע?)



בניית עצי החלטה (אלגוריתם ID3)

לולאה:

- ❖ מצא את המאפיין הטוב ביותר X_i ושים אותו בצומת.
- ❖ עבור כל ערך של X_i צור קשת לכל אפשרות וצומת היוצאת ממנה
- ❖ מיין את הדוגמאות ב-train-set לצמתים החדשים (מהשלב הקודם)
- ❖ אם שילוב המאפיין והערך שלו, מוביל להחלטה טובה מספיק, צומת זה הוא עלה ומייצג החלטה.
- ❖ אם לא נוכל לקבל "החלטה טובה", נשאיר צומת זה כעלה
- ❖ אחרת, בצע את אותו תהליך עבור הצומת הזה

תרגיל 5 - בחירת התכונה לצומת

סיווג	גיל גדול מ-30	צבע
YES	כן	שחור
NO	לא	לבן
NO	לא	צהוב
YES	כן	שחור
YES	כן	צהוב
YES	כן	לבן
YES	כן	צהוב
NO	לא	שחור
NO	לא	שחור
YES	כן	לבן
YES	כן	לבן
NO	לא	שחור
YES	כן	צהוב
NO	לא	שחור
YES	כן	צהוב
NO	לא	שחור
YES	כן	צהוב
YES	כן	לבן
YES	כן	לבן
NO	לא	צהוב
NO	לא	צהוב

נתון ה-train-set הבא,
עם 2 מאפיינים:
צבע (שחור/לבן/צהוב)
גיל (גדול מ-30/קטן או
שווה ל-30)
וקטגוריה: סיווג
(Yes/No)

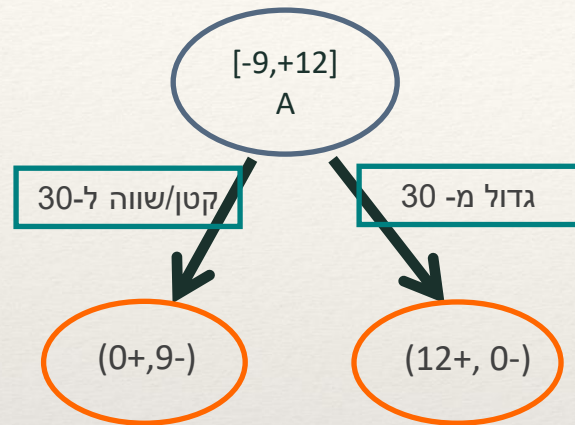
תרגיל 5 - בחירת התכונה לצומת

- ❖ נתונה קבוצת האימון בה 12 דוגמאות מסווגות כחיובי ו-9 דוגמאות כשלילי.
- ❖ נתונה תכונה A (מייצגת "גיל") לה שני פיצולים אפשריים:
- ❖ "אם גדול מ-30" – הסיווג הוא YES, "אם קטן/שווה ל-30" הסיווג הוא NO.
- ❖ נסמן זאת $(0+, 9-)$ ו- $(12+, 0-)$
- ❖ נתונה תכונה B ("צבע") לה שלושה פיצולים אפשריים:
- ❖ "אם שחור – אנו נשארים עם קבוצה של 2 חיוביים וחמישה שליליים",
- ❖ "אם לבן – נשארים עם קבוצה של 5 חיוביים ואחד שלילי",
- ❖ "אם צהוב – נשארים עם קבוצה של חמישה חיוביים ושלושה שליליים"
- ❖ $(2+, 5-)$, $(5+, 1-)$ ו- $(5+, 3-)$

איזו תכונה עדיפה כצומת הבא בעץ?

