Prediction models

Seminar Data Science for Economics

Madina Kurmangaliyeva
m.kurmangaliyeva@uvt.nl

Spring 2021

Tilburg University

The proportion of class k observations in node m

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

The proportion of class k observations in node m

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

Number of obs in node m

The proportion of class k observations in node m

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

Number of obs in node m

Indicator whether obs i belongs to class k

The proportion of class k observations in node m

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

Number of obs in node m

Indicator whether obs i belongs to class k

Only those obs that fall into region m

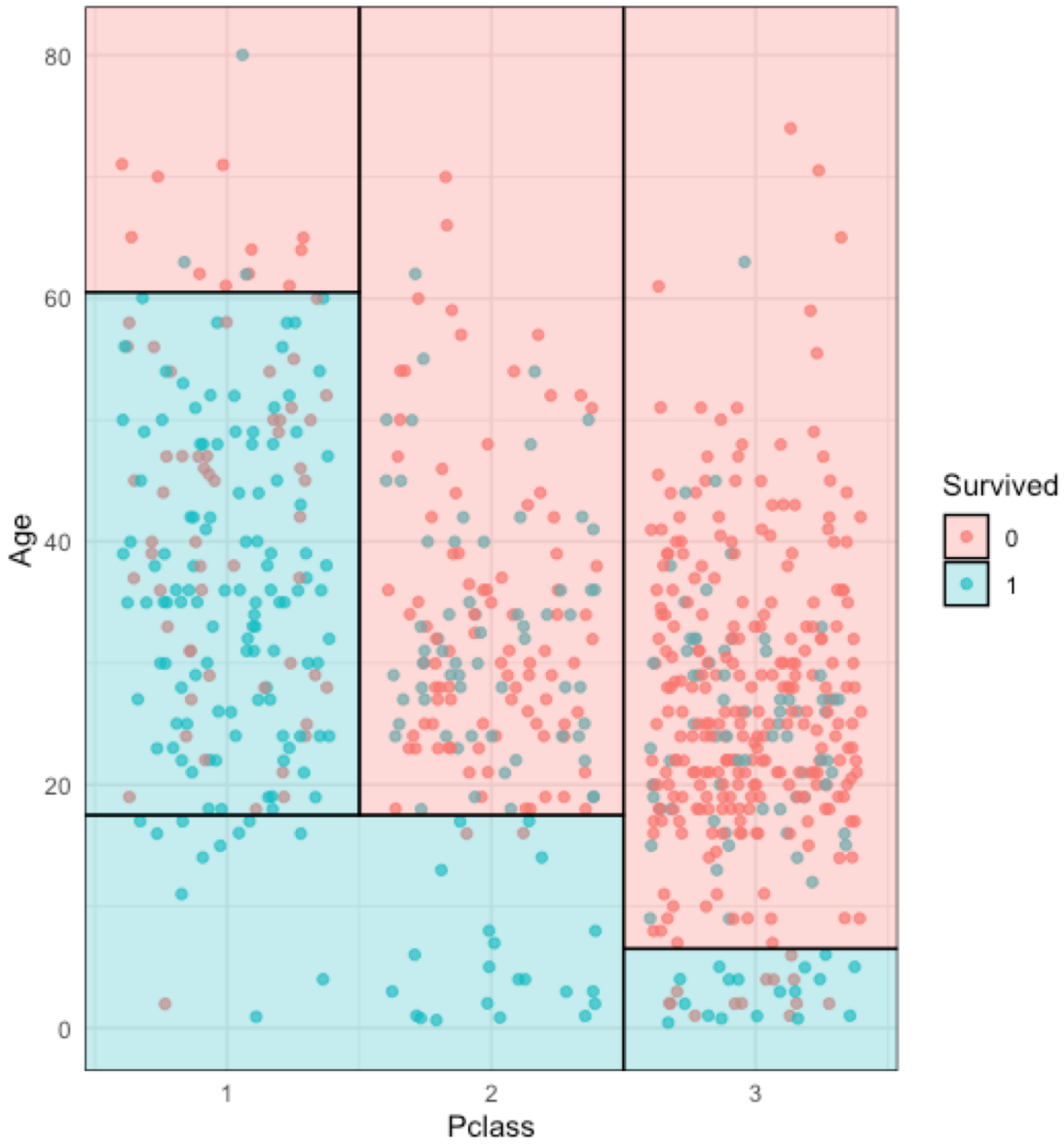
The proportion of class k observations in node m

