Question 2

a)

Data breakdown:

Forecast: 5x Sunny, 4x Overcast, 5x Rain

Temp: 4x Hot, 6x Mild, 4x Cool

Humid: 7x High, 7x Normal

~~Windy: 6x True, 8x False~~

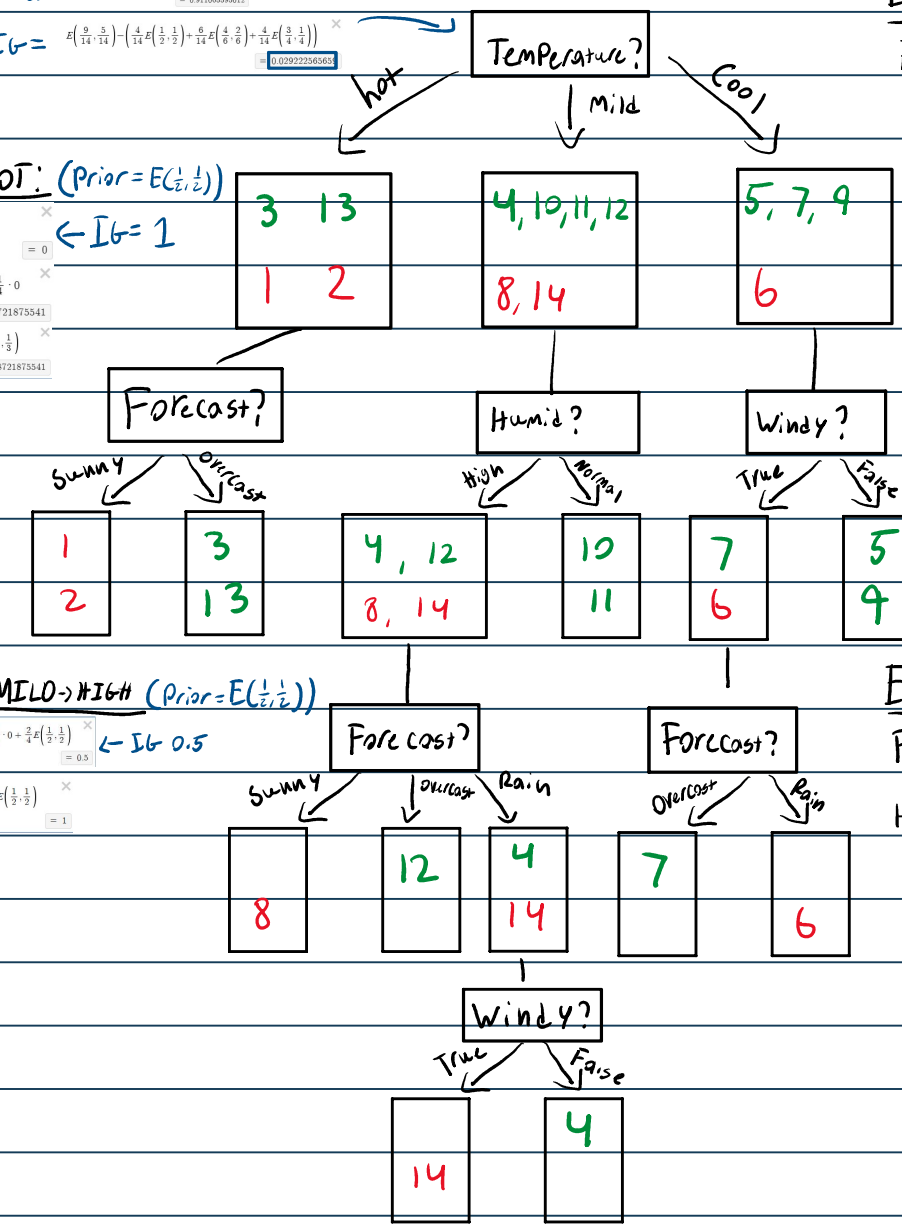
Picnic: 9x Yes, 5x No

#	Forecast	Temperature	Humidity	Windy	Picnic?
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rain	mild	high	false	yes
5	rain	cool	normal	false	yes
6	rain	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rain	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rain	mild	high	true	no

Initial Entropy $E\left(\frac{9}{14}, \frac{5}{14}\right) = 0.940285958671$

$E_{new} = \left(\frac{4}{14}E\left(\frac{1}{2},\frac{1}{2}\right)+\frac{6}{14}E\left(\frac{4}{6},\frac{2}{6}\right)+\frac{4}{14}E\left(\frac{3}{4},\frac{1}{4}\right)\right) = 0.911063393012$

$IG = E\left(\frac{9}{14},\frac{5}{14}\right)-\left(\frac{4}{14}E\left(\frac{1}{2},\frac{1}{2}\right)+\frac{6}{14}E\left(\frac{4}{6},\frac{2}{6}\right)+\frac{4}{14}E\left(\frac{3}{4},\frac{1}{4}\right)\right) = 0.029222565659$



Entropy New HOT: (Prior = $E(\frac{1}{2},\frac{1}{2})$)

