Assignment: Random Forests

a) From the below data, please create (manual) a random forest with T=3, OOB rate = 20% (10 points)

Data

Sample M	Gene #1	Gene #2	Gene #3	Class
1	-4	1	6	-1
2	3	-1	6	1
3	6	1	12	1
4	1	0	2	-1
5	0	1	0	-1

Tree 1

Out of Bag

Sample M	Gene #1	Gene #2	Class
2	3	-1	1

Bootstrapping

Random Vector : D1

Sample M	Gene #1	Gene #2	Class
1	-4	1	-1
1	-4	1	-1
3	6	1	1
2	3	-1	1
5	0	1	-1

2 samples are class 1.
3 samples are class -1.

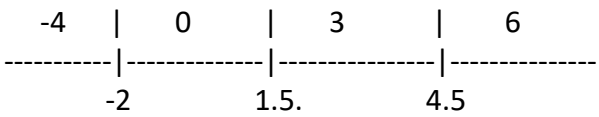
$Gini(D_t) = 1 - \Sigma_{j=1}^{L} P_j^2$

$= 1 - [\left(\frac{2}{5}\right)^2 + (\frac{3}{5})^2]$

$= 1 - \frac{13}{25}$

$= \frac{12}{25}$

Gini Gain (Gene #1)



$$GiniGain(A_1, D_t) = Gini(D_t) - [\frac{|D_{left}|}{|D_t|} Gini(D_{left}) + \frac{|D_{right}|}{|D_t|} Gini(D_{right})]$$

Split No.1 Split at (-2)

2 samples to the left
3 samples to the right

Gini Gain $(A_1, split(-2))$

$= \frac{12}{25} - [\frac{2}{5}(1 - [(\frac{0}{2})^2 + (\frac{2}{2})^2]) + \frac{3}{5}(1 - [(\frac{2}{3})^2 + (\frac{1}{3})^2])]$

$= \frac{12}{25} - [\frac{2}{5}(1-1) + \frac{3}{5}(1 - [(\frac{5}{9})])]$

$= \frac{12}{25} - [0 + \frac{3}{5}[(\frac{4}{9})]]$

$= \frac{12}{25} - [\frac{12}{45}]$

$= \frac{12(45-25)}{1{,}125}$

$= \frac{240}{1{,}125}$

$= 0.21$

Split No.2 Split at (1.5)

3 samples to the left
2 samples to the right

Gini Gain $(A_1, split(1.5))$

$= \frac{12}{25} - [\frac{3}{5}(1 - [(\frac{0}{3})^2 + (\frac{3}{3})^2]) + \frac{2}{5}(1 - [(\frac{2}{2})^2 + (\frac{0}{2})^2])]$

$= \frac{12}{25} - [\frac{3}{5}(1-1) + \frac{2}{5}(1 - 1)]$

$= \frac{12}{25} - [0]$

$= \frac{12}{25}$

= 0.48

Split No.3 Split at (4.5)

4 samples to the left
1 samples to the right

Gini Gain $(A_1, split(4.5))$

$= \frac{12}{25} - [\frac{4}{5}(1 - [(\frac{1}{4})^2 + (\frac{3}{4})^2]) + \frac{1}{5}(1 - [(\frac{1}{1})^2 + (\frac{0}{1})^2])]$

$= \frac{12}{25} - [\frac{4}{5}(1 - (\frac{10}{16})\ ) + \frac{1}{5}(1 - 1)]$

$= \frac{12}{25} - [\frac{4}{5}(\ (\frac{6}{16})\ )]$

$= \frac{12}{25} - [\frac{6}{20}]$

$= \frac{240-150}{500}$

$= \frac{90}{500}$

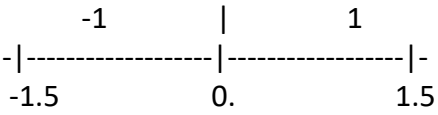
= 0.18

Gene #1	Split No.	1	2	3
	Position	-2	1.5	4.5
	Gini Gain	0.21	0.48	0.18

Gini Gain (Gene #2)

Random Vector : D1

Sample M	Gene #1	Gene #2	Class
1	-4	1	-1
1	-4	1	-1
3	6	1	1
2	3	-1	1
5	0	1	-1



Split No.1 Split at (-1.5)

0 samples to the left
5 samples to the right

Gini Gain $(A_2, split(-2))$

$= \frac{12}{25} - [\frac{0}{5}(1 - [(\frac{0}{0})^2 + (\frac{0}{0})^2]) + \frac{5}{5}(1 - [(\frac{2}{5})^2 + (\frac{3}{5})^2])]$

