REGRESSÃO

MANOELA KOHLER

Prof.manoela@ica.ele.puc-rio.br

TÓPICOS

R

Análise exploratória

Pré-processamento

- Balanceamento
- *Outliers*
- *Missing values*
- Normalização
- Seleção de atributos (Filtros, *Wrappers*, PCA)

Associação:

- *Apriori*
- *FP-Growth*
- *Eclat*

Classificação:

- Regressão logística
- *Support Vector Machine* (SVM)
- Árvores de Decisão
- *Random Forest*
- ~~Redes Neurais~~
- *K nearest neighbors*

Regressão

- Regressão linear simples
- Regressão linear múltipla
- Regressão não linear simples
- Regressão não linear múltipla

Agrupamento

- *Particionamento (K-means, K-medoids)*
- *Hierárquico (DIANA, AGNES)*
- *Densidade (DBSCAN)*

Séries Temporais

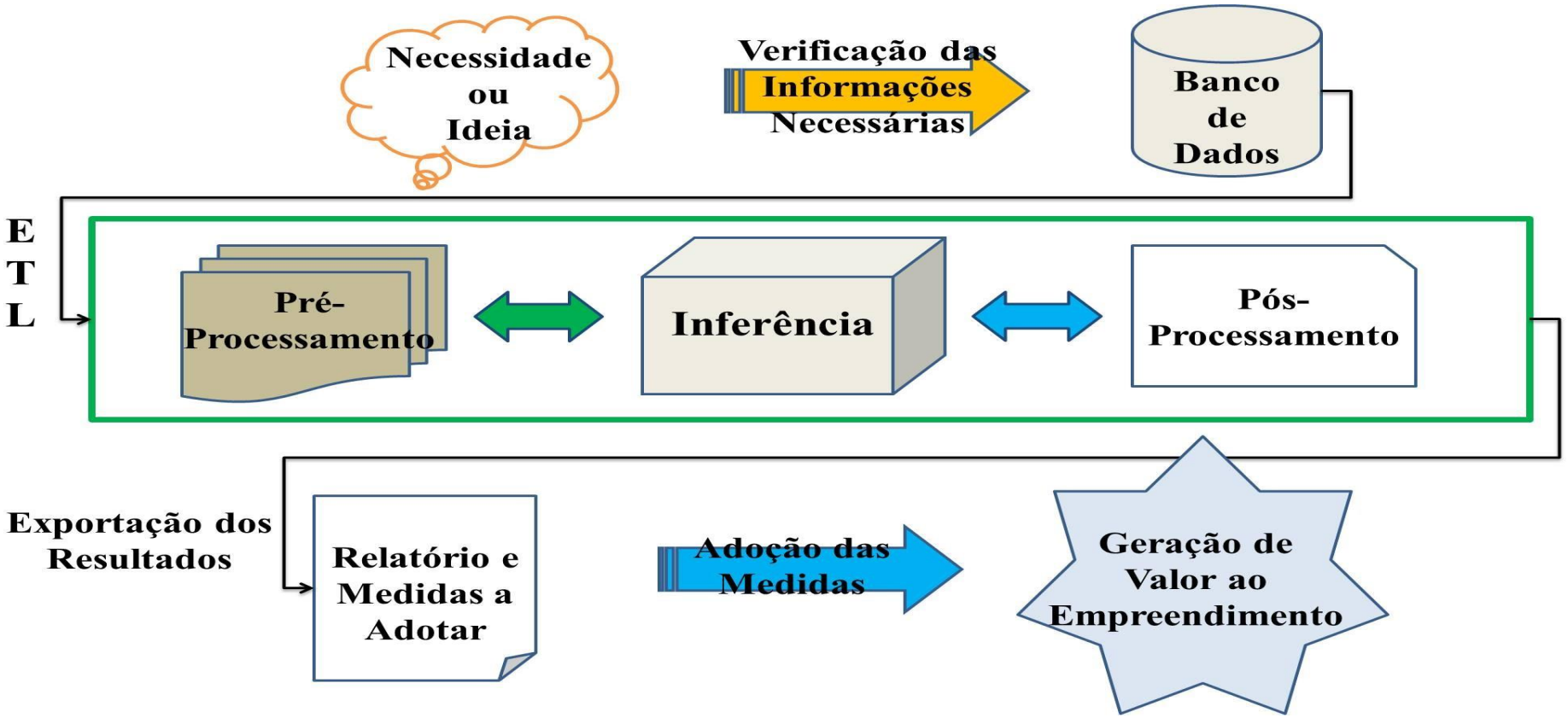
- Naive
- Média Móvel
- Amortecimento exponencial
- Auto-regressivo integrados de média móvel
- Auto regressivo não linear

BI Master

Recapitulação

ETAPAS DE UM PROJETO DE DATA MINING

ESQUEMA BÁSICO DE UM PROJETO DE DM



AGRUPAMENTO (CLUSTERIZAÇÃO)

Aprendizado não supervisionado: Agrupamento

- K-means (Clusterização baseada em Particionamento)
- Clusterização Hierárquica
- Clusterização baseada em Densidade

MÉTODOS DE CLUSTERIZAÇÃO

- **Particionamento:** Constrói várias partições e as avalia usando algum critério.
- **Hierárquico:** Cria uma decomposição hierárquica dos objetos usando algum critério.
- **Baseado em densidade:** Fundamenta-se em funções de conectividade e de densidade.