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

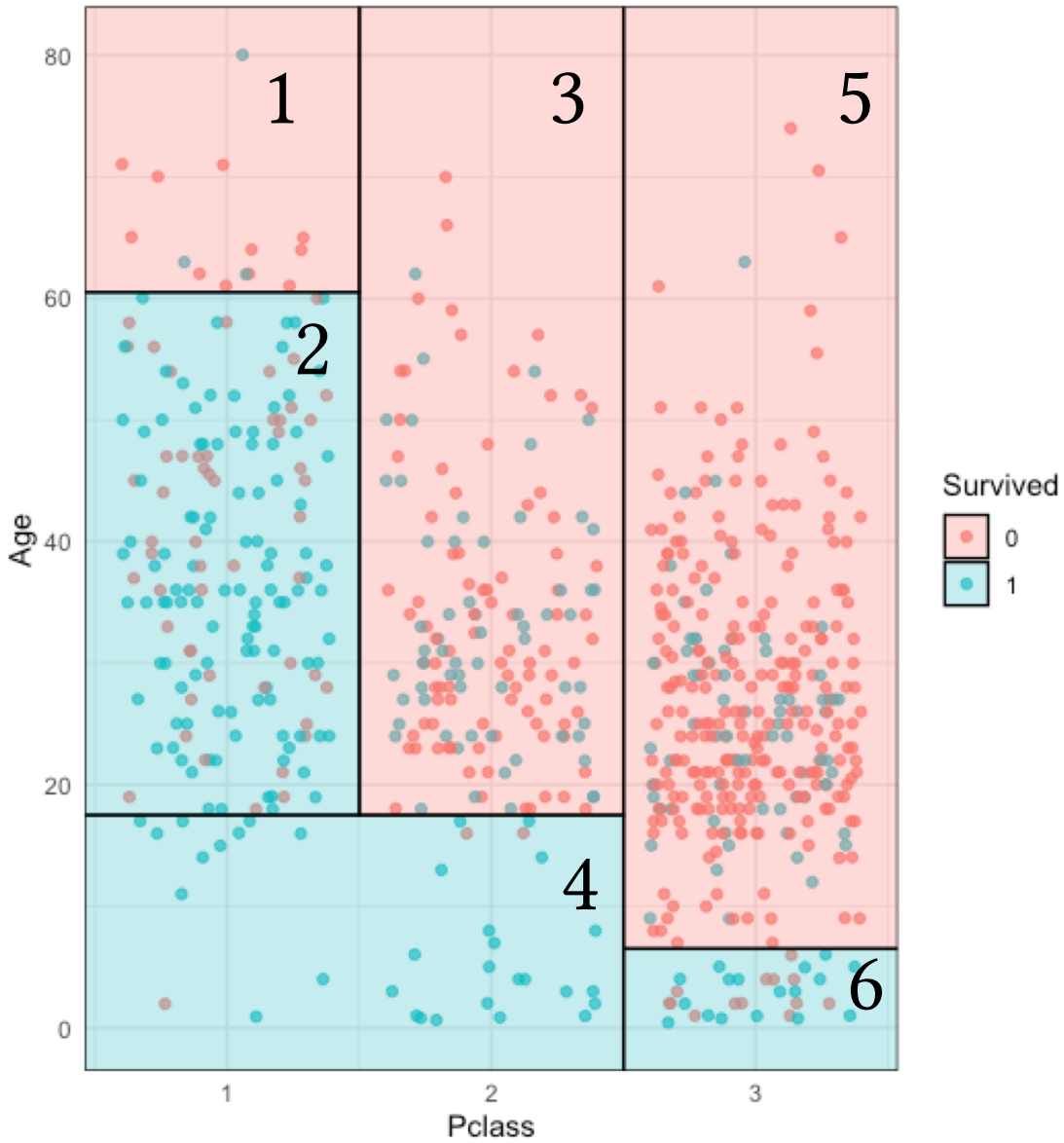
Number of obs in node m

Indicator whether obs i belongs to class k

Only those obs that fall into region m



