INFS 4203 / 7203 Data Mining

Week 10: Building Decision Trees using Gini index

&

Calculating Performance metrics using Confusion matrix

Week 10: Q1

In the dataset given in the next slide, There are 14 instances of golf playing decisions based on attributes- outlook, temperature, humidity and wind factors.

Using the provided dataset, construct a decision tree that will determine if a person will be classified as Yes or No to play golf on that particular day. Use the GINI index based splitting criterion to construct the decision tree.

Formulas required -

Gini Index- Stores the sum of squared probabilities for each class.

$$\textbf{Gini} = 1 - \Sigma \ \text{square}(P_i)$$

where value of i ranges from 1 to total no. of classes

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes

Week 10: Q1

- **Outlook –**

Outlook	Yes	NO	No. of instances
Sunny	2	3	5
Overcast	4	0	4
Rain	3	2	5

Gini(Outlook = Sunny) = 1 – sq(2/5) – sq(3/5) = 1 – 0.16-0.36 = 0.48

Gini(Outlook = Overcast) = 1-sq(4/4) – sq(0/4) = 1-1-0 = 0

Gini(Outlook = Rain) = 1- sq(3/5) – sq(2/5) = 1- 0.36-0.16 = 0.48

Then, we will calculate weighted sum of gini indexes for outlook feature-

Gini_index(Outlook) = (5/14) x 0.48 + (4/14) x 0 + (5/14) x 0.48 = 0.171 + 0 + 0.171 = **0.342**

Week 10: Q1

Temperature-

Temperature	Yes	No	No. of instances
Hot	2	2	4
Cool	3	1	4
Mild	4	2	6

Gini(Temp = Hot) = 1- sq(2/4) – sq(2/4) = 1 - 0.25- 0.25 = 0.5
Gini(Temp = Cool) = 1- sq(3/4) – sq(1/4) = 1 – 0.5625 – 0.0625 = 0.375
Gini(Temp= Mild) = 1- sq(4/6) – sq(2/6) = 1- 0.444 – 0.111 = 0.445

Calculating weighted sum of gini index for temperature feature-
Gini_index(Temp) = (4/14) x 0.5 + (4/14) x 0.375 + (6/14) x 0.445 = 0.142 + 0.107 + 0.190 = **0.439**

Week 10: Q1

Humidity-

Humidity	Yes	No	No. of instances
High	3	4	7
Normal	6	1	7

Gini(Humidity=High) = 1 – sq(3/7) – sq(4/7) = 1 – 0.183 – 0.326 = 0.489

Gini(Humidity=Normal) = 1 – sq(6/7) – sq(1/7) = 1 – 0.734 – 0.02 = 0.244

Weighted sum of gini index for humidity feature-
Gini_index(Humidity) = (7/14) x 0.489 + (7/14) x 0.244 = **0.367**

Week 10: Q1

WIND-

Wind	Yes	No	No. of instances
Weak	6	2	8
Strong	3	3	6

Gini(Wind=Weak) = 1 – sq(6/8) – sq(2/8)= 1 – 0.5625 – 0.062 = 0.375

Gini(Wind=Strong) = 1 – sq(3/6) – sq(3/6) = 1 – 0.25 – 0.25 = 0.5

Weighted sum of gini index for wind feature-
Gini_index(Wind) = (8/14) x 0.375 + (6/14) x 0.5 = **0.428**

Week 10: Q1

Which one to choose?

Feature	Gini index
Outlook	0.342
Temperature	0.439
Humidity	0.367
Wind	0.428

The lower the Gini index cost, the better.
Hence, we choose Outlook feature for the first split of our tree as it has the lowest Gini index!