ESTUDO DE CASO

ESTUDO DE CASO

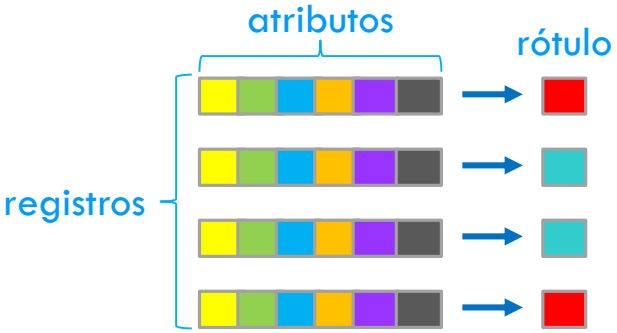
Clientes de um Shopping

- Ganho Anual
- Traço de Gastos

Machine Learning

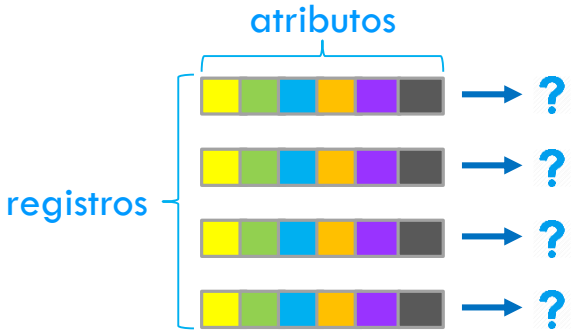
Supervisionado

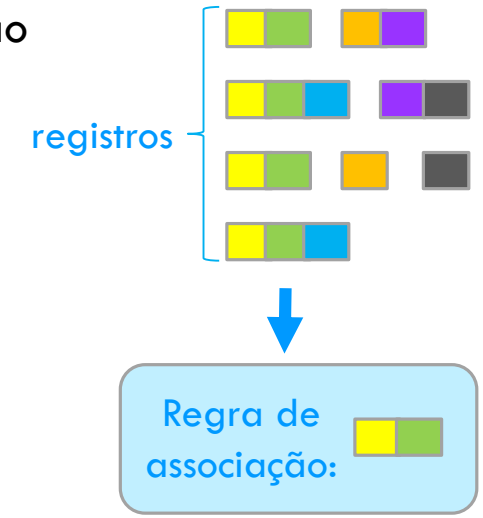
Classificação
Regressão
Previsão de séries temporais



Não Supervisionado

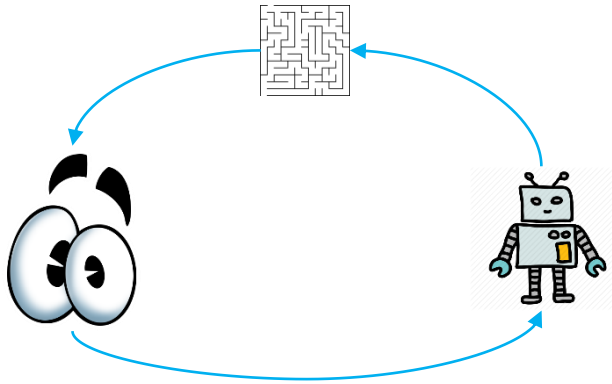
Agrupamento
Associação





Reforço

Aprendizado através da interação de agentes com um ambiente

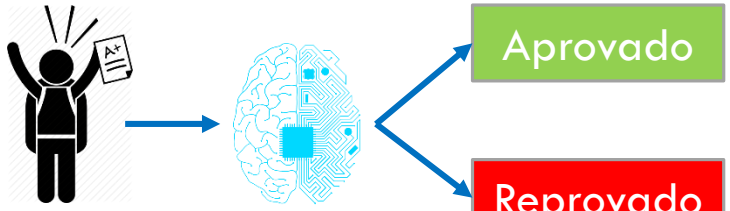


SUPERVISIONADO

- Aproximador: função mapeia entradas e saída.

Classificação

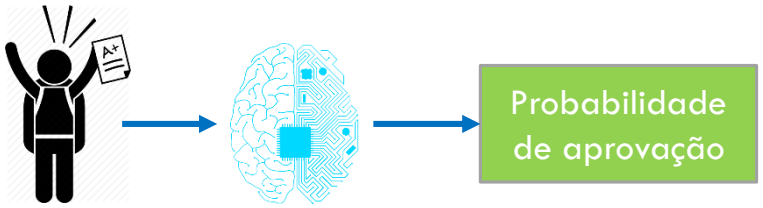
Rótulo é categórico.



Dados do estudante

Regressão

Rótulo é contínuo.



Dados do estudante

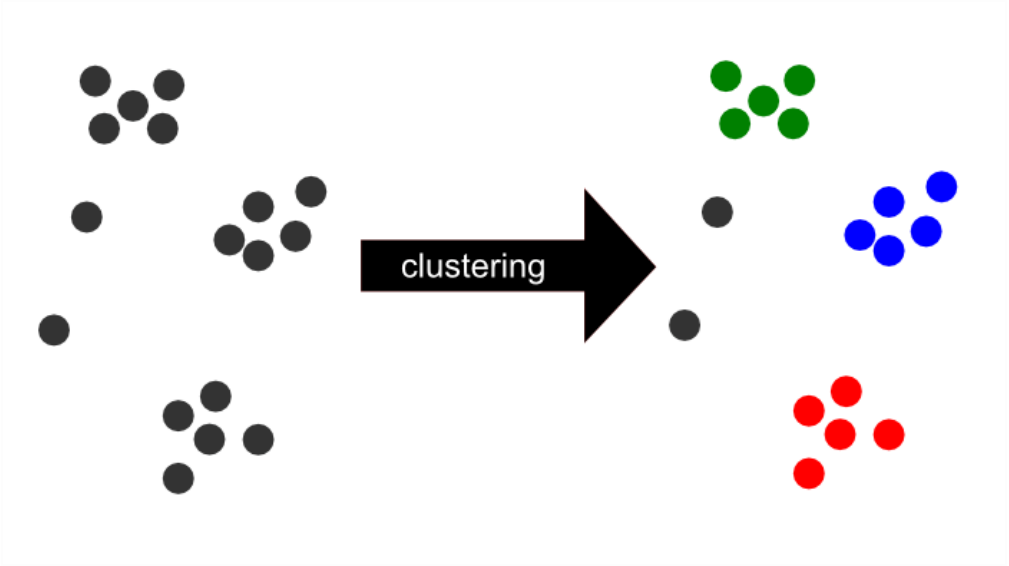
Previsão de Séries

Rótulo é contínuo e dependente do tempo.