The proportion of class
Survived observations in node *1*

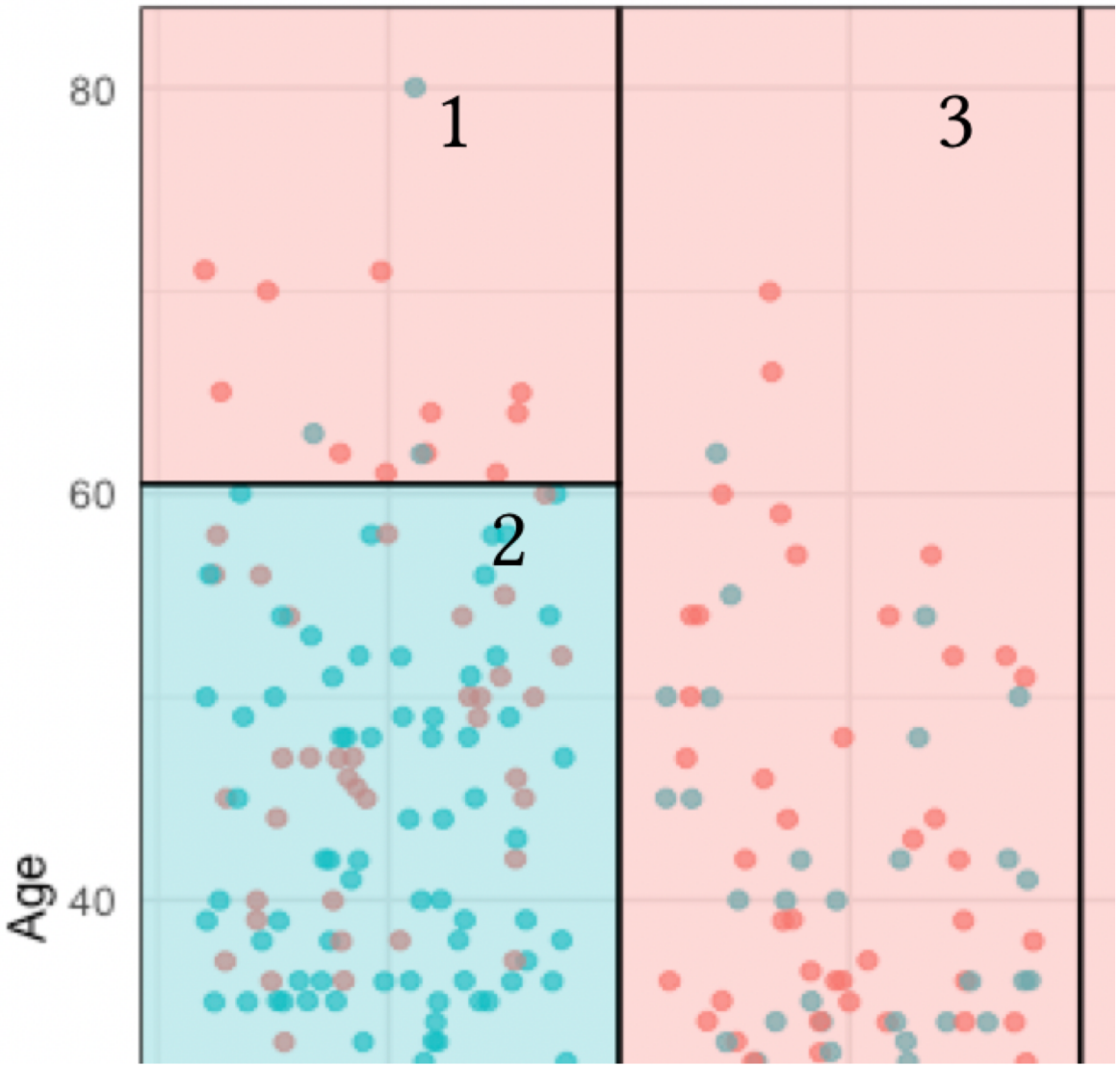


The proportion of class *Survived* observations in node *1*

$$\hat{p}_{1,Surv} = \frac{1}{14} \sum_{x_i \in R_m} I(y_i = Survived)$$

