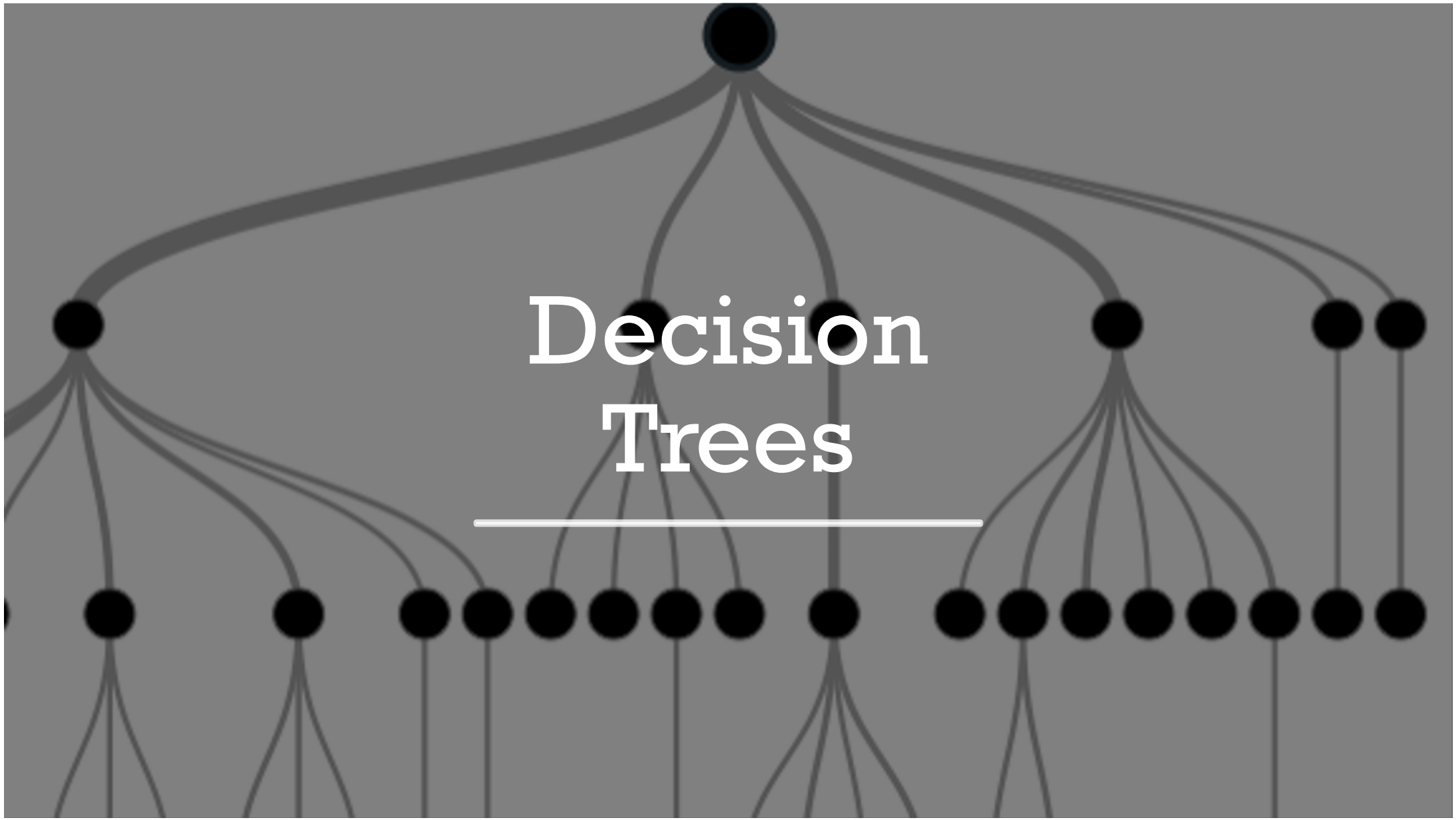
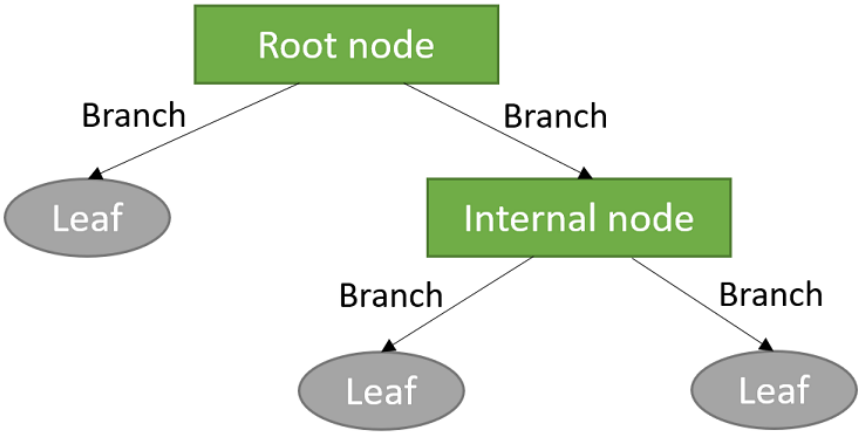

Decision
Trees

Contents

What is a *decision tree*?

A decision tree is a kind of **supervised learning** algorithm used for classification and regression problems. It creates a tree-like model of decisions built by **recursively splitting** the data into subsets based on the values of the features.



