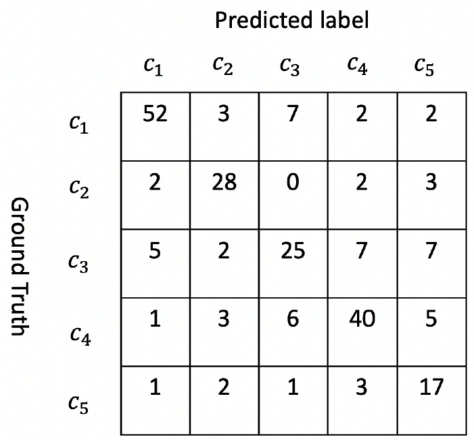
CS 5402 - HW 2
Will Weidler
2/26/25

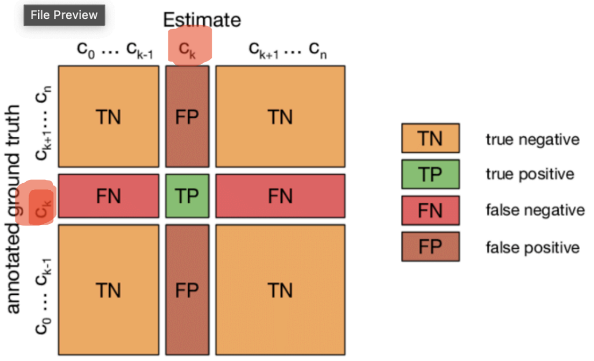
Q1. Given the confusion matrix for 5 classes as below:



		Predicted label				
		c_1	c_2	c_3	c_4	c_5
Ground Truth	c_1	52	3	7	2	2
	c_2	2	28	0	2	3
	c_3	5	2	25	7	7
	c_4	1	3	6	40	5
	c_5	1	2	1	3	17

What is the Accuracy and F1-score for class c_1, c_2, c_3, c_4, c_5?

NOTES:



○ Recall = $\frac{TP}{FN+TP}$

○ Precision = $\frac{TP}{FP+TP}$

○ F-Measure = $2 * \frac{Precision*Recall}{Precision+Recall}$

○ Accuracy = $\frac{TN+TP}{TN+FP+FN+TP}$

ANSWERS:

c_1:

TN = 28+0+2+3+2+25+7+7+3+6+40+5+2+1+3+17 = 151

FN = 3+7+2+2 = 14

FP = 2+5+1+1 = 9

TP = 52

Accuracy = $\frac{151+52}{151+9+14+52}$ **= 0.8982300885**

Recall = $\frac{52}{14+52}$ = 0.7878787879

Precision = $\frac{52}{9+52}$ = 0.8524590164

F-1 Score = $2 * \frac{0.8524590164*0.7878787879}{0.8524590164+0.7878787879}$ **= 0.8188976378**

c_2:

TN = 52+7+2+2+5+25+7+7+1+6+40+5+1+1+3+17 = 181

FN = 2+0+2+3 = 7

FP = 3+2+3+2 = 10

TP = 28

Accuracy = $\frac{181+28}{181+10+7+28}$ **= 0.9247787611**

Recall = $\frac{28}{7+28}$ = 0.8

Precision = $\frac{28}{10+28}$ = 0.7368421053

F-1 Score = $2 * \frac{0.7368421053*0.8}{0.7368421053+0.8}$ **= 0.7671232877**

c_3:

TN = 52+3+2+2+2+28+2+3+1+3+40+5+1+2+3+17 = 166

FN = 5+2+7+7 = 21

FP = 7+0+6+1 = 14

TP = 25

Accuracy = $\frac{166+25}{166+14+21+25}$ **= 0.8451327434**

Recall = $\frac{25}{21+25}$ = 0.5434782609

Precision = $\frac{25}{14+25}$ = 0.641025641

F-1 Score = $2 * \frac{0.641025641*0.5434782609}{0.641025641+0.5434782609}$ **= 0.5882352941**

c_4:

TN = 52+3+7+2+2+28+0+3+5+2+25+7+1+2+1+17 = 157

FN = 1+3+6+5 = 15

FP = 3+7+2+2 = 14

TP = 40

Accuracy = $\frac{157+40}{157+14+15+40}$ **= 0.8716814159**

Recall = $\frac{40}{15+40}$ = 0.7272727273

Precision = $\frac{40}{14+40}$ = 0.7407407407

F-1 Score = $2 * \frac{0.7407407407*0.7272727273}{0.7407407407+0.7272727273}$ **= 0.7339449541**

c_5:

TN = 52+3+7+2+2+28+0+2+5+2+25+7+1+3+6+40 = 185

FN = 1+2+1+3 = 7

FP = 2+3+7+5 = 17

TP = 17

Accuracy = $\frac{185+17}{185+17+7+17}$ **= 0.8938053097**

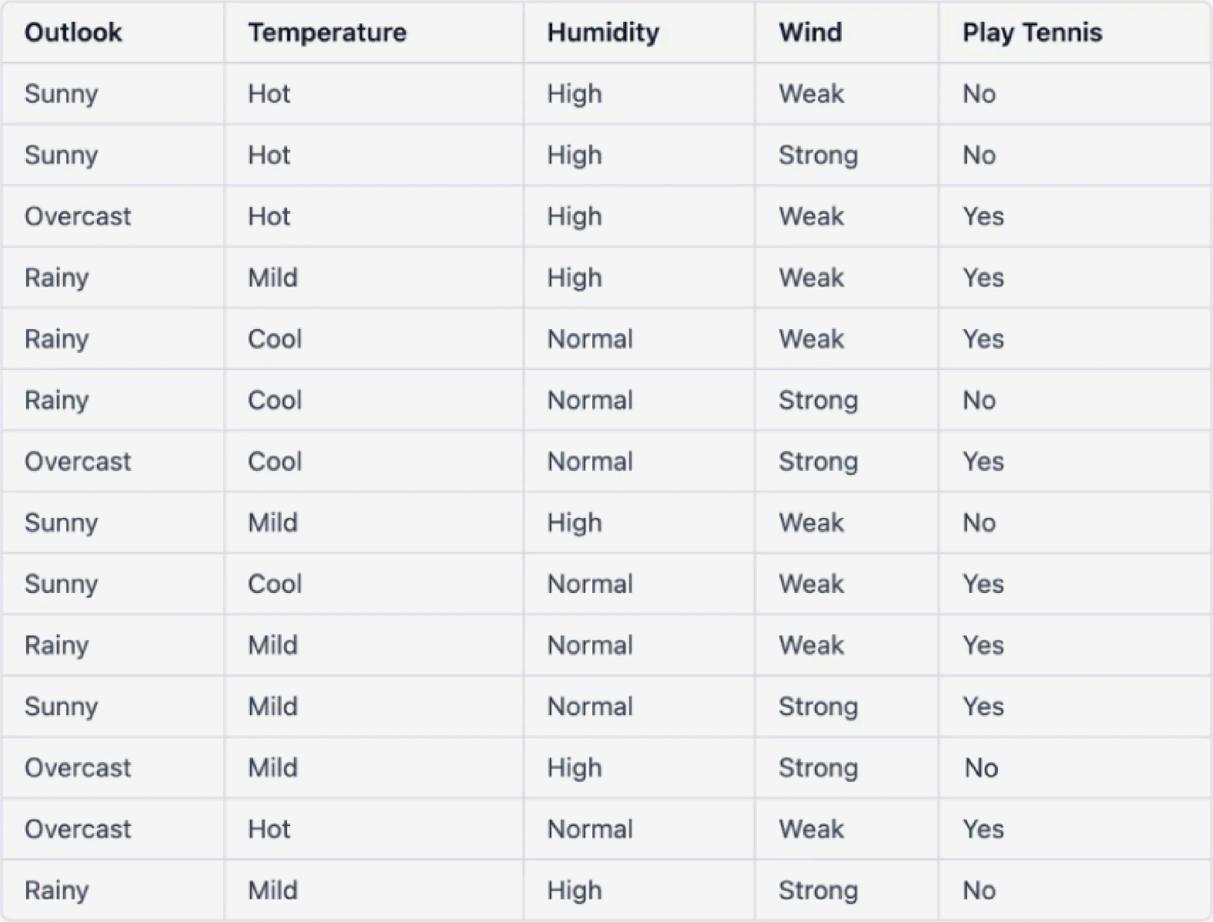
Recall = $\frac{17}{7+17}$ = 0.7083333333

