Utiliser le gradient boosting pour prédire une variable continue

↳ Gradient boost for regression

Gradient boost for regression ≠ linear regression

1er étape : $\boxed{\text{Moyenne des valeurs}}$

↳ Création d'un premier arbre

↳ Amélioration de l'arbre

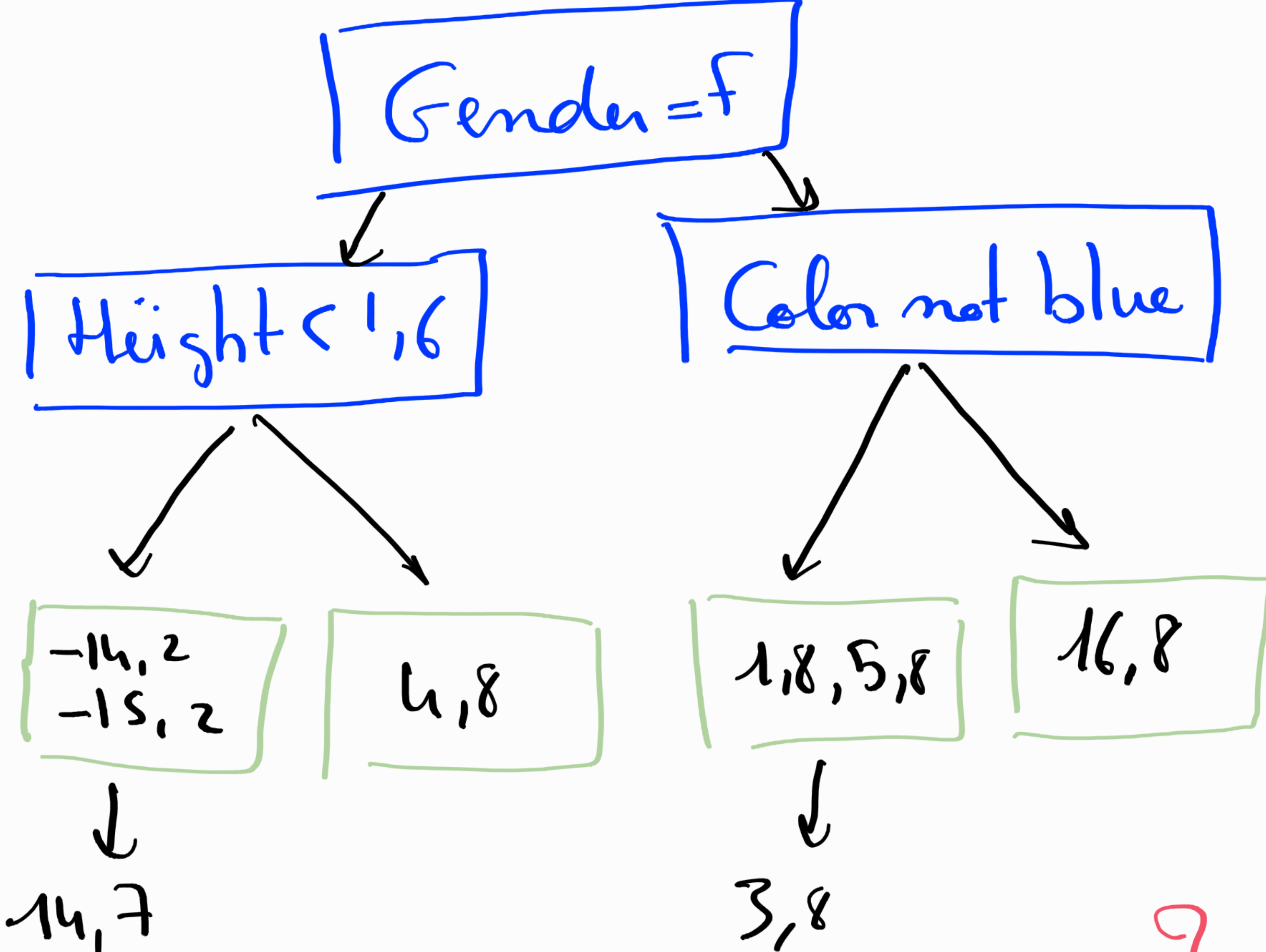
↳ 1) Nombre d'arbre demandé
2) Meilleur modèle trouvé

Height	Favorite color	Gender	Weight
1,6	Blue	m	88
1,6	Green	f	76
1,5	Blue	f	56
1,8	Red	m	73
1,5	Green	m	77
1,4	Blue	f	57

$$\boxed{\overline{w} = 71,2}$$

Residual

-16,8	4,8	-15,2	1,8	5,8	-14,2



Predicted weight : $\boxed{71,2 + 16,8 = 88}$?

No! The model fits the training data too well

⟹ low Biais
hight Variance

$\Rightarrow$ Learning rate

↳ average weight $\boxed{71,2}$ + Learning rate.

$0 <$ Learning rate < 1

fix it at : 0,1

Predicted weight : $71,2 + (0,1 \times 16,8) = 72,9$

$\Rightarrow 72,9 \neq 88$

↳ Let's built an other tree

↳ Recalcul the residual with the learning rate supposition.