Types of decision trees

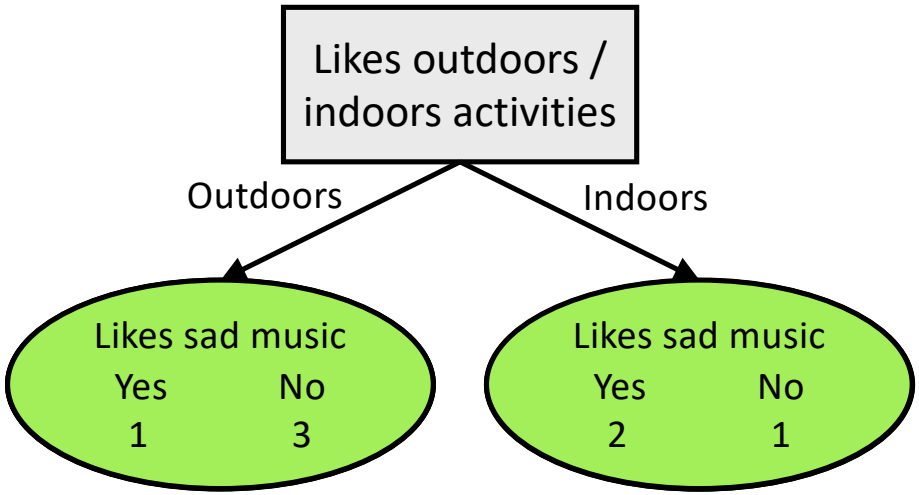
- **Classification tree:** the target variable is discrete. The algorithm identifies in which of the possible classes the target variable is most likely to fall.

- **Regression tree:** the response variable is continuous. The tree is used to predict its value.

Classification trees

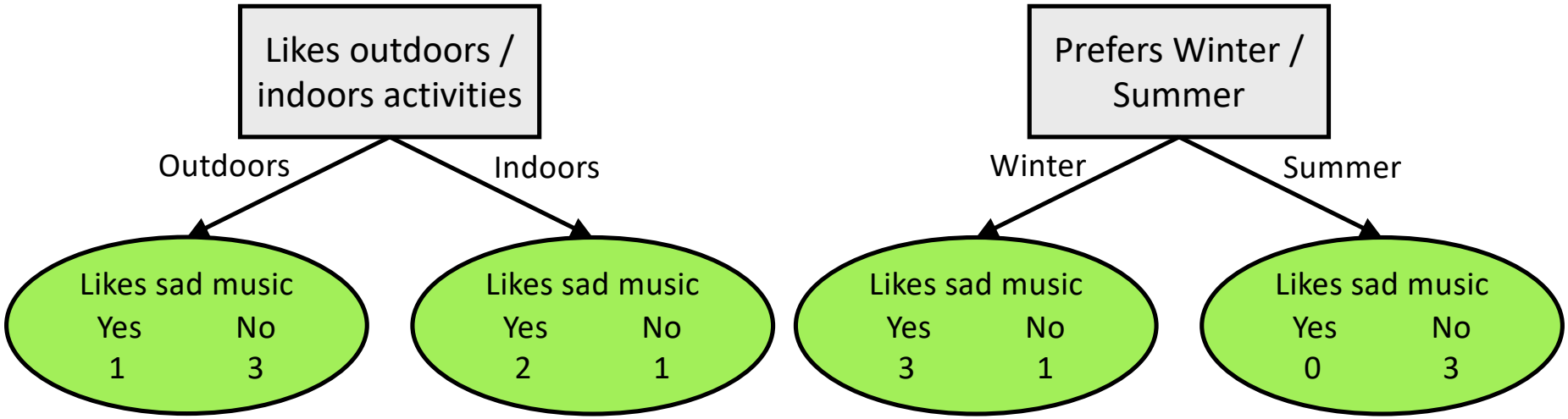
Likes outdoor/indoor activities (O/I)	Prefers Winter/Summer (W/S)	Age	Likes happy / sad music (H/S)
O	S	7	H
O	W	12	H
I	S	18	S
I	S	35	S
O	S	38	S
O	W	50	H
I	W	83	H

Classification trees

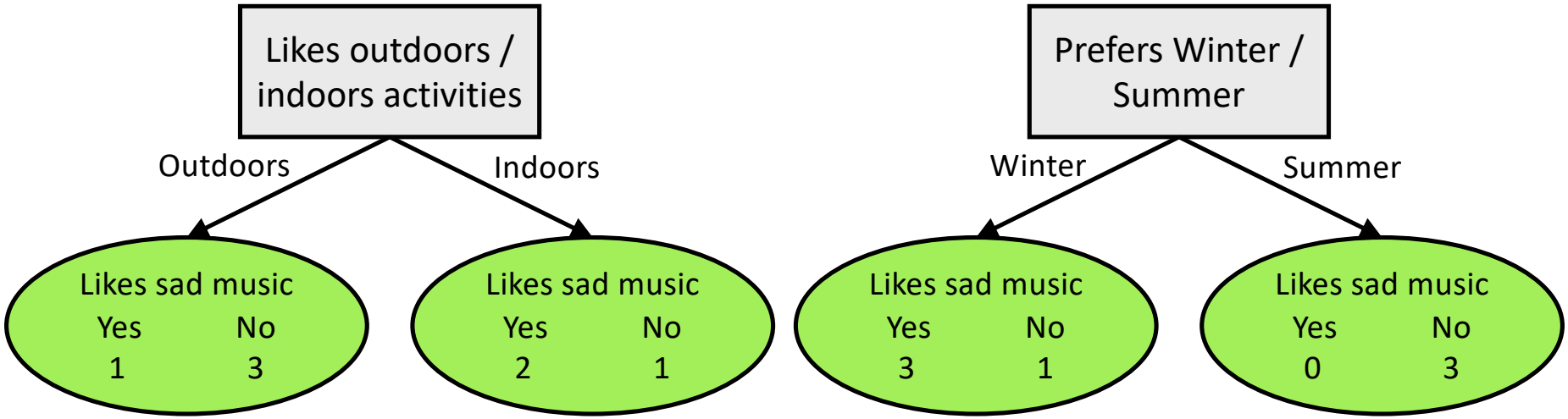


Likes outdoor/indoor activities (O/I)	Prefers Winter/Summer (W/S)	Age	Likes happy / sad music (H/S)
O	S	7	H
O	W	12	H
I	S	18	S
I	S	35	S
O	S	38	S
O	W	50	H
I	W	83	H