$$= \frac{3}{14}$$

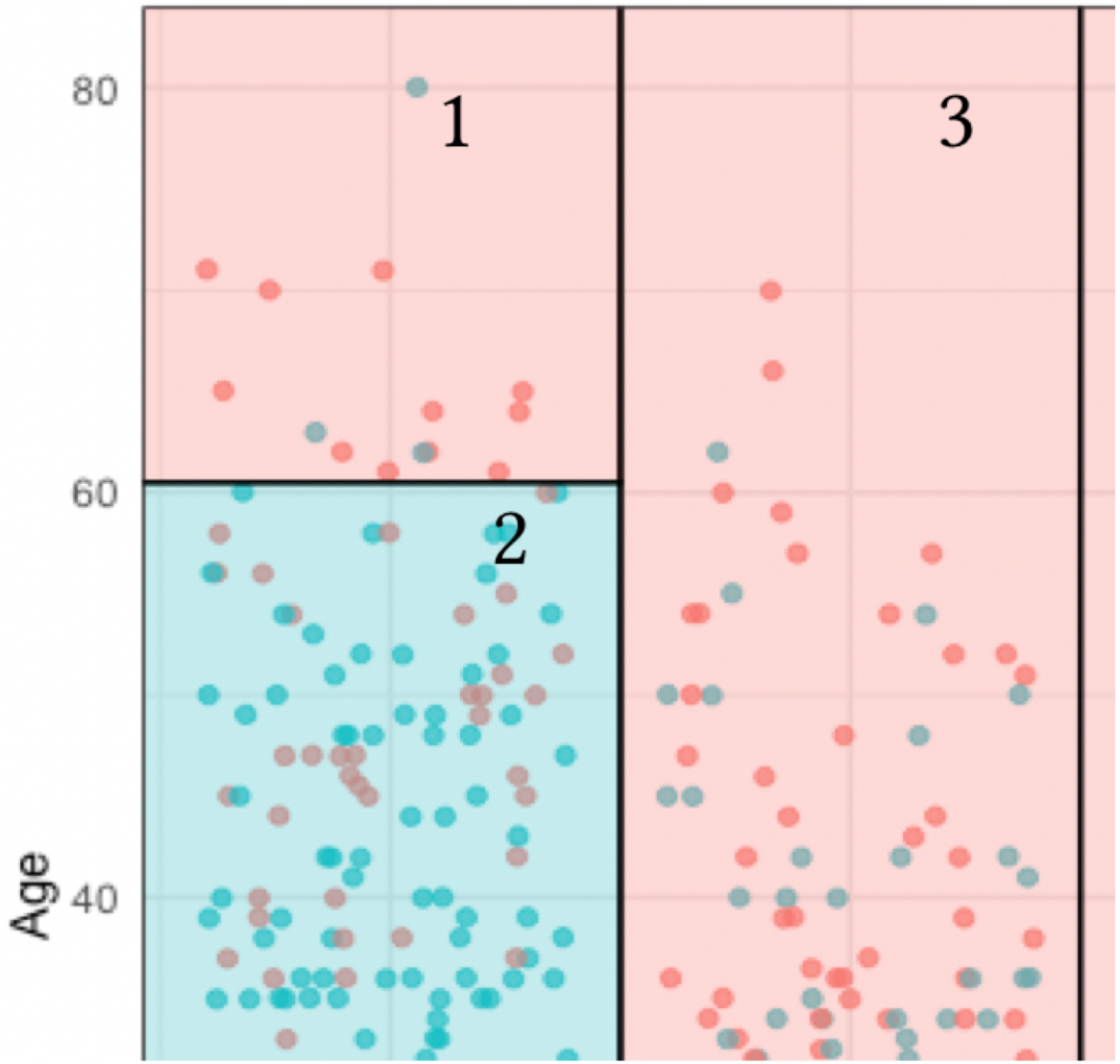
$$\hat{p}_{1,Died} = \frac{11}{14}$$



Misclassification error

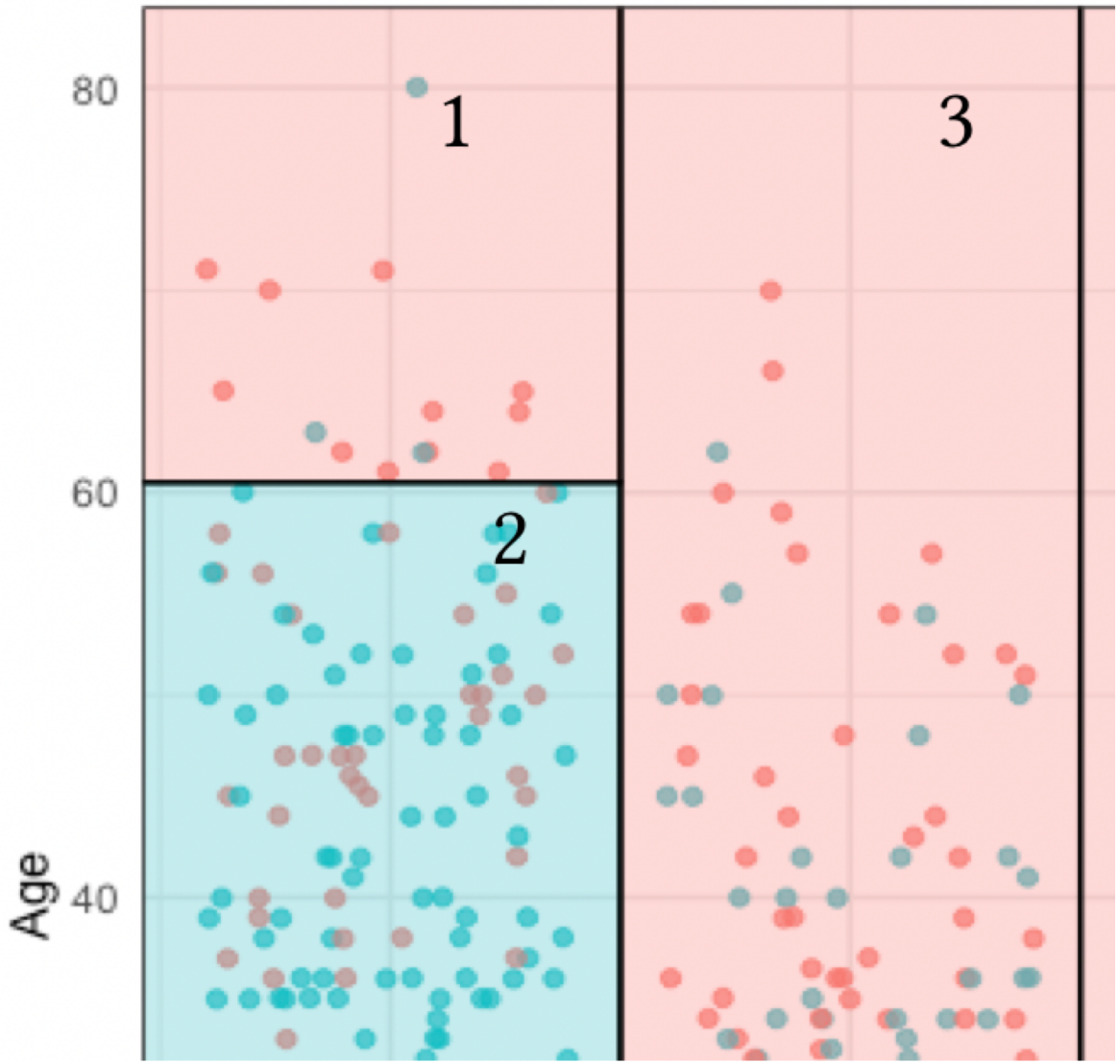
$$E_m = 1 - max_k(\hat{p}_{mk})$$