תרגיל 5 - בחירת התכונה לצומת – פתרון בעזרת IG

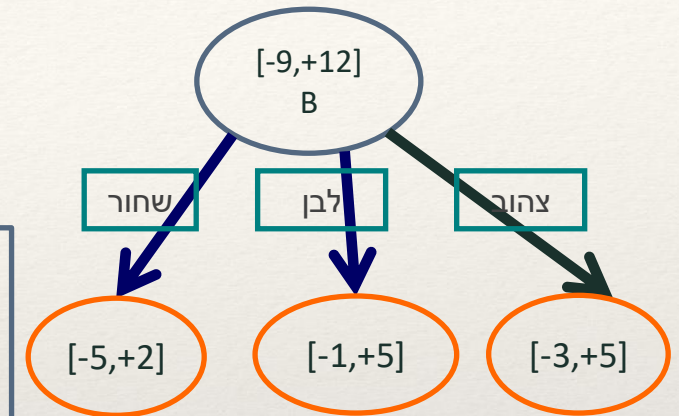
נחשב את ההסתברויות ...



$$\begin{aligned} p(\text{סיווג} = \text{Yes} | \text{age} \leq 30) &= \frac{0}{9} = 0 \\ p(\text{סיווג} = \text{No} | \text{age} \leq 30) &= \frac{9}{9} = 1 \\ p(\text{סיווג} = \text{Yes} | \text{age} > 30) &= \frac{12}{12} = 1 \\ p(\text{סיווג} = \text{No} | \text{age} > 30) &= \frac{0}{12} = 0 \end{aligned}$$

ראשית נחשב הסתברויות ...

$$\begin{aligned} p(\text{סיווג} = \text{Yes} | \text{color} = \text{black}) &= \frac{2}{7} \approx 0.286 \\ p(\text{סיווג} = \text{No} | \text{color} = \text{black}) &= \frac{5}{7} \approx 0.714 \\ p(\text{סיווג} = \text{Yes} | \text{color} = \text{white}) &= \frac{5}{6} \approx 0.833 \\ p(\text{סיווג} = \text{No} | \text{color} = \text{white}) &= \frac{1}{6} \approx 0.166 \\ p(\text{סיווג} = \text{Yes} | \text{color} = \text{yellow}) &= \frac{5}{8} = 0.625 \\ p(\text{סיווג} = \text{No} | \text{color} = \text{yellow}) &= \frac{3}{8} = 0.375 \end{aligned}$$



$$\begin{aligned} p(\text{סיווג} = \text{Yes}) &= \frac{12}{21} \approx 0.57 \\ p(\text{סיווג} = \text{No}) &= \frac{9}{21} \approx 0.43 \end{aligned}$$

תרגיל 5 - בחירת התכונה לצומת – פתרון בעזרת IG

נחשב את האנטרופיה של המחלקה ...

חישבנו:

$$p(\text{סיווג} = \text{Yes}) \approx 0.57$$

$$p(\text{סיווג} = \text{No}) \approx 0.43$$

נחשב את האנטרופיה של המחלקה ...

$$\begin{aligned} H(Y) &\approx -0.57 \cdot \log_2(0.57) - 0.43 \cdot \log_2(0.43) = \\ &= -0.57 \cdot (-0.807) - 0.43 \cdot (-1.22) = 0.4611 + 0.5228 = \mathbf{0.98} \end{aligned}$$

תרגיל 5 - בחירת התכונה לצומת – פתרון בעזרת IG

חישוב:

$$p(\text{סיווג} = \text{Yes} | \text{age} \leq 30) = 0$$

$$p(\text{סיווג} = \text{No} | \text{age} \leq 30) = 1$$

$$p(\text{סיווג} = \text{Yes} | \text{age} > 30) = 1$$

$$p(\text{סיווג} = \text{No} | \text{age} > 30) = 0$$

$$p(\text{סיווג} = \text{No} | \text{color} = \text{white}) \approx 0.166$$

$$p(\text{סיווג} = \text{No} | \text{color} = \text{yellow}) = 0.625$$

$$p(\text{סיווג} = \text{No} | \text{color} = \text{yellow}) = 0.375$$

$$p(\text{סיווג} = \text{Yes} | \text{color} = \text{black}) \approx 0.286$$

$$p(\text{סיווג} = \text{No} | \text{color} = \text{black}) \approx 0.714$$

$$p(\text{סיווג} = \text{Yes} | \text{color} = \text{white}) \approx 0.833$$

נחשב את האנטרופיה המותנת $H(Y|B)$...