Classification trees



Classification trees



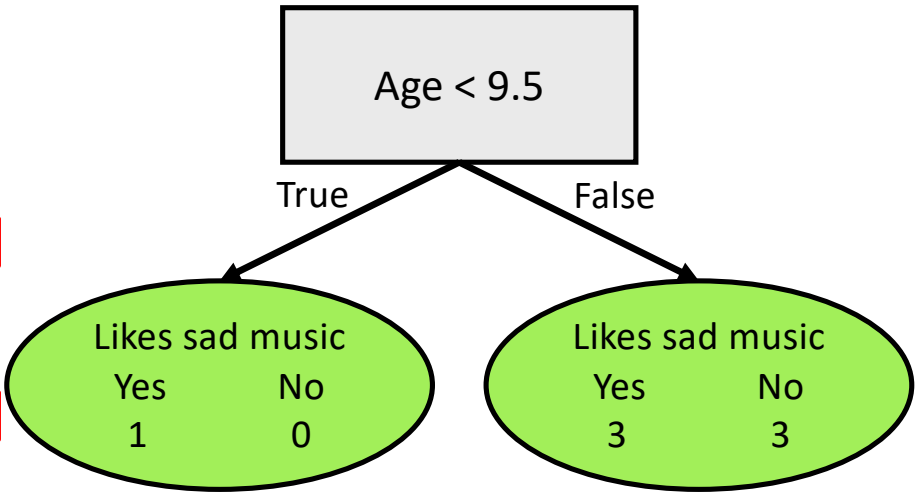
$G = 0.375$ $G = 0.444$ $G = 0.375$ $G = 0$

$G_W = 0.405$ $G_W = 0.214$

$G_W = 1 - \left(\frac{4}{4+3}\right)^2 - \left(\frac{3}{4+3}\right)^2 = 0.245$ $G_W = \sum_{i=1}^{N} \left(\frac{n_i}{n}\right)^2 \cdot G = 0.405$

Classification trees

Age		Likes happy / sad music (H/S)	
7	9.5	S	$G_W = 0.429$
12	15	S	$G_W = 0.343$
18	26.5	H	$G_W = 0.476$
35	36.5	H	$G_W = 0.476$
38	44	H	$G_W = 0.343$
50	66.5	S	$G_W = 0.429$
83		S	