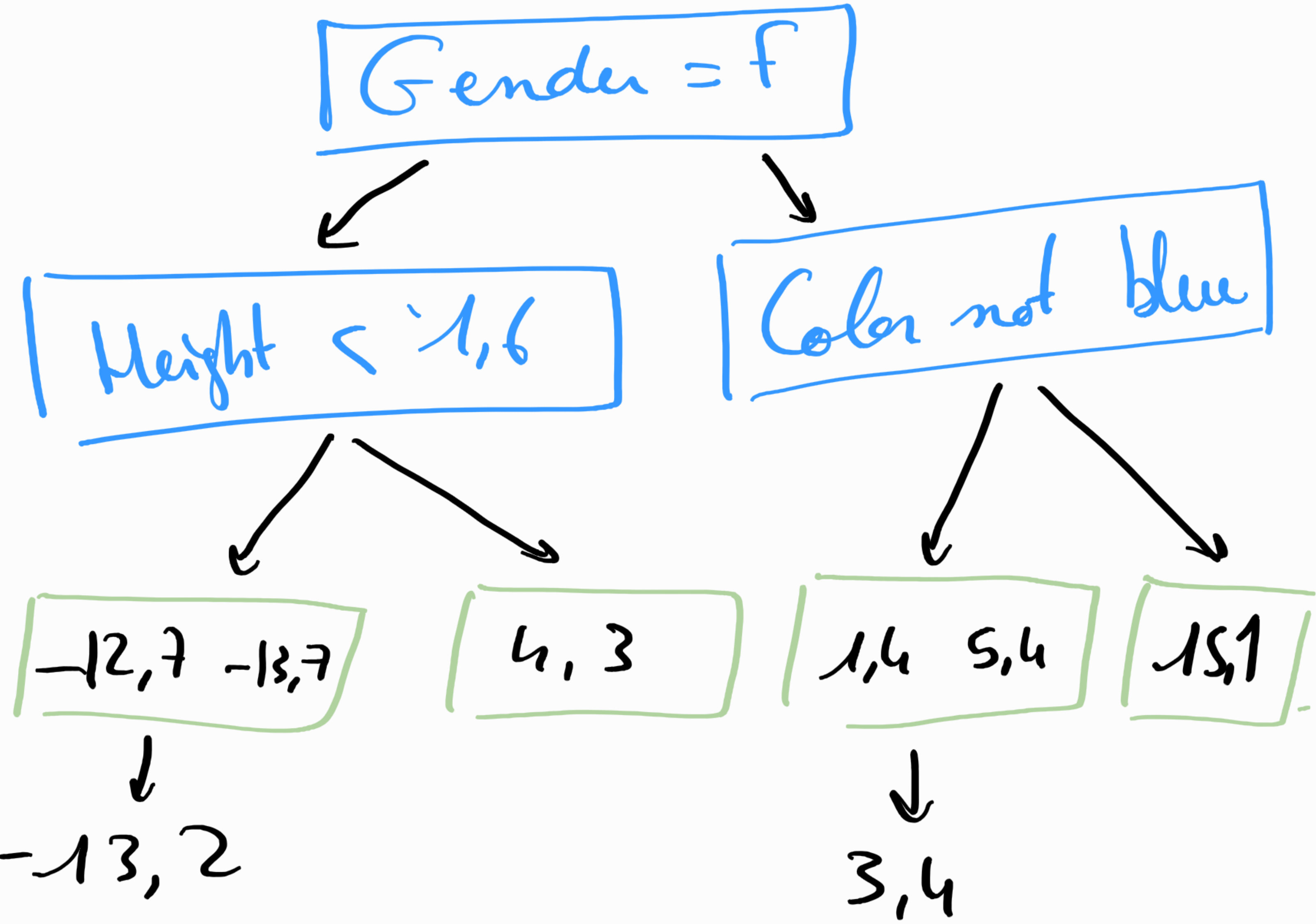
New Residual

15,1	4,3	-13,7	1,4	5,4	-12,7

$\hookrightarrow 88 - (71,2 + 0,1 \times 16,8) = 15,1$

→ The new residuals are a bit smaller than before. We are getting closer and closer from the right prediction

↳ Let's build the tree with the new residual



Now we use the average weight plus the 1st tree plus the 2nd tree

$$\Longrightarrow 71,2 + 0,1 \times 16,8 + 0,1 \times 15,1$$

$$= 71,4$$ → The new prediction

New residual

13,6	3,9	-12,4	1,1	5,1	-11,4

Another small step through the good prediction

The a third tree

↳ skip the details

⟹ Prediction of the first weight is 70

Conclusion:

$$71,2 + \underline{0,1} \times \text{Tree}_1$$

$$+ \underline{0,1} \times \text{Tree}_2$$

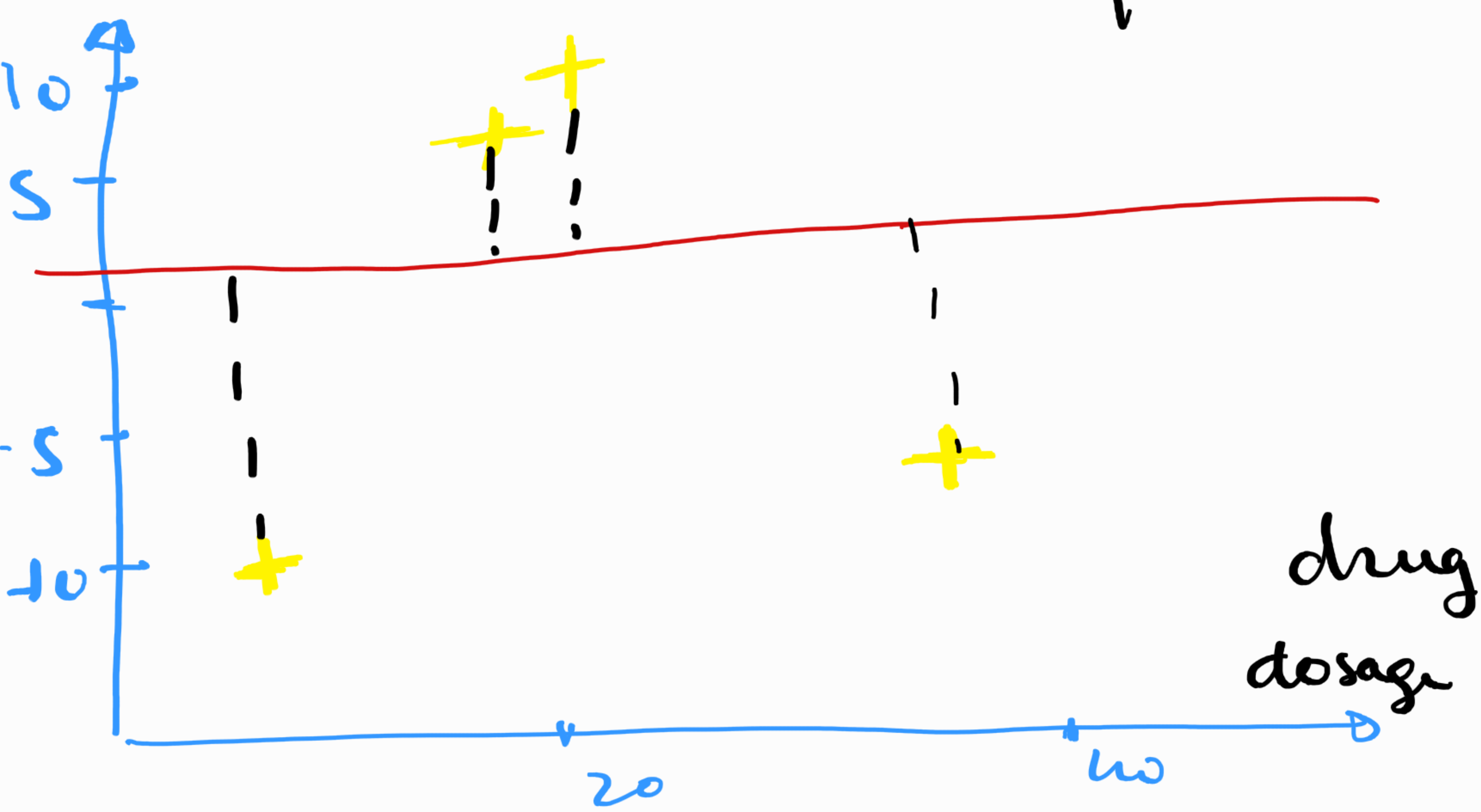
$$+ \underline{0,1} \times \text{Tree}_3$$

