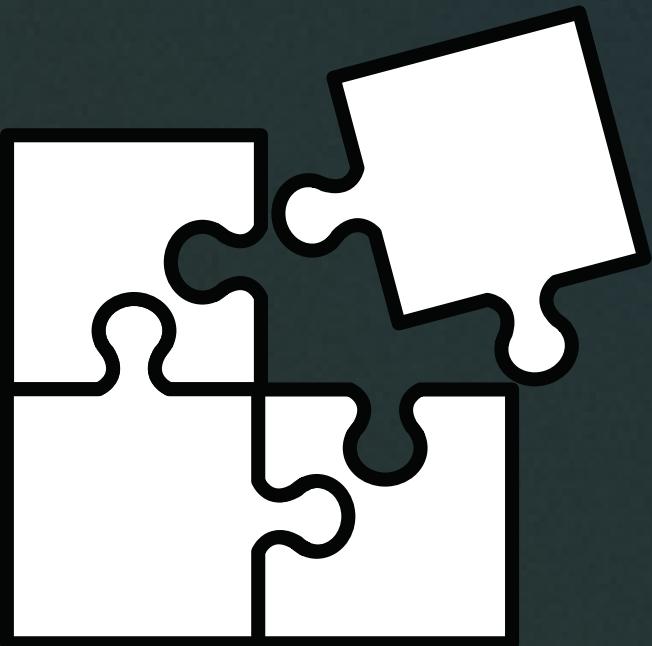




# ÁRVORE DE DECISÃO

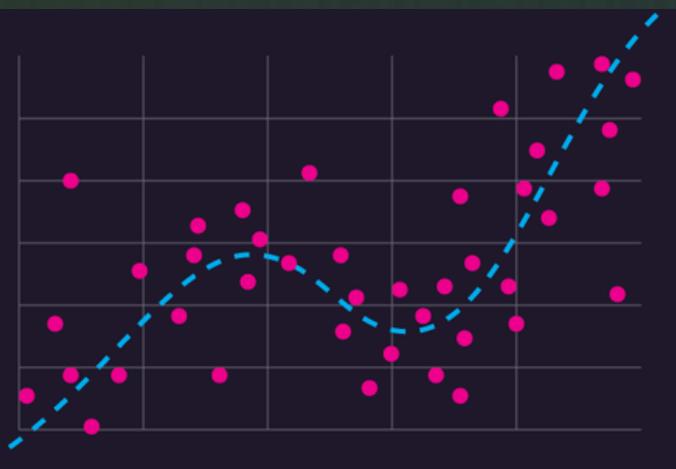
Módulo: MACHINE LEARNING I



**Conseguimos explicar melhor os resultados no nosso modelo de detecção de câncer desenvolvido nas últimas aulas?**



**Vamos desenvolver juntos um modelo de detecção multivariável de 5 defeitos na industria de aço.**



**Vamos construir um modelo de regressão para preços de casas um pouco diferente das aulas anteriores.**

# PROGRAMAÇÃO DA AULA



O que é uma **árvore de decisão**;

# PROGRAMAÇÃO DA AULA

- O que é uma árvore de decisão;
- Rapida aplicação;

# PROGRAMAÇÃO DA AULA

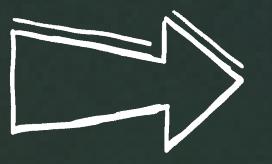
- O que é uma **árvore de decisão**;
- **Rápida aplicação**;
- **Como exatamente funciona uma árvore?**

# PROGRAMAÇÃO DA AULA

- O que é uma **árvore de decisão**;
- Rapida **aplicação**;
- Como exatamente **funciona uma árvore?**;
- Árvores de **Classificação**;

# PROGRAMAÇÃO DA AULA

- O que é uma **árvore de decisão**;
- **Rápida aplicação**;
- **Como exatamente funciona uma árvore?**;
- **Árvores de Classificação**;
- **Árvores de Regressão**;



O QUE É UMA ÁRVORE DE DECISÃO.

# O QUE É UMA ÁRVORE DE DECISÃO

Sol?

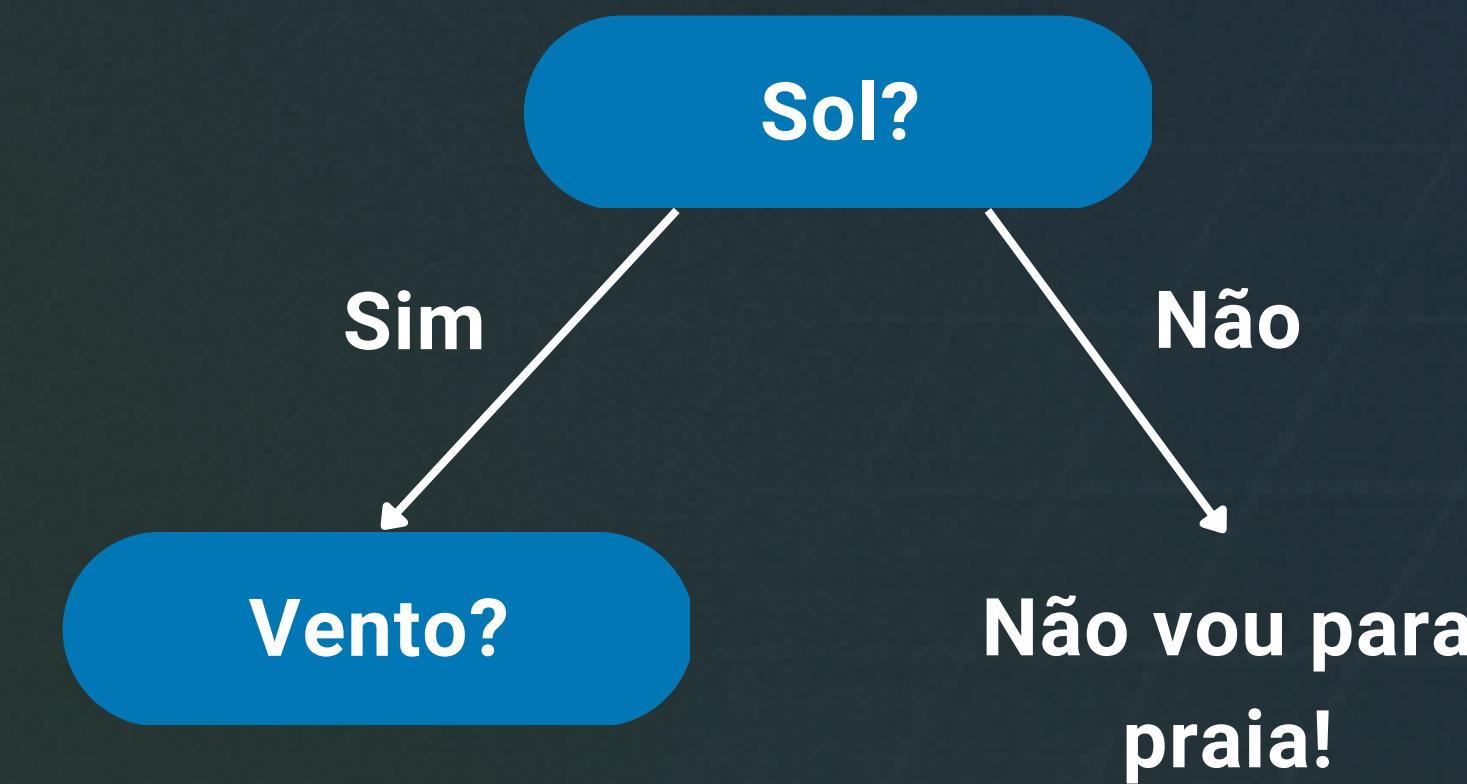
# O QUE É UMA ÁRVORE DE DECISÃO

Sol?