$$H(Y|B = \text{black}) \approx -0.286 \cdot \log_2(0.286) - 0.714 \cdot \log_2(0.714) \approx \mathbf{1.002}$$

$$H(Y|B = \text{white}) \approx -0.833 \cdot \log_2(0.833) - 0.166 \cdot \log_2(0.166) \approx \mathbf{0.645}$$

$$H(Y|B = \text{yellow}) \approx -0.625 \cdot \log_2(0.625) - 0.375 \cdot \log_2(0.375) \approx \mathbf{0.954}$$

$$P(B = \text{black}) = \frac{7}{21} \approx 0.333 \quad P(B = \text{white}) = \frac{6}{21} \approx 0.286 \quad P(B = \text{yellow}) = \frac{8}{21} \approx 0.381$$

$$H(Y|B) \approx 0.333 \cdot 1.002 + 0.286 \cdot 0.645 + 0.381 \cdot 0.954 \approx \mathbf{0.882}$$

נחשב את האנטרופיה המותנת $H(Y|A)$...

$$H(Y|A > 30) = -1 \cdot \log_2(1) - 0 \cdot \log_2(0) = \mathbf{0}$$

$$H(Y|A \leq 30) = -0 \cdot \log_2(0) - 1 \cdot \log_2(1) = \mathbf{0}$$

$$H(Y|A) = p(A \leq 30) \cdot 0 + p(A > 30) \cdot 0 = \mathbf{0}$$

תרגיל 5 - בחירת התכונה לצומת – פתרון בעזרת IG

חישבנו:

$$H(Y|A) = 0$$

$$H(Y|B) \approx 0.882$$

$$H(Y) = 0.98$$

נחשב את האנטרופיה המותנת $IG(Y|B)$...

$$IG(Y|B) = H(Y) - H(Y|B) \approx 0.98 - 0.882 = \mathbf{0.098}$$

נחשב את האנטרופיה המותנת $IG(Y|A)$...

$$IG(Y|A) = H(Y) - H(Y|A) = \mathbf{0.98}$$

ולכן נבחר ב-A כתכונה בצומת הבא

Confusion matrix:

	Predicted Yes	Predicted No
Actual Yes	True Positive (TP)	False Negative (FN)
Actual No	False Positive (FP)	True Negative (TN)

$$\text{accuracy} = \frac{\#correct\ predictions = \#TP + \#TN}{\#test\ instances = \#TP + \#TN + \#FP + \#FN}$$

$$\text{Error (rate)} = 1 - \text{accuracy} = \frac{\#incorrect\ predictions = \#FP + \#FN}{\#test\ instances = \#TP + \#TN + \#FP + \#FN}$$

תרגיל 6 - שיערוך – קורונה

בדיקה חדשה לגילוי קורונה נוסתה על 500 איש. מתוכם
400 בריאים ו 100 חולים – והתקבלו הנתונים הבאים:

Calculate the accuracy ...

	סווג כחולה	סווג כלא חולה
חולה בפועל	95	5
לא חולה בפועל	15	385

תרגיל 6 - שיערוך – קורונה

בדיקה חדשה לגילוי קורונה נוסתה על 500 איש. מתוכם 400 בריאים ו 100 חולים – והתקבלו הנתונים הבאים:

Calculate the accuracy ...

Accuracy =

$$\frac{(95+385)}{(95+5+15+385)} = \frac{480}{(500)} \sim 0.96$$

	סוג בחולה	סוג כלא חולה
חולה בפועל	95	5
לא חולה בפועל	15	385

תרגיל 7 – שיערוך – The Titanic dataset

Titanic - British passenger liner that sank in the North Atlantic Ocean in 1912 after striking an iceberg



Dataset - used to predict passenger survival status.

Calculate the accuracy ...

	סווג כשרד	סווג כלא שרד
שרד בפועל	150	30
לא שרד בפועל	15	300

תרגיל 7 – שיערוך – The Titanic dataset

Titanic - British passenger liner that sank in the North Atlantic Ocean in 1912 after striking an iceberg



Dataset - used to predict passenger survival status.

Accuracy =

$$\frac{(150+300)}{(150+300+30+15)} = \frac{450}{(495)} \sim 0.91$$

	סווג כשרד	סווג כלא שרד
שרד בפועל	150	30
לא שרד בפועל	15	300

שאלות?

נראה בשבוע הבא 😊