NÃO SUPERVISIONADO

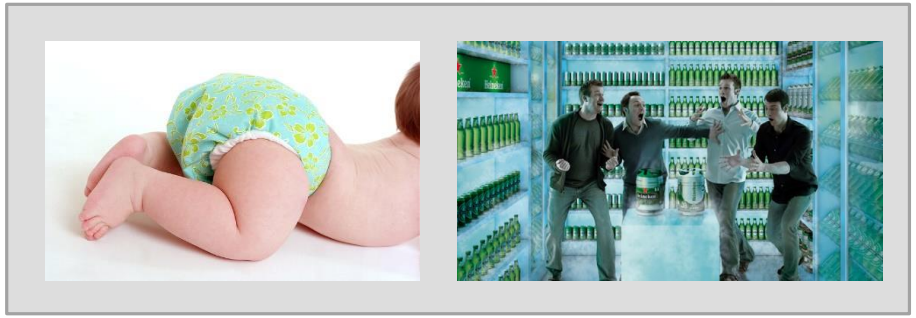
Agrupamento

Descoberta de semelhanças e grupos entre registros.



Associação

Descoberta de relações entre variáveis.



REGRESSÃO

CORRELAÇÃO

Indica a força e a direção do relacionamento entre dois atributos;

Trata-se de uma medida da relação entre dois atributos, embora correlação não implique causalidade:

- Duas variáveis podem estar altamente correlacionadas e não existir relação de causa e efeito entre elas.

Ela permite verificar se é possível ajustar um modelo que expresse a mencionada relação;

Esse é o objetivo da análise de regressão.

REGRESSÃO

Existe uma série de técnicas voltada para a modelagem e a investigação de relações entre dois ou mais atributos.

Exemplo:

- Na análise de correlação linear, o objetivo é determinar o grau de relacionamento entre duas variáveis.
- Já na análise de regressão linear, o objetivo é determinar o modelo que expressa esta relação (equação de regressão), a qual é ajustada aos dados.

REGRESSÃO

Para que serve?

Podemos usar esse modelo para predizer o valor de y para um dado valor de x

- Realizar previsões sobre o comportamento futuro de algum fenômeno da realidade.
- Neste caso extrapola-se para o futuro as relações de causa-efeito – já observadas no passado – entre as variáveis.

REGRESSÃO

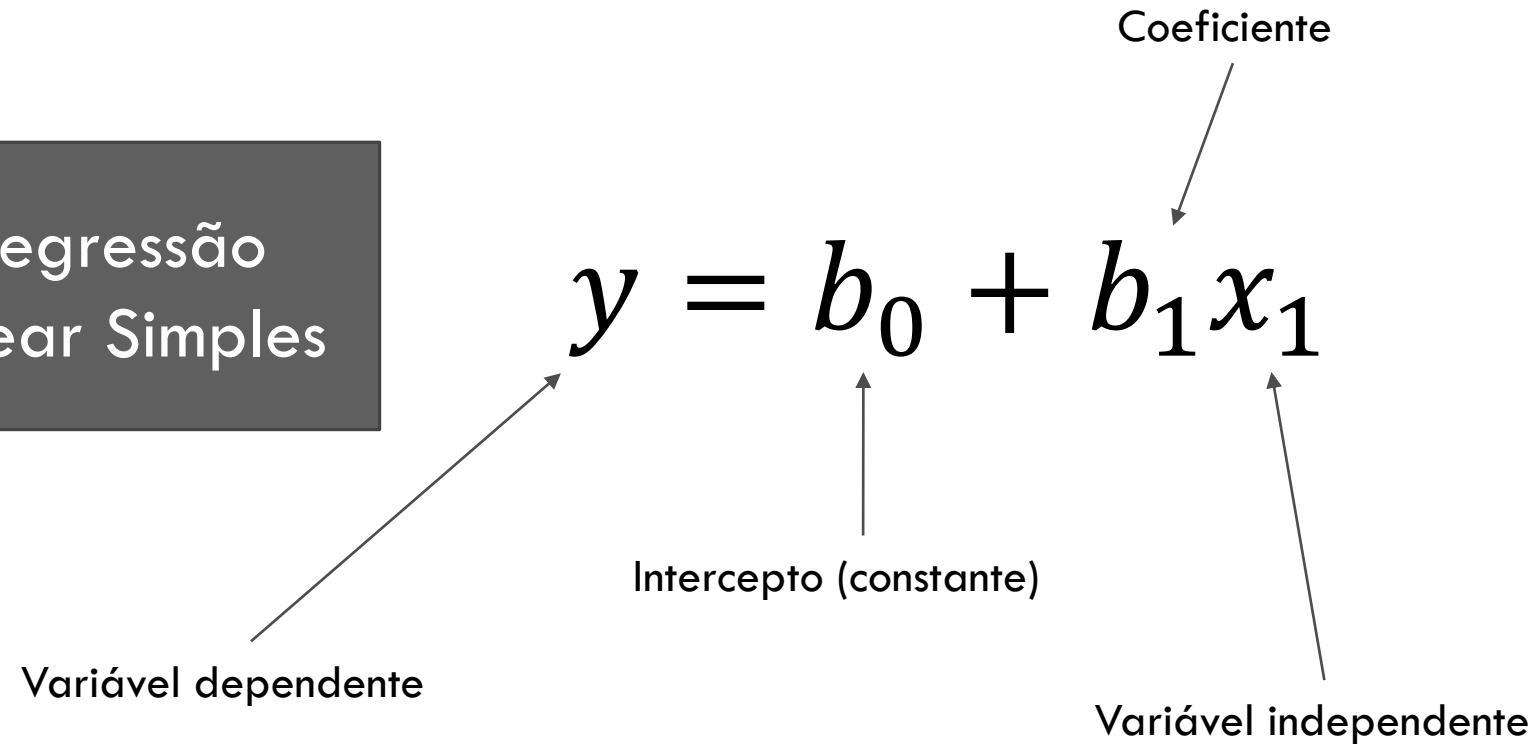
A análise de regressão compreende quatro tipos básicos de modelos:

- Linear simples;
- Linear Múltipla;
- Não linear simples;
- Não linear múltipla.