The proportion of observations in node m, which do not belong to the majority class k



Misclassification error in node 1

$$E_1 = \frac{3}{14}$$



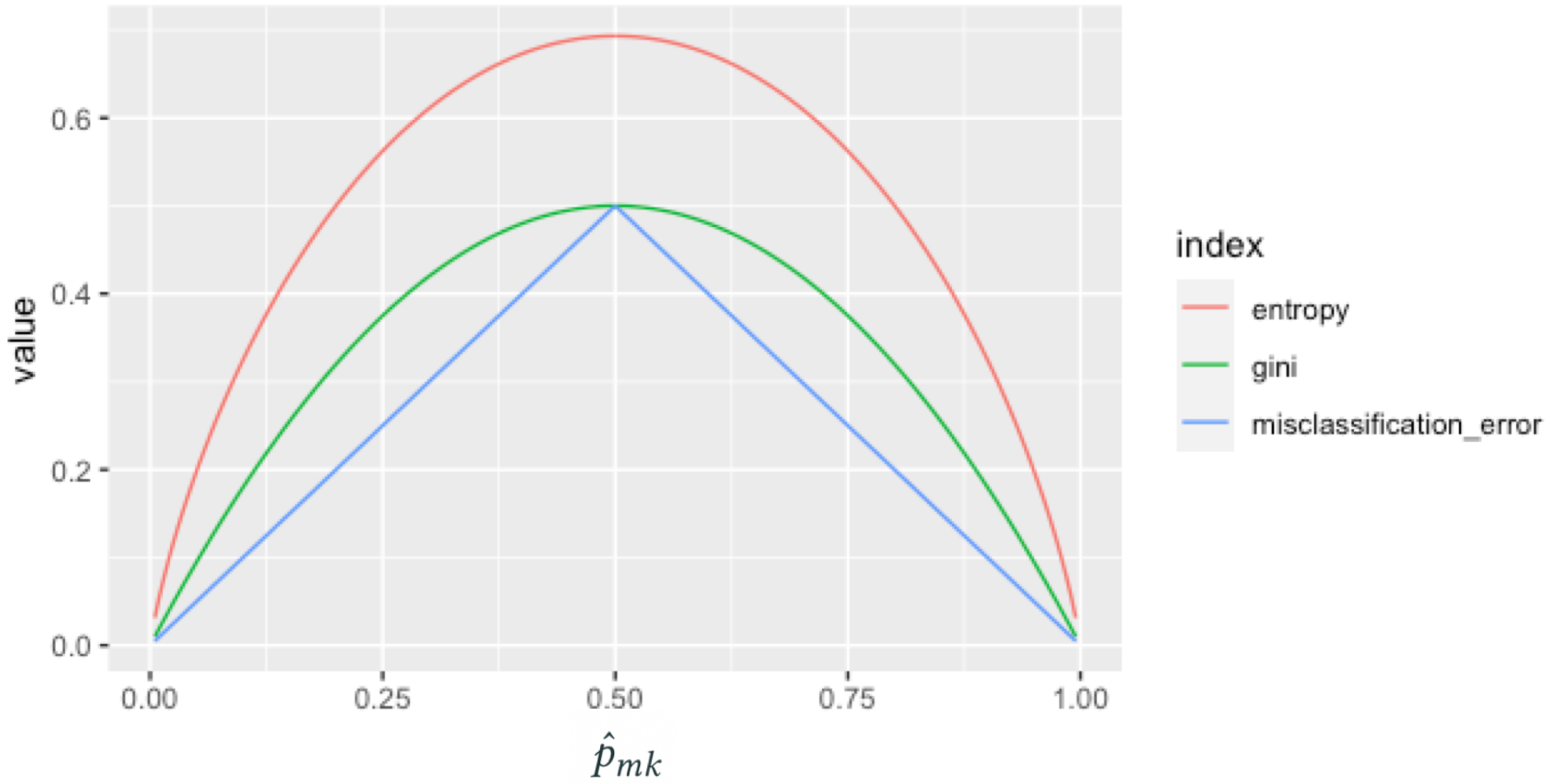
Gini index:

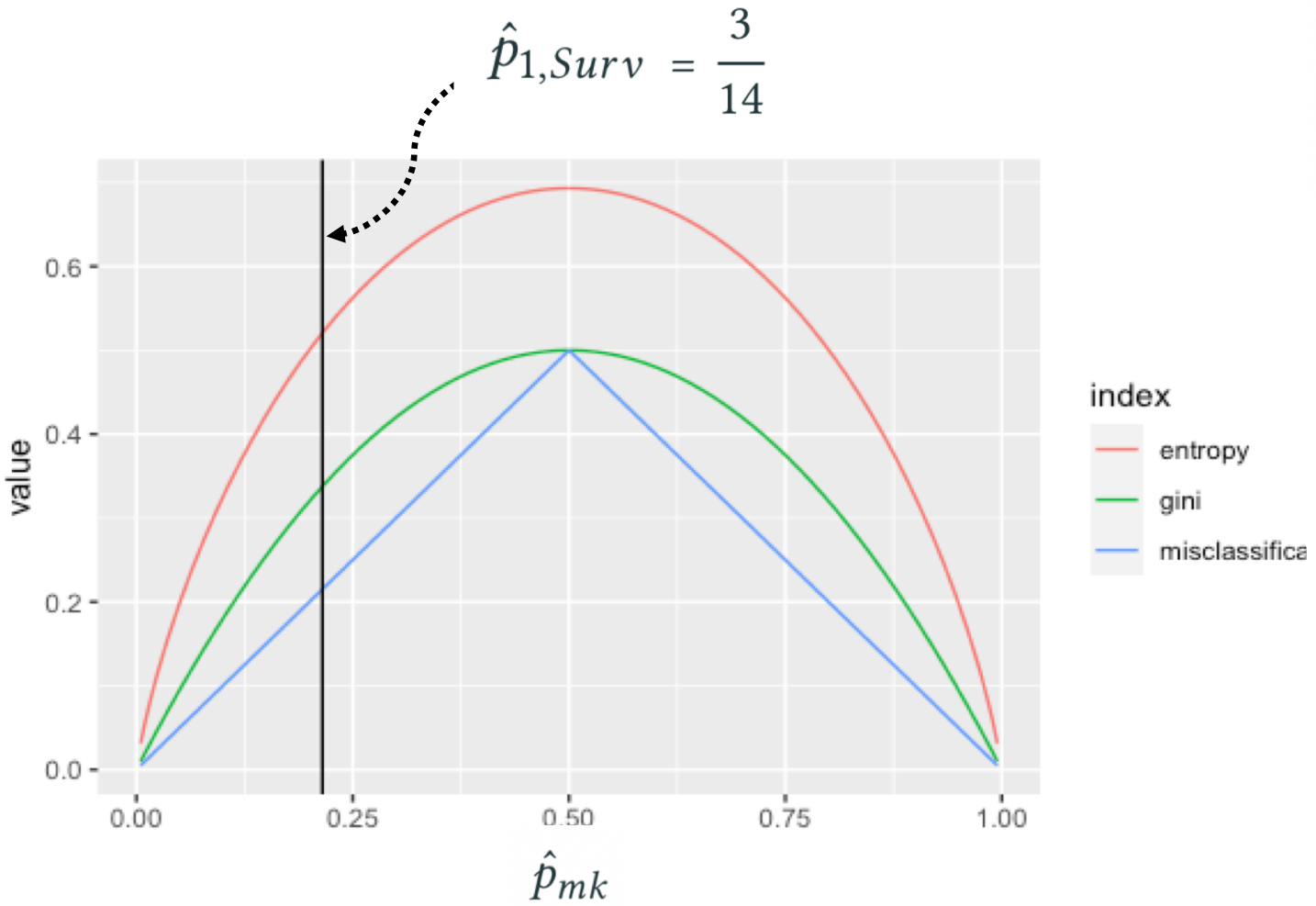
$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