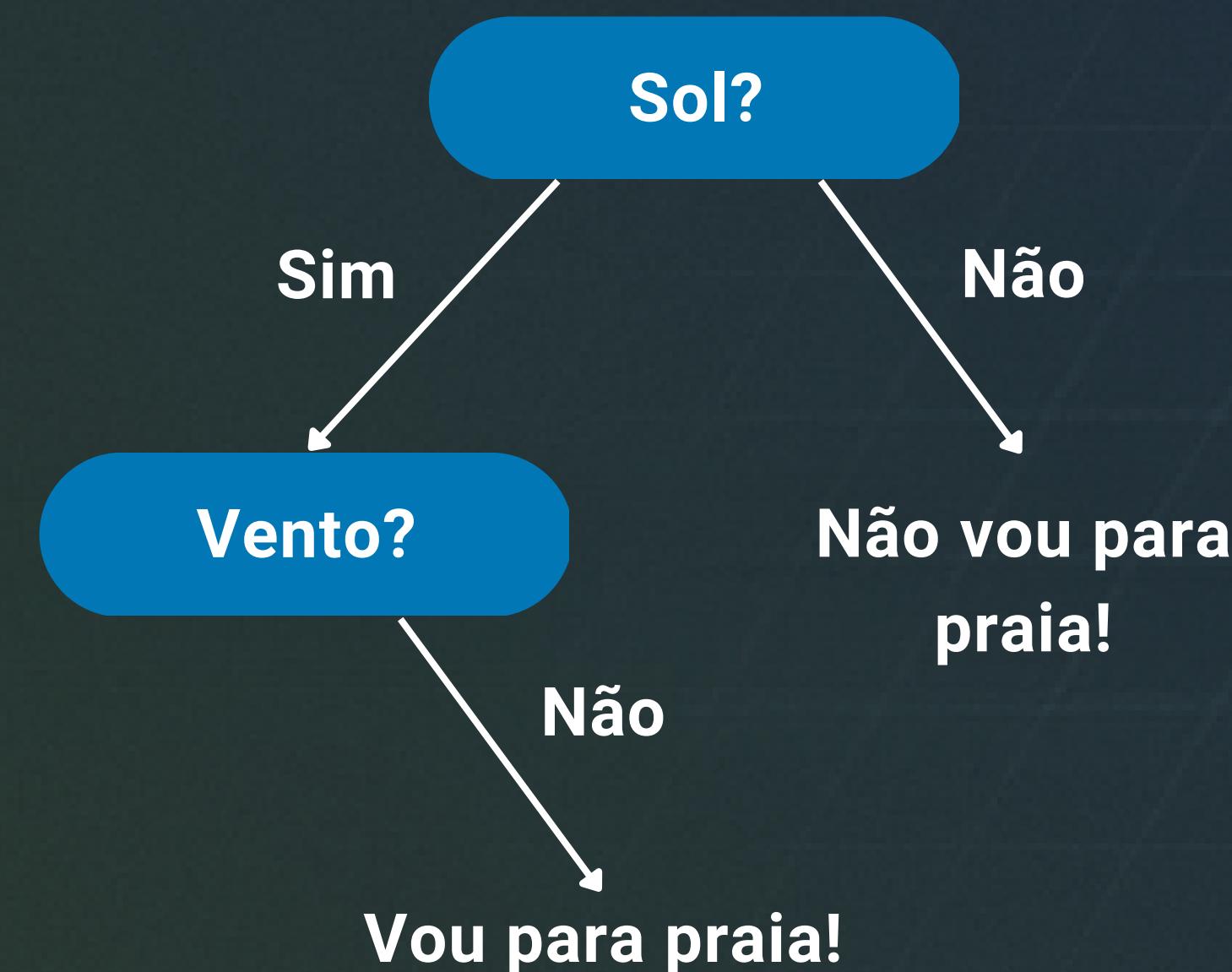
Não

**Não vou para  
praia!**

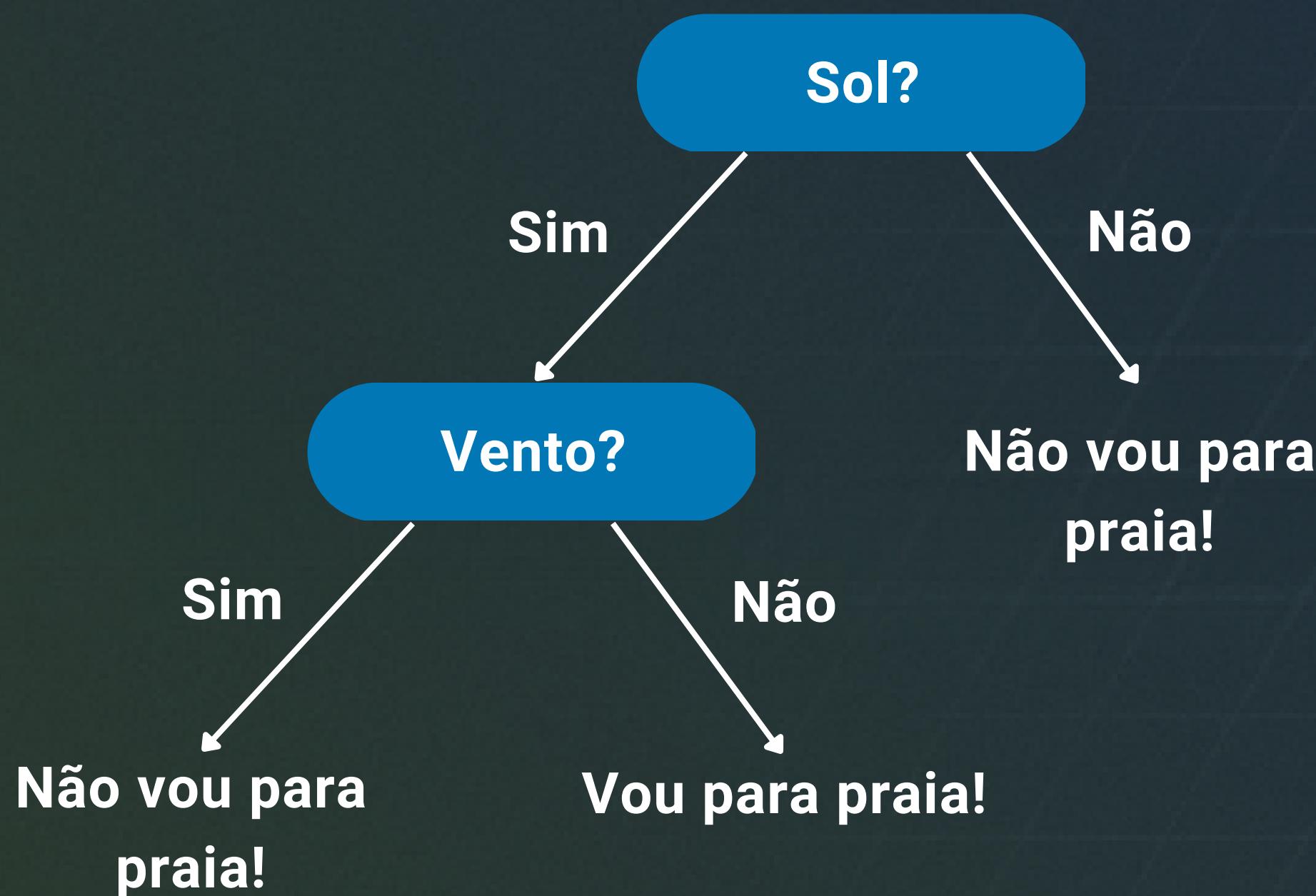
# O QUE É UMA ÁRVORE DE DECISÃO



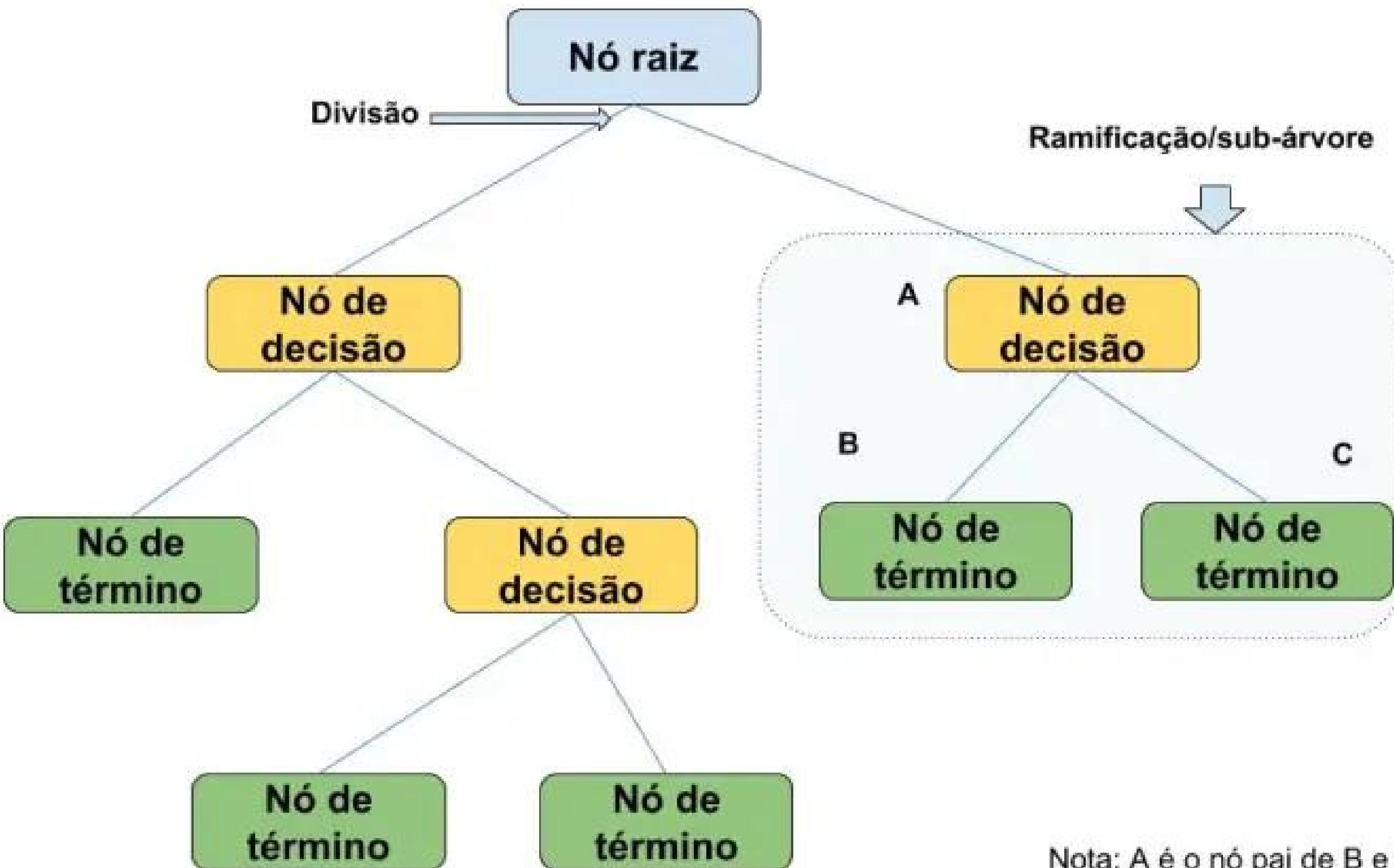
# O QUE É UMA ÁRVORE DE DECISÃO



# O QUE É UMA ÁRVORE DE DECISÃO



# ESTRUTURA DA ÁRVORE



# PORQUE USAR UMA ÁRVORE:

- Fácil **interpretabilidade**;
- Auxilia na **exploração de dados**;
- Traz um **apelo visual**;

# PORQUE USAR UMA ÁRVORE:

- Fácil **interpretabilidade**;
- Auxilia na **exploração de dados**;
- Traz um **apelo visual**;

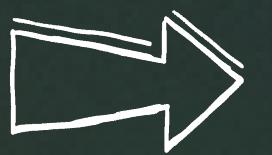


Mas **cuidado** com o risco de **overfitting!**



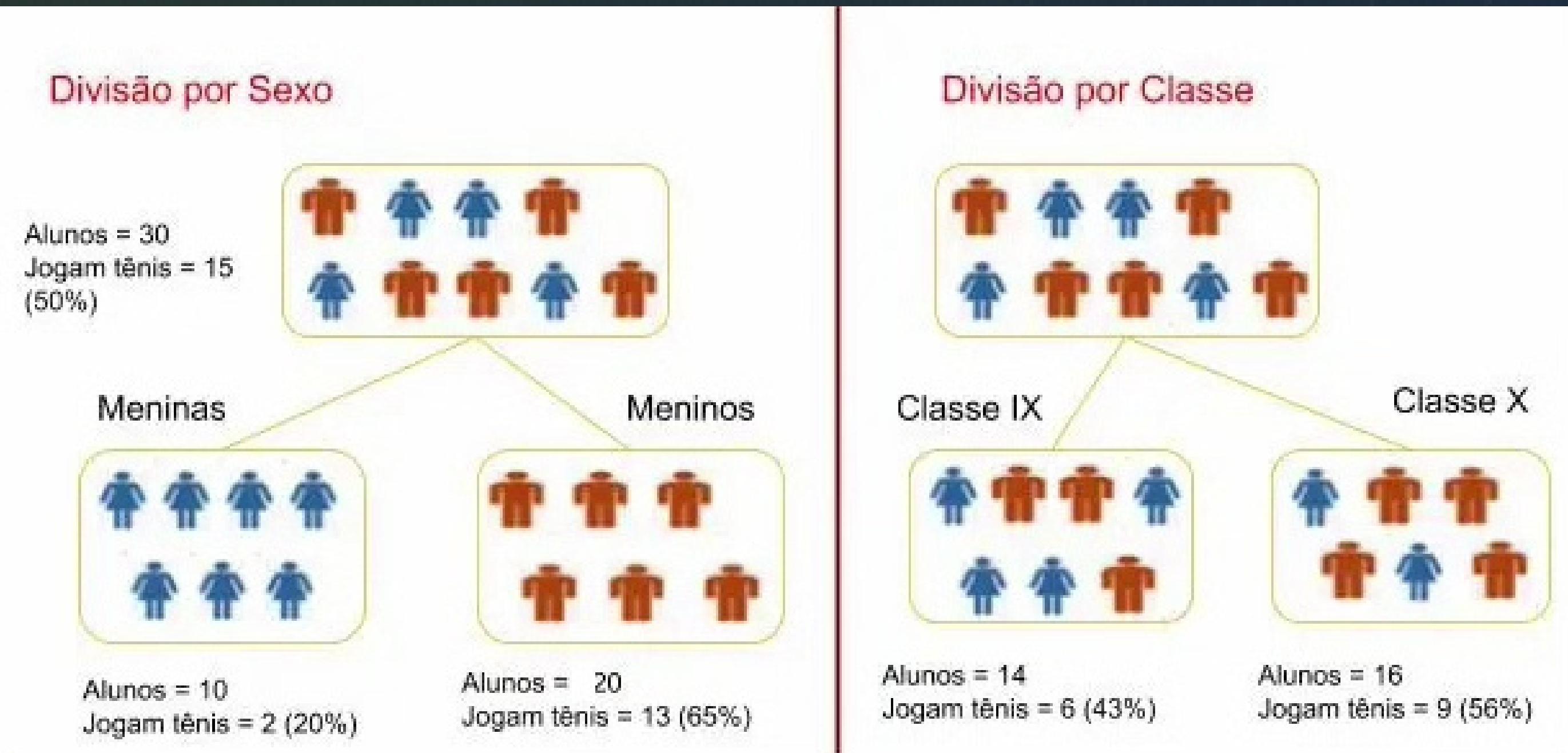
# RÁPIDA APLICAÇÃO.





**COMO EXATAMENTE FUNCIONA UMA  
ÁRVORE?**

# O QUE ACONTECE REALMENTE COM A ÁRVORE?



# O QUE ACONTECE REALMENTE COM A ÁRVORE?

13 Meninos de 20 jogam  
tênis ->  $2/10 = 65\%$

		Sexo	
		Meninas	Meninos
jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
	2	8	13
turma	10	20	7
	IX	X	
jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
6	8	9	7
14	16		

# O QUE ACONTECE REALMENTE COM A ÁRVORE?

13 Meninos de 20 jogam  
tênis ->  $2/10 = 65\%$

2 Meninas de 10 jogam  
tênis ->  $2/10 = 20\%$

		Sexo			
		Meninas	Meninos		
	jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis	
	2	8	13	7	
		10	20		
		turma			
		IX	X		
		jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
		6	8	9	7
		14	16		

# O QUE ACONTECE REALMENTE COM A ÁRVORE?

13 Meninos de 20 jogam  
tênis  $\rightarrow 2/10 = 65\%$

2 Meninas de 10 jogam  
tênis  $\rightarrow 2/10 = 20\%$

9 alunos da turma X de 16  
jogam tênis  $\rightarrow 9/16 = 56\%$

		Sexo		
		Meninas	Meninos	
	jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
	2	8	13	7
	10		20	
		turma		
		IX	X	
	jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
	6	8	9	7
	14		16	

# O QUE ACONTECE REALMENTE COM A ÁRVORE?

13 Meninos de 20 jogam  
tênis  $\rightarrow 2/10 = 65\%$

2 Meninas de 10 jogam  
tênis  $\rightarrow 2/10 = 20\%$

9 alunos da turma X de 16  
jogam tênis  $\rightarrow 9/16 = 56\%$

6 alunos da turma IX de 14  
jogam tênis  $\rightarrow 6/14 = 43\%$

		Sexo		
		Meninas	Meninos	
	jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
	2	8	13	7
	10		20	

		turma		
		IX	X	
	jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
	6	8	9	7
	14		16	

# O QUE ACONTECE REALMENTE COM A ÁRVORE?

13 Meninos de 20 jogam  
tênis  $\rightarrow 2/10 = 65\%$

2 Meninas de 10 jogam  
tênis  $\rightarrow 2/10 = 20\%$

9 alunos da turma X de 16  
jogam tênis  $\rightarrow 9/16 = 56\%$

6 alunos da turma IX de 14  
jogam tênis  $\rightarrow 6/14 = 43\%$

		Sexo		
		Meninas	Meninos	
	jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
	2	8	13	7
	10		20	

		turma		
		IX	X	
	jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
	6	8	9	7
	14		16	

# CRITÉRIO DE GINI

A **impureza de Gini** mede o quanto "impuras" são as folhas das árvores construídas após as **quebras nos nós**.

O coeficiente é dado por:

$$Gini(D) = 1 - \sum p_i^2$$

Onde  $p_i$  são as **proporções** de separação do **target** em cada quebra.

# CRITÉRIO DE GINI

Sexo			
Meninas		Meninos	
jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
2	8	13	7
10	20		

- $G(\text{meninas}) = 1 - \left( \frac{2}{10}^2 + \frac{8}{10}^2 \right) = 0.319$

- $G(\text{meninos}) = 1 - \left( \frac{13}{20}^2 + \frac{7}{20}^2 \right) = 0.454$

# CRITÉRIO DE GINI

Sexo			
Meninas		Meninos	
jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
2	8	13	7

$$\bullet \quad G(\text{pós-divisão}) = \frac{10}{30} \times G(\text{meninas}) + \frac{20}{30} \times G(\text{meninos}) = 0.33 \times 0.319 + 0.66 \times 0.454 = 0.40491$$

$$\bullet \quad G(\text{meninas}) = 1 - \left( \frac{2}{10}^2 + \frac{8}{10}^2 \right) = 0.319$$

$$\bullet \quad G(\text{meninos}) = 1 - \left( \frac{13}{20}^2 + \frac{7}{20}^2 \right) = 0.454$$

# CRITÉRIO DE GINI

		turma			
		IX			X
		jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
6				9	
			8		7

$$\bullet \quad G(IX) = 1 - \left( \frac{6}{14}^2 + \frac{8}{14}^2 \right) = 0.489$$



$$\bullet \quad G(X) = 1 - \left( \frac{9}{16}^2 + \frac{7}{16}^2 \right) = 0.492$$



# CRITÉRIO DE GINI

turma			
IX	X		
jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
6	8	9	7
14	16		

- $$G(\text{pós-divisão}) = \frac{14}{30} \times G(\text{IX}) + \frac{16}{30} \times G(\text{X}) = 0.46 \times 0.489 + 0.53 \times 0.492 = 0.4857$$

- $$G(\text{IX}) = 1 - \left( \frac{6}{14}^2 + \frac{8}{14}^2 \right) = 0.489$$

- $$G(\text{X}) = 1 - \left( \frac{9}{16}^2 + \frac{7}{16}^2 \right) = 0.492$$

# CRITÉRIO DE GINI

- $$G(\text{pós-divisão}) = \frac{10}{30} \times G(\text{meninas}) + \frac{20}{30} \times G(\text{meninos}) = 0.33 \times 0.319 + 0.66 \times 0.454 = 0.40491$$

- $$G(\text{pós-divisão}) = \frac{14}{30} \times G(\text{IX}) + \frac{16}{30} \times G(\text{X}) = 0.46 \times 0.489 + 0.53 \times 0.492 = 0.4857$$

O menor grau de impureza é representado como o menor valor de Gini e representa a melhor divisão e quebra na árvore.

Sexo

Meninas	Meninos
jogam tênis	NÃO jogam tênis
2	8
13	7

turma	
IX	X
jogam tênis	NÃO jogam tênis
6	8
14	9
16	7

# CRITÉRIO DE ENTROPIA

O **critério de entropia** quantifica o grau de desordem de um sistema:

O critério é dado por:

$$E = - \sum p_i \log_2 p_i$$

Onde  $p_i$  são as **proporções** de separação do **target** em cada quebra.