Forecast: $\frac{1}{2}\cdot 0+\frac{1}{2}\cdot 0 = 0$ ← IG = 1

Humid: $\frac{3}{4}E\left(\frac{1}{3},\frac{2}{3}\right)+\frac{1}{4}\cdot 0 = 0.688721875541$

Windy: $\frac{1}{4}\cdot 0+\frac{3}{4}\cdot E\left(\frac{2}{3},\frac{1}{3}\right) = 0.688721875541$

Entropy New MILD: (Prior = $E(\frac{2}{3},\frac{1}{3})$)

Forecast: $\frac{3}{6}E\left(\frac{2}{3},\frac{1}{3}\right)+\frac{2}{6}E\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{6}\cdot 0 = 0.792481250361$

Humid: $\frac{4}{6}E\left(\frac{1}{2},\frac{1}{2}\right)+\frac{2}{6}\cdot 0 = 0.666666666667$ ← IG = 0.25

Windy: $\frac{3}{6}E\left(\frac{2}{3},\frac{1}{3}\right)+\frac{3}{6}E\left(\frac{2}{3},\frac{1}{3}\right) = 0.918295834054$

Entropy New COOL (Prior = $E(\frac{3}{4},\frac{1}{4})$)

Forecast: $\frac{2}{4}E\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{4}\cdot 0+\frac{1}{4}\cdot 0 = 0.5$

Humid: $\frac{4}{4}E\left(\frac{3}{4},\frac{1}{4}\right) = 0.811278124459$

Windy: $\frac{2}{4}\cdot E\left(\frac{1}{2},\frac{1}{2}\right)+\frac{2}{4}\cdot 0 = 0.5$ ← IG = 0.31

Chosen over Forecast since we had 2 false in 1 category instead of true split by 1 instance

Entropy New MILD→HIGH (Prior = $E(\frac{1}{2},\frac{1}{2})$)

Forecast: $\frac{1}{4}\cdot 0+\frac{1}{4}\cdot 0+\frac{2}{4}E\left(\frac{1}{2},\frac{1}{2}\right) = 0.5$ ← IG 0.5

Windy: $\frac{2}{4}E\left(\frac{1}{2},\frac{1}{2}\right)+\frac{2}{4}E\left(\frac{1}{2},\frac{1}{2}\right) = 1$

Entropy New COOL→TRUE (Prior = $E(\frac{1}{2},\frac{1}{2})$)

Forecast: $\frac{1}{2}\cdot 0+\frac{1}{2}\cdot 0 = 0$ ← IG = 1

Humid: $\frac{2}{2}E\left(\frac{1}{2},\frac{1}{2}\right) = 1$

Question 2 Pg.2

b) Init Entropy: $E\left(\frac{9}{14}, \frac{5}{14}\right) = 0.940285958671$

#	Forecast	Temperature	Humidity	Windy	Picnic?
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rain	mild	high	false	yes
5	rain	cool	normal	false	yes
6	rain	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rain	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rain	mild	high	true	no

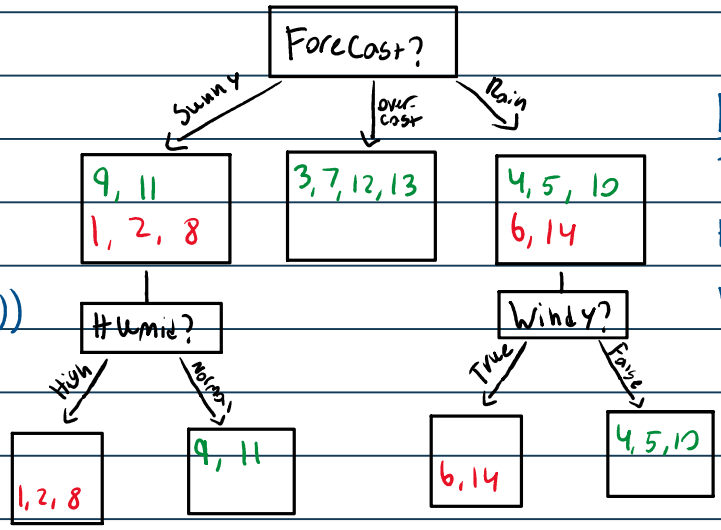
New Entropies

Forecast: $\left(\frac{5}{14}E\left(\frac{2}{5}, \frac{3}{5}\right) + \frac{4}{14}\cdot 0 + \frac{5}{14}E\left(\frac{3}{5}, \frac{2}{5}\right)\right) = 0.693536138896$ ← IG = 0.25

Temp: $\left(\frac{4}{14}E\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{6}{14}E\left(\frac{4}{6}, \frac{2}{6}\right) + \frac{4}{14}E\left(\frac{3}{4}, \frac{1}{4}\right)\right) = 0.911063393012$

Humid: $\frac{7}{14}E\left(\frac{3}{7}, \frac{4}{7}\right) + \frac{7}{14}E\left(\frac{6}{7}, \frac{1}{7}\right) = 0.788450457308$