$G = 0$

$G = 0.5$

$G_W = 0.429$

Classification trees

Likes outdoors / indoors activities

$G_W = 0.405$

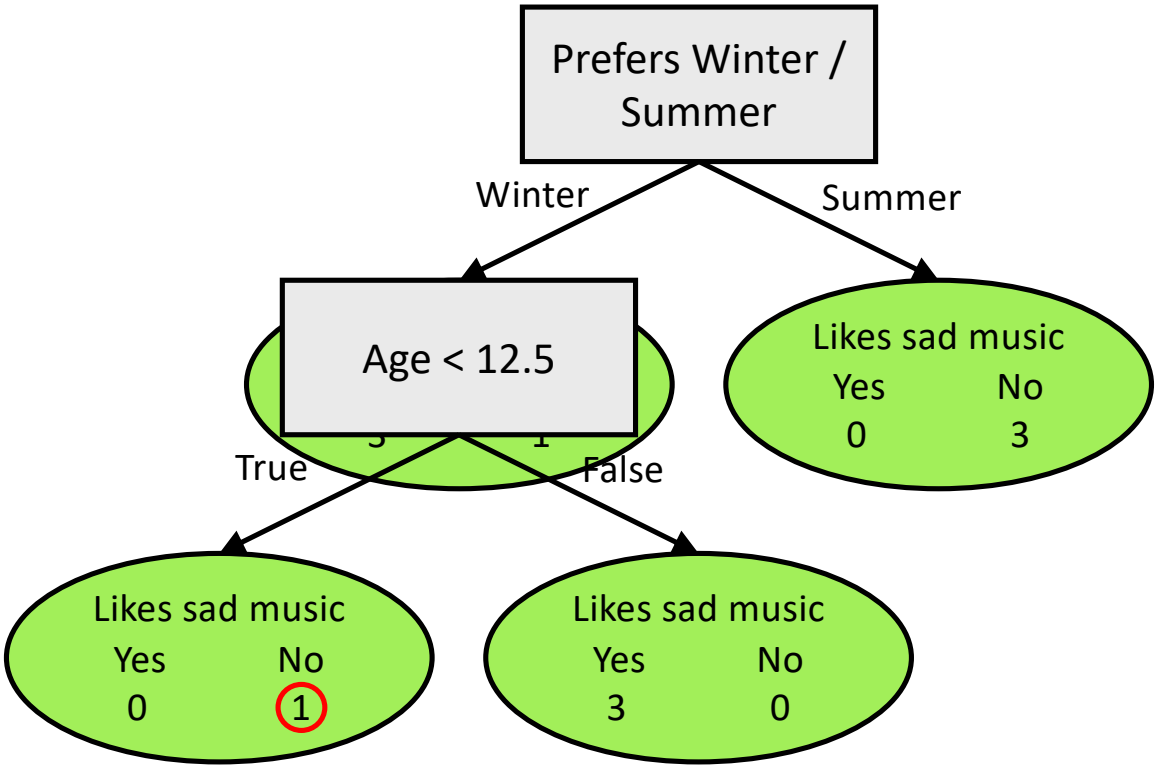
Prefers Winter / Summer

$G_W = 0.214$

Age < 9.5

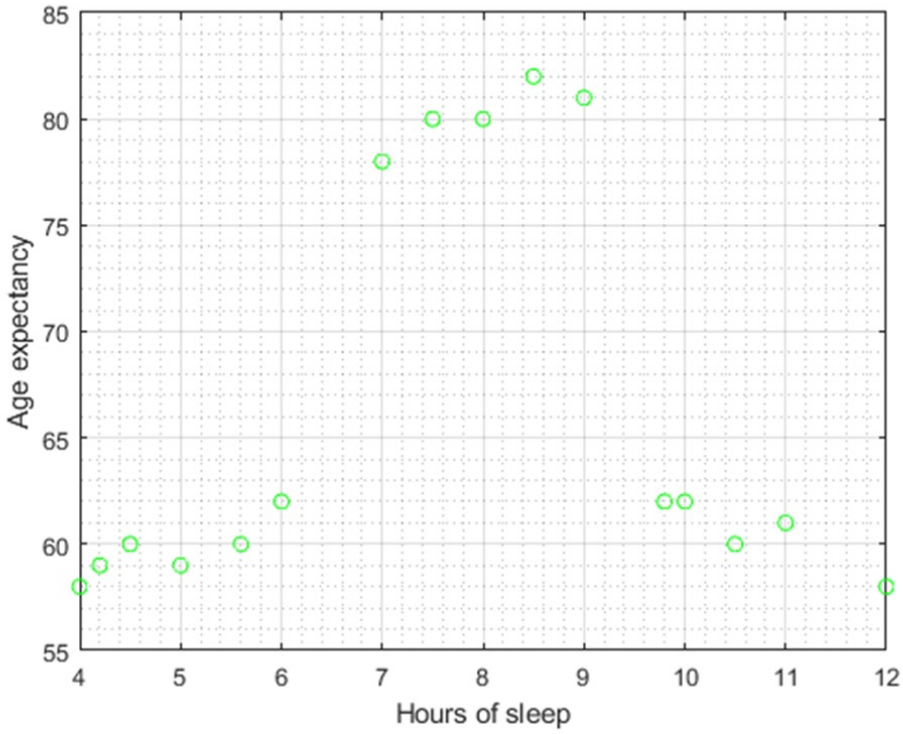
$G_W = 0.343$

Classification trees



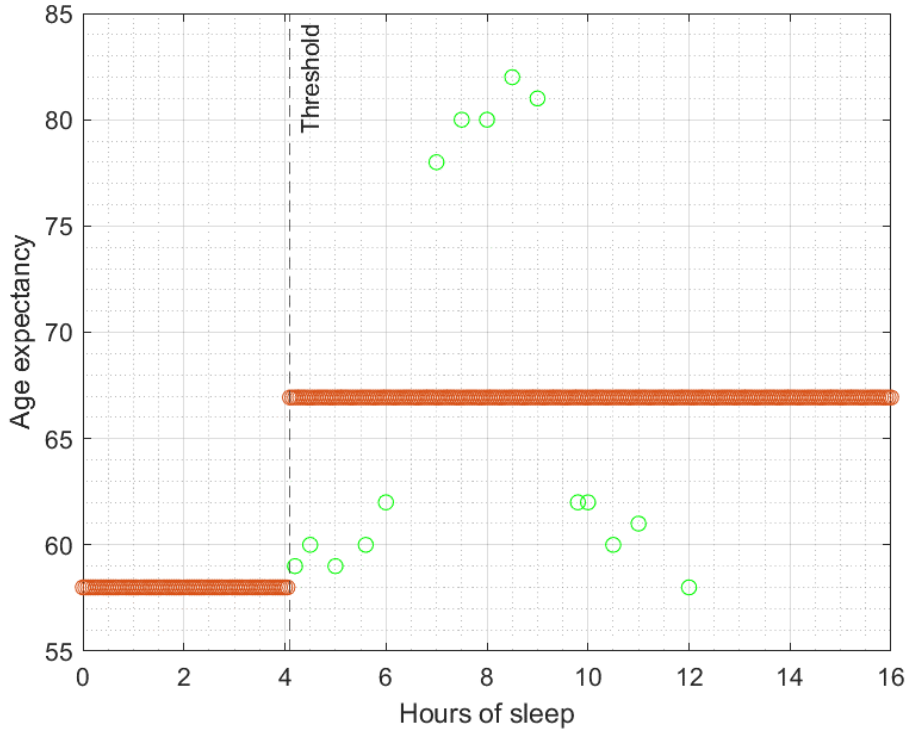
Regression trees

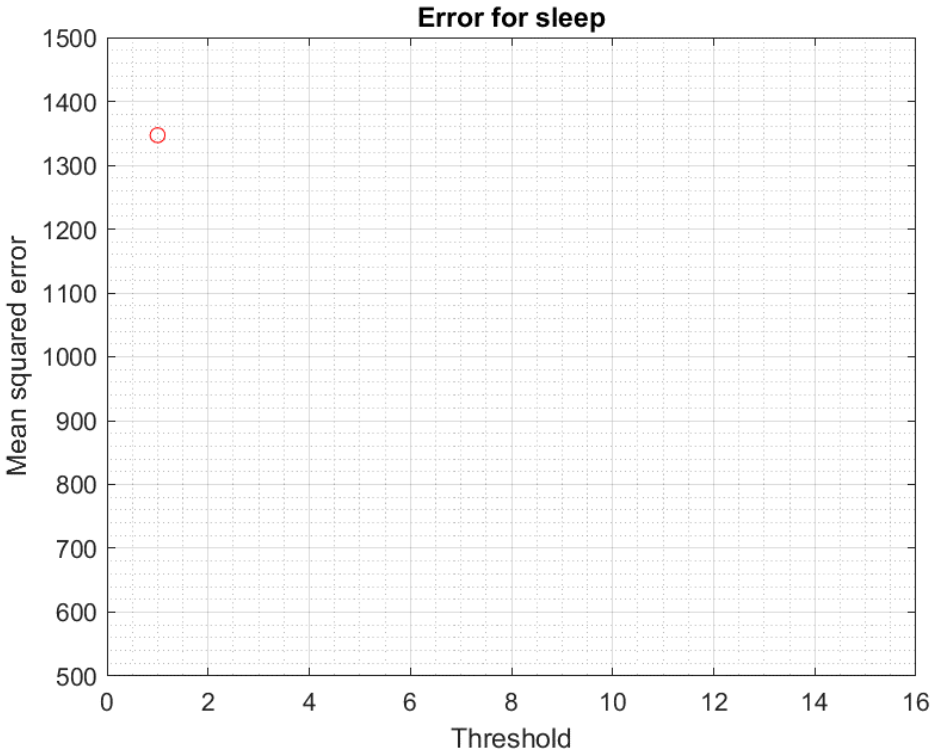
- They are used when it is not interesting to approximate the distribution to a continuous function (linear, quadratic, exponential…), since there are **sudden changes** under certain **thresholds**.
- The **output** is **continuous** so there are no classes.