# CRITÉRIO DE ENTROPIA

Sexo			
Meninas		Meninos	
jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
2	8	13	7
10	20	10	7

- $E(\text{meninas}) = -1 \times \left( \frac{2}{10} \log_2 \frac{2}{10} + \frac{8}{10} \log_2 \frac{8}{10} \right) = 0.721$

- $E(\text{meninos}) = -1 \times \left( \frac{13}{20} \log_2 \frac{13}{20} + \frac{7}{20} \log_2 \frac{7}{20} \right) = 0.934$

# CRITÉRIO DE ENTROPIA

Sexo			
Meninas	Meninos	Meninas	Meninos
jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
2	8	13	7

- $E(\text{pós-divisão}) = \frac{10}{30} \times E(\text{meninas}) + \frac{20}{30} \times E(\text{meninos}) = 0.863$
- $E(\text{meninas}) = -1 \times \left( \frac{2}{10} \log_2 \frac{2}{10} + \frac{8}{10} \log_2 \frac{8}{10} \right) = 0.721$
- $E(\text{meninos}) = -1 \times \left( \frac{13}{20} \log_2 \frac{13}{20} + \frac{7}{20} \log_2 \frac{7}{20} \right) = 0.934$

# CRITÉRIO DE ENTROPIA

turma

IX

X

jogam tênis NÃO jogam tênis jogam tênis NÃO jogam tênis

6

8

9

7

$$\bullet E(IX) = -1 \times \left( \frac{6}{14} \log_2 \frac{6}{14} + \frac{8}{14} \log_2 \frac{8}{14} \right) = 0.985$$

$$\bullet E(X) = -1 \times \left( \frac{9}{16} \log_2 \frac{9}{16} + \frac{7}{16} \log_2 \frac{7}{16} \right) = 0.988$$

# CRITÉRIO DE ENTROPIA

		turma			
		IX			X
jogam tênis		NÃO jogam tênis	jogam tênis	NÃO jogam tênis	
6		8	9	7	
	14			16	

$$\bullet E(\text{pós-divisão}) = \frac{10}{30} \times E(\text{IX}) + \frac{20}{30} \times E(\text{X}) = 0.986$$

$$\bullet E(\text{IX}) = -1 \times \left( \frac{6}{14} \log_2 \frac{6}{14} + \frac{8}{14} \log_2 \frac{8}{14} \right) = 0.985$$

$$\bullet E(\text{X}) = -1 \times \left( \frac{9}{16} \log_2 \frac{9}{16} + \frac{7}{16} \log_2 \frac{7}{16} \right) = 0.988$$

# CRITÉRIO DE ENTROPIA

- $E(\text{pós-divisão}) = \frac{10}{30} \times E(\text{meninas}) + \frac{20}{30} \times E(\text{meninos}) = 0.863$

- $E(\text{pós-divisão}) = \frac{10}{30} \times E(\text{IX}) + \frac{20}{30} \times E(\text{X}) = 0.986$

O menor grau de impureza é representado como o menor valor de entropia e representa a melhor divisão e quebra na árvore.

Sexo			
Meninas		Meninos	
jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
2	8	13	7
10		20	

turma			
IX		X	
jogam tênis	NÃO jogam tênis	jogam tênis	NÃO jogam tênis
6	8	9	7
14		16	

# O QUE ACONTECE REALMENTE COM A ÁRVORE?

- $G(\text{pós-divisão}) = \frac{10}{30} \times G(\text{meninas}) + \frac{20}{30} \times G(\text{meninos}) = 0.33 \times 0.319 + 0.66 \times 0.454 = 0.40491$

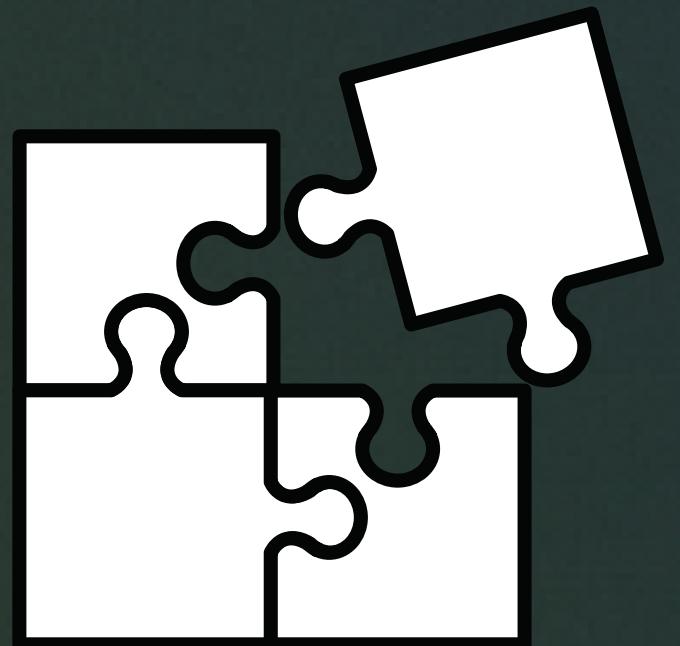
- $E(\text{pós-divisão}) = \frac{10}{30} \times E(\text{meninas}) + \frac{20}{30} \times E(\text{meninos}) = 0.863$

		<b>Sexo</b>	
		Meninas	Meninos
	jogam tênis	NÃO jogam tênis	jogam tênis
	2	8	13
	10		20

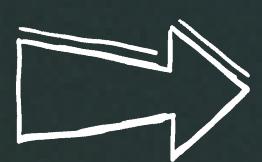
- $G(\text{pós-divisão}) = \frac{14}{30} \times G(\text{IX}) + \frac{16}{30} \times G(\text{X}) = 0.46 \times 0.489 + 0.53 \times 0.492 = 0.4857$

- $E(\text{pós-divisão}) = \frac{10}{30} \times E(\text{IX}) + \frac{20}{30} \times E(\text{X}) = 0.986$

		<b>turma</b>	
		IX	X
	jogam tênis	NÃO jogam tênis	jogam tênis
	6	8	9
	14		16



Conseguimos explicar melhor os resultados no nosso modelo de detecção de câncer desenvolvido nas últimas aulas?

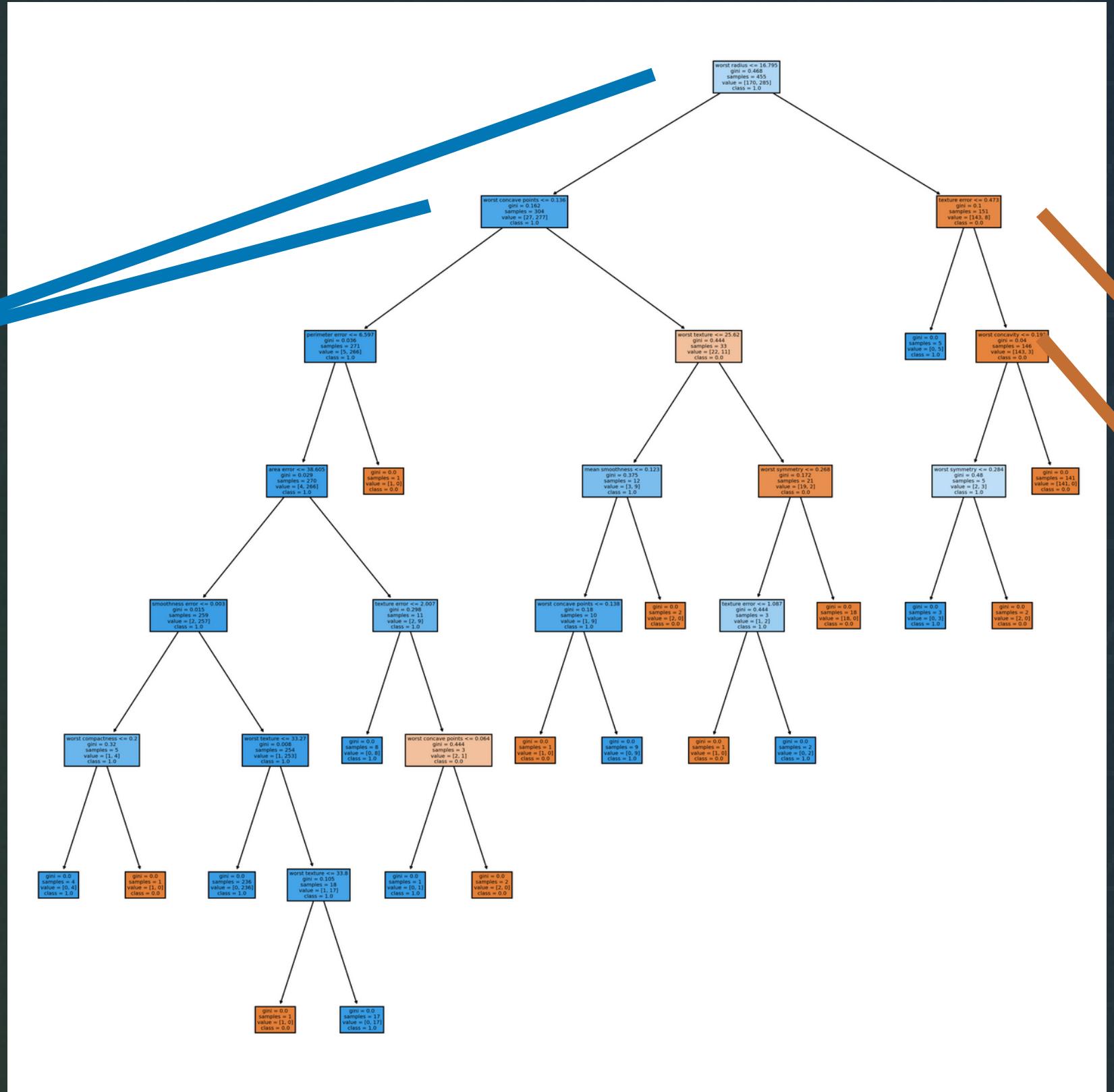


# ÁRVORES DE CLASSIFICAÇÃO.



# ÁRVORES DE CLASSIFICAÇÃO

Escala de Azul  
representa a  
classe 1 -  
probabilidade  
maior de ter  
câncer



# ÁRVORES DE CLASSIFICAÇÃO

Condição de divisão da folha.

Na primeira folha:

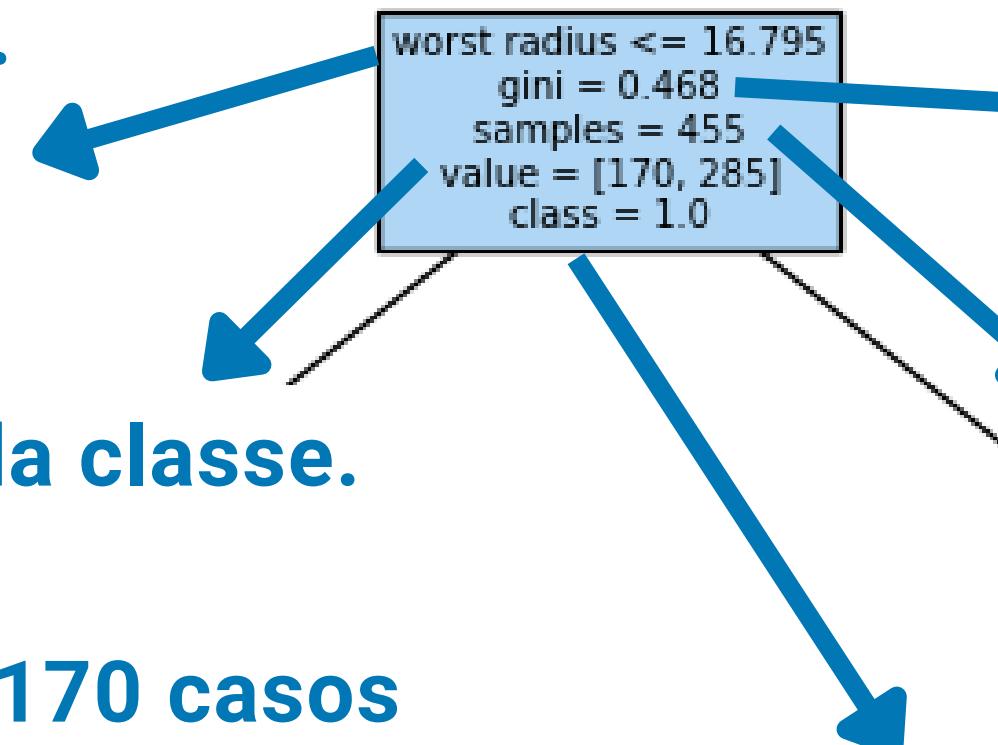
"worst radius <= 16.795"

Número de amostras de cada classe.

Na primeira folha:

"value = [170, 285]", ou seja, 170 casos de baixa chance de câncer e 285 casos de alta chance de câncer

```
worst concave points <= 0.136
gini = 0.162
samples = 304
value = [27, 277]
class = 1.0
```



Valor do Criterio de Gini para a folha. Na primeira folha: "gini = 0.468"

Número de amostras na folha.

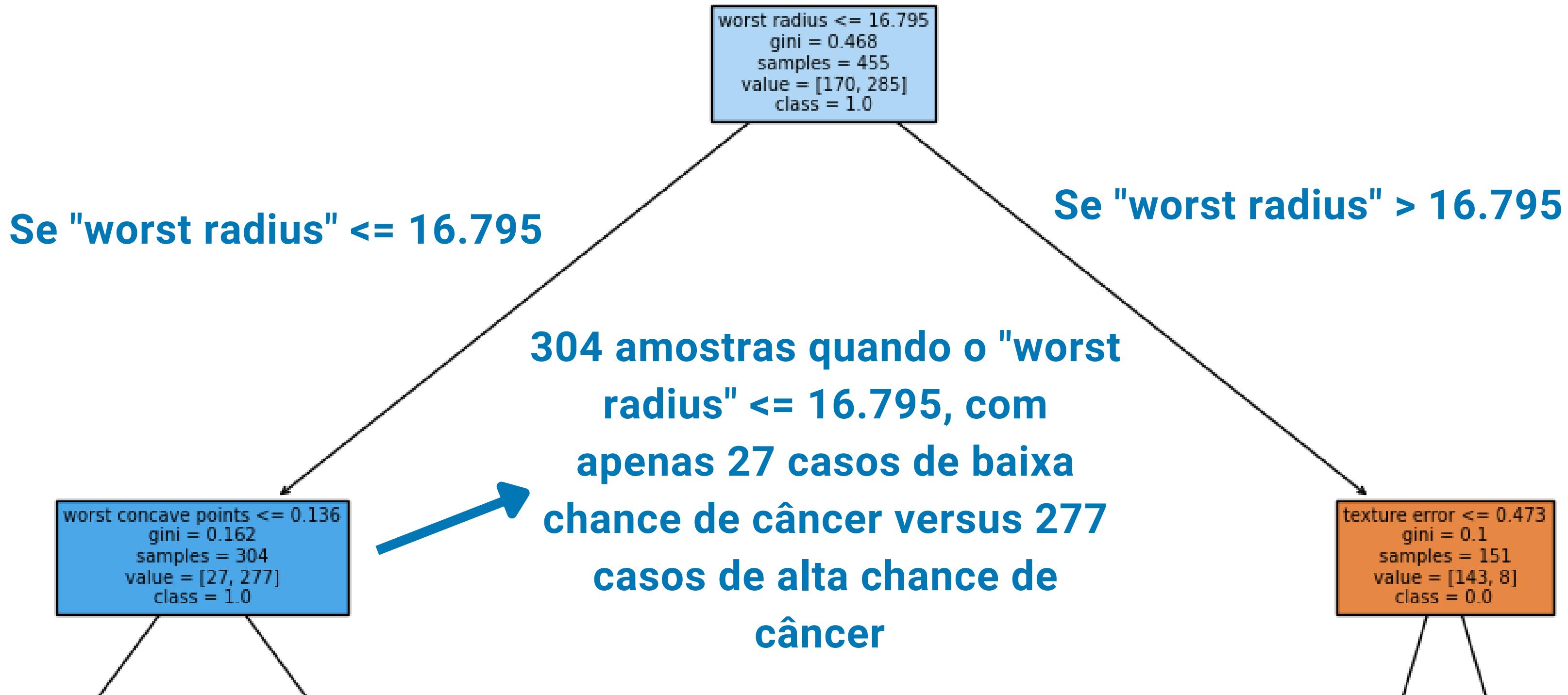
Na primeira folha:

"samples = 455"

Classe predominante na folha. Na primeira folha: "class = 1.0", ou seja, mais casos de alta chance de câncer.