$= \frac{12}{25} - [0 + 1(1 - [(\frac{13}{25})])]$

$= \frac{12}{25} - [0 + (\frac{12}{25})]$

$= \frac{12}{25} - [\frac{12}{25}]$

= 0

Split No.2 Split at (0)

1 samples to the left
4 samples to the right

Gini Gain $(A_2, split(0))$

$= \frac{12}{25} - [\frac{1}{5}(1 - [(\frac{1}{1})^2 + (\frac{0}{1})^2]) + \frac{4}{5}(1 - [(\frac{1}{4})^2 + (\frac{3}{4})^2])]$

$= \frac{12}{25} - [\frac{1}{5}(1 - 1) + \frac{4}{5}(1 - [(\frac{10}{16})])]$

$= \frac{12}{25} - [\frac{1}{5} + \frac{4}{5}(\frac{6}{16})]$

$= \frac{12}{25} - [\frac{1}{5} + \frac{6}{20}]$

$= \frac{12}{25} - [\frac{10}{20}]$

$= \frac{48+50}{100}$

= 0.98

Split No.3 Split at (1.5)

5 samples to the left
0 samples to the right

Gini Gain $(A_2, split(1.5))$

$= \frac{12}{25} - [\frac{5}{5}(1 - [(\frac{2}{5})^2 + (\frac{3}{5})^2]) + \frac{0}{5}(1 - [(\frac{0}{0})^2 + (\frac{0}{0})^2])]$

$= \frac{12}{25} - [1\left(1 - \left[(\frac{13}{25})\right]\right) + 0]$

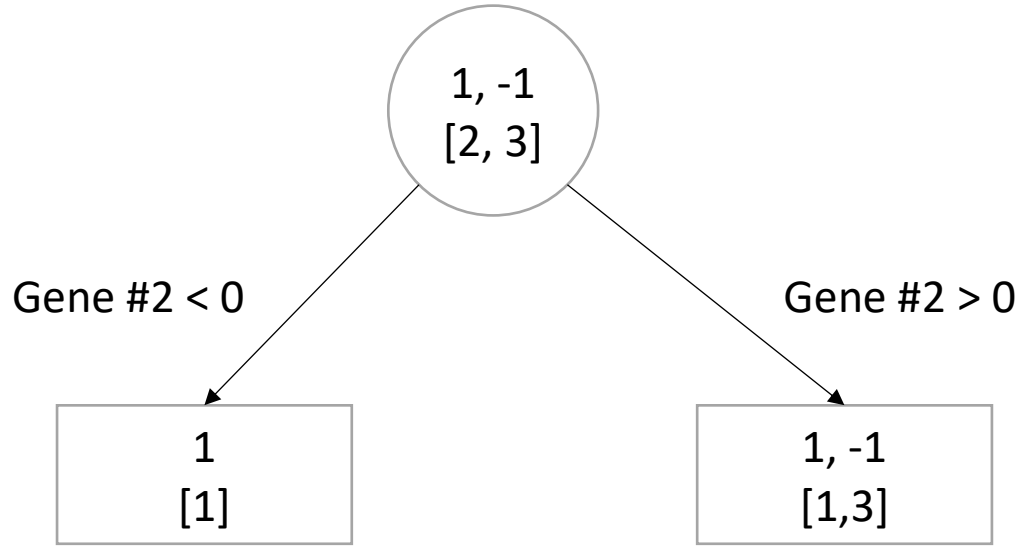
$= \frac{12}{25} - [(\frac{12}{25})]$

= 0

Gene #2	Split No.	1	2	3
	Position	-1.5	0	1.5
	Gini Gain	0	0.98*	0

Tree 1

Gene #2, Split No. 2 has highest Gini.



