

Exercise 1

Build a decision tree with Misclassification Error; compute confusion matrix and various metrics on the train & test

Training set

Altezza	Peso	Età	Sesso	Malattia
Bassa	Alto	Giovane	F	No
Bassa	Basso	Giovane	F	Si
Bassa	Basso	Anziano	M	No
Bassa	Medio	Giovane	M	Si
Bassa	Alto	Giovane	M	Si
Alta	Medio	Anziano	F	No
Bassa	Alto	Giovane	F	No
Alta	Basso	Anziano	M	Si
Alta	Basso	Anziano	M	Si
Bassa	Medio	Anziano	M	Si

Test set

Altezza	Peso	Età	Sesso	Malattia
Alta	Alto	Anziano	F	No
Bassa	Alto	Anziano	F	No
Bassa	Basso	Giovane	F	No
Alta	Basso	Giovane	M	Si
Alta	Medio	Anziano	M	Si
Alta	Basso	Giovane	M	No
Alta	Medio	Anziano	F	No
Alta	Medio	Anziano	F	No

Compute Gain for all possible splits of the root

New grid

Altezza	Bassa	Alta
class=Si	4	2
class=No	3	1

$$\begin{aligned} \text{ME}(\text{Altezza}) &= 7/10 * \text{ME}(\text{Altezza}=\text{Bassa}) + \\ & 3/10 * \text{ME}(\text{Altezza}=\text{Alta}) \\ &= 4 / 10 = \mathbf{0.40} \end{aligned}$$

$$\text{GAIN}(\text{Altezza}) = \text{ME}(\text{root}) - \text{ME}(\text{Altezza}) = 0$$

New grid

Peso	Basso	Medio	Alto
class=Si	3	2	1
class=No	1	1	2

$$\text{ME}(\text{Peso}) = 3 / 10 = \mathbf{0.30}$$

$$\text{GAIN}(\text{Peso}) = \text{ME}(\text{root}) - \text{ME}(\text{Peso}) = 0.1$$

New grid

Sesso	F	M
class=Si	1	5
class=No	3	1

$$\text{ME}(\text{Sesso}) = 2 / 10 = \mathbf{0.20}$$

$$\text{GAIN}(\text{Età}) = \text{ME}(\text{root}) - \text{ME}(\text{Età}) = 0.20$$