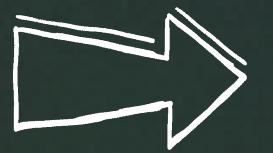
```
texture error <= 0.473
gini = 0.1
samples = 151
value = [143, 8]
class = 0.0
```

# ÁRVORES DE CLASSIFICAÇÃO



# ALGUNS HIPERPARÂMETROS

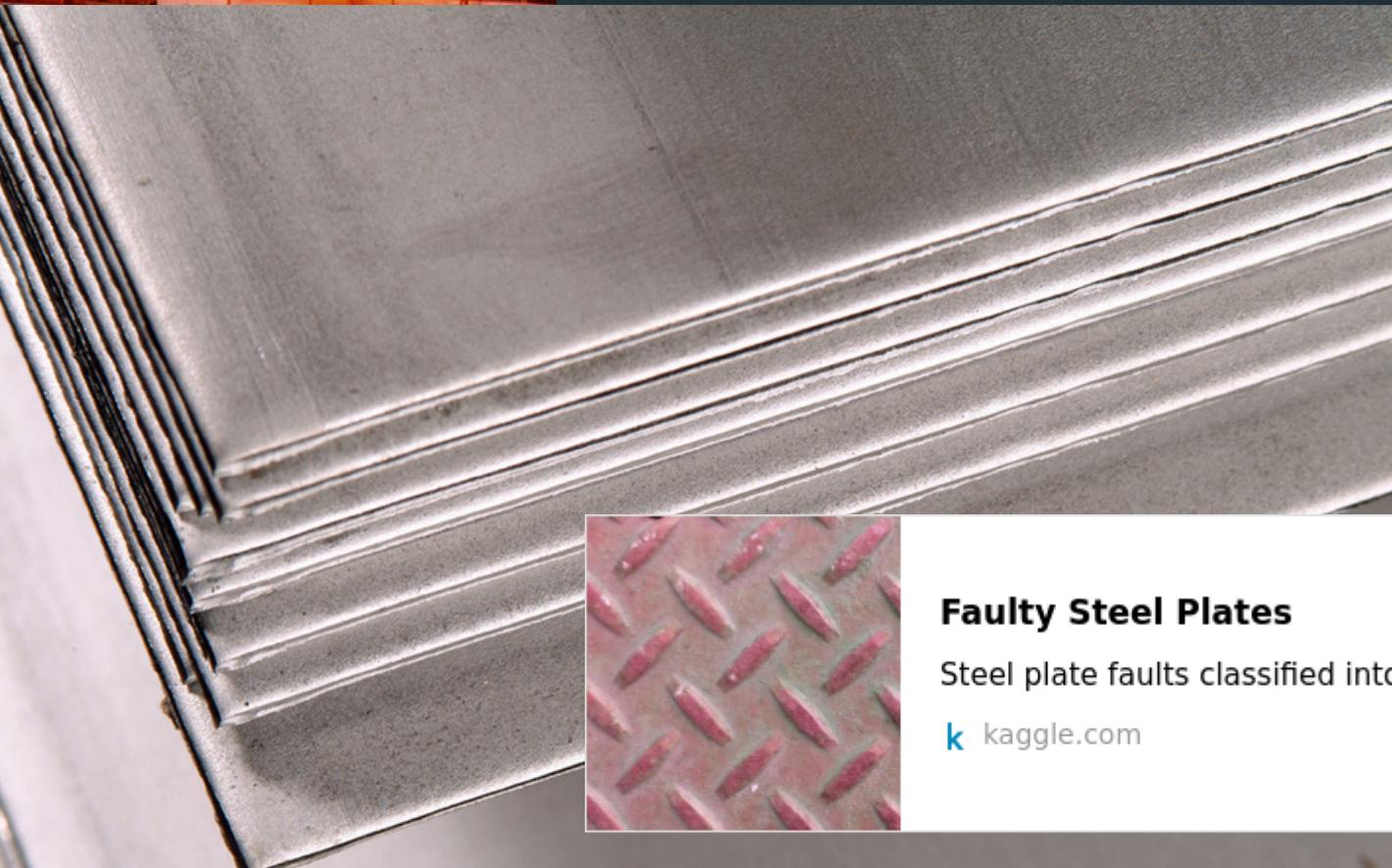
- **max\_depth**: Maxima profundidade;
- **min\_samples\_split**: minimo de amostras para realizar um split;
- **min\_samples\_leaf**: minimo de amostras necessarias para ter uma folha;
- **min\_impurity\_decrease**: menor grau de impureza para realizar um split.



**ANTES DE CONTINUAR, VAMOS  
VOLTAR E TREINAR UM  
MODELO AINDA MELHOR**



# Modelo de detecção multivariável de 5 defeitos na indústria de aço.



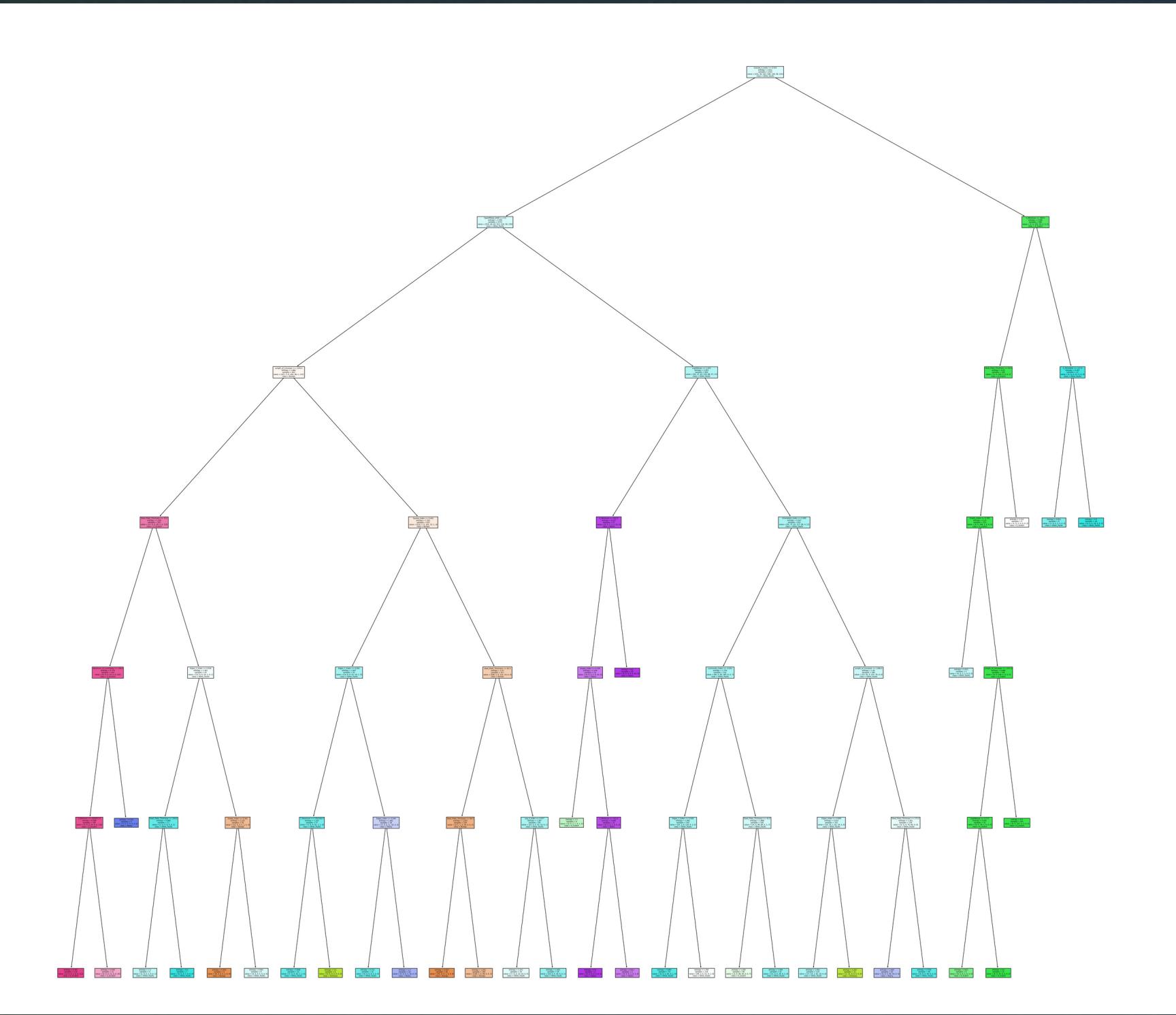
## Faulty Steel Plates

Steel plate faults classified into seven types

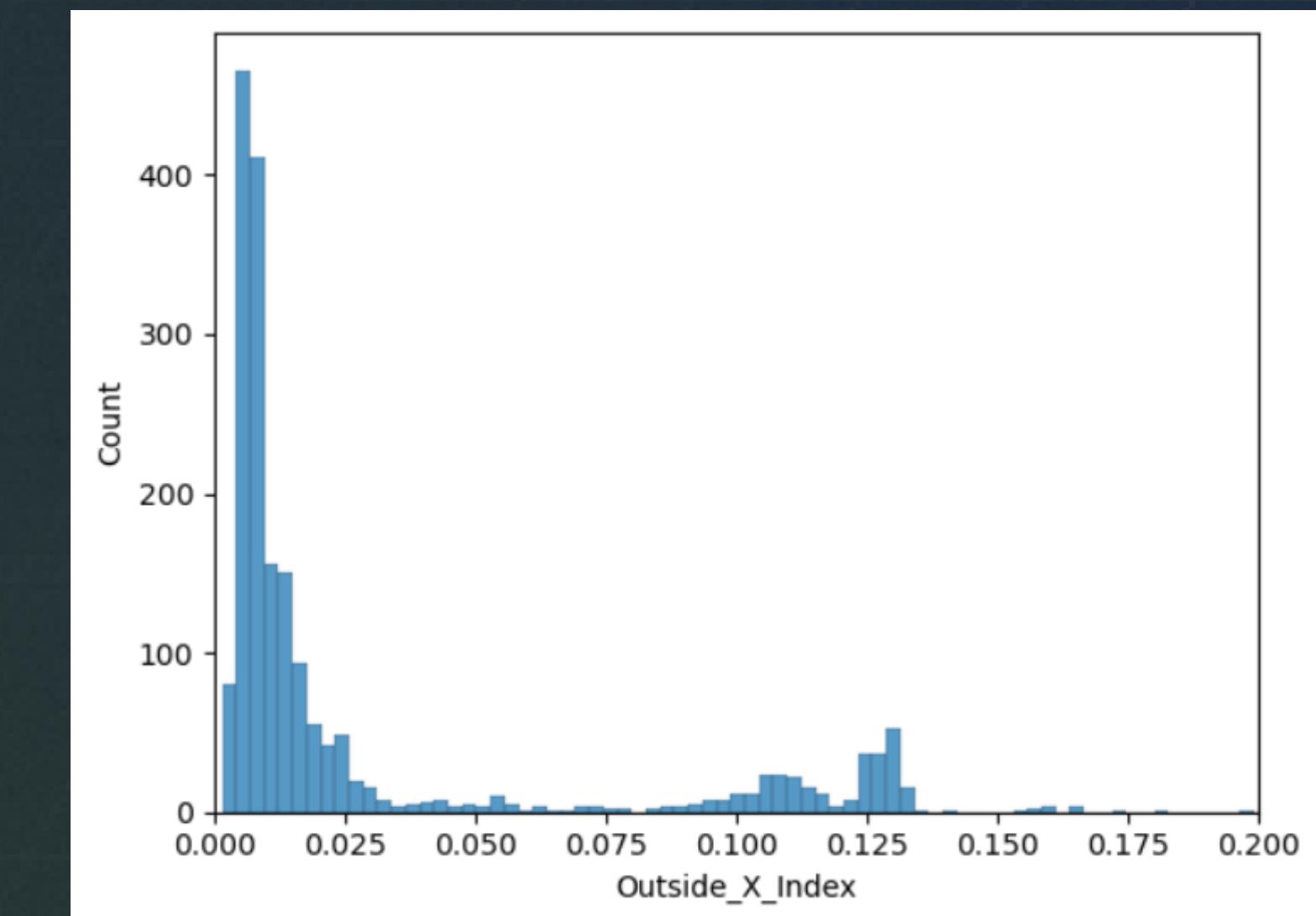
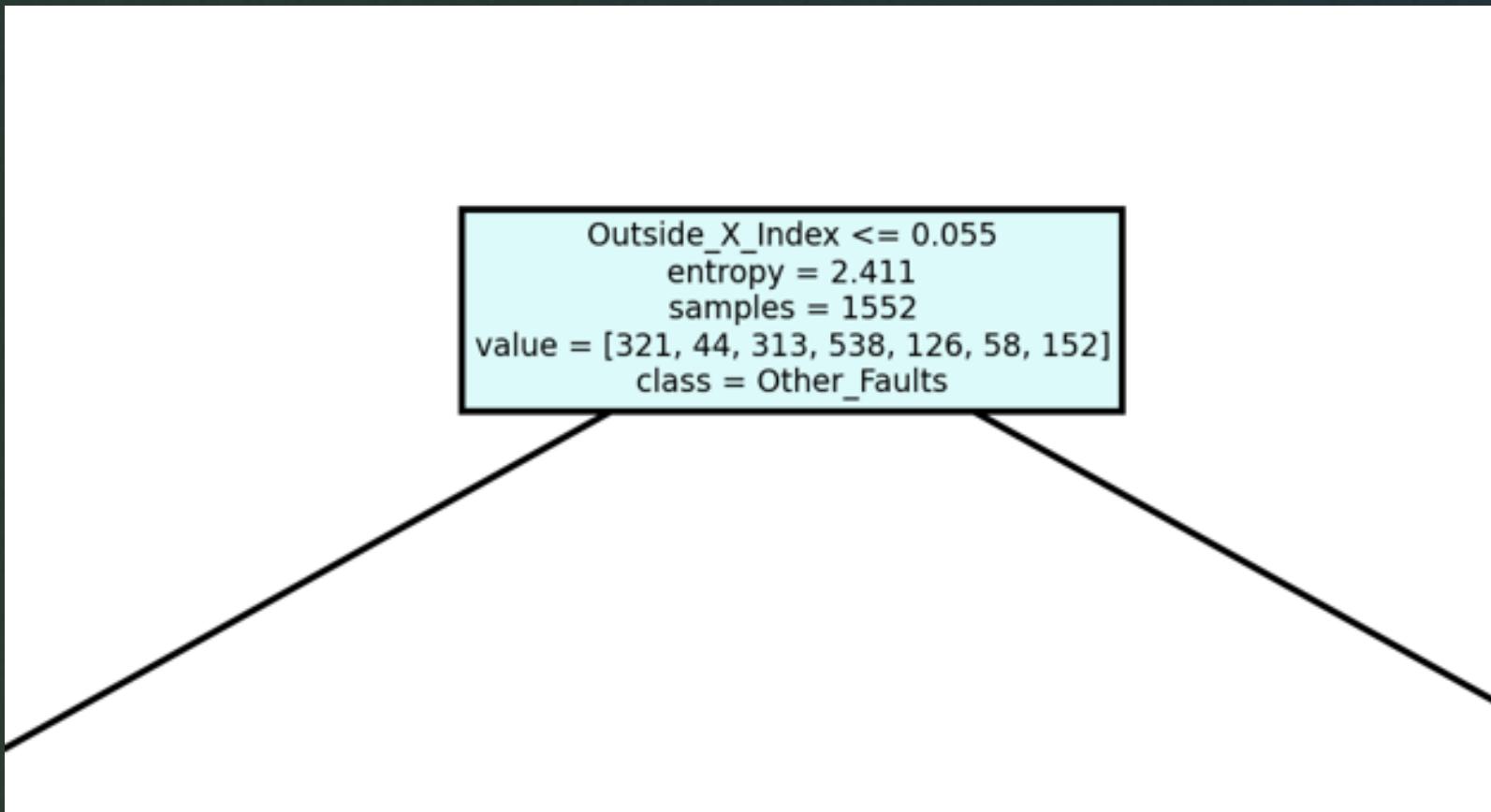
[kaggle.com](https://www.kaggle.com)