Regression trees

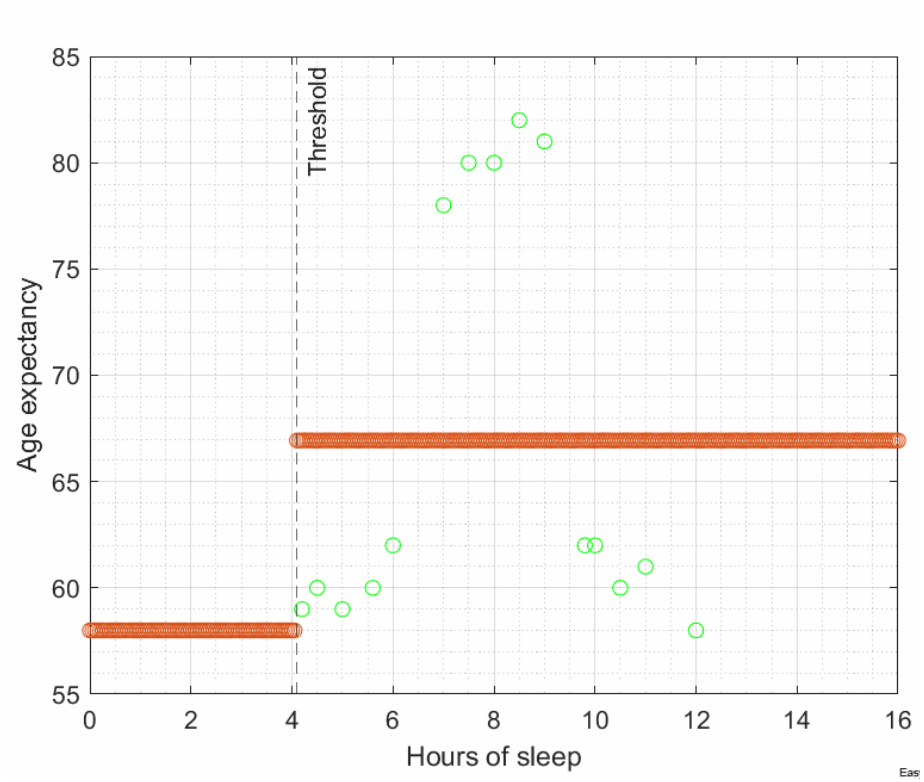
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