Windy: $\frac{6}{14}E\left(\frac{3}{6}, \frac{3}{6}\right) + \frac{8}{14}E\left(\frac{6}{8}, \frac{2}{8}\right) = 0.892158928262$



Entropy SUNNY → ... (Prior = $E\left(\frac{2}{5}, \frac{3}{5}\right)$)

Temp: $\frac{2}{5}\cdot 0 + \frac{2}{5}E\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{5}\cdot 0 = 0.4$

Humid: $\frac{3}{5}\cdot 0 + \frac{2}{5}\cdot 0 = 0$ ← IG = 0.97

Windy: $\frac{2}{5}E\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{3}{5}E\left(\frac{1}{3}, \frac{2}{3}\right) = 0.950977500433$

Entropy RAIN → ... (Prior = $E\left(\frac{3}{5}, \frac{2}{5}\right)$)

Temp: $\frac{3}{5}E\left(\frac{2}{3}, \frac{1}{3}\right) + \frac{2}{5}E\left(\frac{1}{2}, \frac{1}{2}\right) = 0.950977500433$

Humid: $\frac{2}{5}E\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{3}{5}E\left(\frac{2}{3}, \frac{1}{3}\right) = 0.950977500433$

Windy: $\frac{2}{5}\cdot 0 + \frac{3}{5}\cdot 0 = 0$ ← IG = 0.97

Question 3

3. (20 points) Given the following dataset containing three attributes and one class, use a Naïve Bayes classifier to determine the class (Yes/No) of Stolen for a *Red Domestic SUV*. Note: You do not need to write/use code for this, just use the table to calculate probabilities.

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Domestic	No
10	Red	Sports	Imported	Yes

We set-up the problem and then calculate the individual components from the data:

Note: I define "case" = Red Domestic SUV to reduce writing

I start by writing our data in a more useful form:

$P(Y = Yes) = 5/10$

$P(Y = No) = 5/10$

Color	Yes	No
Red	3/5	2/5
Yellow	2/5	3/5

Type	Yes	No
Sports	4/5	2/5
SUV	1/5	3/5

Origin	Yes	No
Dom	2/5	4/5
Imp	3/5	1/5

Now we can calculate the probabilities:

$P(Yes \mid case) = P(Yes)\,P(Red \mid Yes)\,P(Dom \mid Yes)\,P(SUV \mid Yes)$

$= (5/10)(3/5)(2/5)(1/5) = 0.024$

$P(No \mid case) = P(No)\,P(Red \mid No)\,P(Dom \mid No)\,P(SUV \mid No)$

$= (5/10)(2/5)(4/5)(3/5) = 0.096$

We use the values to get our relative % probabilities:

$$P(Yes \mid case) = \frac{0.024}{0.024 + 0.096} = 20\%$$

$$P(No \mid case) = \frac{0.096}{0.024 + 0.096} = 80\%$$

Since "not stolen" has the highest probability we classify the case as that

Question 4

Short Text A: the carbon atom is the foundation of life on earth

c	Physics	Biology	Chemistry
$p(c)$	0.35	0.40	0.25
$p(\text{atom}\|c)$	0.1	0.01	0.2
$p(\text{carbon}\|c)$	0.005	0.03	0.05
$p(\text{proton}\|c)$	0.05	0.001	0.05
$p(\text{life}\|c)$	0.001	0.1	0.008
$p(\text{earth}\|c)$	0.005	0.006	0.003

Atom = True
Carbon = True
~~Proton = False~~
Life = True
Earth = True

$P(y=\text{Physics}\,|\,1,1,0,1,1) \propto (0.35)(0.1)(0.005)(0.95)(0.001)(0.005) = 8.31 \cdot 10^{-12}$

$P(y=\text{Biology}\,|\,1,1,0,1,1) \propto (0.40)(0.01)(0.03)(0.999)(0.1)(0.006) = \boxed{7.19 \cdot 10^{-8}}$ ← Highest

$P(y=\text{Chem}\,|\,1,1,0,1,1) \propto (0.25)(0.2)(0.05)(0.95)(0.008)(0.003) = 5.70 \cdot 10^{-8}$

$\Rightarrow$ **Biology**

Short Text B: the carbon atom contains 12 protons

Atom = True
Carbon = True
~~Proton = True~~
Life = False
Earth = False

$P(y=\text{Physics}\,|\,1,1,1,0,0) \propto (0.35)(0.1)(0.005)(0.05)(0.999)(0.995) = 8.70 \cdot 10^{-6}$

$P(y=\text{Biology}\,|\,1,1,1,0,0) \propto (0.40)(0.01)(0.03)(0.001)(0.9)(0.994) = 1.07 \cdot 10^{-7}$

$P(y=\text{Chem}\,|\,1,1,1,0,0) \propto (0.25)(0.2)(0.05)(0.05)(0.992)(0.997) = \boxed{1.24 \cdot 10^{-4}}$ ← Highest

$\Rightarrow$ **Chemistry**