Precision = $\frac{17}{17+17}$ = 0.5

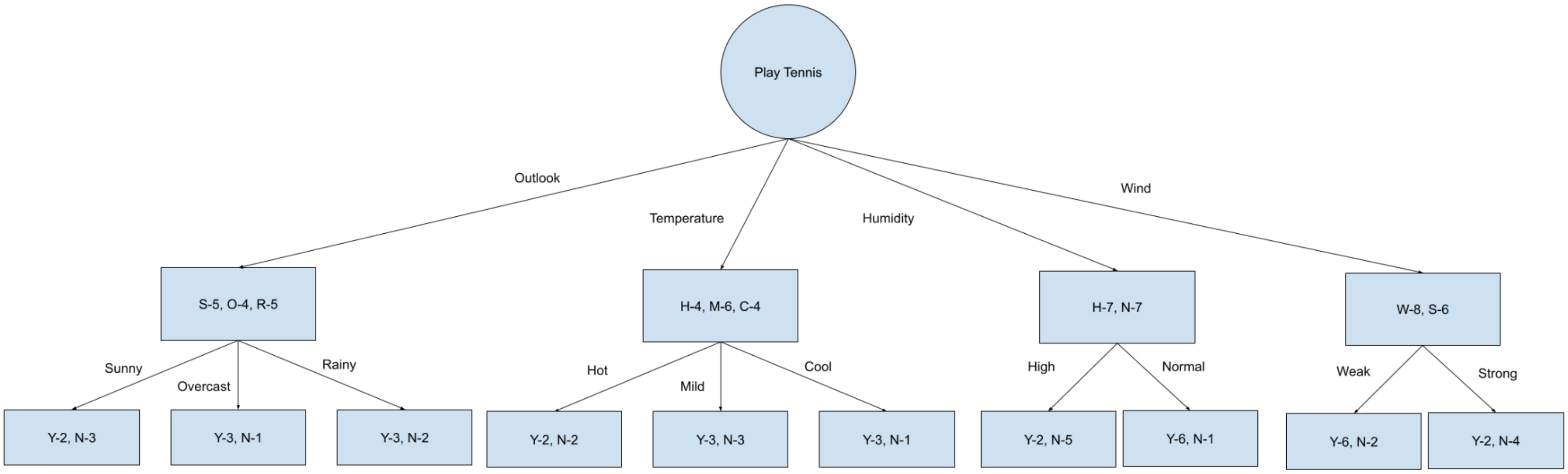
F-1 Score = $2 * \frac{0.7407407407*0.7272727273}{0.7407407407+0.7272727273}$ **= 0.5862068965**

Q2. Given the following sample dataset that represents whether or not to play tennis based on different features:

- Calculate the Information Gain for each feature: $G(D, Outlook)$, $G(D, Temperature)$, $G(D, Humidity)$, $G(D, Wind)$.
- Based on the Information Gain for each feature, which feature should be used first to split the dataset when creating a decision tree?



Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	No
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No



OUTLOOK:

$$H(Y) = -\sum_{y\in Y} p(y)log_2 p(y) = -\left(\frac{8}{14}log_2\frac{8}{14} + \frac{6}{14}log_2\frac{6}{14}\right) = 0.985228136034$$

Sunny:

$$P(Y|Sunny) = \frac{2}{5}$$

$$P(N|Sunny) = \frac{3}{5}$$

$$H(Y|Sunny) = -\left(\frac{2}{5}log_2\frac{2}{5} + \frac{3}{5}log_2\frac{3}{5}\right) = 0.970950594455$$

Overcast:

$$P(Y|Overcast) = \frac{3}{4}$$

$$P(N|Overcast) = \frac{1}{4}$$

$$H(Y|Overcast) = -\left(\frac{3}{4}log_2\frac{3}{4} + \frac{1}{4}log_2\frac{1}{4}\right) = 0.811278124459$$

Rainy:

$$P(Y|Rainy) = \frac{3}{5}$$

$$P(N|Rainy) = \frac{2}{5}$$

$$H(Y|Rainy) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.970950594455$$

$$H(Y|Outlook) = -\sum_{y\in Y}\sum_{o\in O} p(y, O)log_2 p(y|O) =$$

$$\frac{5}{14}H(Y|Sunny) + \frac{4}{14}H(Y|Overcast) + \frac{5}{14}H(Y|Rainy) =$$

$$\frac{5}{14}(0.97095059445) + \frac{4}{14}(0.811278124459) + \frac{5}{14}(0.970950594455)$$

$$= 0.925329888742$$

G(D, Outlook) = H(Y) - H(Y|Outlook) =

0.985228136034 - 0.925329888742 = 0.059898247292

TEMPERATURE:

H(Y) = $-\sum_{y\in Y} p(y)log_2 p(y)$ = $-$ ($\frac{8}{14}log_2\frac{8}{14} + \frac{6}{14}log_2\frac{6}{14}$) = 0.985228136034

Hot:

P(Y|Hot) = $\frac{2}{4}$

P(N|Hot) = $\frac{2}{4}$

H(Y|Hot) = $-$ ($\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}$) = 1

Mild:

P(Y|Mild) = $\frac{3}{6}$

P(N|Mild) = $\frac{3}{6}$

H(Y|Mild) = $-$ ($\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}$) = 1

Cool:

P(Y|Cool) = $\frac{3}{4}$

P(N|Cool) = $\frac{1}{4}$

H(Y|Cool) = $-$ ($\frac{3}{4}log_2\frac{3}{4} + \frac{1}{4}log_2\frac{1}{4}$) = 0.811278124459

H(Y|Temperature) = $-\sum_{y\in Y}\sum_{o\in O} p(y, O)log_2 p(y|O)$ =

$\frac{4}{14}H(Y|Hot) + \frac{6}{14}H(Y|Mild) + \frac{4}{14}H(Y|Cool) =$

$\frac{5}{14}(1) + \frac{4}{14}(1) + \frac{5}{14}(0.811278124459)$

= 0.946079464131

G(D, Temperature) = H(Y) - H(Y|Temperature) =
0.985228136034 - 0.946079464131 = 0.039148671903

HUMIDITY:

H(Y) = $-\sum_{y\in Y} p(y)log_2 p(y)$ = $-$ ($\frac{8}{14}log_2\frac{8}{14} + \frac{6}{14}log_2\frac{6}{14}$) = 0.985228136034

High:

P(Y|High) = $\frac{2}{7}$

P(N|High) = $\frac{5}{7}$

H(Y|High) = $-$ ($\frac{2}{7}log_2\frac{2}{7} + \frac{5}{7}log_2\frac{5}{7}$) = 0.863120568567

Normal:

P(Y|Normal) = $\frac{6}{7}$

P(N|Normal) = $\frac{1}{7}$

H(Y|Normal) = $-\ (\frac{6}{7}log_2\frac{6}{7} + \frac{1}{7}log_2\frac{1}{7})$ = 0.591672778582

H(Y|Humidity) = $-\sum_{y\in Y}\sum_{o\in O} p(y, O)log_2p(y|O)$ =

$\frac{1}{2}H(Y|High)\ +\frac{1}{2}H(Y|Normal)\ =$

$\frac{1}{2}(0.863120568567)\ +\frac{1}{2}(0.591672778582)$

= 0.727396673574

G(D, Humidity) = H(Y) - H(Y|Humidity) = 0.985228136034 - 0.727396673574 = 0.25783146246

WIND:

H(Y) = $-\sum_{y\in Y} p(y)log_2p(y)\ \ = -\ (\frac{8}{14}log_2\frac{8}{14} + \frac{6}{14}log_2\frac{6}{14})$ = 0.985228136034

Weak:

P(Y|Weak) = $\frac{6}{8}$

P(N|Weak) = $\frac{2}{8}$

H(Y|Weak) = $-\ (\frac{3}{4}log_2\frac{3}{4} + \frac{1}{4}log_2\frac{1}{4})$ = 0.811278124459

Strong:

P(Y|Strong) = $\frac{2}{6}$

P(N|Strong) = $\frac{4}{6}$

H(Y|Strong) = $-\ (\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3})$ = 0.918295834054

H(Y|Wind) = $-\sum_{y\in Y}\sum_{o\in O} p(y, O)log_2p(y|O)$ =

$\frac{5}{14}H(Y|Weak)\ +\frac{4}{14}H(Y|Strong)\ =$

$\frac{8}{14}(0.811278124459)\ +\frac{6}{14}(0.918295834054)$

= 0.864786979257

G(D, Wind) = H(Y) - H(Y|Wind) = 0.985228136034 - 0.864786979257 = 0.120441156777

Based on the Information Gain for each feature, Humidity should be used first to split the dataset when creating a decision tree.