depth

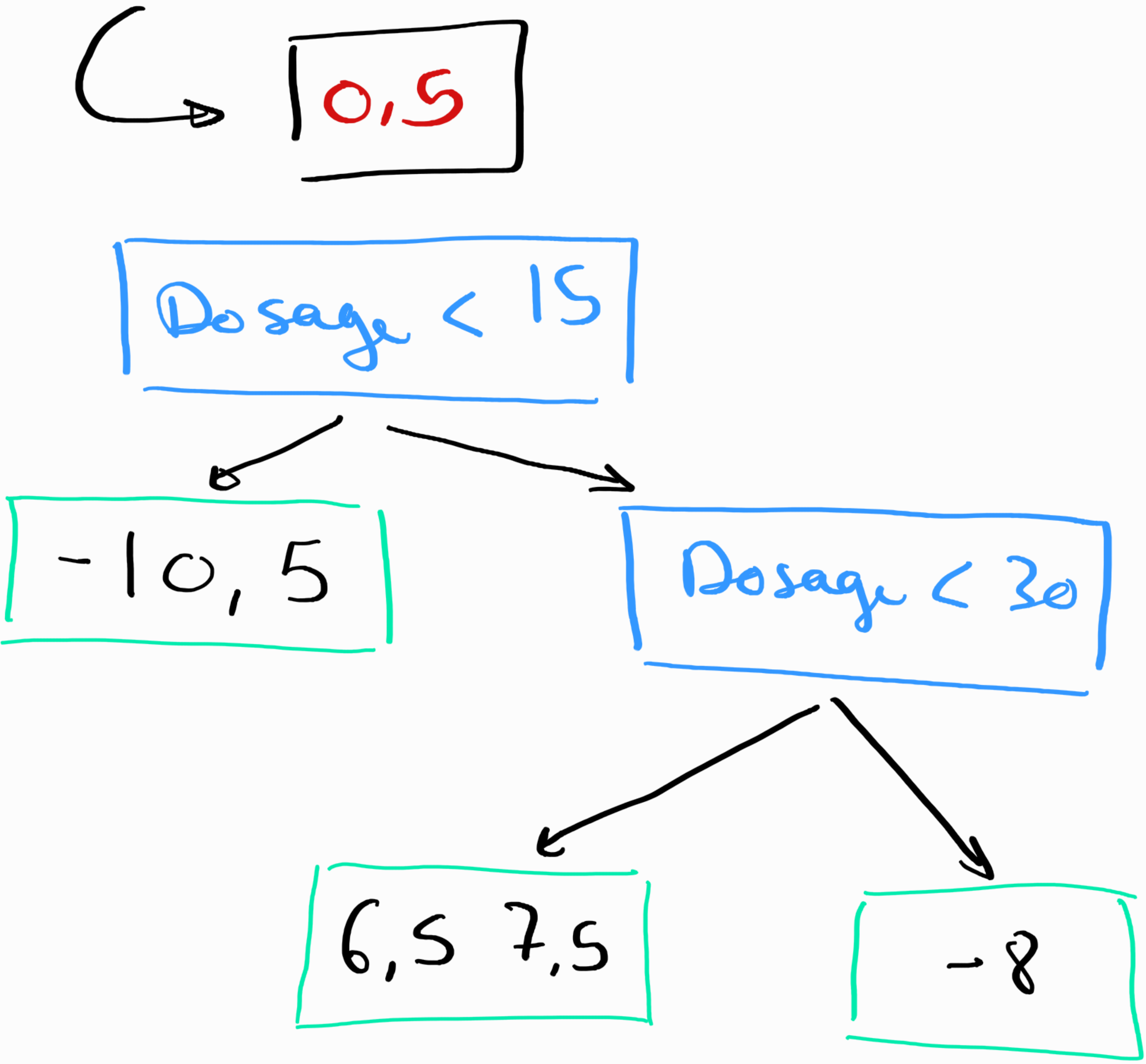
XGboost

→ Regression

→ Classification



Predicted drug Effectiveness



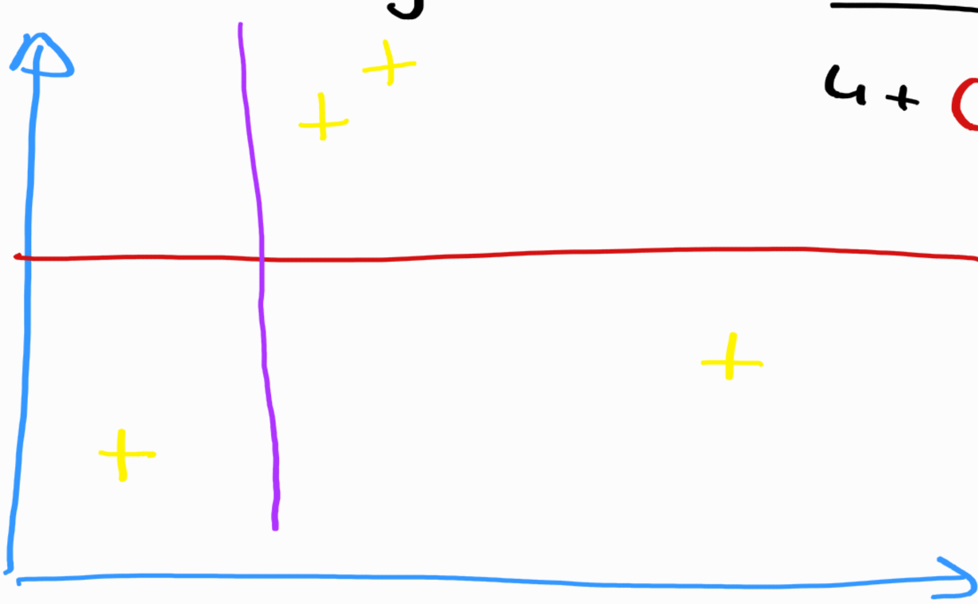
$$\boxed{-10,5 \,/\, 6,5 \,/\, 7,5 \,/\, -7,5}$$

similarité score

$$= \frac{\left(\sum_{1}^{n} \text{Residuals}\right)^2}{n + \lambda}$$

$$\frac{(-10,5 + 6,5 + 7,5 - 7,5)^2}{4 + 0}$$

Similarity score $= \dfrac{(-4)^2}{4+0} = \boxed{4}$



SS = 4

Dosage < 15

-10,5

6,5 7,5 -7,5

$$SS = \frac{(-10,5)^2}{1} = 110,25$$

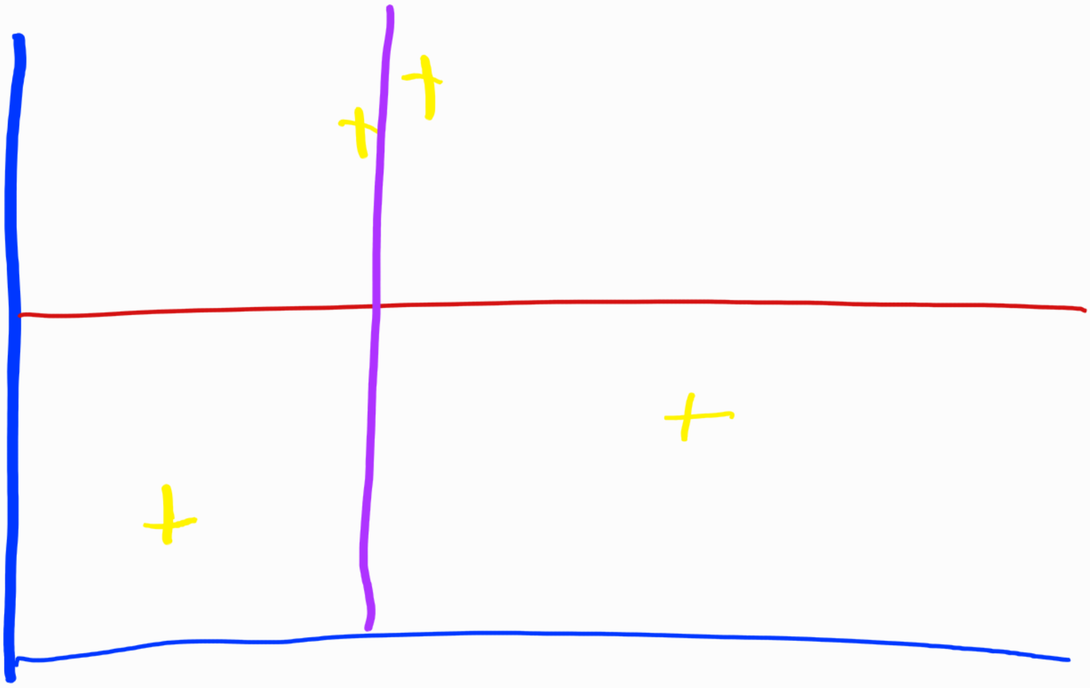
$$SS = \frac{(6,5+7,5-7,5)^2}{3} = 14,08$$