Week 10: Q1

sunny

1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Outlook

overcast

Day	Outlook	Temp.	Humidity	Wind	Decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

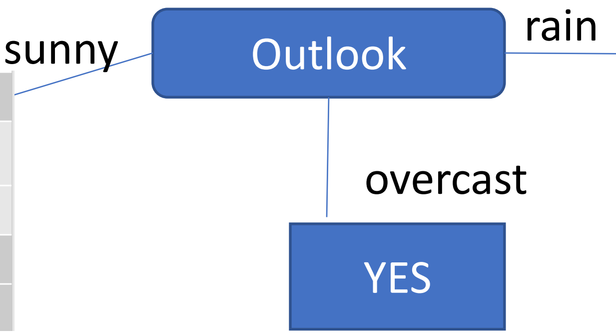
Rain

4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Week 10: Q1

- As all the obs. in overcast have decision “Yes”, the tree can be updated-

1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes



4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Week 10: Q1

- **Next**: Focus on the dataset where outlook = sunny and repeat the steps

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Week 10: Q1

Gini index for Temperature of sunny outlook

Temperature	Yes	No	No. of instances
Hot	0	2	2
Cool	1	0	1
Mild	1	1	2

Gini(Outlook=Sunny and Temp.=Hot) = 1 – sq(0/2) – sq(2/2) = 0

Gini(Outlook=Sunny and Temp.=Cool) = 1 – sq(1/1) – sq(0/1) = 0

Gini(Outlook=Sunny and Temp.=Mild) = 1 – sq(1/2) – sq (1/2)= 1 – 0.25 – 0.25 = 0.5

Gini_index(Outlook=Sunny and Temp.) = (2/5)x0 + (1/5)x0 + (2/5)x0.5 = **0.2**

Week 10: Q1

Gini index for Humidity of sunny outlook

Humidity	Yes	No	No. of instances
High	0	3	3
Normal	2	0	2

Gini(Outlook=Sunny and Humidity=High) = 1 – sq(0/3) – sq(3/3)= 0

Gini(Outlook=Sunny and Humidity=Normal) = 1 – sq(2/2) – sq(0/2) = 0

Gini_index(Outlook=Sunny and Humidity) = (3/5)x0 + (2/5)x0 = **0**

Week 10: Q1

Gini index for wind of sunny outlook

WIND	Yes	No	No of instances
weak	1	2	3
strong	1	1	2

Gini(Outlook=Sunny and Wind=Weak) = 1 – (1/3)2 – (2/3)2 = 0.266

Gini(Outlook=Sunny and Wind=Strong) = 1- (1/2)2 – (1/2)2 = 0.2

Gini_index(Outlook=Sunny and Wind) = (3/5)x0.266 + (2/5)x0.2 = **0.466**

Week 10: Q1

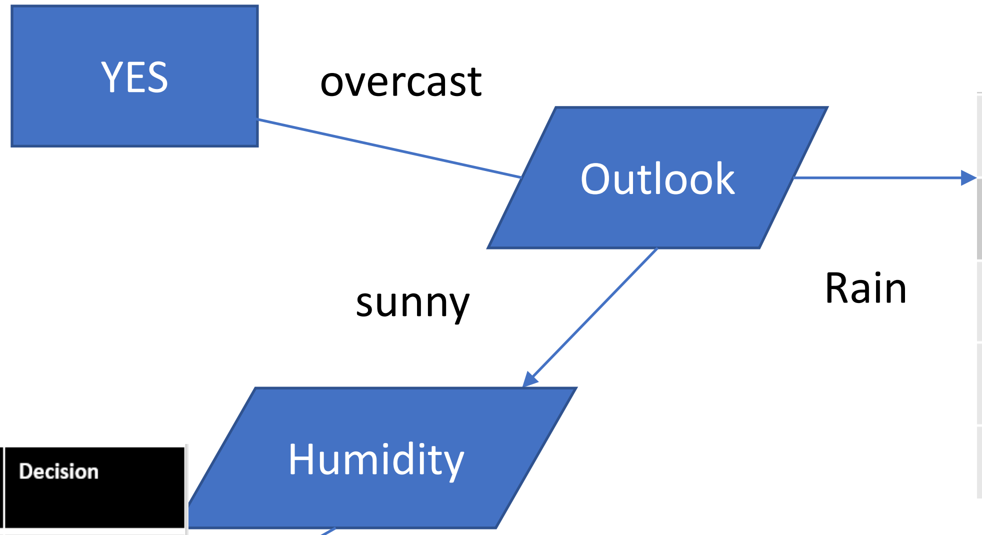
Which one to choose?

Feature	Gini Index
Temperature	0.2
Humidity	0
Wind	0.466

we choose Humidity feature for the next split of our tree as it has the lowest Gini index.