Tree 2

Out of Bag

Sample M	Gene #1	Gene #3	Class
4	1	2	-1

Random Vector : D2

Sample M	Gene #1	Gene #3	Class
2	3	6	1
2	3	6	1
5	0	0	-1
3	6	12	1
1	-4	6	-1

3 samples are class 1.
2 samples are class -1.

$$Gini(D_t) = 1 - \Sigma_{j=1}^{L} P_j^2$$

$$= 1 - [\left(\frac{3}{5}\right)^2 + (\frac{2}{5})^2]$$

$$= 1 - \frac{13}{25}$$

$$= \frac{12}{25}$$

Gini Gain (Gene #1)

-4	0	3	6
-2 1.5 4.5

Split No.1 Split at (-2)

1 samples to the left
4 samples to the right

Gini Gain $(A_1, split(-2))$

$= \frac{12}{25} - [\frac{1}{5}(1 - [(\frac{0}{1})^2 + (\frac{1}{1})^2]) + \frac{4}{5}(1 - [(\frac{3}{4})^2 + (\frac{1}{4})^2])]$

$= \frac{12}{25} - [\frac{2}{5}(1-1) + \frac{3}{5}(1 - [(\frac{10}{16})])]$

$= \frac{12}{25} - [0 + \frac{3}{5}[(\frac{6}{16})]]$

$= \frac{12}{25} - [\frac{18}{80}]$

$= \frac{960-450}{2000}$

$= \frac{510}{2000}$

= 0.255

Split No.2 Split at (1.5)

2 samples to the left
3 samples to the right

Gini Gain $(A_1, split(1.5))$

$= \frac{12}{25} - [\frac{2}{5}(1 - [(\frac{0}{2})^2 + (\frac{2}{2})^2]) + \frac{3}{5}(1 - [(\frac{3}{3})^2 + (\frac{0}{3})^2])]$

$= \frac{12}{25} - [\frac{2}{5}(1-1) + \frac{3}{5}(1 - 1)]$

$= \frac{12}{25} - [0]$

$= \frac{12}{25}$

= 0.48

Split No.1 Split at (4.5)

4 samples to the left
1 samples to the right

Gini Gain $(A_1, split(4.5))$

$= \frac{12}{25} - [\frac{4}{5}(1 - [(\frac{2}{4})^2 + (\frac{2}{4})^2]) + \frac{1}{5}(1 - [(\frac{1}{1})^2 + (\frac{0}{1})^2])]$

$= \frac{12}{25} - [\frac{4}{5}(1 - [(\frac{2}{4})\ ]) + \frac{1}{5}(1 - 1)]$

$= \frac{12}{25} - [\frac{4}{5}[(\frac{1}{2})\ ] + 0]$

$= \frac{12}{25} - [\frac{2}{5}]$

$= \frac{2}{25}$

= 0.08

Gene #1	Split No.	1	2	3
	Position	-2	1.5	4.5
	Gini Gain	0.255	0.48*	0.08

Gini Gain (Gene #3)

Random Vector : D2

Sample M	Gene #1	Gene #3	Class
2	3	6	1
2	3	6	1
5	0	0	-1
3	6	12	1
1	-4	6	-1

```
0     |     6     |     12
------|-----------|-----------
      3           9
```

Split No.1 Split at (3)

1 samples to the left
4 samples to the right

Gini Gain $(A_2, split(3))$

$= \frac{12}{25} - [\frac{1}{5}(1 - [(\frac{0}{1})^2 + (\frac{1}{1})^2]) + \frac{4}{5}(1 - [(\frac{3}{4})^2 + (\frac{1}{4})^2])]$

$= \frac{12}{25} - [\frac{2}{5}(1-1) + \frac{3}{5}(1 - [(\frac{10}{16})])]$

$= \frac{12}{25} - [0 + \frac{3}{5}[(\frac{6}{16})]]$

$= \frac{12}{25} - [\frac{18}{80}]$

$= \frac{960-450}{2000}$

$= \frac{510}{2000}$

= 0.255

Split No.2 Split at (9)

4 samples to the left
1 samples to the right

Gini Gain $(A_2, split(3))$

$= \frac{12}{25} - [\frac{4}{5}(1 - [(\frac{2}{4})^2 + (\frac{2}{4})^2]) + \frac{1}{5}(1 - [(\frac{1}{1})^2 + (\frac{0}{1})^2])]$