$$Gain = Left_{ss} + Right_{ss} - Root_{ss}$$

$$= 110,25 + 14,08 - 4$$

$$\boxed{Gain_1 = 120,33}$$

Let's see the gain for an other threshold



-10,5	6,5	7,5	-7,5

-10,5	6,5

7,5	-7,5

SS = 8

SS = 0

$$\text{Gain} = 8 + 0 - 4$$

$$\boxed{\text{Gain}_2 = 4}$$

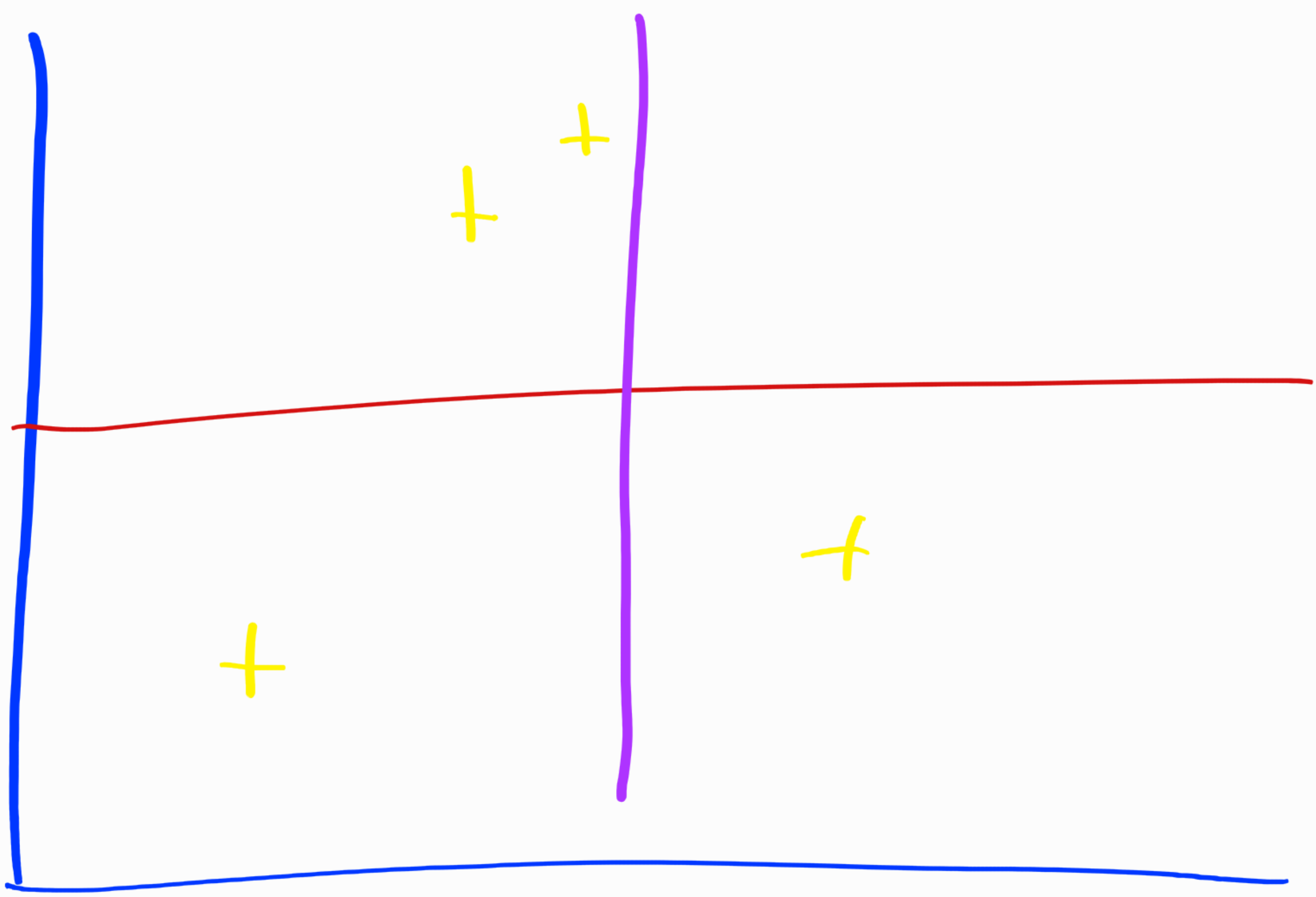
$$Gain_1 = 120,33$$

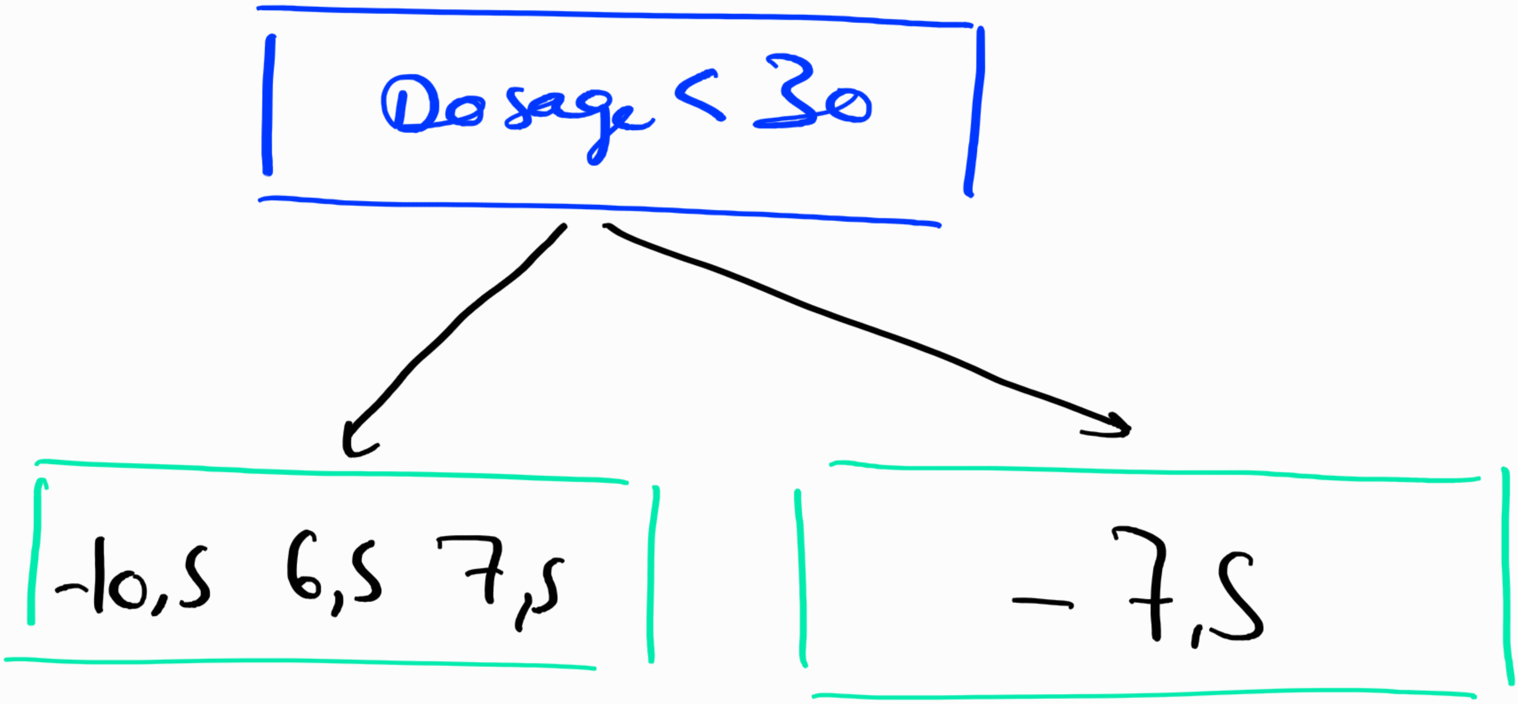
$$Gain_2 = 4$$

$$\Rightarrow Gain_1 > Gain_2$$