Week 10: Q1

- Updated tree-

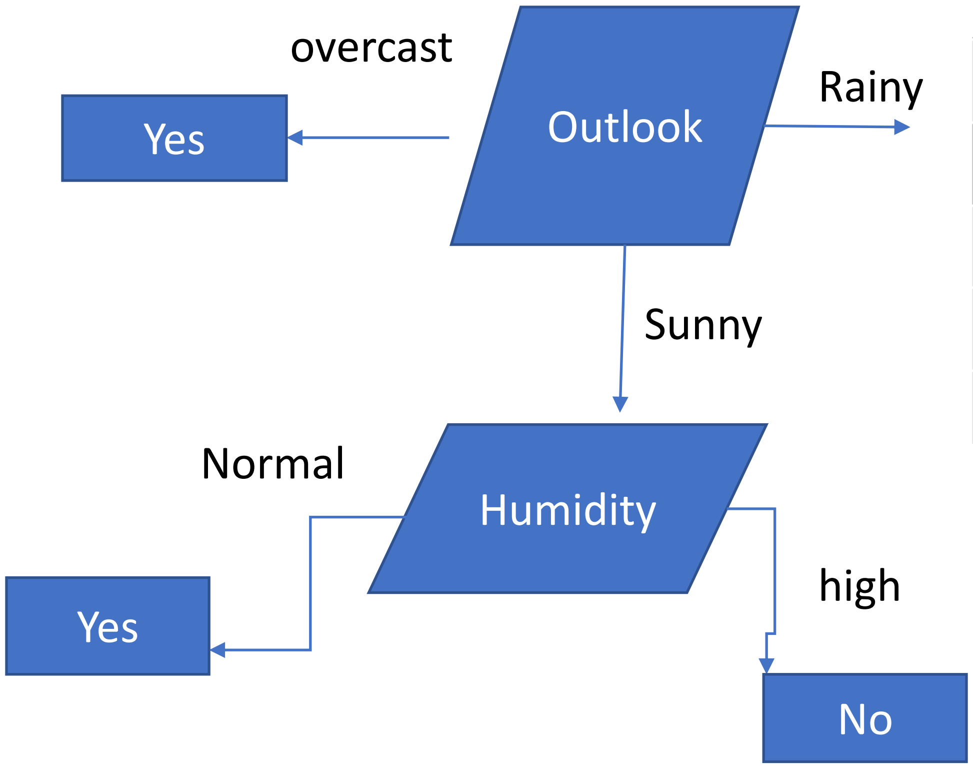


4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Temp.	Humidity	Wind	Decision
Hot	High	Weak	No
Hot	High	Strong	No
Mild	High	Weak	No
Cool	Normal	Weak	Yes
Mild	Normal	Strong	Yes