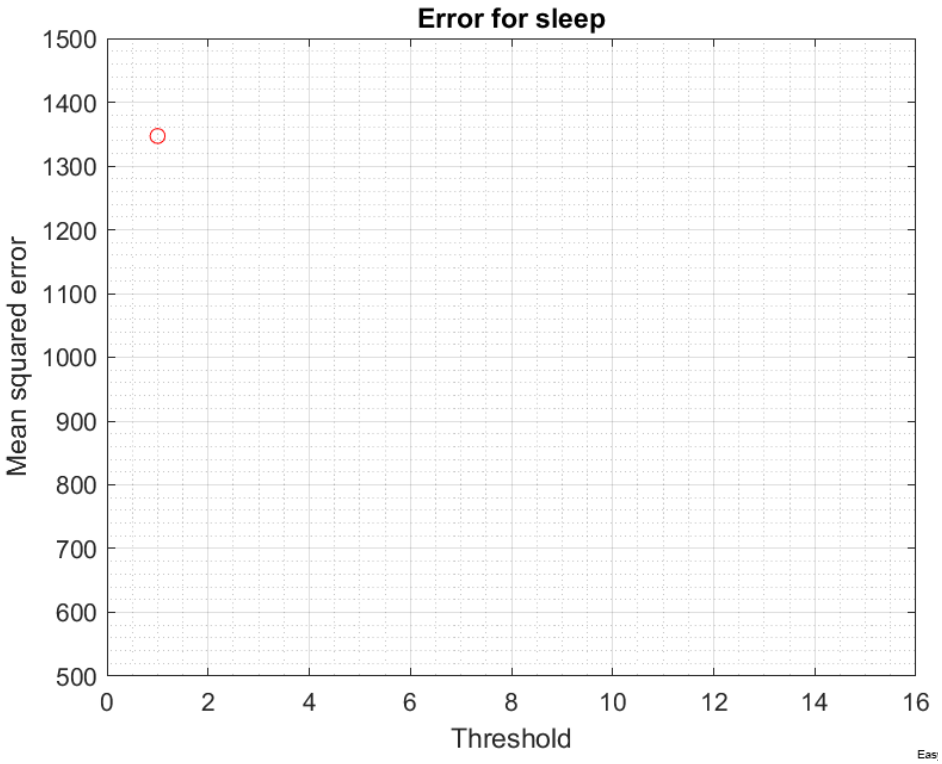




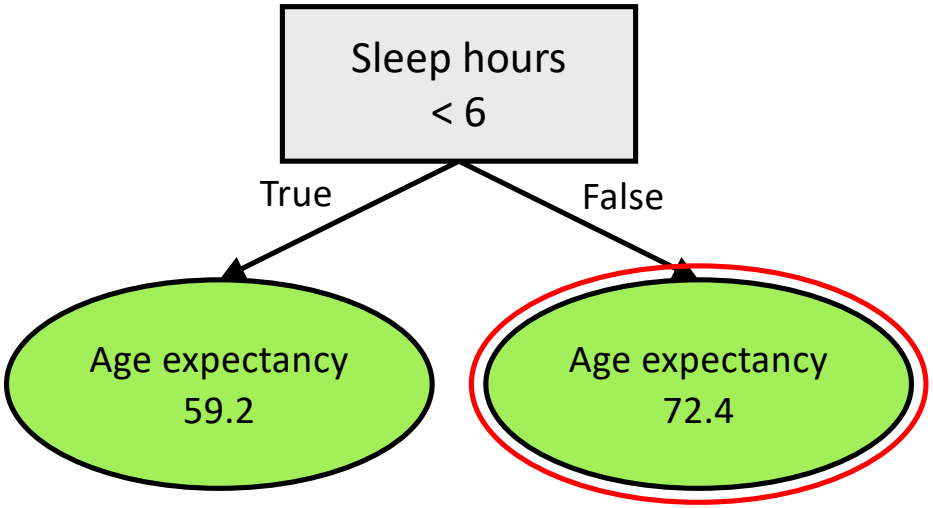
Regression trees

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$



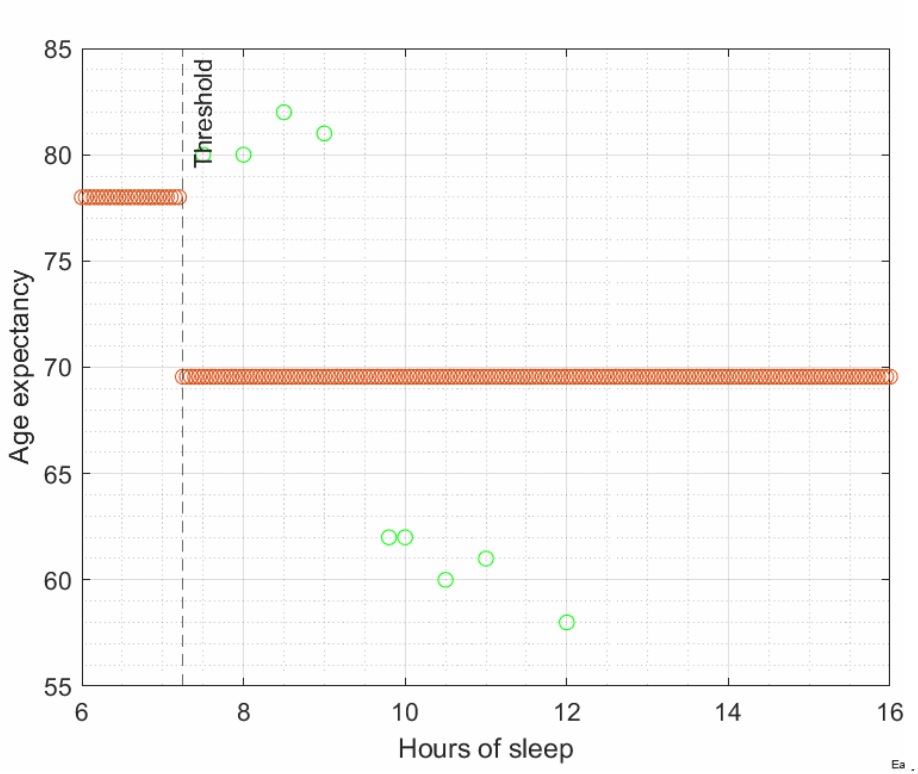


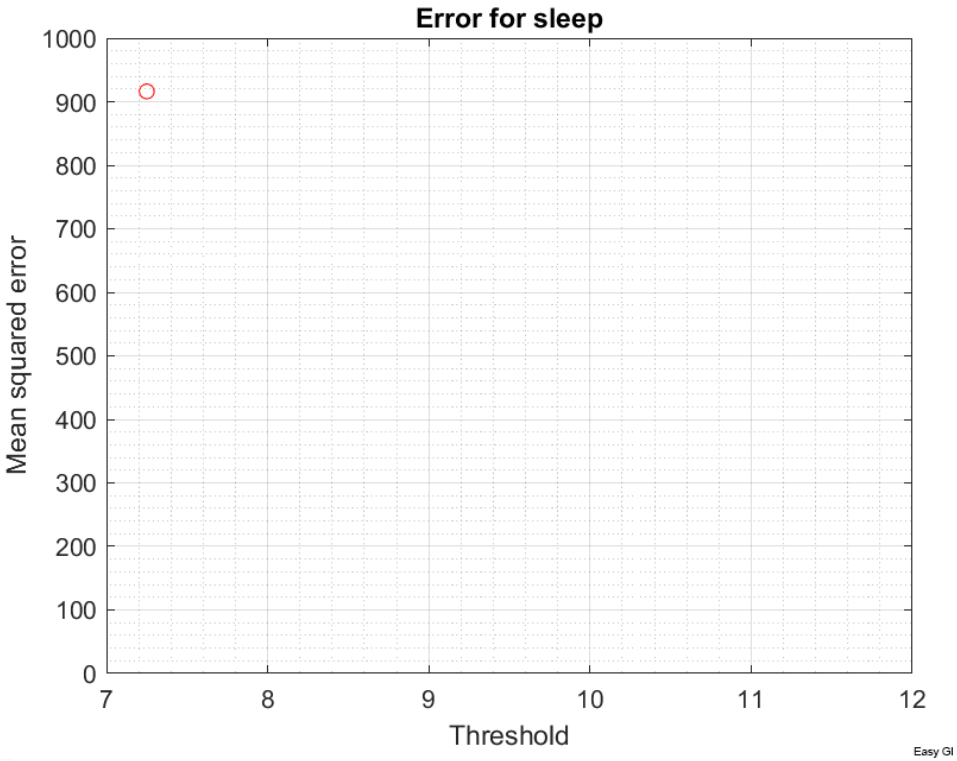
Regression trees



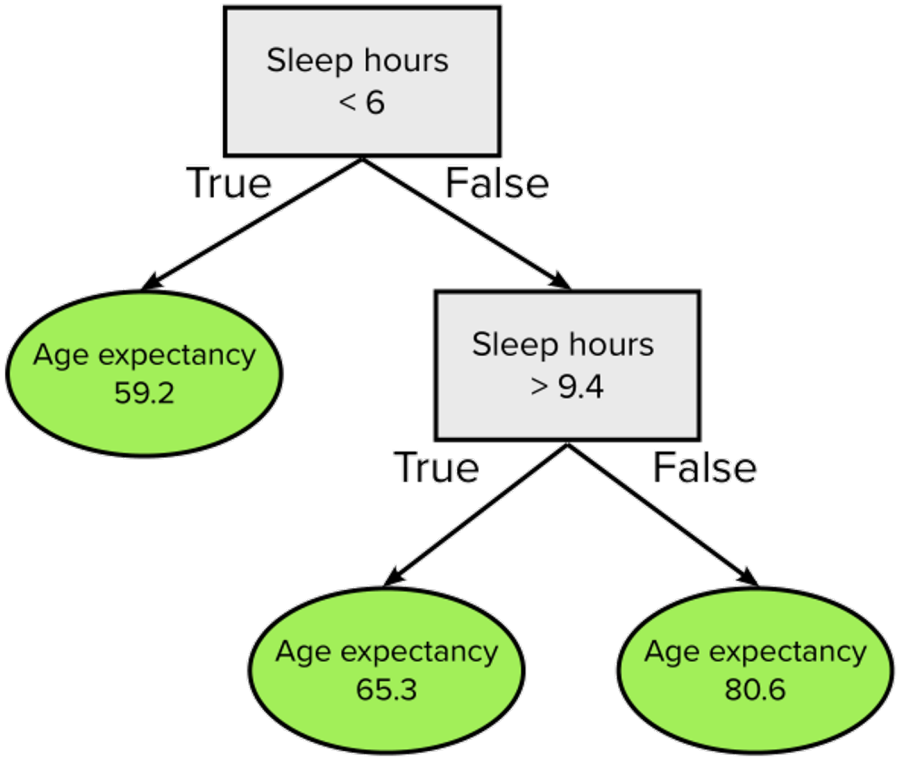
Regression trees

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

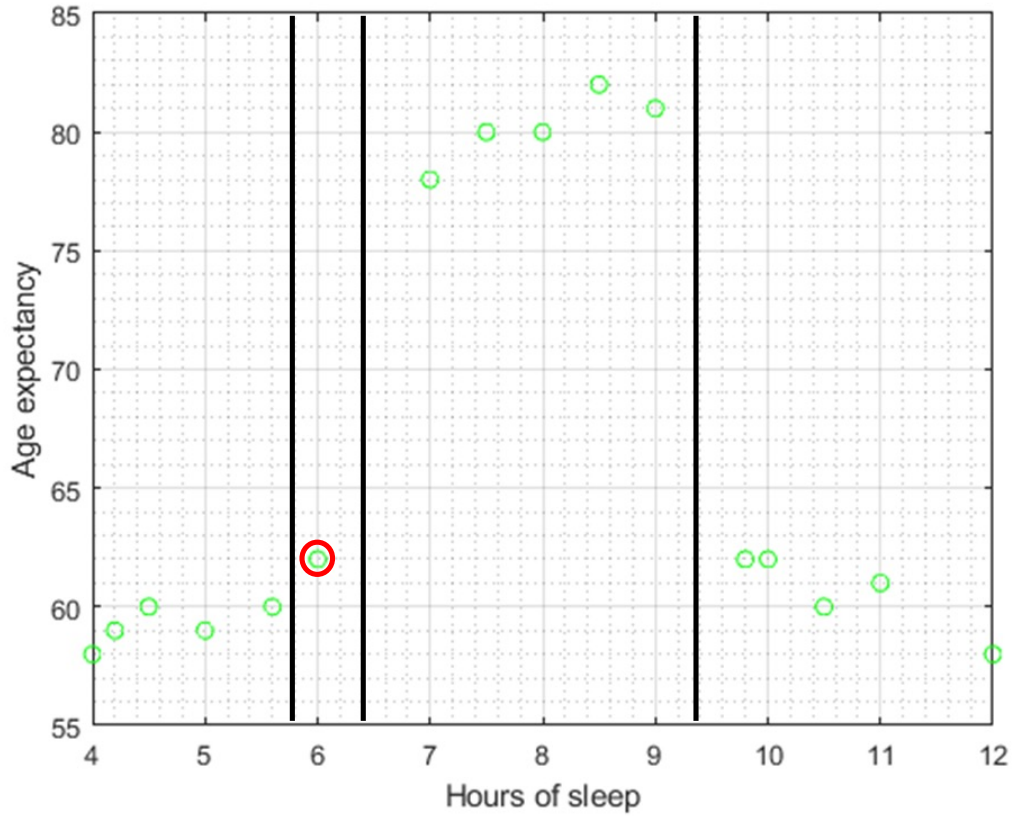




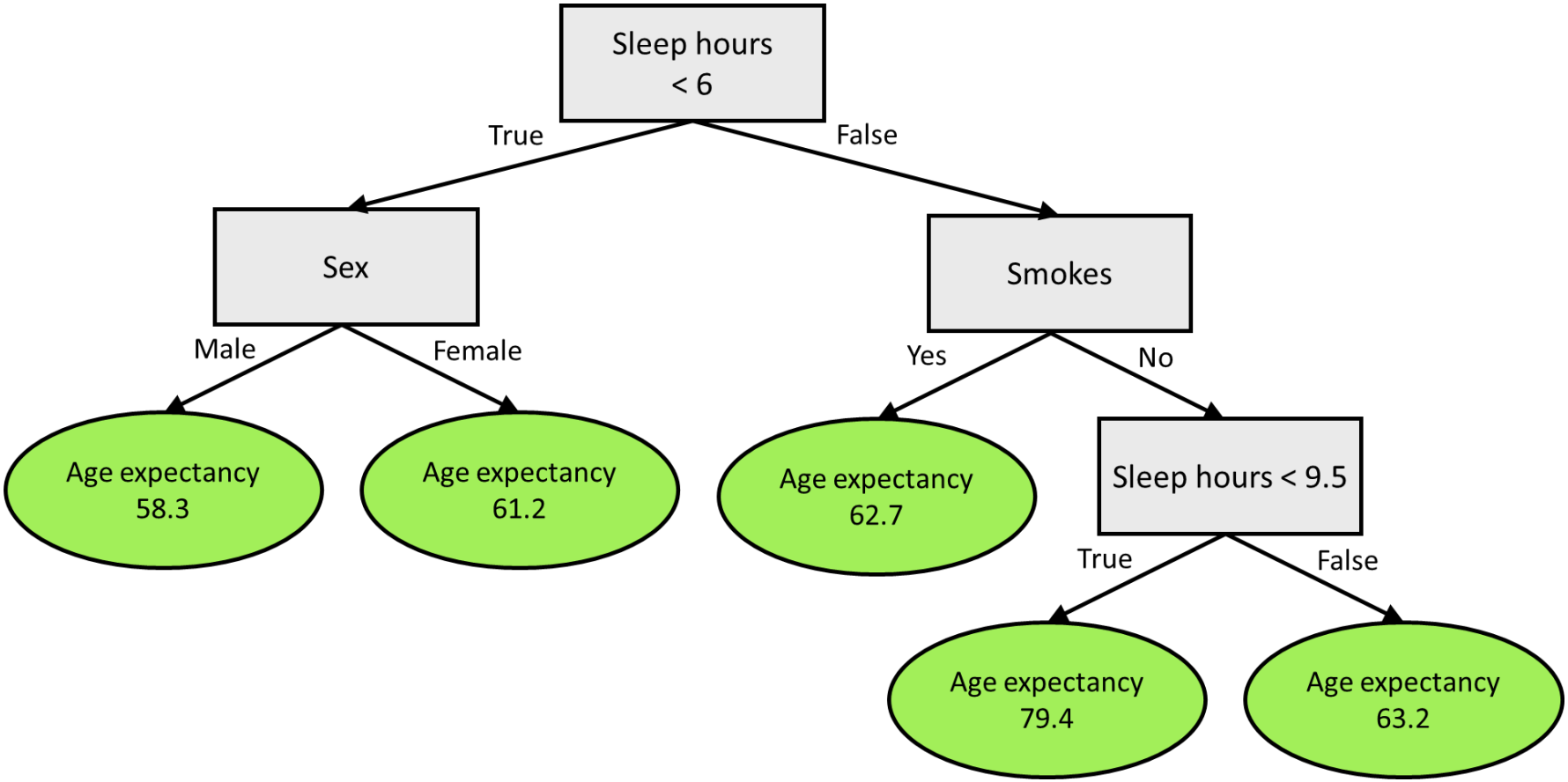
Regression trees



Regression trees



Regression trees



Advantages & disadvantages

Advantages

- Decision trees are highly interpretable.
- It is easy to know how the decision is taken, just follow the tree.
- Easy to check for ethical biases.
- Good for high level features.

Disadvantages

- Unstable algorithm: small changes in the training data can lead to big changes on the model.

Music applications

Feature #	Description
1	Relative amplitude of the first histogram peak
2	Relative amplitude of the second histogram peak
3	Ratio between the amplitudes of the second peak and the first peak
4	Period of the first peak in bpm
5	Period of the second peak in bpm
6	Overall histogram sum (beat strength)
7	Spectral centroid mean