$$120,33 > 4$$

Gain$_1$ is better than Gain$_2$





$SS_1 = 4{,}08$ $\qquad$ $SS_2 = 56{,}25$

$$Gain_3 = 4{,}08 + 56{,}25 - 4 = 56{,}33$$

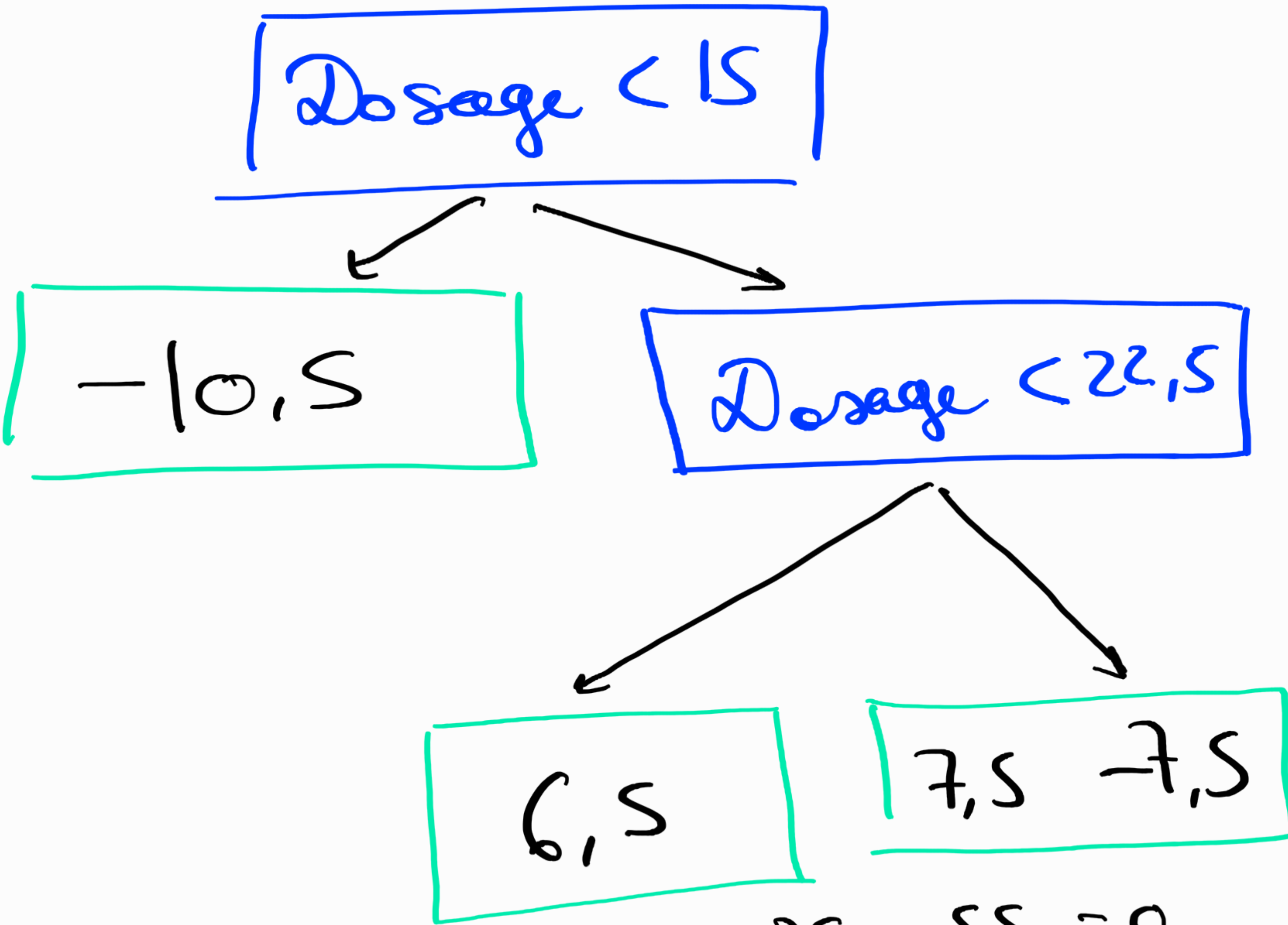
$$\boxed{Gain_3 = 56{,}33}$$

$$Gain_2 < Gain_3 < Gain_1$$

$\Longrightarrow$ $4 < 56{,}33 < 120{,}33$

Dosage < 15 is better

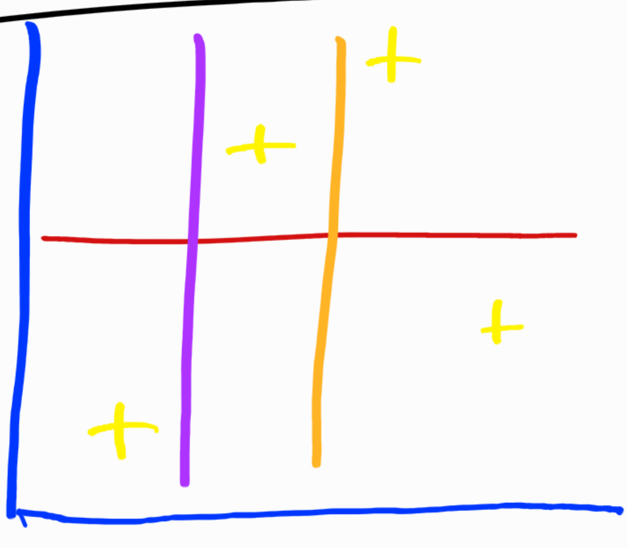
$\Longrightarrow$ a splitting the obs.