Week 10: Q1

- Updated based on humidity-



4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Week 10: Q1

- **Next**: Focus on the dataset where outlook = Rain and repeat the steps

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Week 10: Q1

Gini index for Temperature of Rain outlook

Temperature	Yes	No	No of instances
Cool	1	1	2
Mild	2	1	3

Gini(Outlook=Rain and Temp.=Cool) = 1 – sq(1/2) – sq(1/2) = 0.5

Gini(Outlook=Rain and Temp.=Mild) = 1 – sq(2/3) – sq(1/3) = 0.444

Gini_index(Outlook=Rain and Temp.) = (2/5)x0.5 + (3/5)x0.444 = **0.466**

Week 10: Q1

Gini index for Humidity of Rain outlook

Humidity	Yes	No	No of instances
High	1	1	2
Normal	2	1	3

Gini(Outlook=Rain and Humidity=High) = 1 – sq(1/2) – sq(1/2) = 0.5

Gini(Outlook=Rain and Humidity=Normal) = 1 – sq(2/3)– sq(1/3) = 0.444

Gini_index(Outlook=Rain and Humidity) = (2/5)x0.5 + (3/5)x0.444 = **0.466**

Week 10: Q1

Gini index for Wind of Rain outlook

Wind	Yes	No	No of instances
Weak	3	0	3
Strong	0	2	2

Gini(Outlook=Rain and Wind=Weak) = 1 – sq(3/3) – sq(0/3) = 0

Gini(Outlook=Rain and Wind=Strong) = 1 – sq(0/2) – sq(2/2) = 0

Gini_index(Outlook=Rain and Wind) = (3/5)x0 + (2/5)x0 = **0**

Week 10: Q1

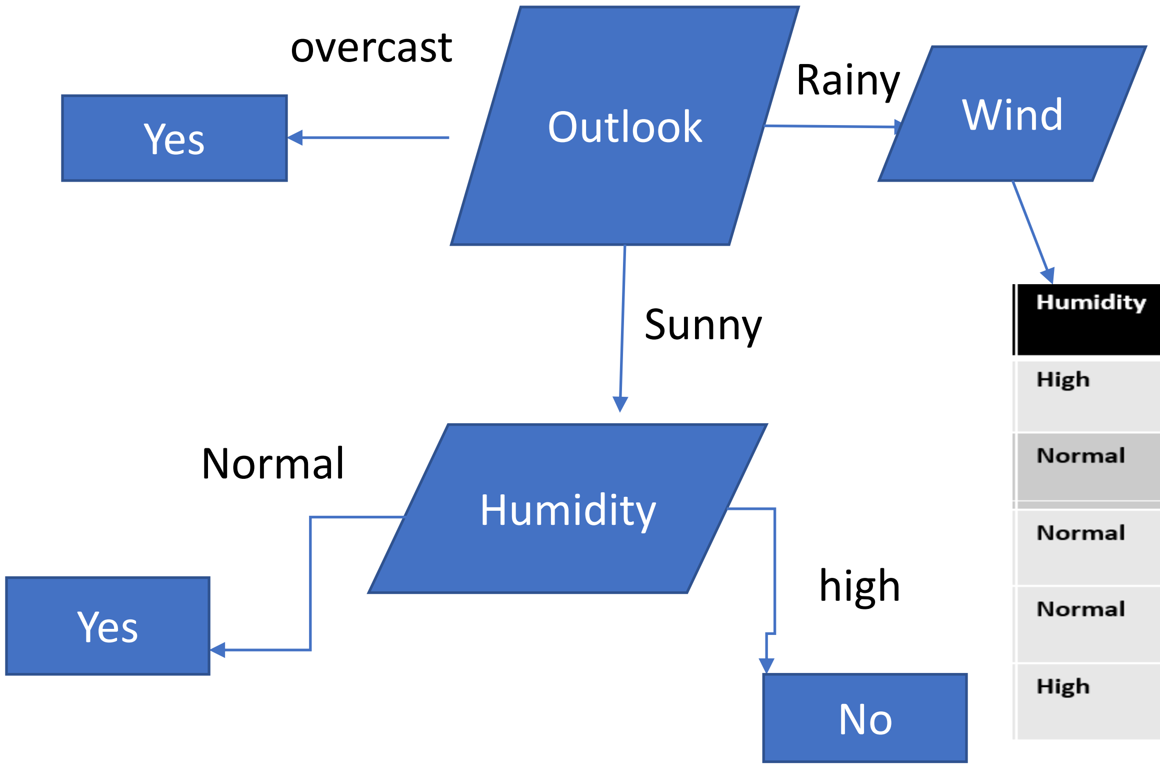
Which one to choose?

Feature	Gini index
Temperature	0.466
Humidity	0.466
Wind	0

we choose **Wind** feature for the next split of our tree as it has the lowest Gini index.