8	Spectral rolloff mean
9	Spectral flow mean
10	Zero crossing rate mean
11	Standard deviation for spectral centroid
12	Standard deviation for spectral rolloff
13	Standard deviation for spectral flow
14	Standard deviation for zero crossing rate
15	Low energy
16	First MFCC mean
17	Second MFCC mean
18	Third MFCC mean
19	Fourth MFCC mean
20	Fifth MFCC mean
21	Standard deviation for first MFCC
22	Standard deviation for second MFCC
23	Standard deviation for third MFCC
24	Standard deviation for fourth MFCC
25	Standard deviation for fifth MFCC
26	The overall sum of the histogram (pitch strength)
27	Period of the maximum peak of the unfolded histogram
28	Amplitude of maximum peak of the folded histogram
29	Period of the maximum peak of the folded histogram
30	Pitch interval between the two most prominent peaks of the folded histogram

Table 8: Confusion Matrix for Model 1

Genre	a	b	c	d	e	f	g	h	i	j
Tango = a	60	0	0	0	0	0	0	0	0	0
Bachata = b	0	54	0	4	2	0	0	0	0	0
Bolero = c	1	0	33	0	5	3	7	9	2	0
Merengue = d	0	2	0	46	1	2	0	4	1	4
Salsa = e	0	4	0	3	41	3	1	3	3	2
Forró = f	0	3	4	4	6	21	9	9	4	0
Pagode = g	0	1	4	1	1	3	37	8	2	3
Sertanejo = h	0	0	8	0	6	1	3	29	7	6
Gaúcha = i	4	1	3	1	3	2	4	7	26	9