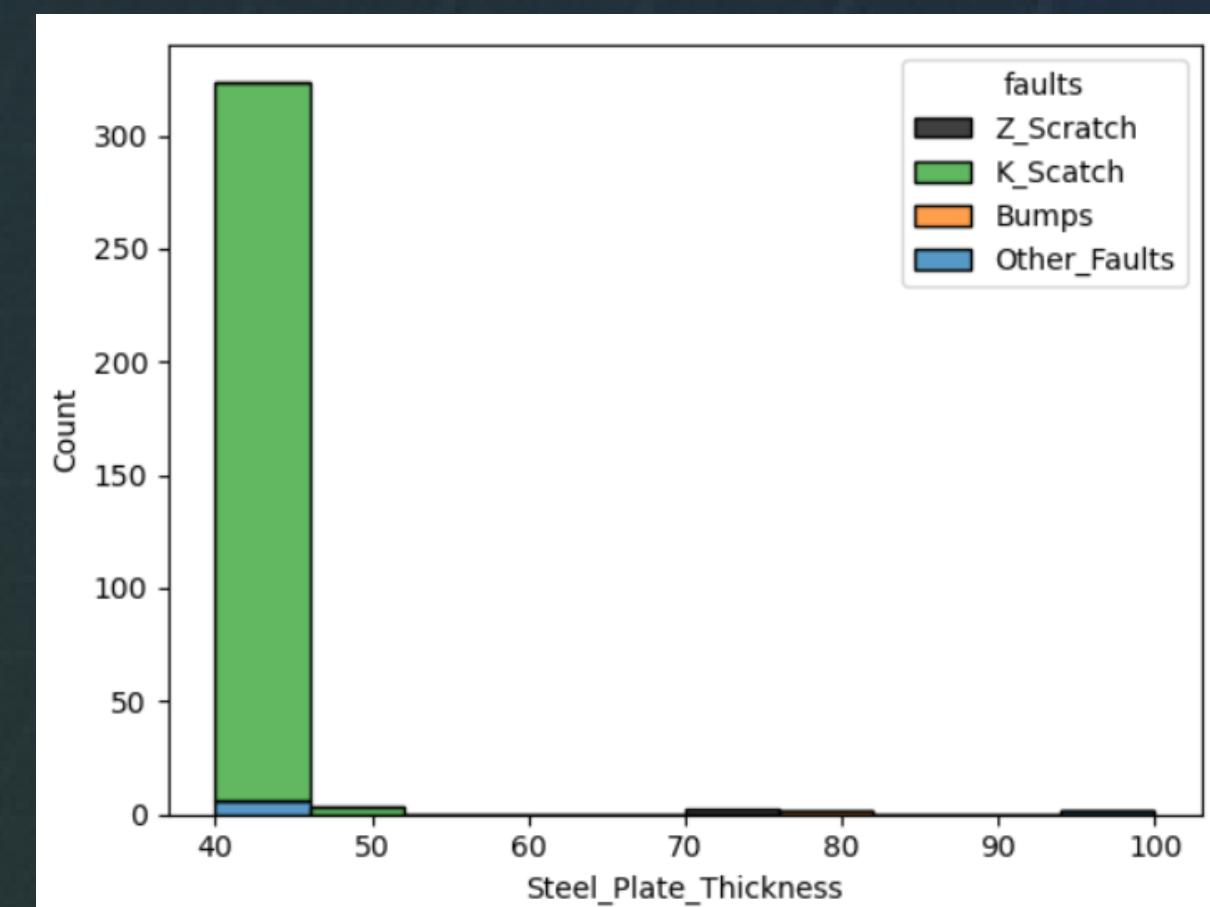
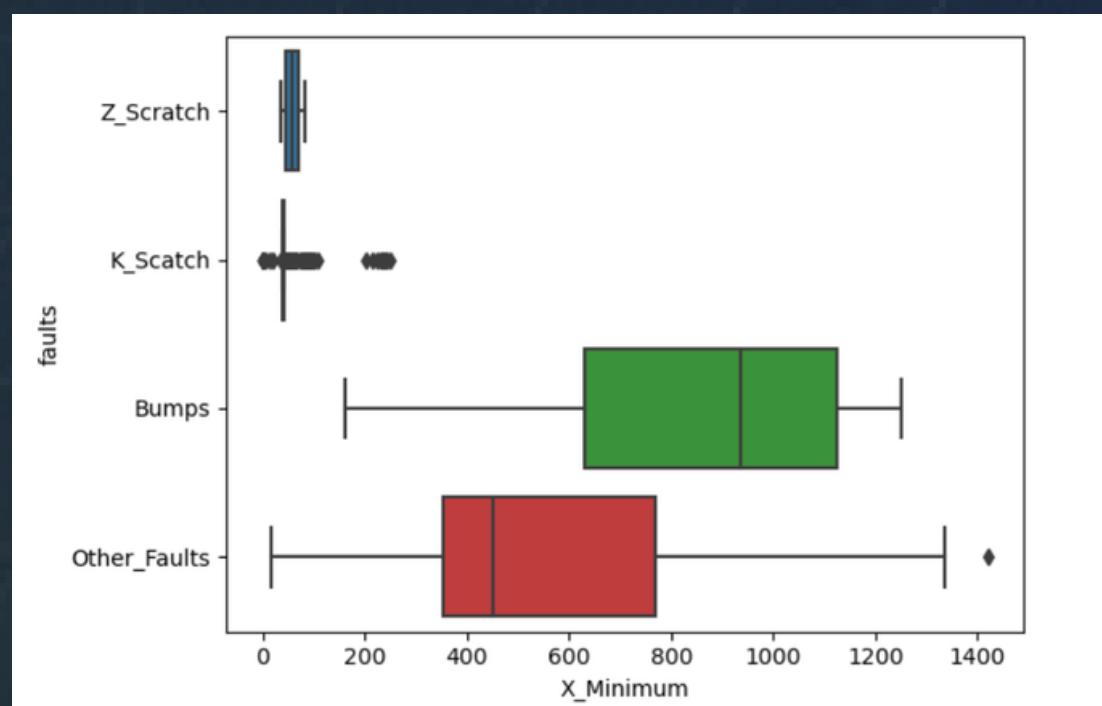
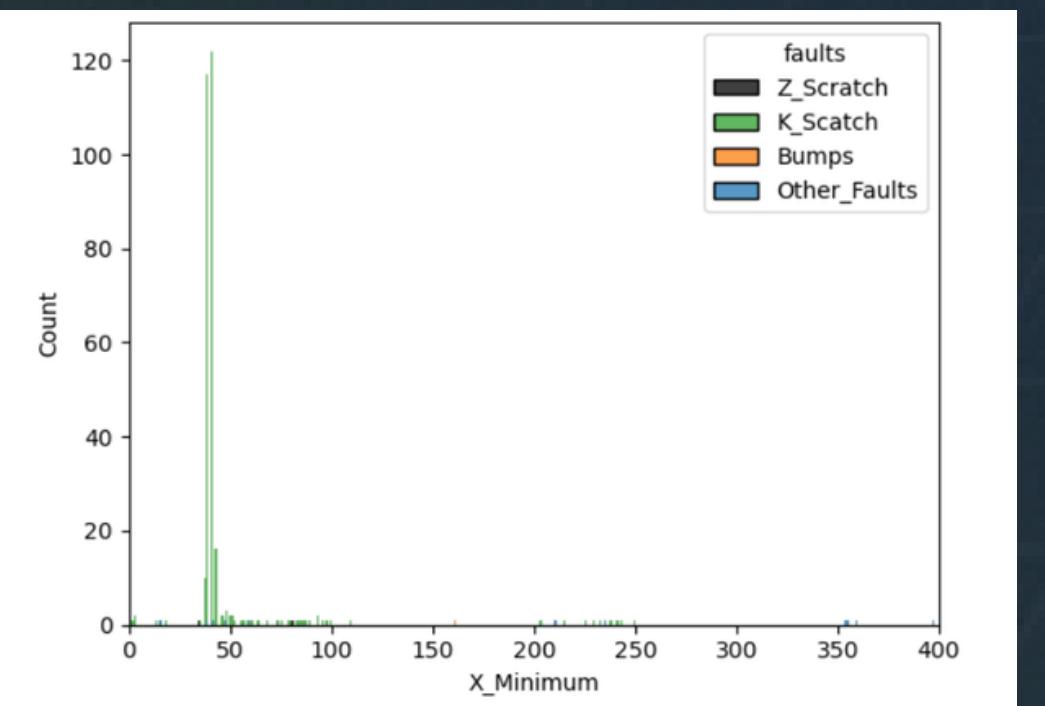
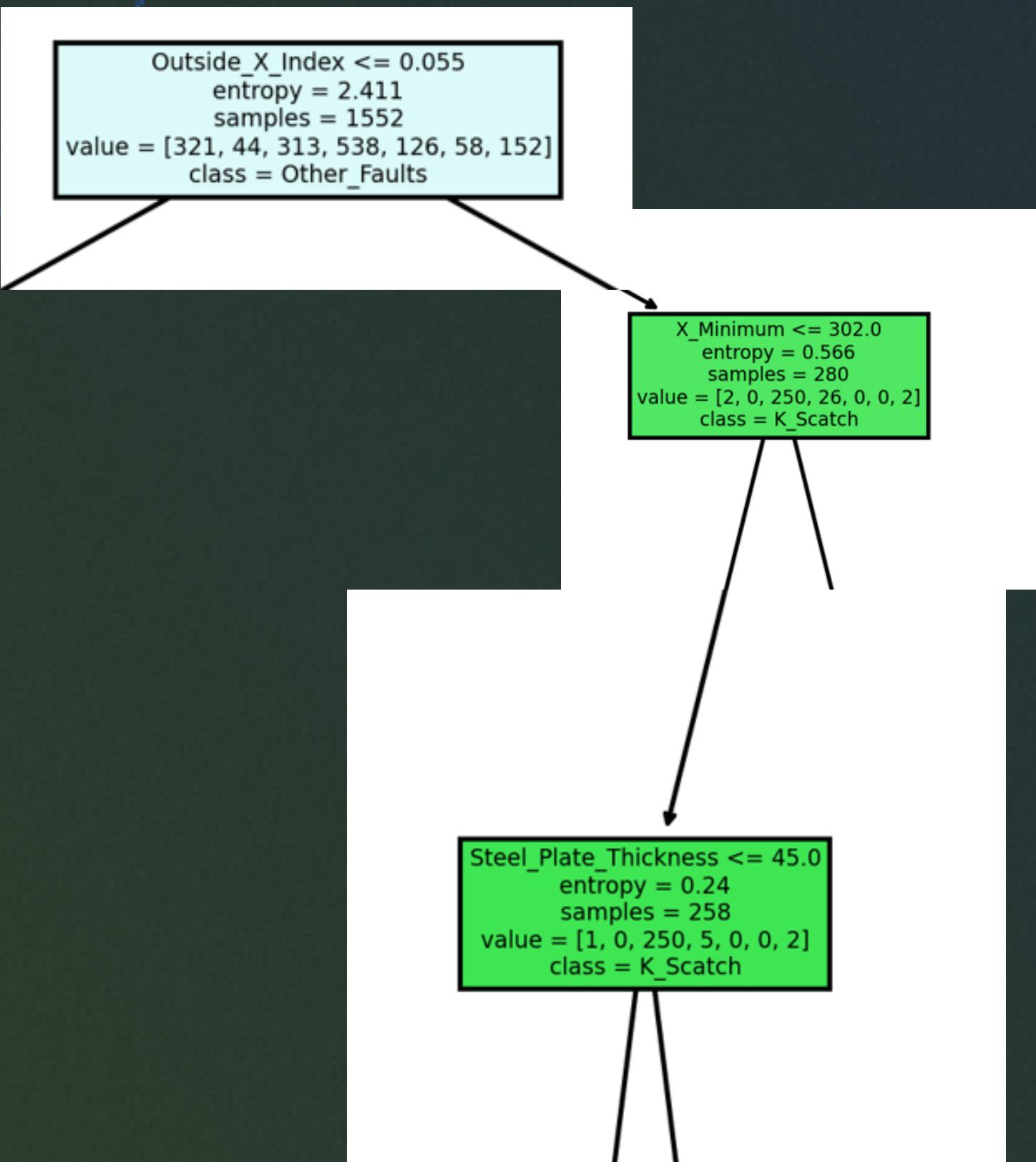
# Visualização da Árvore