Regressão Linear Simples

REGRESSÃO

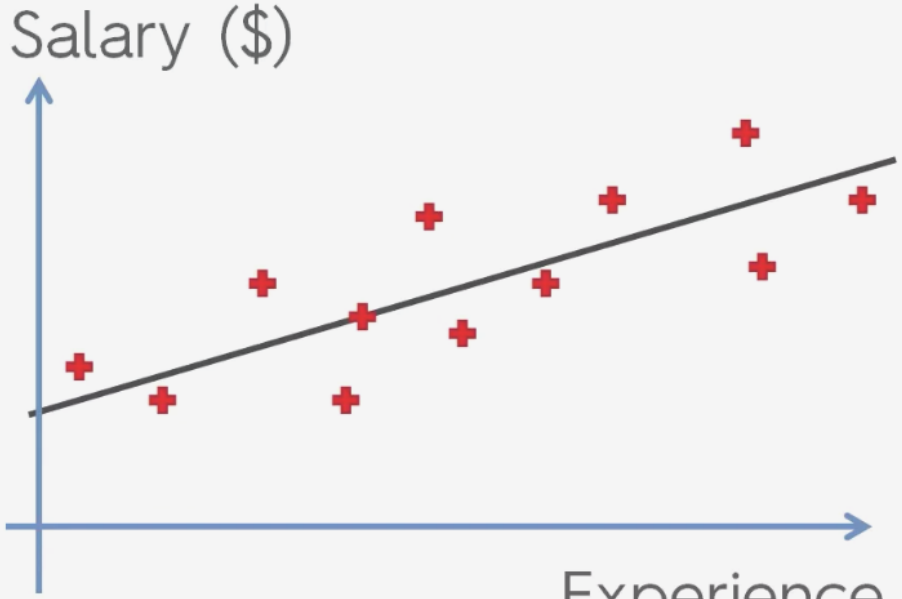
Regressão Linear Simples



REGRESSÃO

REGRESSÃO

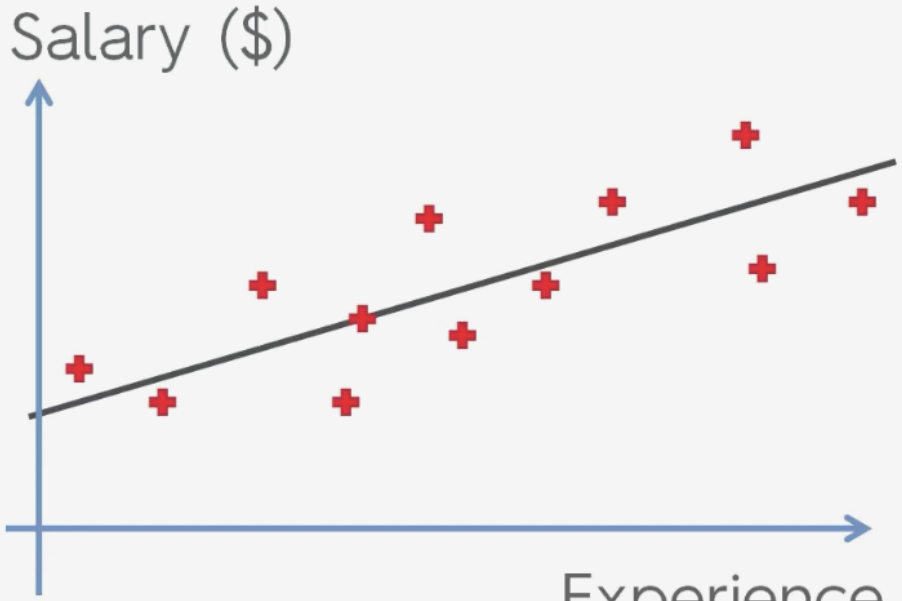
Simple Linear Regression:



$$y = b_0 + b_1 * x$$

REGRESSÃO

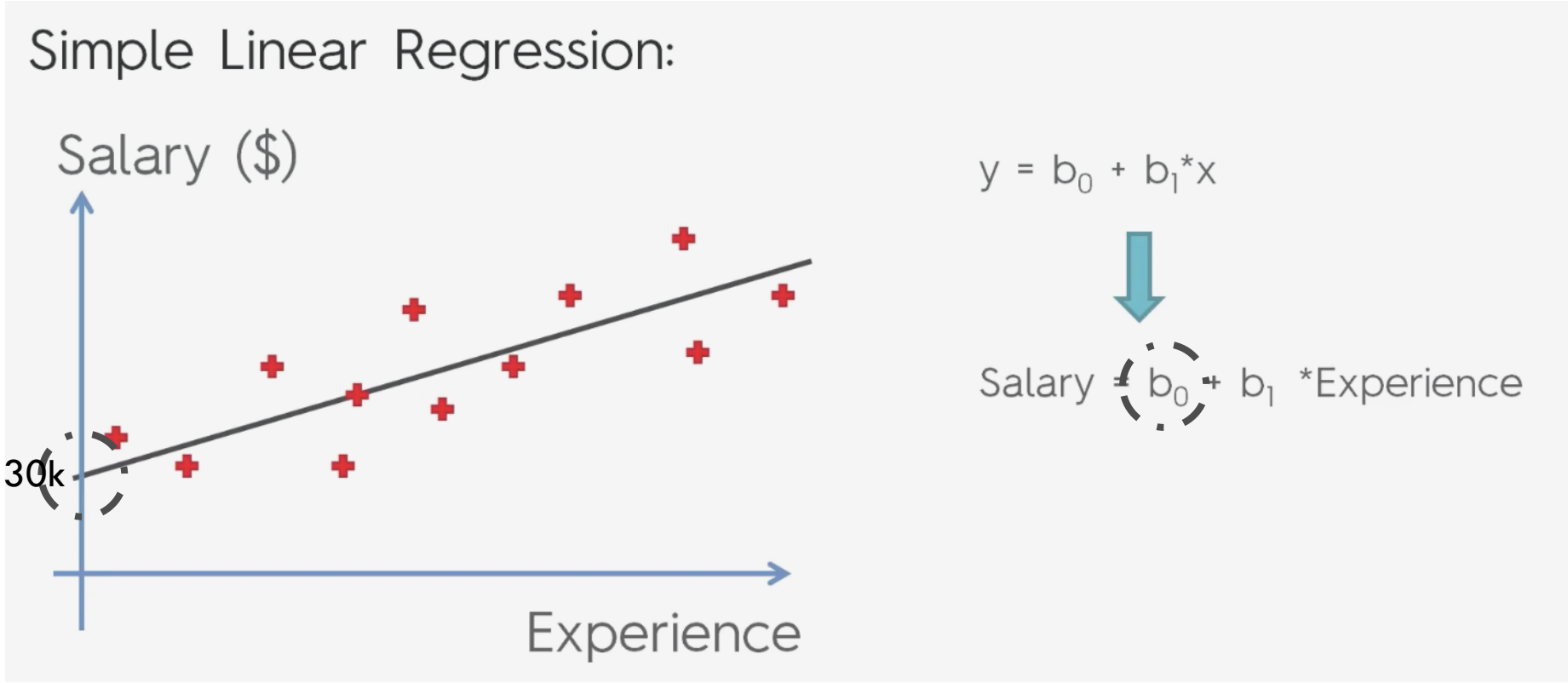
Simple Linear Regression:



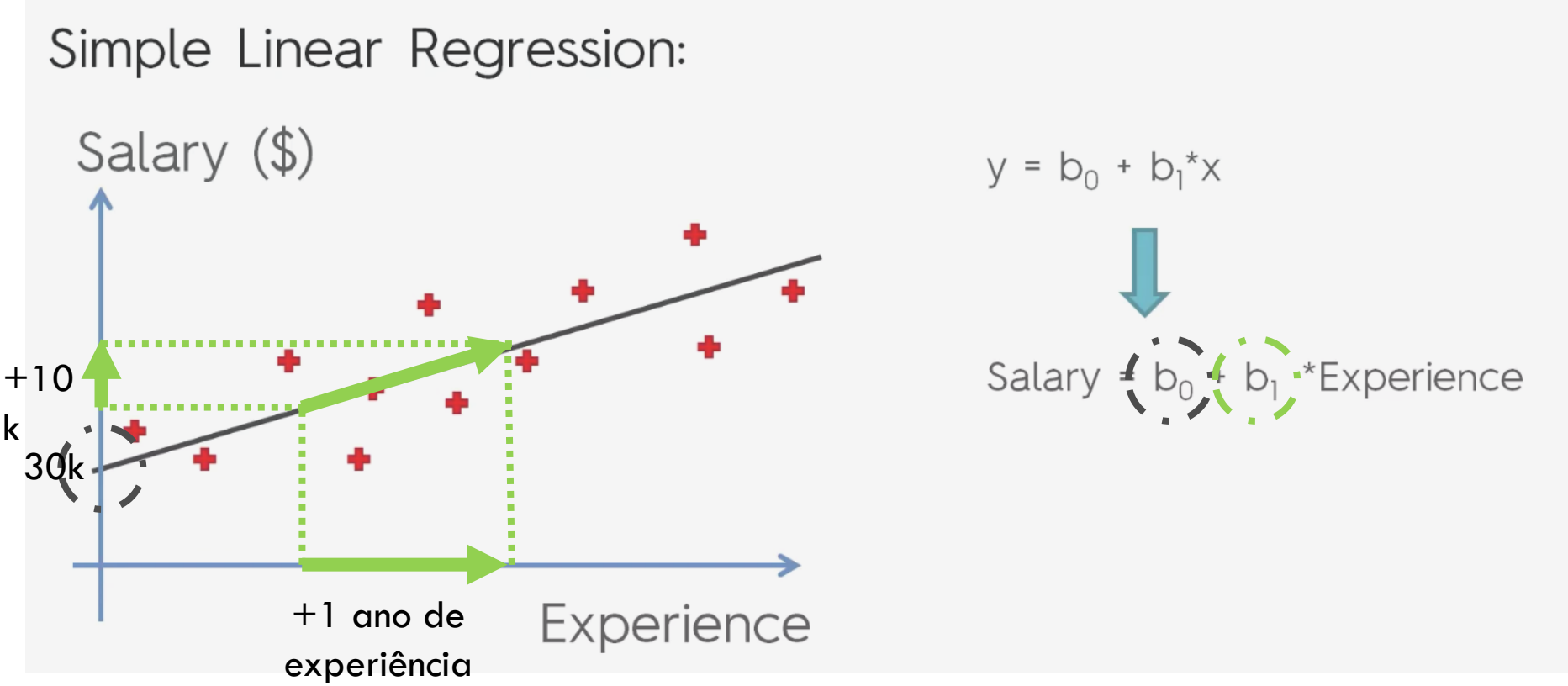
$y = b_0 + b_1 * x$

$\text{Salary} = b_0 + b_1 * \text{Experience}$

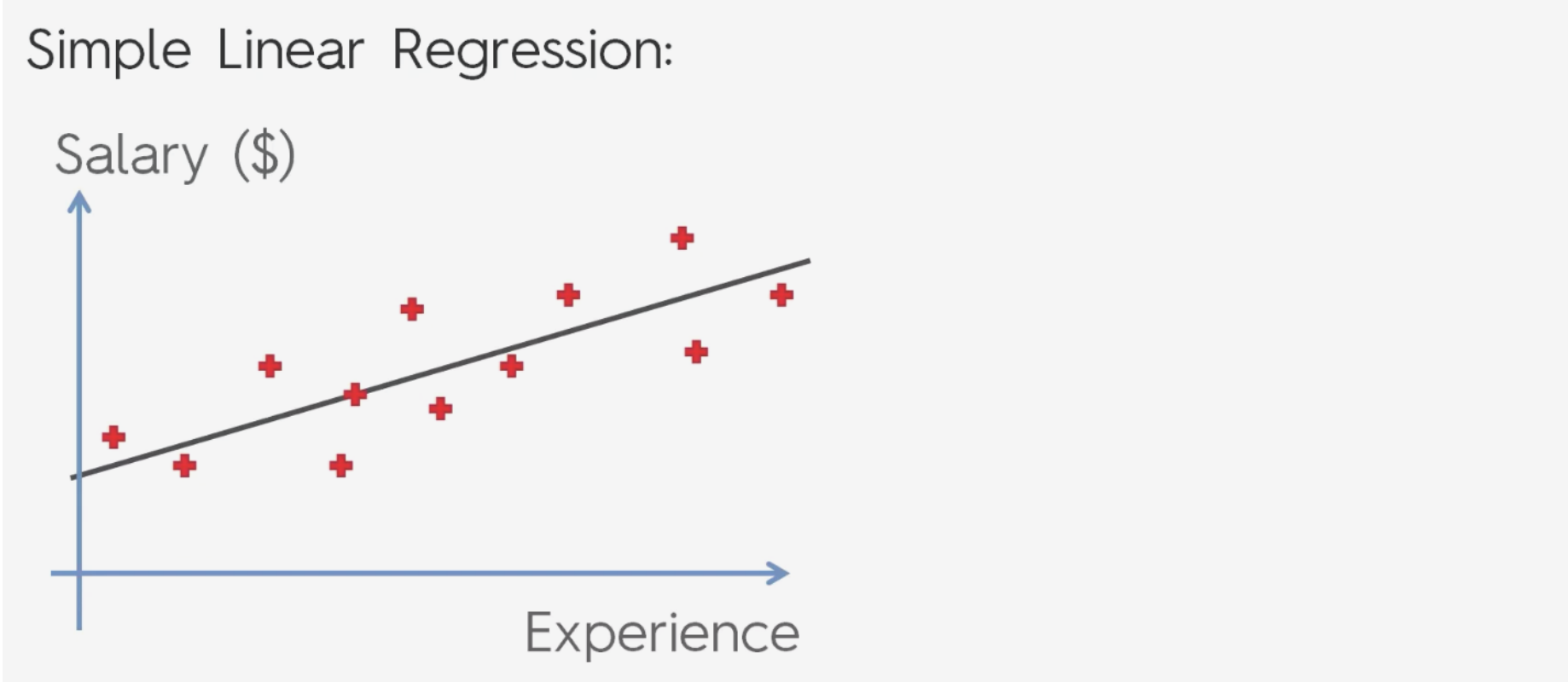
REGRESSÃO


Simple Linear Regression:
Salary ($)
30k
Experience
y = b0 + b1*x
Salary = b0 + b1 *Experience