$= \frac{12}{25} - [\frac{4}{5}(1 - [(\frac{2}{4}) \ ]) + \frac{1}{5}(1 - 1)]$

$= \frac{12}{25} - [\frac{4}{5}[(\frac{1}{2}) \ ] + 0]$

$= \frac{12}{25} - [\frac{2}{5}]$

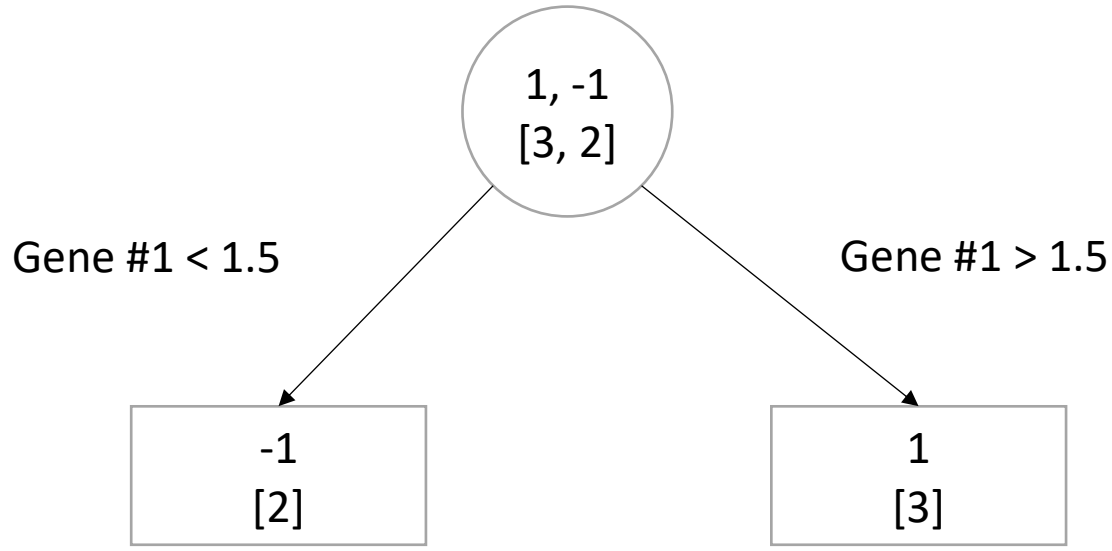
$= \frac{2}{25}$

= 0.08

Gene #3	Split No.	1	2
	Position	3	9
	Gini Gain	0.255	0.08

Tree 2

Gene #1, Split No. 2 has highest Gini.



Tree 3

Out of Bag

Sample M	Gene #2	Gene #3	Class
5	1	0	-1

Random Vector : D3

Sample M	Gene #2	Gene #3	Class
2	-1	6	1
2	-1	6	1
1	1	6	-1
3	1	12	1
1	1	6	-1

3 samples are class 1.
2 samples are class -1.

$$Gini(D_t) = 1 - \Sigma_{j=1}^{L} P_j^2$$

$$= 1 - [\left(\frac{3}{5}\right)^2 + (\frac{2}{5})^2]$$

$$= 1 - \frac{13}{25}$$

$$= \frac{12}{25}$$

Gini Gain (Gene #2)

-1	1
0

Split No.1 Split at (0)

2 samples to the left
3 samples to the right

Gini Gain $(A_1, split(0))$

$$= \frac{12}{25} - [\frac{2}{5}(1 - [\left(\frac{2}{2}\right)^2 + (\frac{0}{2})^2]) + \frac{3}{5}(1 - [\left(\frac{1}{3}\right)^2 + (\frac{2}{3})^2])]$$

$$= \frac{12}{25} - [\frac{2}{5}(1-1) + \frac{3}{5}(1 - [\left(\frac{5}{9}\right)])]$$

$$= \frac{12}{25} - [0 + \frac{3}{5}[\left(\frac{5}{9}\right)]]$$

$= \frac{12}{25} - [\frac{1}{3}]$

$= \frac{36-25}{75}$

$= \frac{11}{75}$

= 0.146

Gene #1	Split No.	1
	Position	0
	Gini Gain	0.146

Gini Gain (Gene #3)

Random Vector : D3

Sample M	Gene #2	Gene #3	Class
2	-1	6	1
2	-1	6	1
1	1	6	-1
3	1	12	1
1	1	6	-1

6	12
9

Split No.1 Split at (9)

4 samples to the left
1 samples to the right

Gini Gain $(A_2, split(9))$

$= \frac{12}{25} - [\frac{4}{5}(1 - [(\frac{2}{4})^2 + (\frac{2}{4})^2]) + \frac{1}{5}(1 - [(\frac{1}{1})^2 + (\frac{0}{1})^2])]$

$= \frac{12}{25} - [\frac{4}{5}(1 - [(\frac{2}{4})\ ]) + \frac{1}{5}(1 - 1)]$

$= \frac{12}{25} - [\frac{4}{5}[(\frac{1}{2})\ ] + 0]$

$= \frac{12}{25} - [\frac{2}{5}]$

$= \frac{2}{25}$

= 0.08

Gene #3	Split No.	1
	Position	9
	Gini Gain	0.08

Tree 3

Gene #2, Split No. 1 has highest Gini.