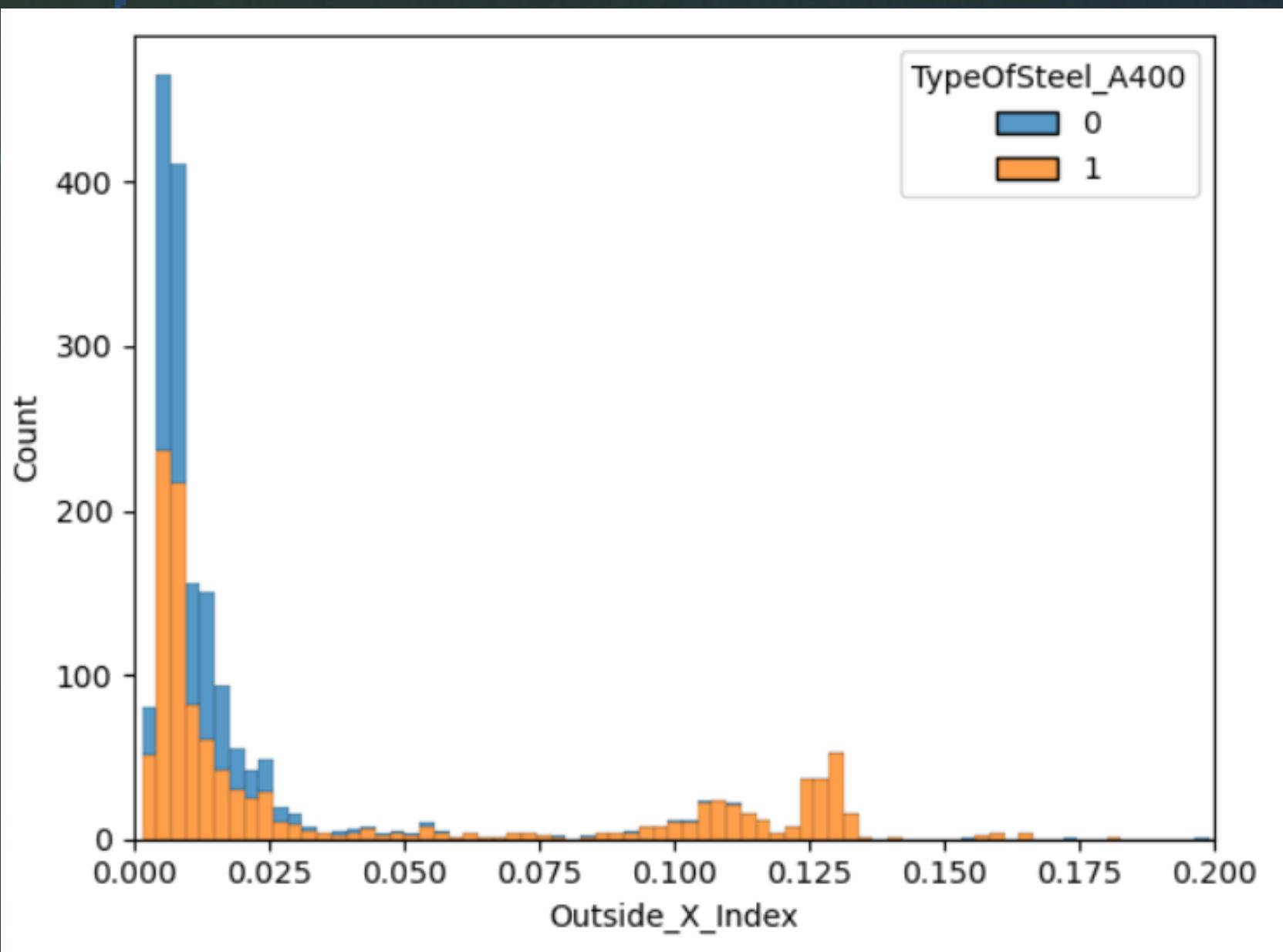
# Validação dos Resultados: Classe "K-Scatch"



# Validação dos Resultados: Classe "K-Scatch"



# Validação dos Resultados: Classe "Stains"

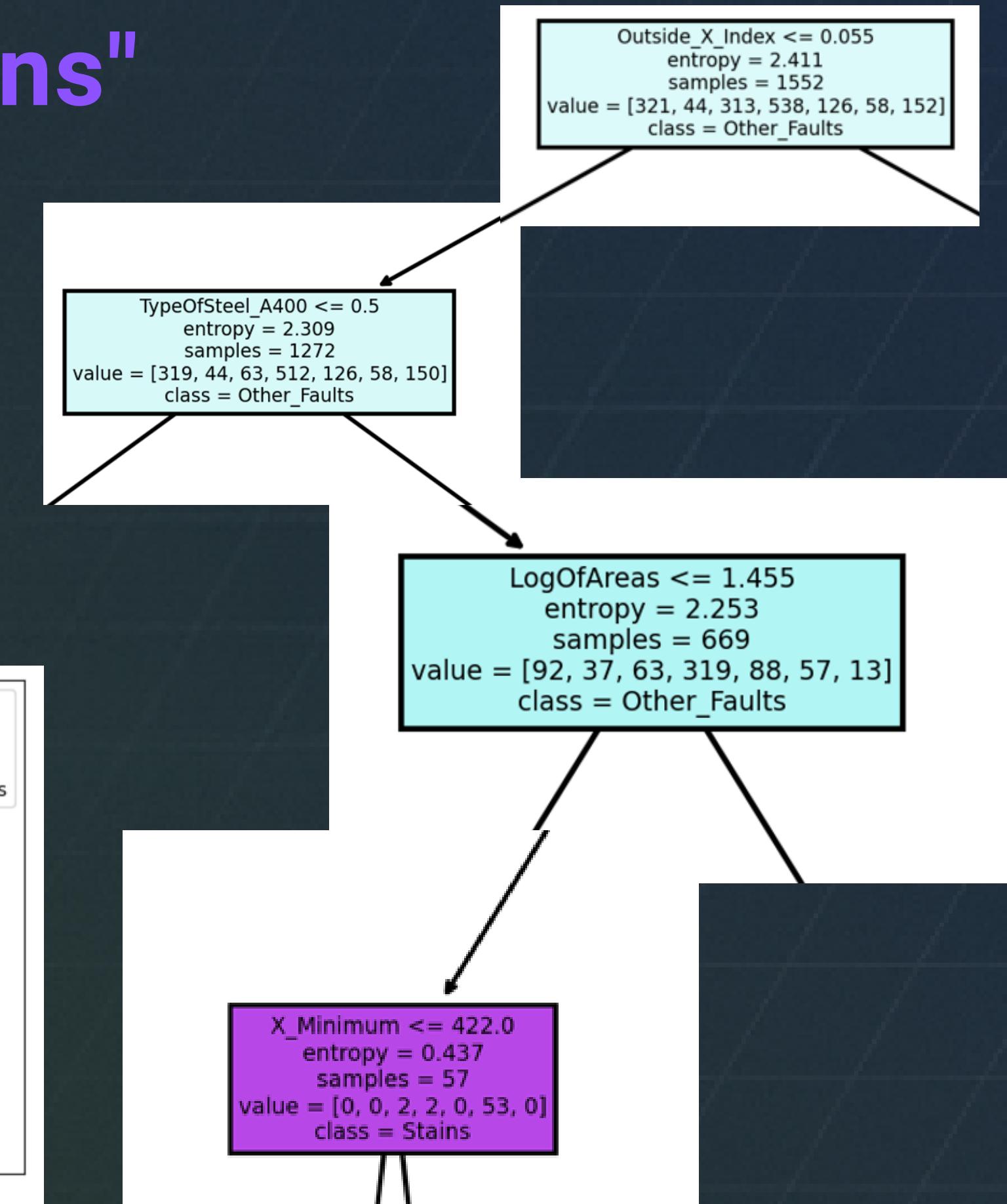
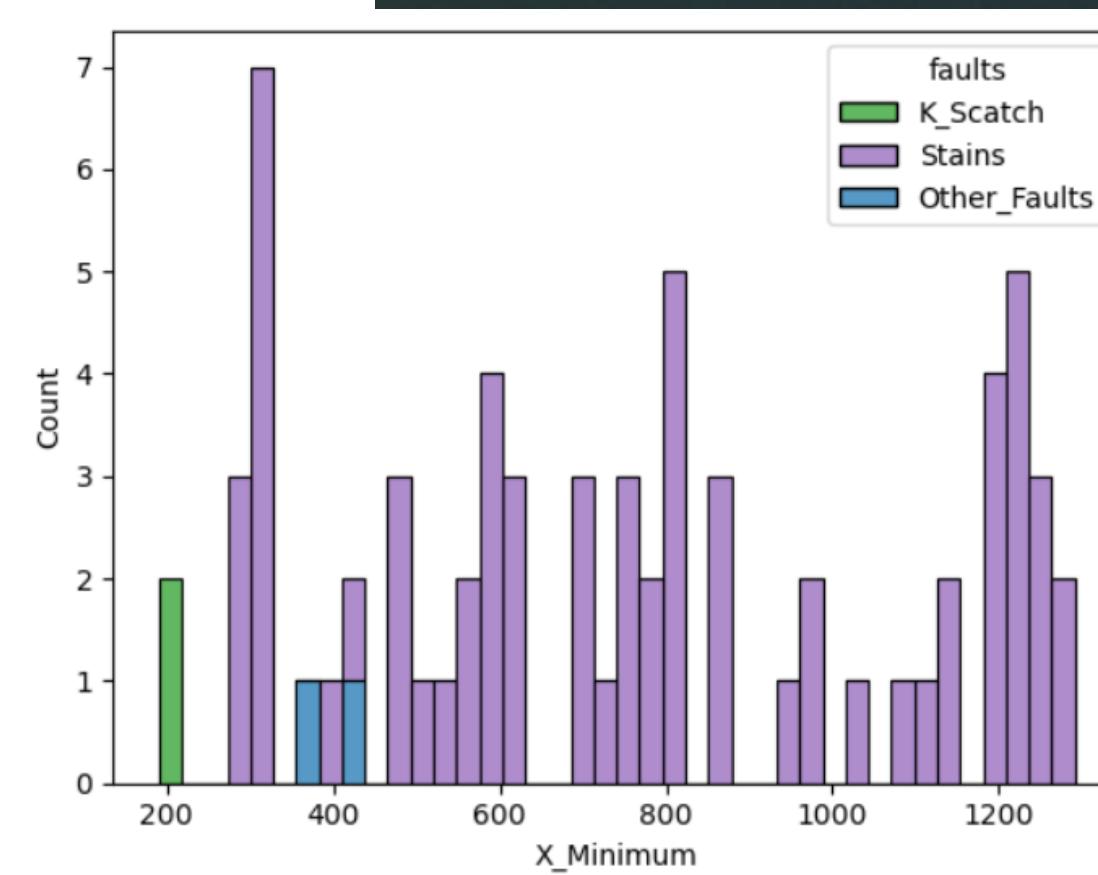
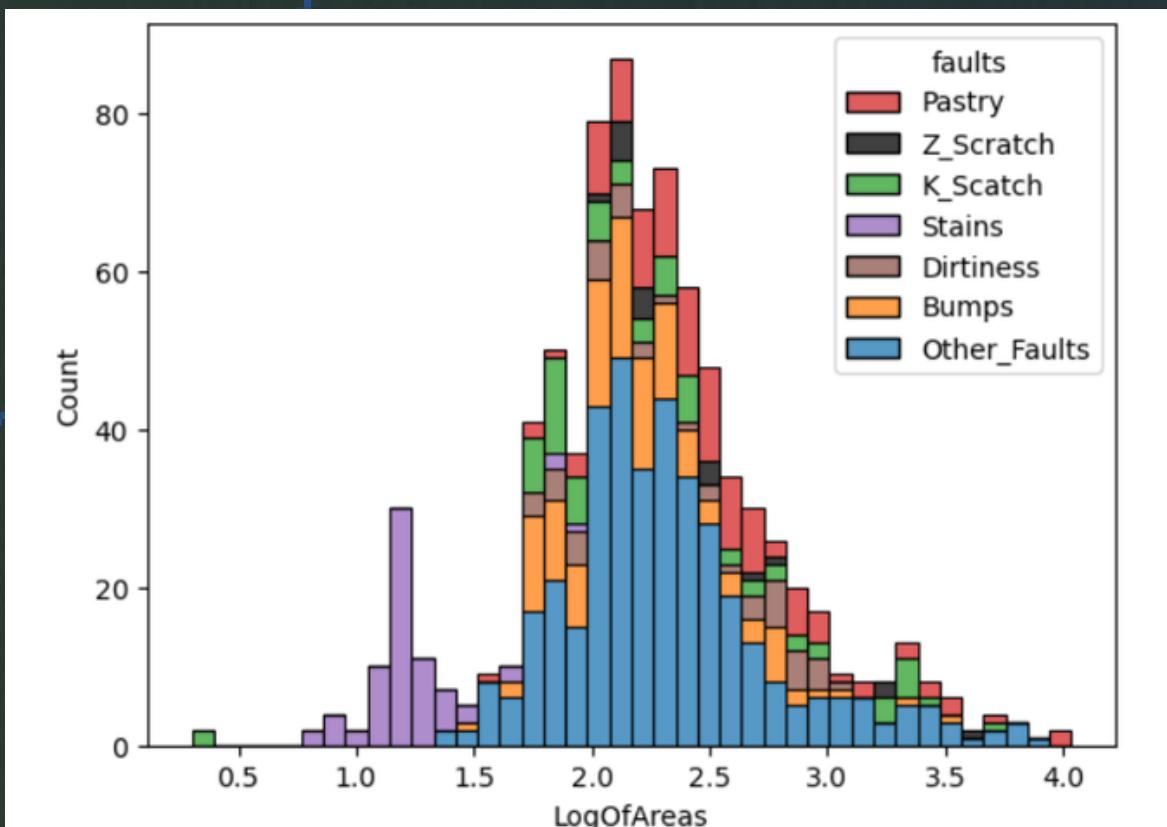


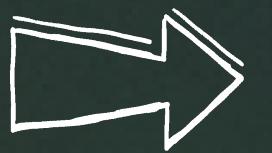
```
Outside_X_Index <= 0.055
entropy = 2.411
samples = 1552
value = [321, 44, 313, 538, 126, 58, 152]
class = Other_Faults
```

```
TypeOfSteel_A400 <= 0.5
entropy = 2.309
samples = 1272
value = [319, 44, 63, 512, 126, 58, 150]
class = Other_Faults
```



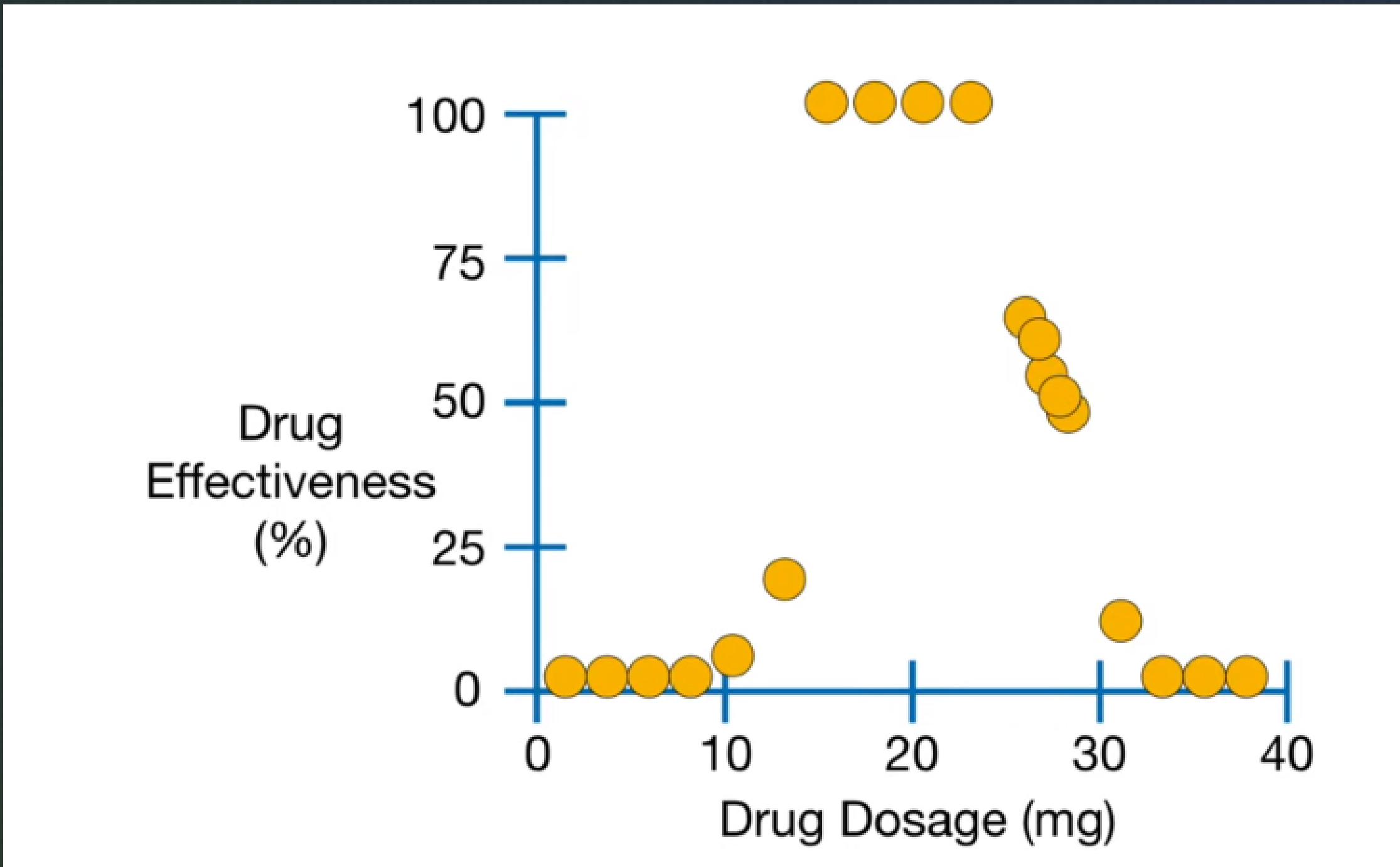
# Validação dos Resultados: Classe "Stains"