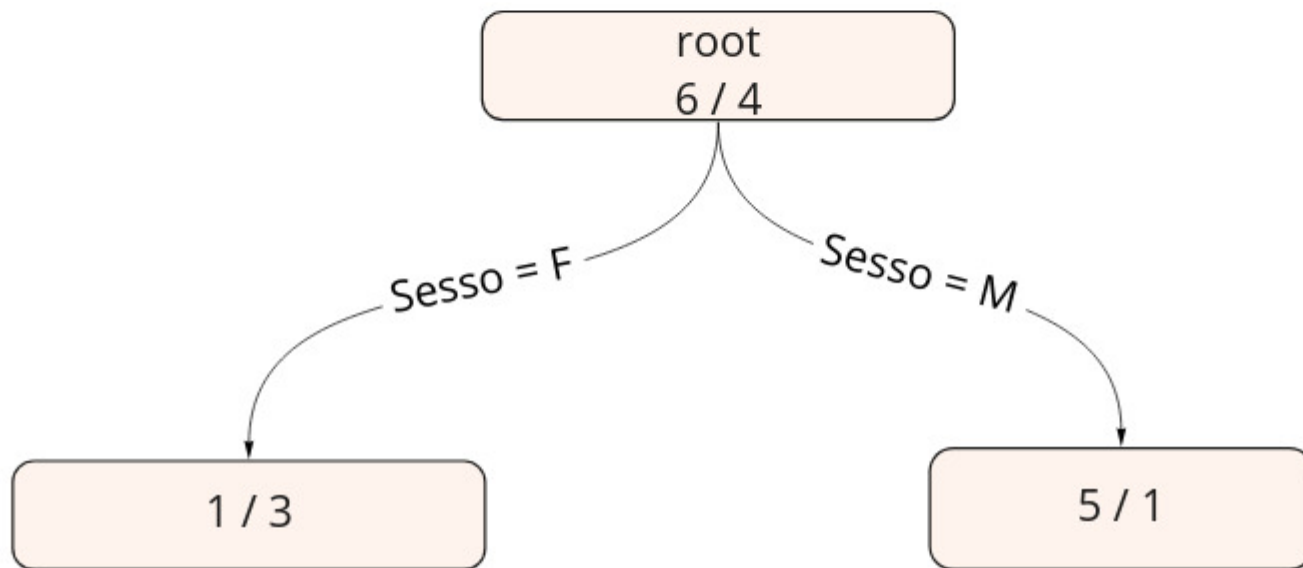
New grid

Età	Giovane	Anziano
class=Si	3	3
class=No	2	2

$$\text{ME}(\text{Età}) = 4 / 10 = \mathbf{0.40}$$

$$\text{GAIN}(\text{Età}) = \text{ME}(\text{root}) - \text{ME}(\text{Età}) = 0$$

Repeat for all nodes and possible splits



Altezza	Peso	Età	Sesso	Malattia
Bassa	Alto	Giovane	F	No
Bassa	Basso	Giovane	F	Si
Bassa	Basso	Anziano	F	No
Bassa	Medio	Giovane	F	No
Alta	Alto	Giovane	F	No
Alta	Medio	Anziano	F	No
Bassa	Alto	Giovane	F	No
Bassa	Basso	Anziano	F	No
Bassa	Basso	Anziano	F	No
Bassa	Basso	Anziano	F	No

Altezza	Peso	Età	Sesso	Malattia
Bassa	Basso	Anziano	F	No
Bassa	Medio	Giovane	F	Si
Bassa	Alto	Giovane	F	Si
Bassa	Basso	Anziano	F	No
Alta	Basso	Anziano	F	Si
Alta	Basso	Anziano	F	Si
Bassa	Medio	Anziano	F	Si

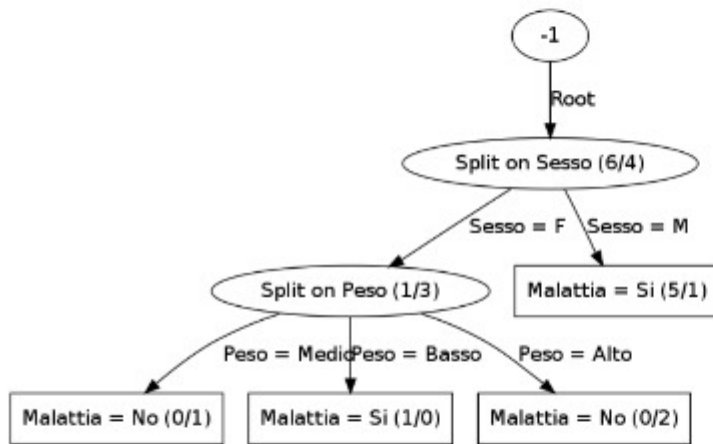
New table

ATTRIBUTE	VALUE 1	VALUE 2
class=Si		
class=No		



Compute performance metrics

Training set



Altezza	Peso	Età	Sesso	Malattia
Bassa	Alto	Giovane	F	No
Bassa	Basso	Giovane	F	Si
Bassa	Basso	Anziano	M	No
Bassa	Medio	Giovane	M	Si
Bassa	Alto	Giovane	M	Si
Alta	Medio	Anziano	F	No
Bassa	Alto	Giovane	F	No
Alta	Basso	Anziano	M	Si
Alta	Basso	Anziano	M	Si
Bassa	Medio	Anziano	M	Si

Test set

Altezza	Peso	Età	Sesso	Malattia
Alta	Alto	Anziano	F	No
Bassa	Alto	Anziano	F	No
Bassa	Basso	Giovane	F	No
Alta	Basso	Giovane	M	Si
Alta	Medio	Anziano	M	Si
Alta	Basso	Giovane	M	No
Alta	Medio	Anziano	F	No
Alta	Medio	Anziano	F	No

New grid

Confusion Matrix on Training set	Predicted = YES	Predicted = NO
Real = YES	6	0
Real = NO	1	3

New grid

Confusion Matrix on Test set	Predicted = YES	Predicted = NO
Real = YES	2	0
Real = NO	2	4

	Training	Test
Accuracy:	0.9	6/8 = 0.75
Precision:	6/7	2/4 = 0.50
Recall:	6/6 = 1.0	2/2 = 1.0
F-measure:	$2 * 6/7 * 1 / (1 + 6/7)$ $2 * 0.5 / 1.5 = 0.66$	