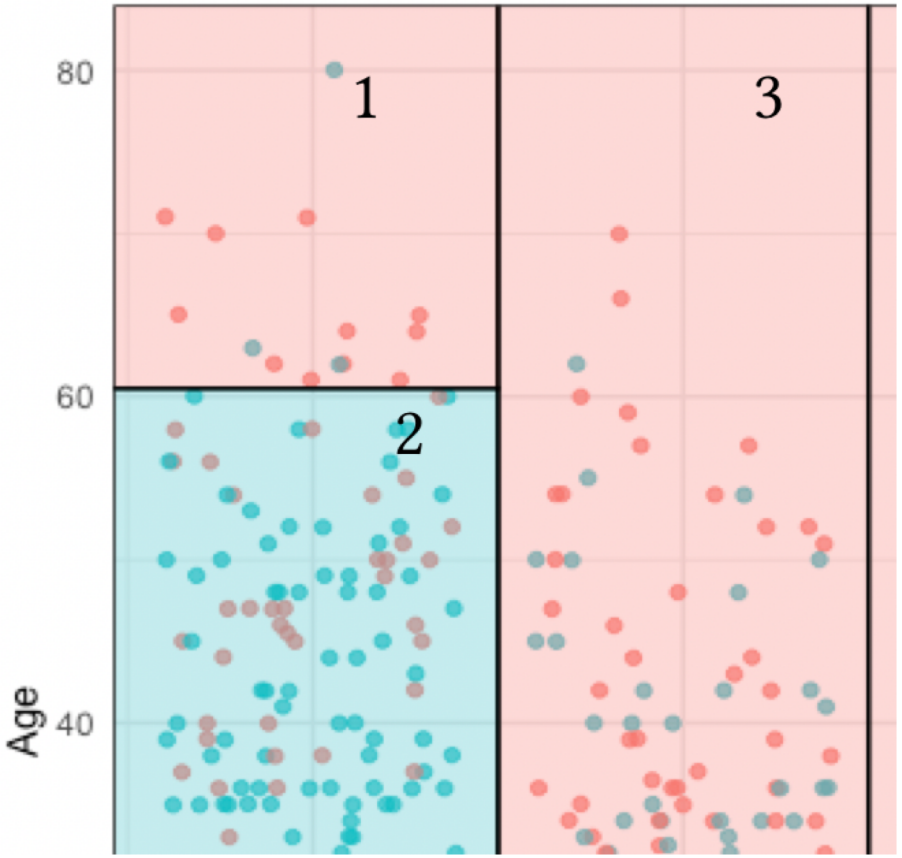
Cross entropy or deviance:

$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} log(\hat{p}_{mk})$$

If there are just two classes (e.g., survived, died)







https://madina-k.github.io/dse_mk2021

Class 1 Class 2

Data = (400, 400)

	Class 1	Class 2
Data =	(400, 400)	

	Node 1	Node 2
A split	(300, 100)	(100, 300)

	Class 1	Class 2
Data =	(400, 400)	

	Node 1	Node 2
A split	(300, 100)	(100, 300)

	Node 1	Node 2
Alternative split	(200, 400)	(200, 0)

	Class 1	Class 2
Data =	(400, 400)	

	Node 1	Node 2
A split	(300, 100)	(100, 300)

Misclassification error

	Node 1	Node 2
Alternative split	(200, 400)	(200, 0)

	Class 1	Class 2
Data =	(400, 400)	

	Node 1	Node 2
A split	(300, 100)	(100, 300)
	0.25	0.25

Misclassification error

	Node 1	Node 2
Alternative split	(200, 400)	(200, 0)

	Class 1	Class 2
Data =	(400, 400)	

	Node 1	Node 2
A split	(300, 100)	(100, 300)
	0.25	0.25

Misclassification error

	Node 1	Node 2
Alternative split	(200, 400)	(200, 0)
	0.33	0

	Class 1	Class 2
Data =	(400, 400)	

	Node 1	Node 2		
A split	(300, 100)	(100, 300)	Overall Misclassification error	0.25
	0.25	0.25		

	Node 1	Node 2		
Alternative split	(200, 400)	(200, 0)	Overall Misclassification error	0.25
	0.33	0		

	Class 1	Class 2
Data =	(400, 400)	

	Node 1	Node 2	
A split	(300, 100)	(100, 300)	Overall
	0.25	0.25	Misclassification error 0.25

Misclassification error index is indifferent between the two splits

	Node 1	Node 2	
Alternative split	(200, 400)	(200, 0)	Overall
	0.33	0	Misclassification error 0.25

	Class 1	Class 2
Data =	(400, 400)	

	Node 1	Node 2
A split	(300, 100)	(100, 300)
	0.375	0.375

Gini index

	Node 1	Node 2
Alternative split	(200, 400)	(200, 0)
	0.44	0

	Class 1	Class 2
Data =	(400, 400)	