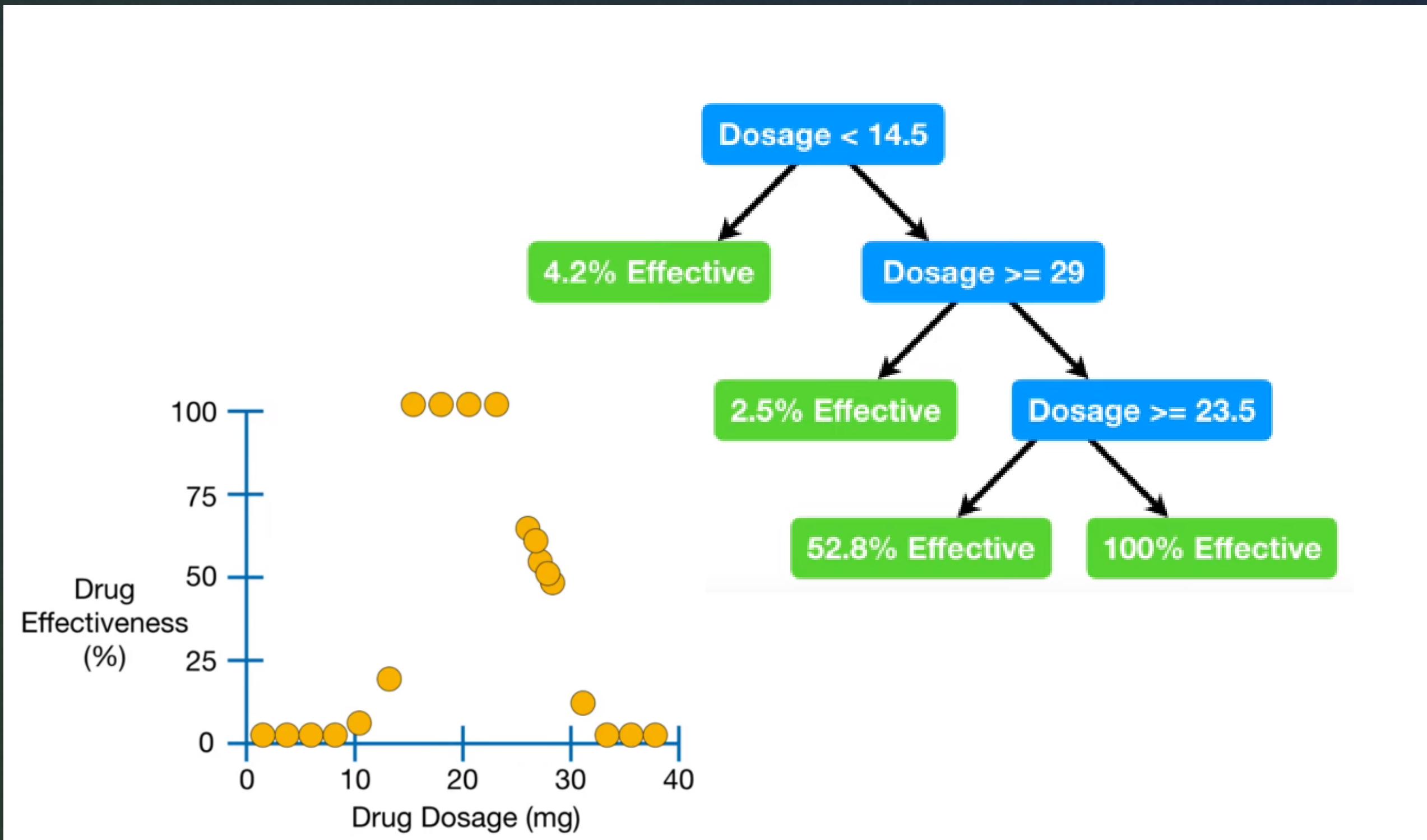


# ÁRVORES DE REGRESSÃO.

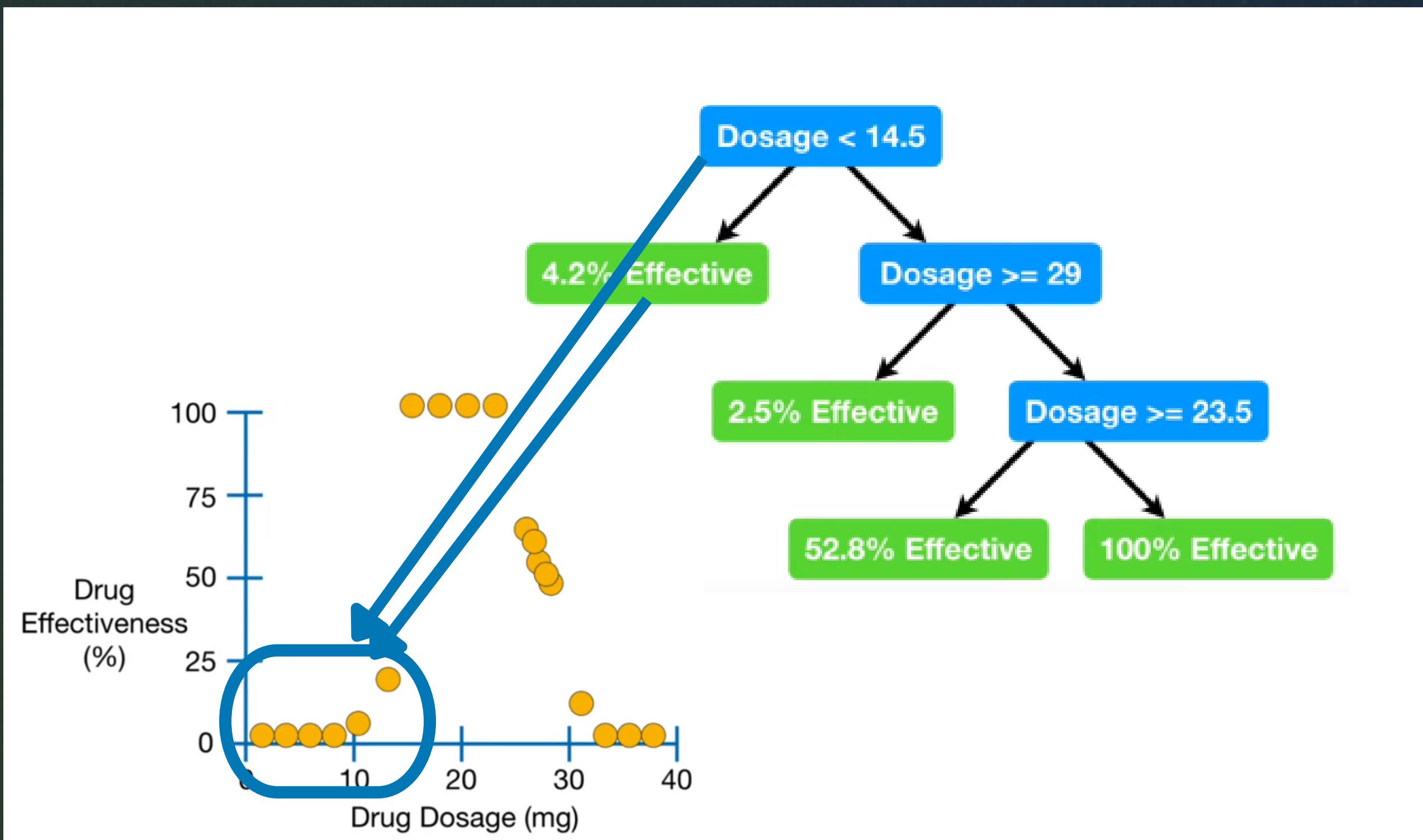
# ÁRVORES DE REGRESSÃO



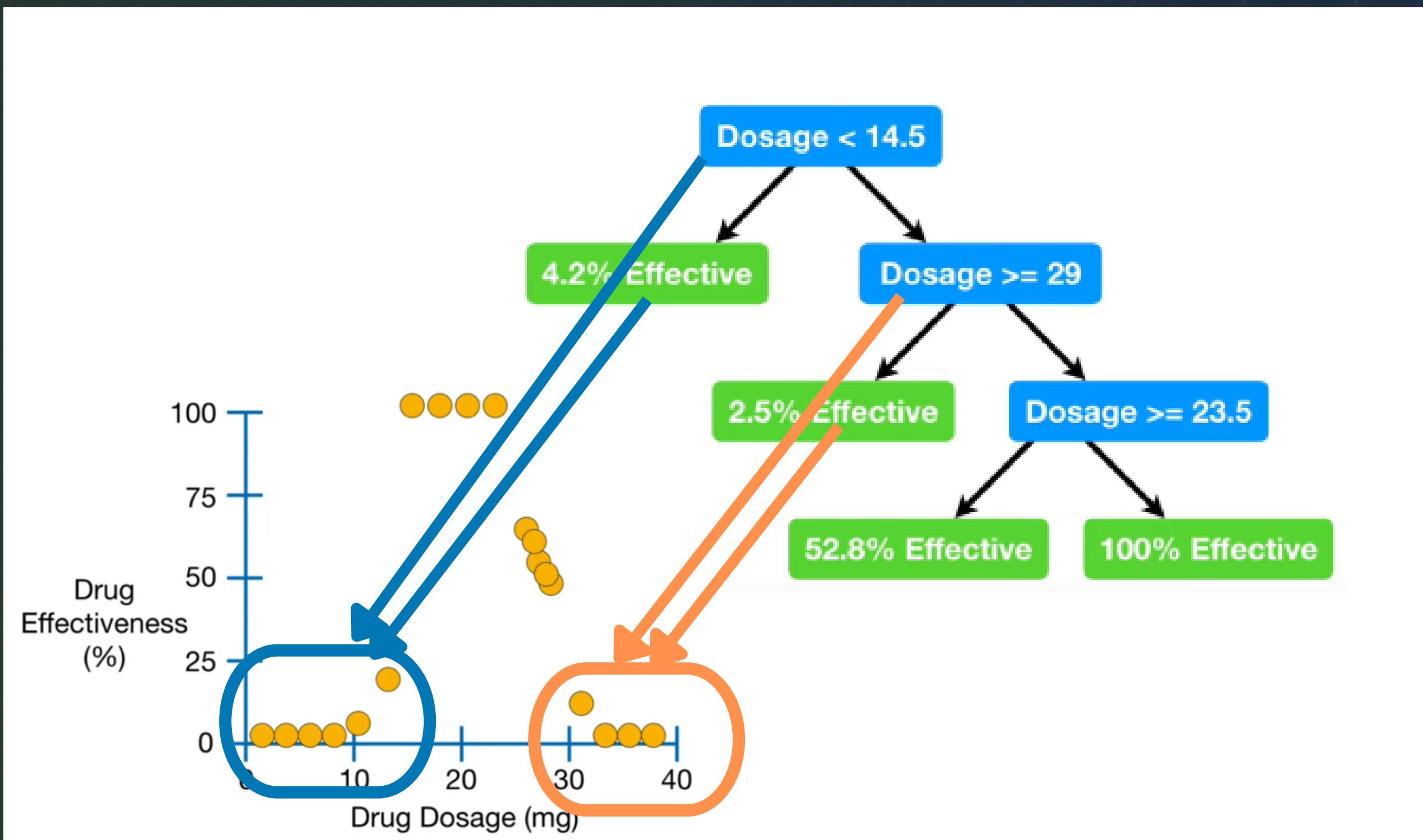
# ÁRVORES DE REGRESSÃO



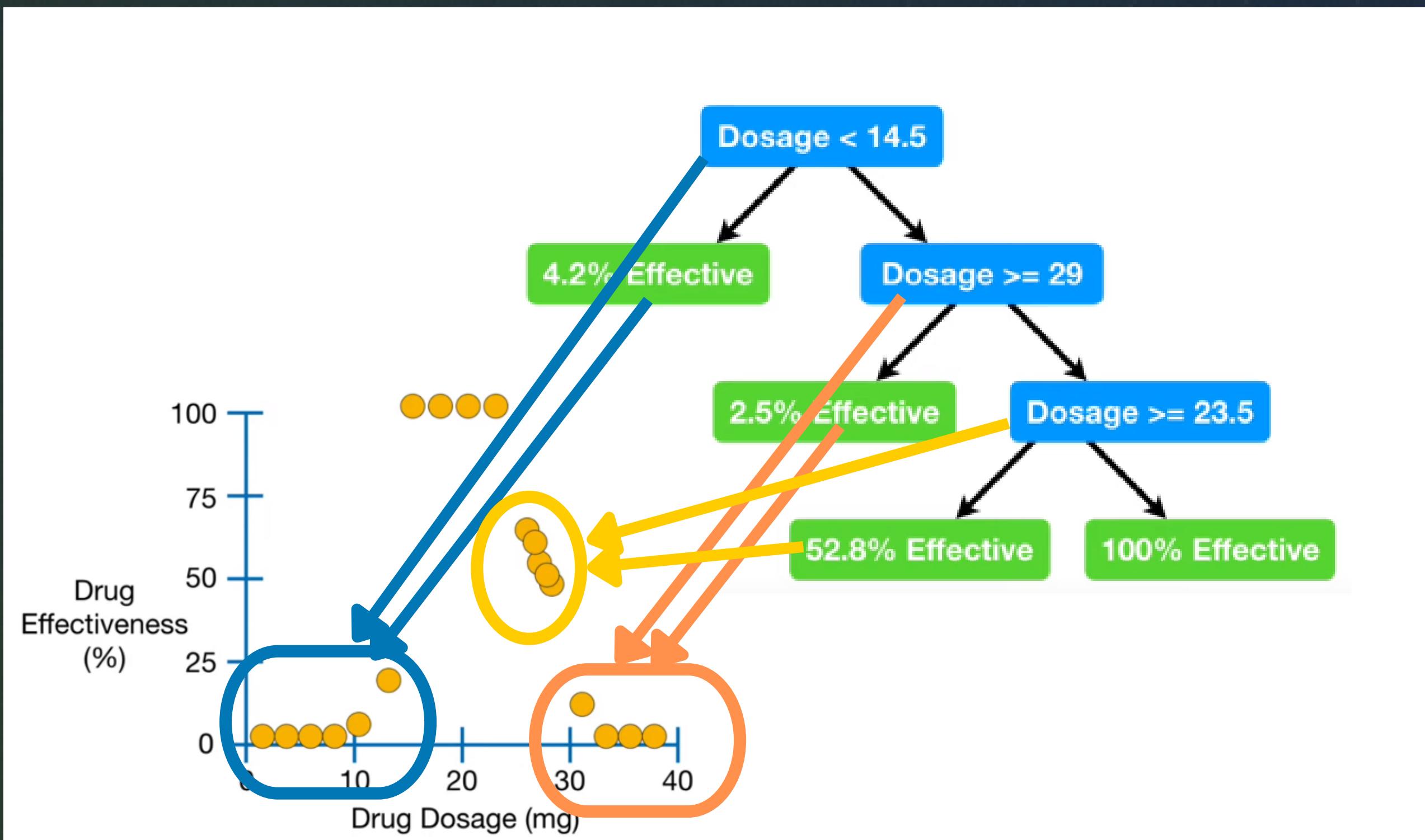
# ÁRVORES DE REGRESSÃO



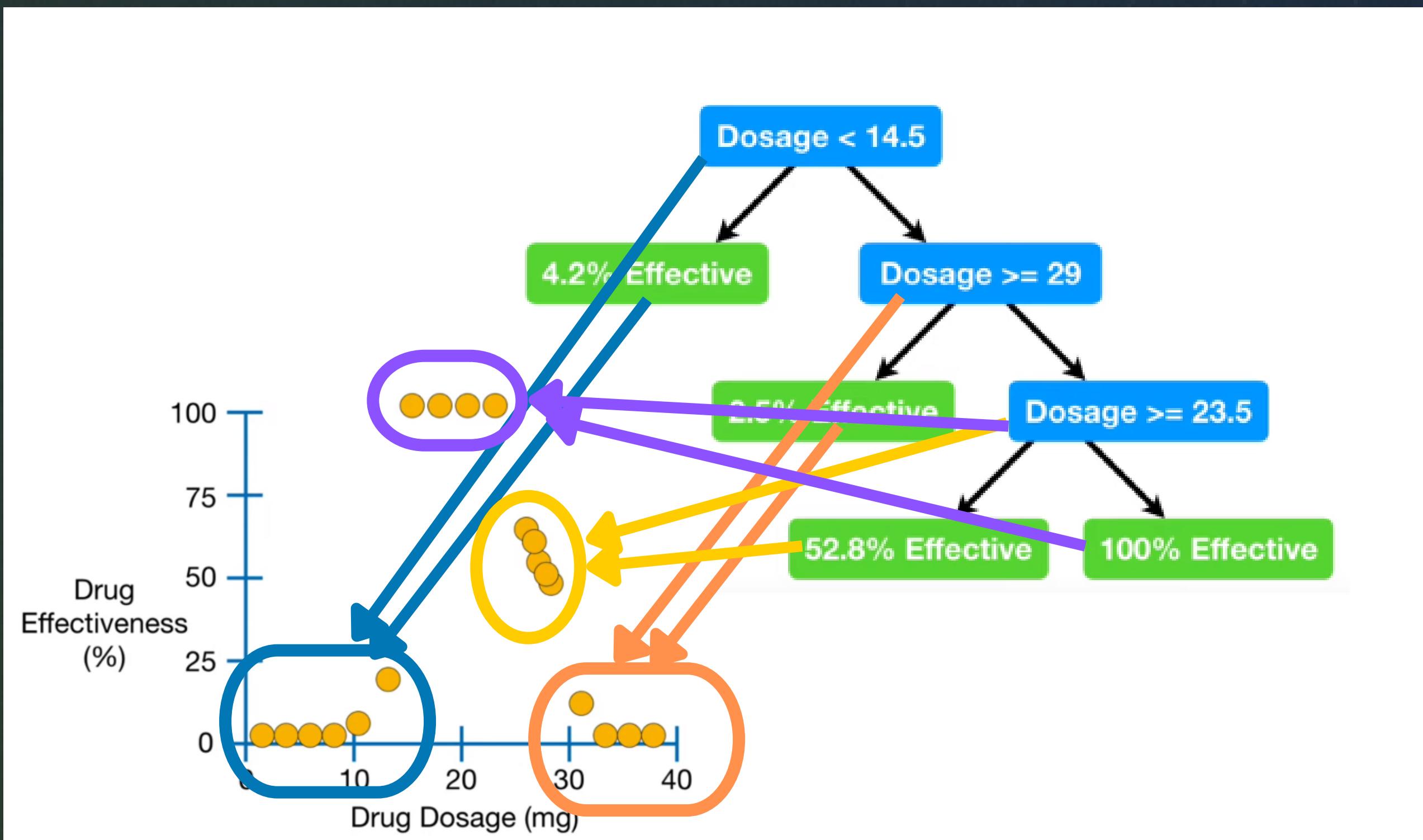
# ÁRVORES DE REGRESSÃO



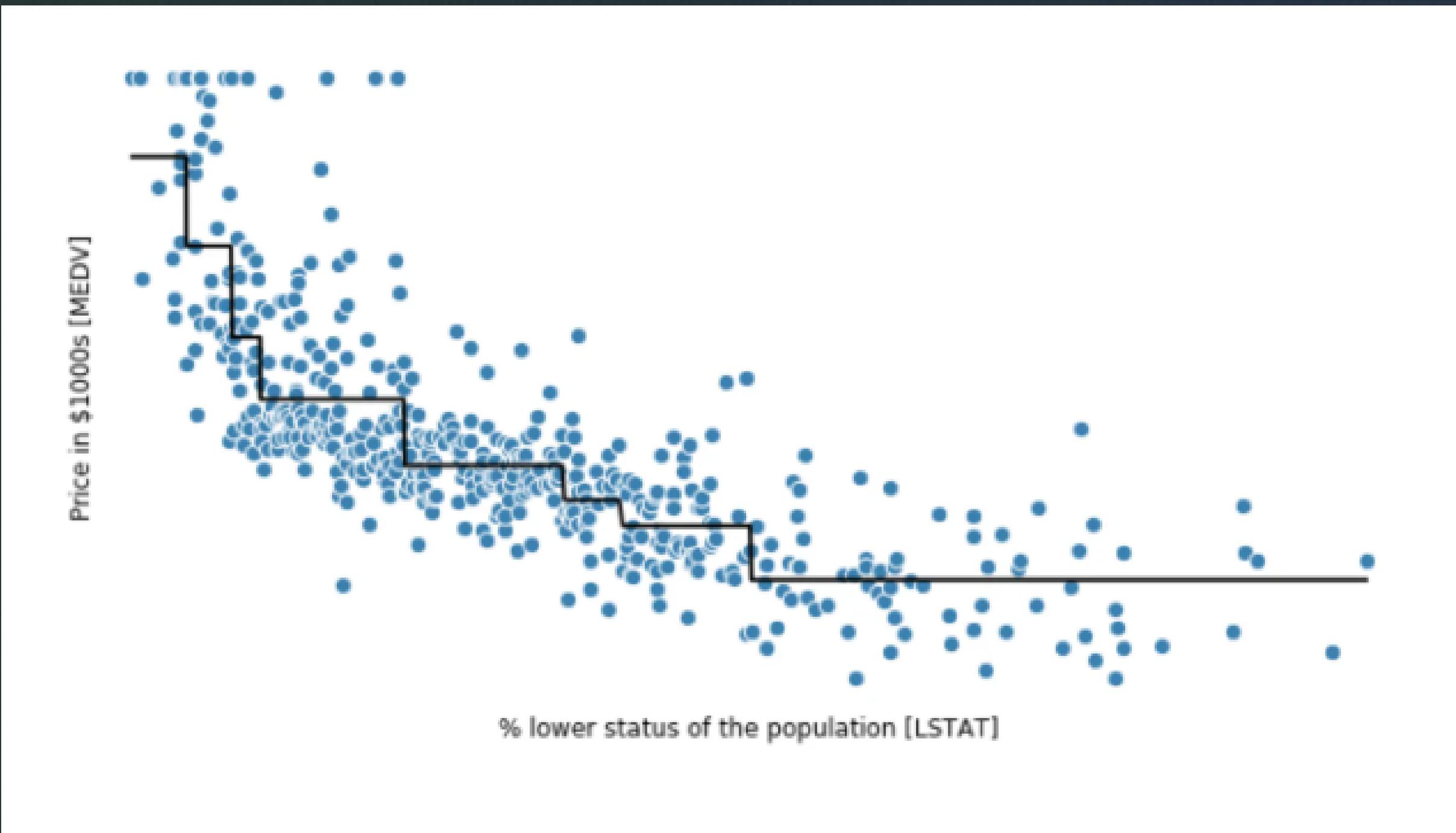