Week 10: Q1

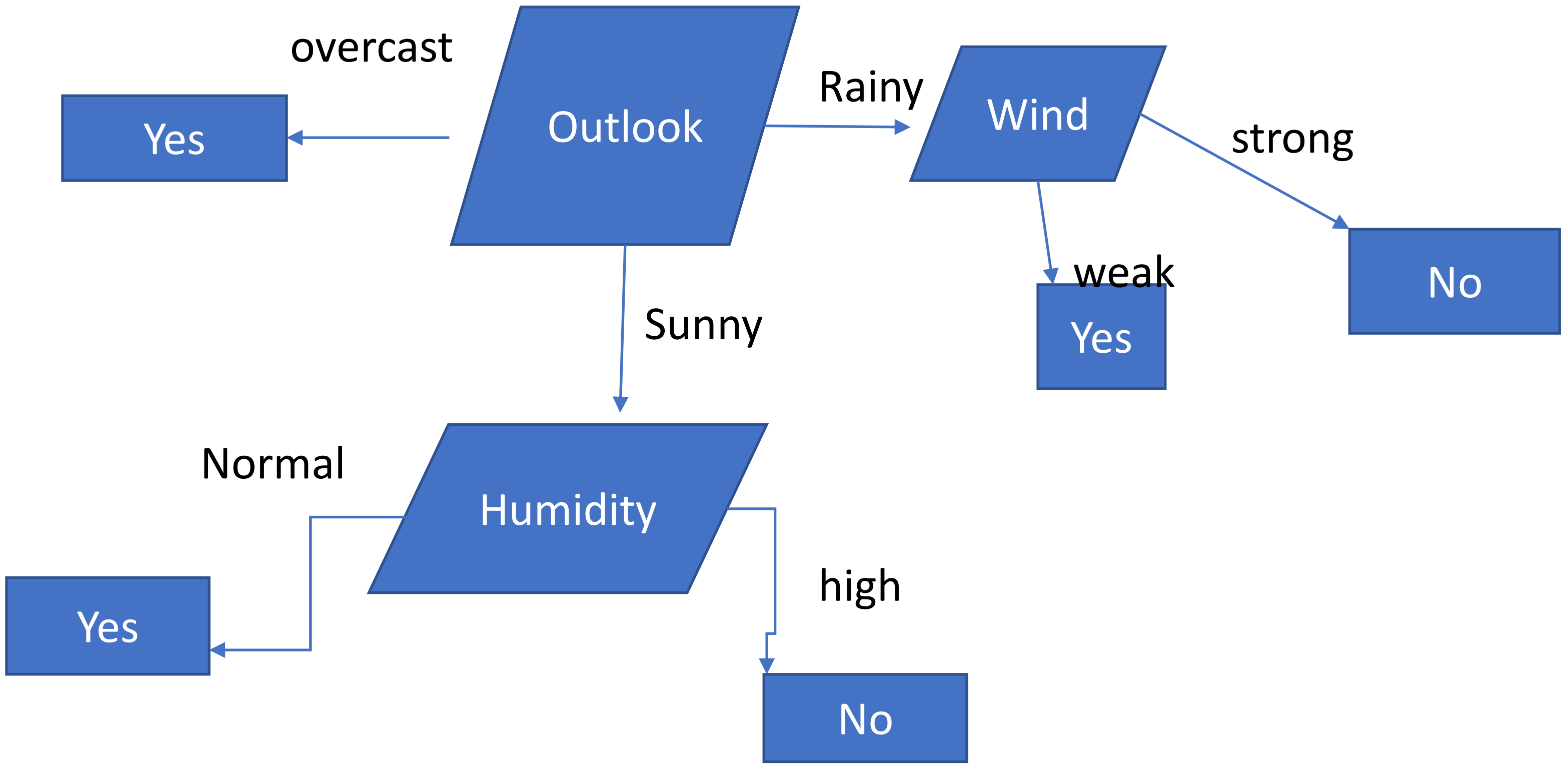
- Updated tree-



Humidity	Temp.	Wind	Decision
High	Mild	Weak	Yes
Normal	Cool	Weak	Yes
Normal	Cool	Strong	No
Normal	Mild	Weak	Yes
High	Mild	Strong	No

Week 10: Q1

- Final Decision Tree-



Week 10: Q1

Final Answer-

- The first decision node is the attribute “Outlook”.
- if the value of outlook = “overcast” the decision of playing the game is classified as “yes”.
- If value of outlook =“sunny”, further splitting is based on the attribute “humidity”, wherein, if humidity = “high” then, decision = “No”, if humidity = “normal”, then decision = “yes”.
- If value of outlook = “Rain”, further splitting is based on the attribute “Wind”, wherein, if wind = “Strong”, then decision is classified “No”, if wind = “weak”, the decision is classified as “Yes”

Week 10: Q2

Calculate Precision, Recall and F1 measurement for both the classifiers below and discuss which one is better ?