Axé = j	1	1	0	5	4	2	3	10	4	30

References

[1] “CART (Classification And Regression Tree) in Machine Learning,” *GeeksforGeeks*, Sep. 23, 2022. *https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/*

[2] scikit learn, “1.10. Decision Trees — scikit-learn 0.22 documentation,” *Scikit-learn.org*, 2009. *https://scikit-learn.org/stable/modules/tree.html*

[3] StatQuest with Josh Starmer. “Decision and Classification Trees, Clearly Explained!!!” *YouTube, Apr. 26, 2021. Available: https://www.youtube.com/watch?v=_L39rN6gz7Y*

[4] StatQuest with Josh Starmer. “Regression Trees, Clearly Explained!!!” *YouTube, Aug. 20, 2019. Available: https://www.youtube.com/watch?v=g9c66TUylZ4*

[5] G. M. Bressan, B. C. F. de Azevedo, and E. Ap. S. Lizzi, “A Decision Tree Approach for the Musical Genres Classification,” *Applied Mathematics & Information Sciences*, vol. 11, no. 6, pp. 1703–1713, Nov. 2017, doi: 10.18576/amis/110617.

[6] C. N. Silla, A. L. Koerich, and C. A. A. Kaestner, “A Feature Selection Approach For Automatic Music Genre Classification,” *International Journal of Semantic Computing*, vol. 03, no. 02, pp. 183–208, Jun. 2009, doi: 10.1142/s1793351x09000719.