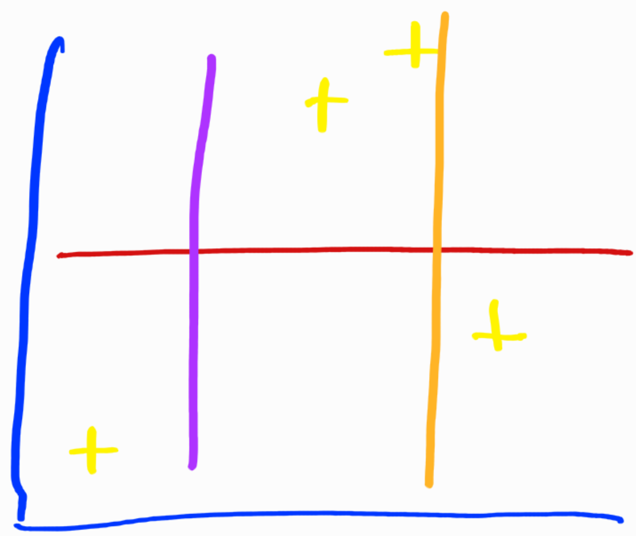


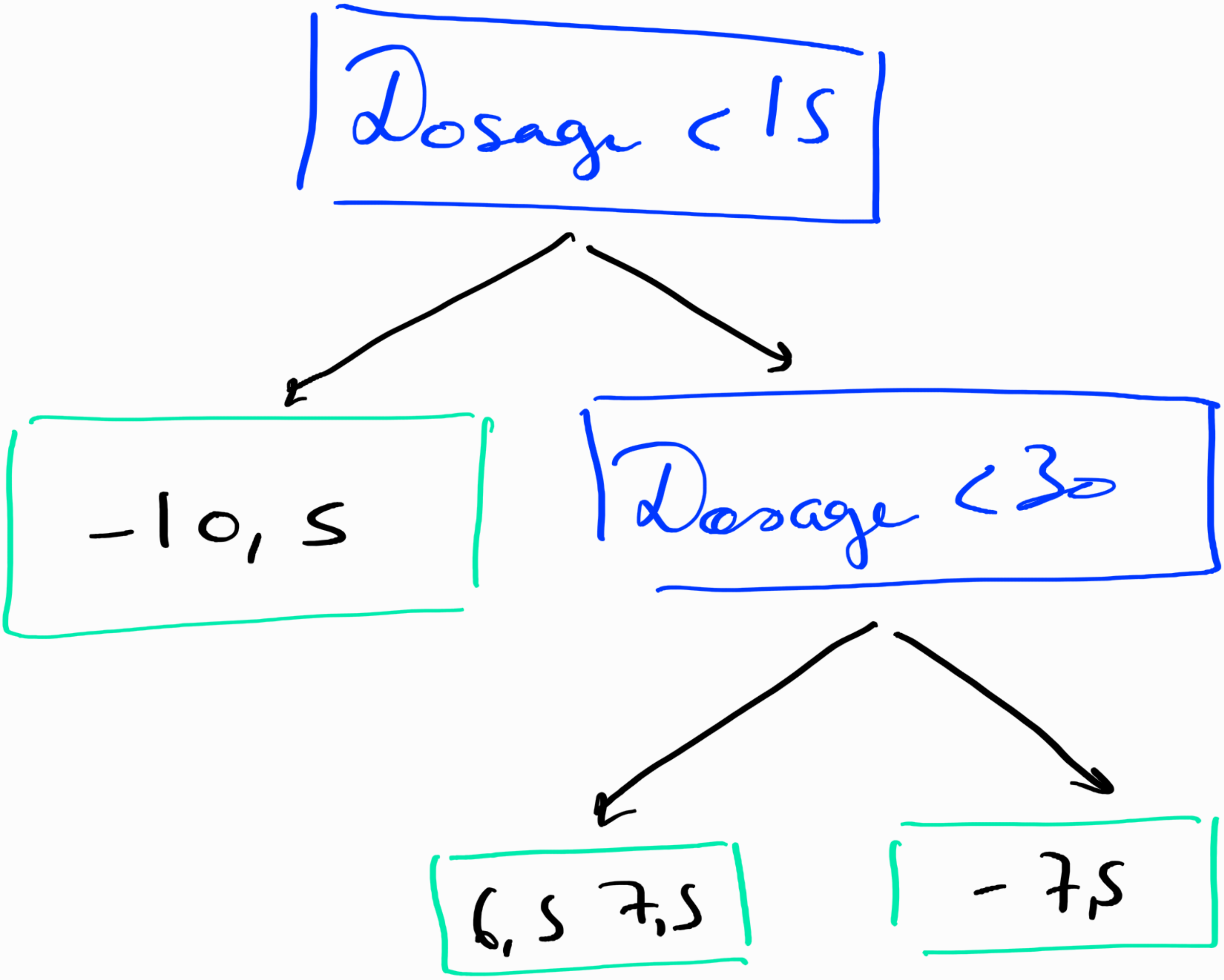
$SS_1 = 42,25$ $SS_2 = 0$

$Gain = 42,25 + 0 - 14,08$

$\boxed{Gain_1 = 28,17}$







$SS_1 = 98$ $SS_2 = 56,25$

$Gain_2 = 98 + 56,25 - 14,02$

$Gain_2 = 140,17$

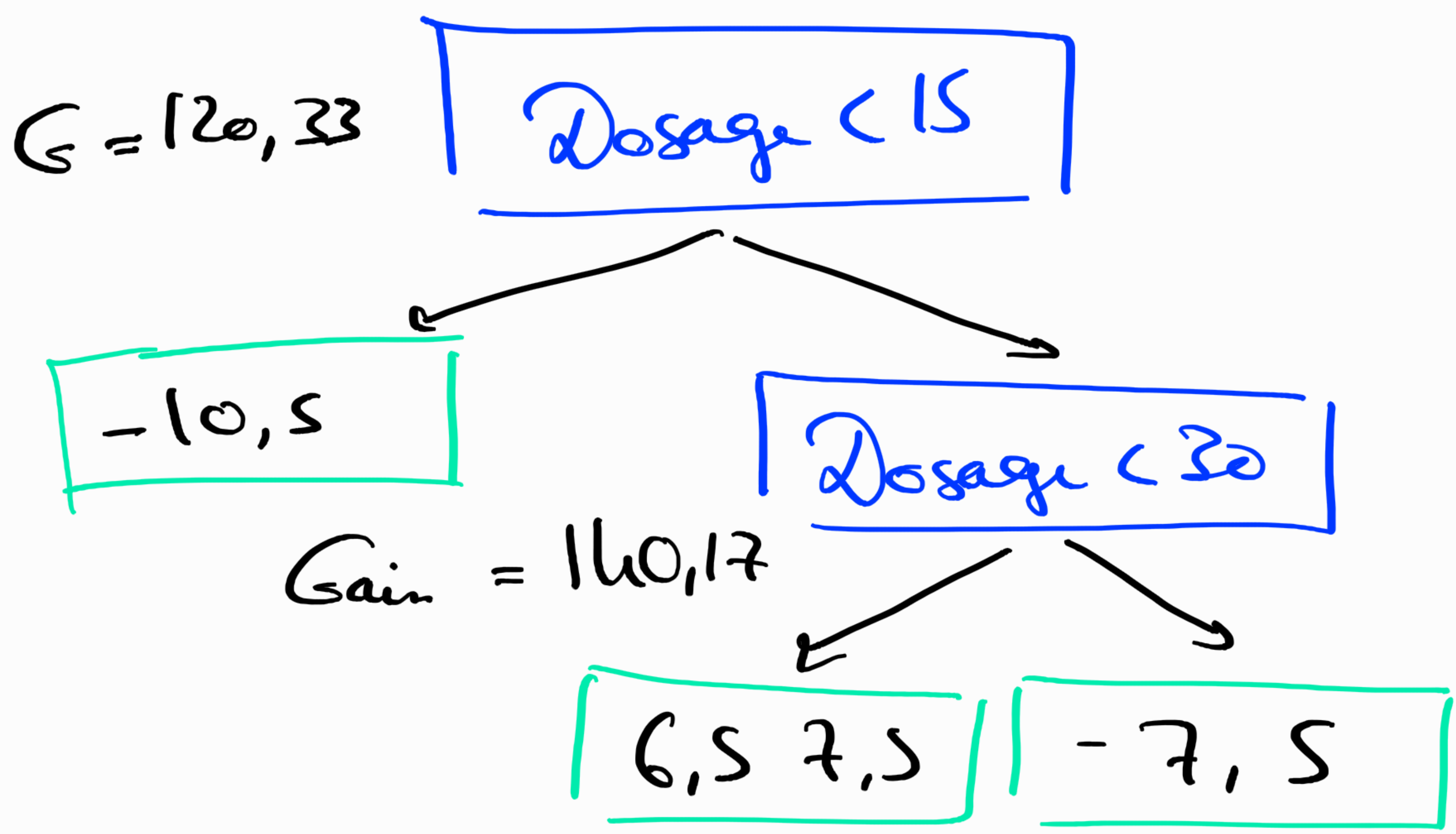
$\Longrightarrow Gain_2 > Gain_1$

$140,17 > 28,17$

$\Longrightarrow$ $Gain_2$ which is Dosage < 30 is the better threshold.

$\boxed{\text{The tree is done}}$

Find tree is:



Let's fix $\gamma = 130$

$\text{Gain} - \gamma < 0$

$\hookrightarrow$ Remove the branch.

$140,17 - 130 > 0$

$\Longrightarrow$ Do not remove the branch

$\gamma = 150$

$\Longrightarrow$ We remove the branch and there no prediction.

Let's fix $\lambda = 1$

Remember λ is a regularization parameter, which means that it is intended to reduce the prediction's sensitivity to individual observation.