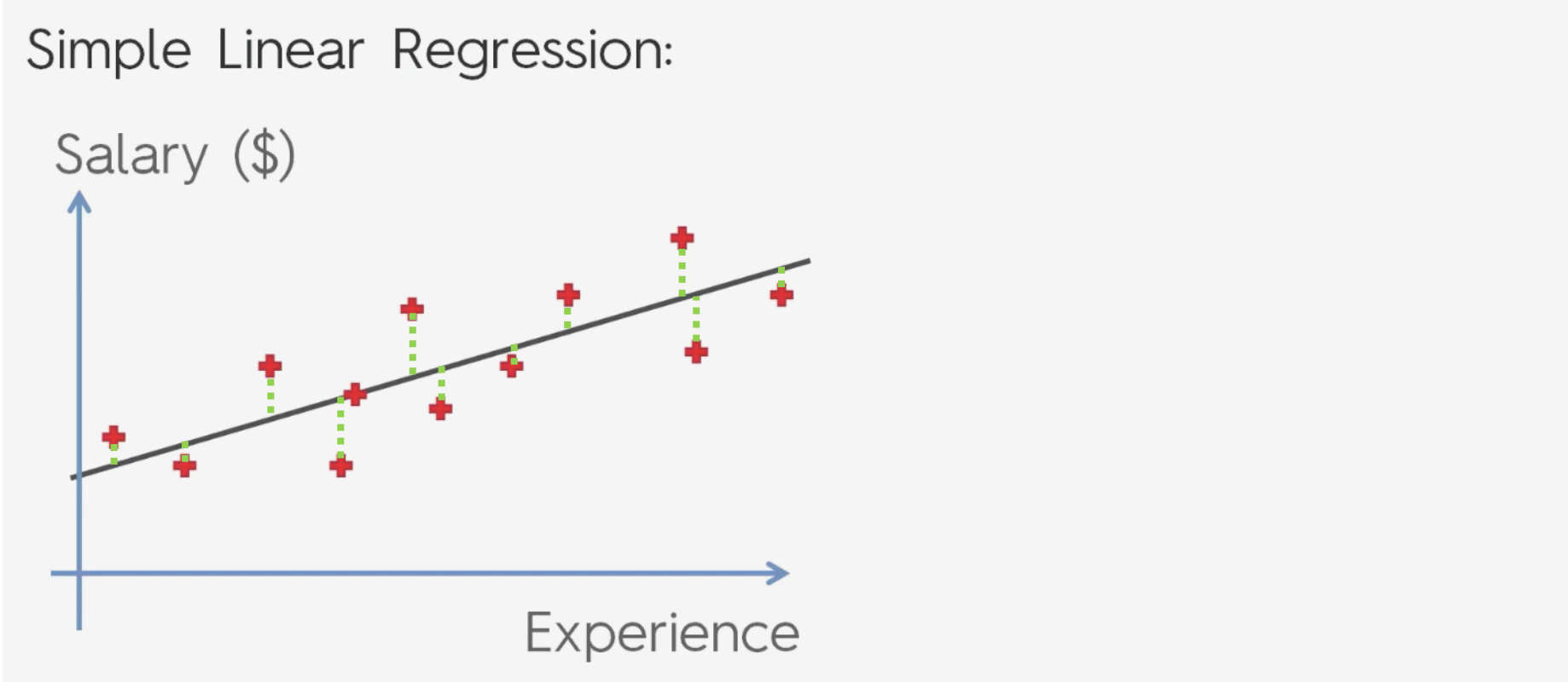
REGRESSÃO


Simple Linear Regression:
Salary ($)
+10
k
30k
+1 ano de
experiência
Experience
y = b0 + b1*x
Salary = b0 + b1 *Experience

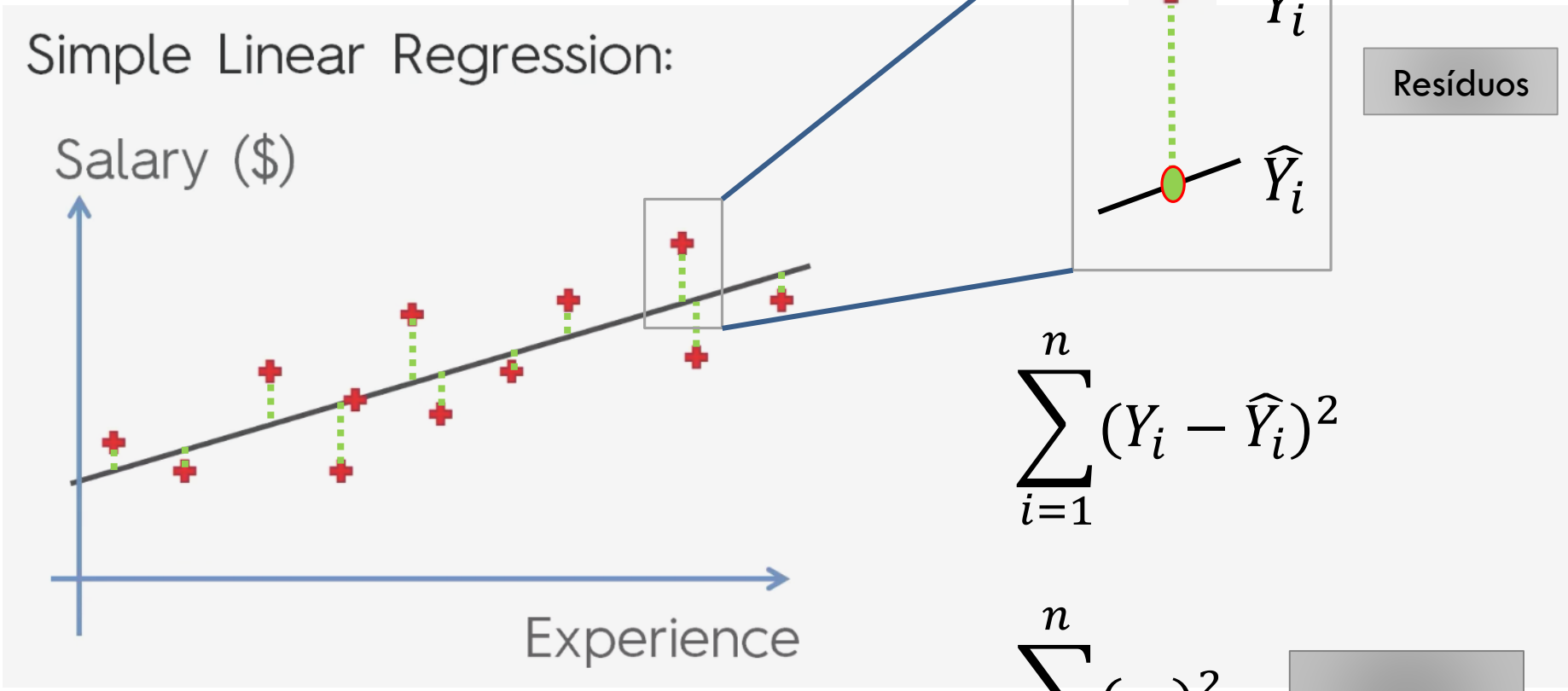
MÍNIMOS QUADRADOS



MÍNIMOS QUADRADOS



MÍNIMOS QUADRADOS



Resíduos

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\sum_{i=1}^{n} (e_i)^2$$

Minimizar

REGRESSÃO

Uma análise de regressão gera uma equação para descrever a relação entre um ou mais preditores e a variável resposta e para predizer novas observações com um valor preditor com precisão maior que o acaso.

A regressão linear geralmente usa o método de estimativa de mínimos quadrados comum que deriva a equação minimizando a soma dos resíduos quadrados.