Classifier 1

(Orange)	Not Blue
(Blue)	Blue
(Blue)	Not Blue
(Pink)	Not Blue
(Blue)	Blue
(Brown)	Blue
(Blue)	Blue
(Blue)	Not Blue
(Blue)	Blue
(Green)	Not Blue

Classifier 2

(Blue)	Not Blue
(Blue)	Blue
(Yellow)	Not Blue
(Blue)	Blue
(Brown)	Not Blue
(Blue)	Blue
(Blue)	Not Blue
(Blue)	Not Blue
(Green)	Not Blue
(Black)	Not Blue

Week 10: Q2

Classifier 1

(Orange)	Not Blue
(Blue)	Blue
(Blue)	Not Blue
(Pink)	Not Blue
(Blue)	Blue
(Brown)	Blue
(Blue)	Blue
(Blue)	Not Blue
(Blue)	Blue
(Green)	Not Blue

Predicted → Actual ↓	**Blue**	**Not Blue**
Blue	True positive	False negative
Not Blue	False positive	True negative

Week 10: Q2

Classifier 1

(Orange)	Not Blue
(Blue)	Blue
(Blue)	Not Blue
(Pink)	Not Blue
(Blue)	Blue
(Brown)	Blue
(Blue)	Blue
(Blue)	Not Blue
(Blue)	Blue
(Green)	Not Blue

Predicted Actual	**Blue**	**Not Blue**
Blue	**4**	False negative
Not Blue	False positive	True negative

Week 10: Q2

Classifier 1

(Orange)	Not Blue
(Blue)	Blue
(Blue)	Not Blue
(Pink)	Not Blue
(Blue)	Blue
(Brown)	Blue
(Blue)	Blue
(Blue)	Not Blue
(Blue)	Blue
(Green)	Not Blue

Predicted → **Actual** ↓	**Blue**	**Not Blue**
Blue	**4**	**2**
Not Blue	False positive	True negative

Week 10: Q2

Classifier 1

(Orange)	Not Blue
(Blue)	Blue
(Blue)	Not Blue
(Pink)	Not Blue
(Blue)	Blue
(Brown)	Blue
(Blue)	Blue
(Blue)	Not Blue
(Blue)	Blue
(Green)	Not Blue

Predicted **Actual**	**Blue**	**Not Blue**
Blue	**4**	**2**
Not Blue	1	True negative

Week 10: Q2

Classifier 1

(Orange)	Not Blue
(Blue)	Blue
(Blue)	Not Blue
(Pink)	Not Blue
(Blue)	Blue
(Brown)	Blue
(Blue)	Blue
(Blue)	Not Blue
(Blue)	Blue
(Green)	Not Blue

Predicted → Actual ↓	**Blue**	**Not Blue**
Blue	**4**	**2**
Not Blue	1	3

Week 10: Q2

Predicted → **Actual** ↓	**Blue**	**Not Blue**
Blue	**4**	**2**
Not Blue	1	3

Precision = TP/(TP+FP) = 4 / (4 +1) = 4/5 = **0.8**

Recall = TP/(TP+FN) = 4 / (4 +2) = 4/6 = **0.67**

F1 = 2* [(Precision *recall)/ (Precision +recall)] = **0.729**

Week 10: Q2

Classifier 2

(Blue)	Not Blue
(Blue)	Blue
(Yellow)	Not Blue
(Blue)	Blue
(Brown)	Not Blue
(Blue)	Blue
(Blue)	Not Blue
(Blue)	Not Blue
(Green)	Not Blue
(Black)	Not Blue

Predicted → Actual ↓	**Blue**	**Not Blue**
Blue	**3**	**3**
Not Blue	0	4

Week 10: Q2

Predicted Actual	Blue	Not Blue
Blue	**3**	**3**
Not Blue	0	4

- Precision = TP/(TP+FP) = 3 / (3 +0) = 1
- Recall = TP/(TP+FN) = 3 / (3+3) = 3/6 = **0.5**
- F1 = 2* [(Precision *recall)/ (Precision +recall)] = **0.67**

Week 10: Q2

- **Which of the two classifiers is better??**
- Precision is better for classifier 2 (1 > 0.8)
- Recall is better for classifier 1 (0.67> 0.5)
- In cases like this, we need a single measure of comparison that trades off between precision and recall ->

F1 – It is the harmonic mean of precision and recall.

Hence, as classifier 1 has higher F1 measurement value than classifier2, it is safe to conclude that **Classifier 1 is better**.