$$SS = \frac{(-10,5 + 6,5 + 7,5 - 7,5)^2}{4 + 1} = \boxed{3,2}$$

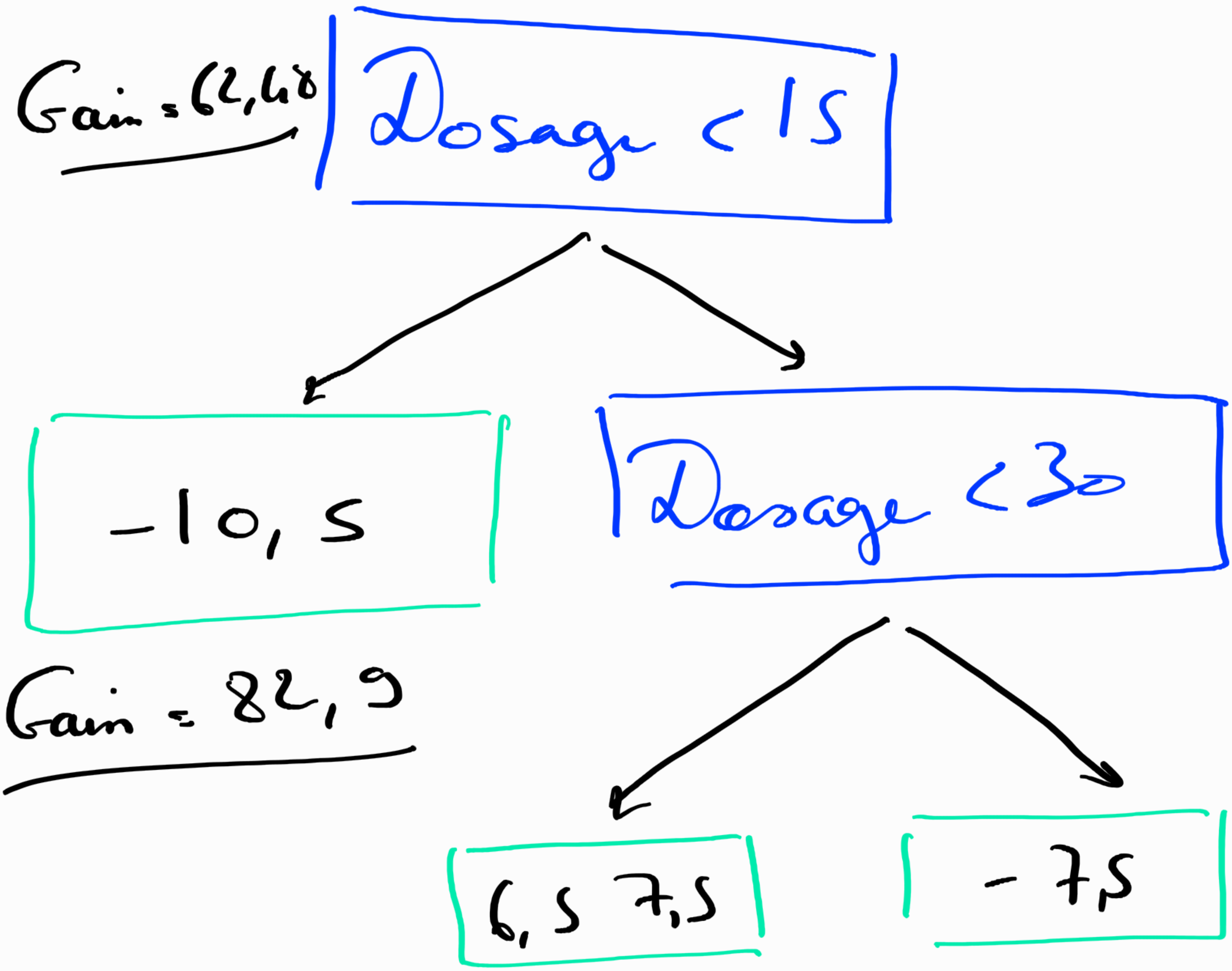
$\boxed{\text{Dosage} < 15}$

$\boxed{-10,5}$ $\boxed{6,5 \quad 7,5 \quad -7,5}$

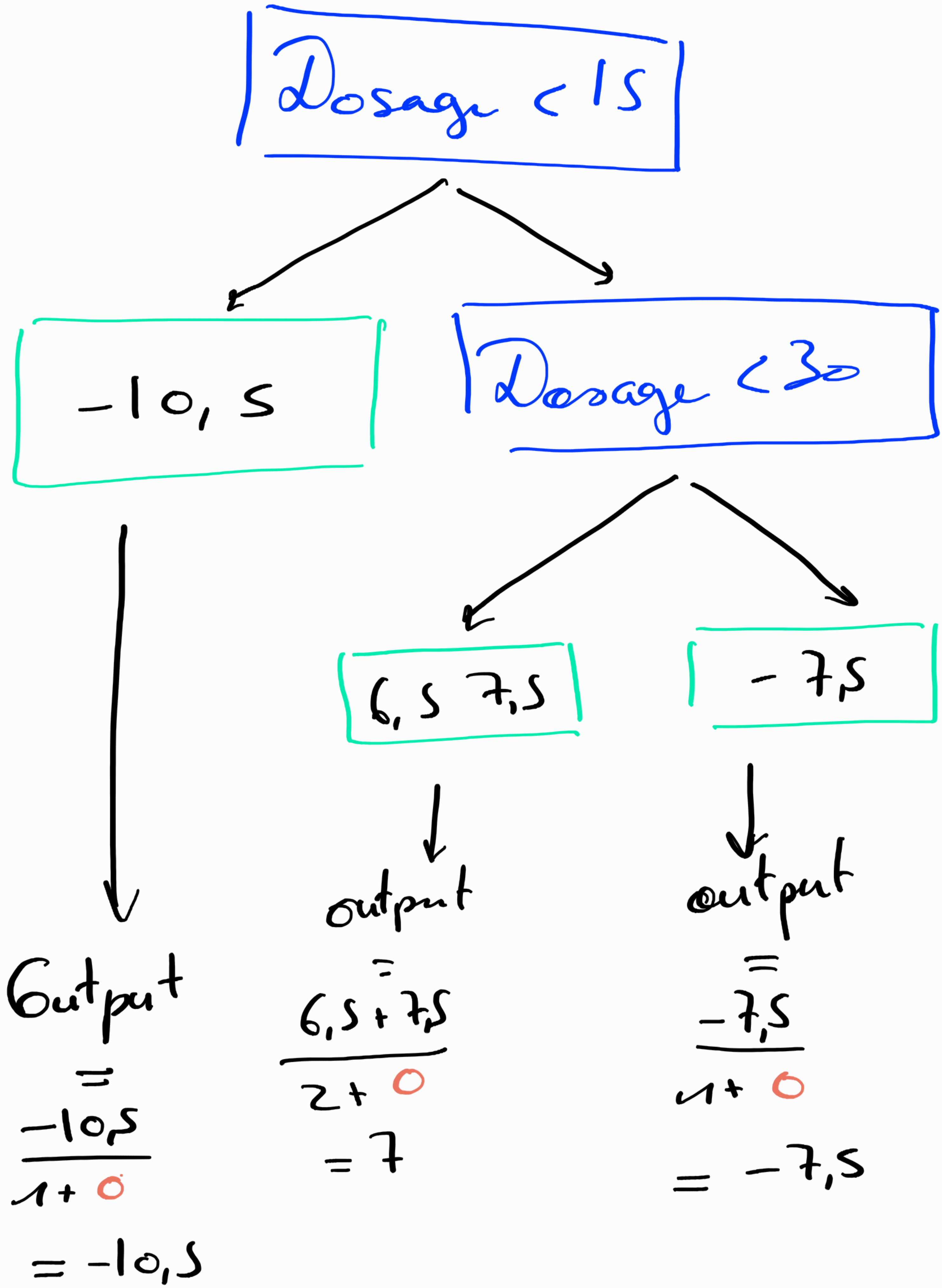
$$SS = \frac{(-10,5)^2}{1 + 1} = 55,12$$

$$SS = \frac{(6,5 + 7,5 - 7,5)^2}{3 + 1} = 10,56$$

$$\text{Gain} = 55{,}12 + 10{,}56 - 3{,}2 = 62{,}48$$

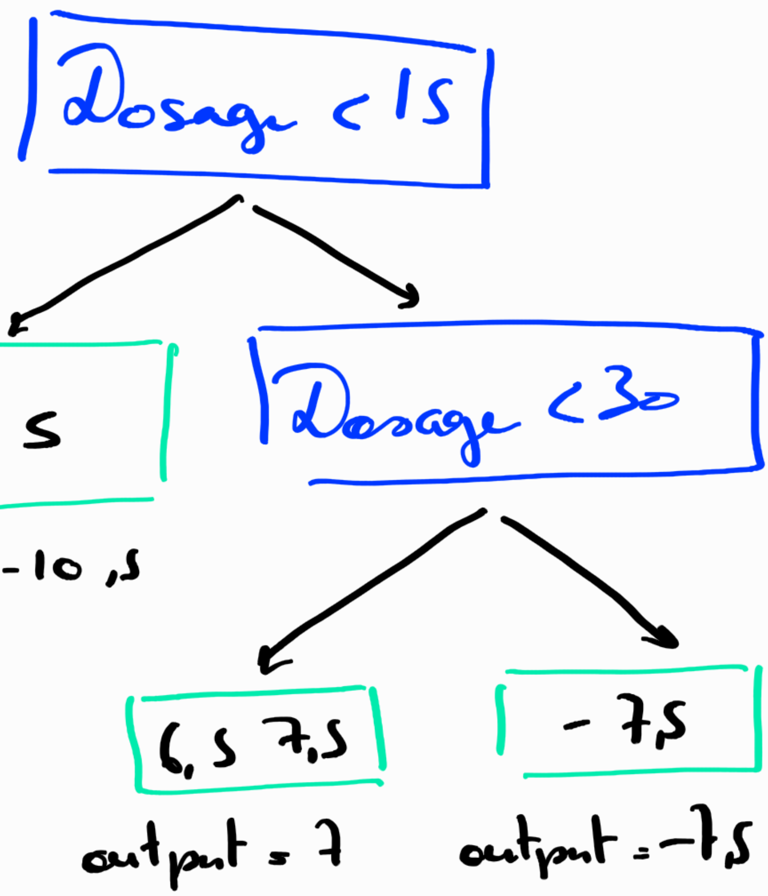


$\lambda > 1 \Rightarrow$ prevent overfitting.



Dosage < 15
-10,5
Dosage < 30
6,5 7,5
-7,5
Output = $\frac{-10,5}{1+0}$ = -10,5
output = $\frac{6,5+7,5}{2+0}$ = 7
output = $\frac{-7,5}{1+0}$ = -7,5

0,5 + Learning rate ×



Protected Drug Effectiveness

Learning rate : ε = 0,3 (default value)

- 0,5 + (0,3 × -10,5) = -2,65
- 0,5 + (0,3 × 7) = 2,6
- 0,5 + (0,3 × -7,5) = -1,75

Continue to build trees with the same learning rate until the residual are super small or we have reached the maximum number