# ÁRVORES DE REGRESSÃO

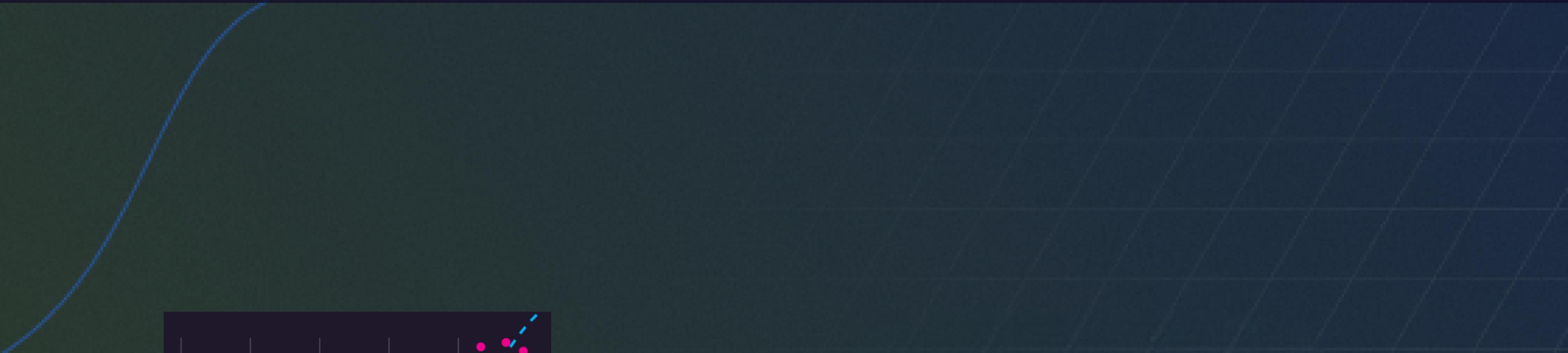


# ÁRVORES DE REGRESSÃO

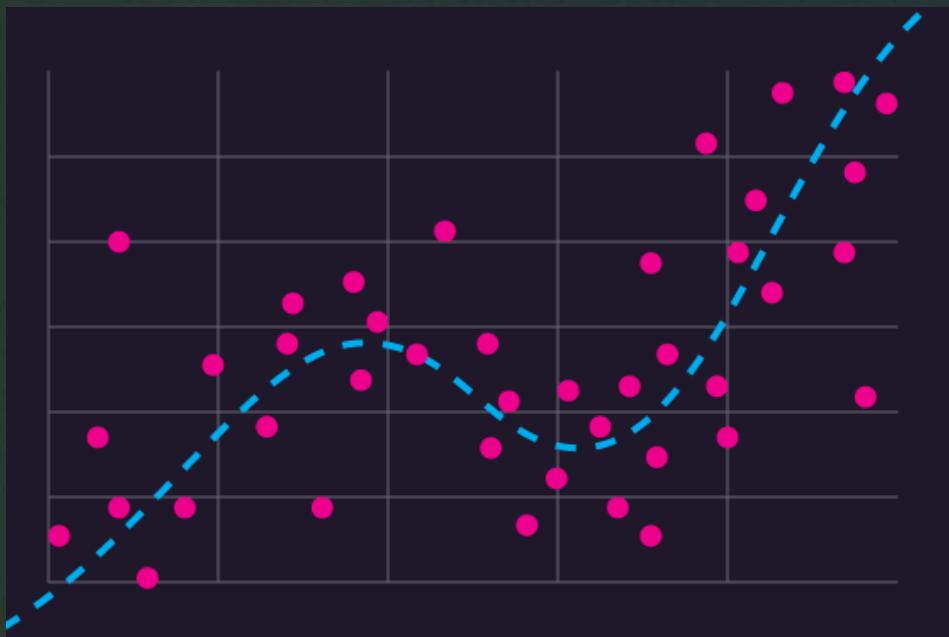


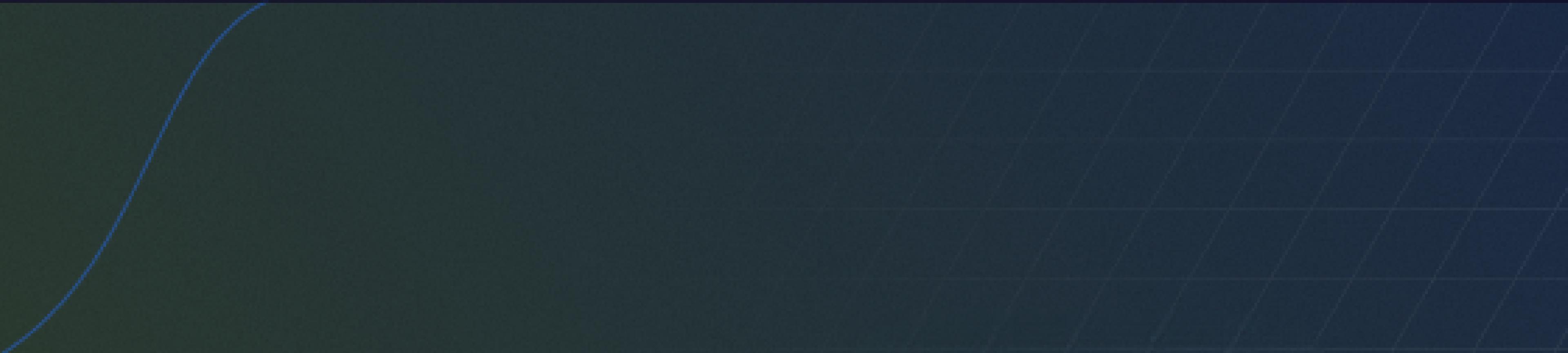
# ÁRVORES DE REGRESSÃO





**Modelo de regressão para preços de casas um pouco diferente das aulas anteriores.**





**MUITO OBRIGADO!**