	Node 1	Node 2		
A split	(300, 100)	(100, 300)	Overall Gini index	0.375
	0.375	0.375		

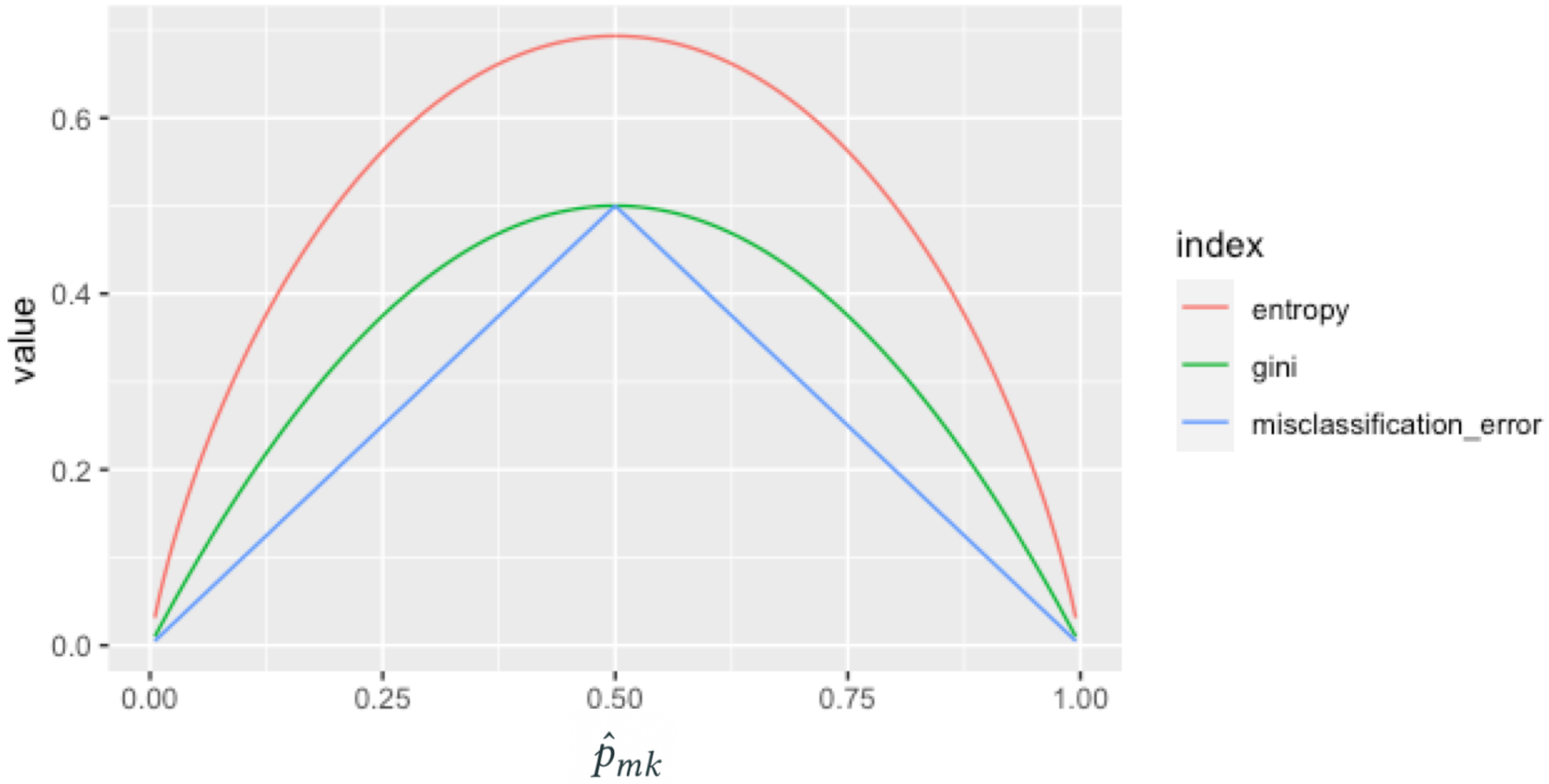
	Node 1	Node 2		
Alternative split	(200, 400)	(200, 0)	Overall Gini index	0.333
	0.44	0		

	Class 1	Class 2
Data =	(400, 400)	

	Node 1	Node 2	
A split	(300, 100)	(100, 300)	Overall Gini index
	0.375	0.375	0.375

Gini index prefers the second split, because it leads to one really pure node

	Node 1	Node 2	
Alternative split	(200, 400)	(200, 0)	Overall Gini index
	0.44	0	0.333


value
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
$\hat{p}_{mk}$
index
entropy
gini
misclassification_error