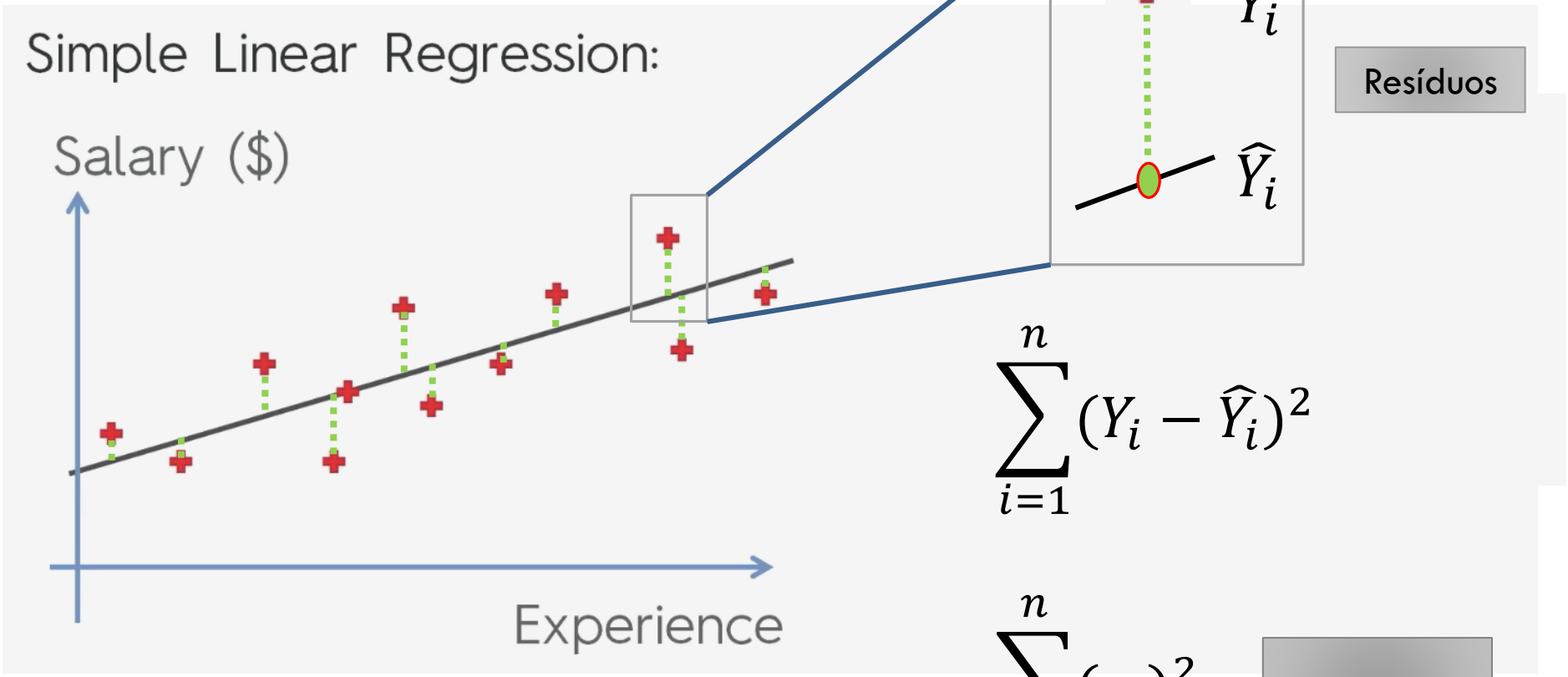
REGRESSÃO LINEAR SIMPLES

A regressão fornece a linha que "melhor" ajusta os dados. Essa linha pode ser usada para:

- Examinar como a variável de resposta muda quando o preditor muda.
- Predizer o valor de uma variável de resposta para qualquer variável preditora.

R Squared

R^2



Resíduos

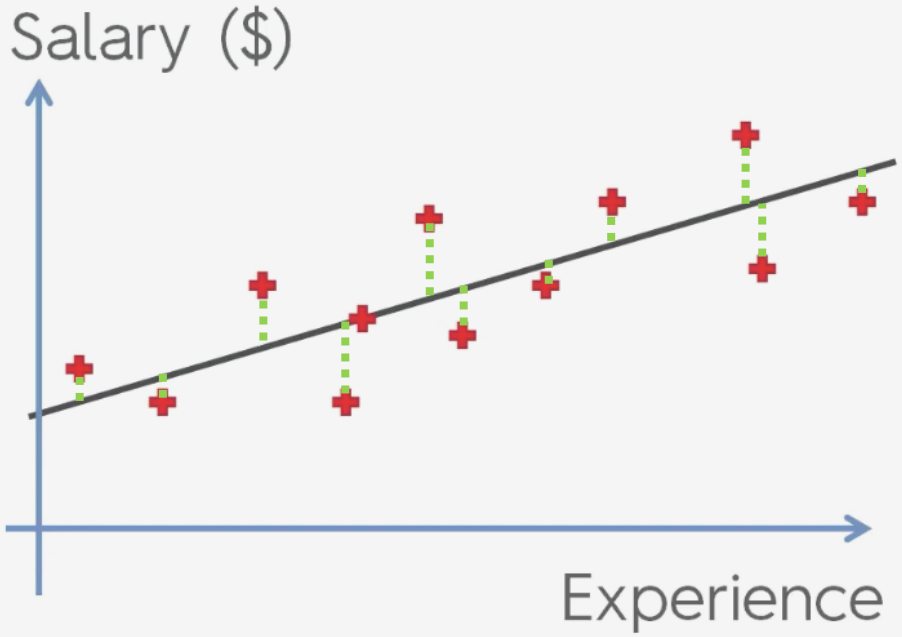
$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\sum_{i=1}^{n} (e_i)^2$$

Minimizar

R^2

Simple Linear Regression:

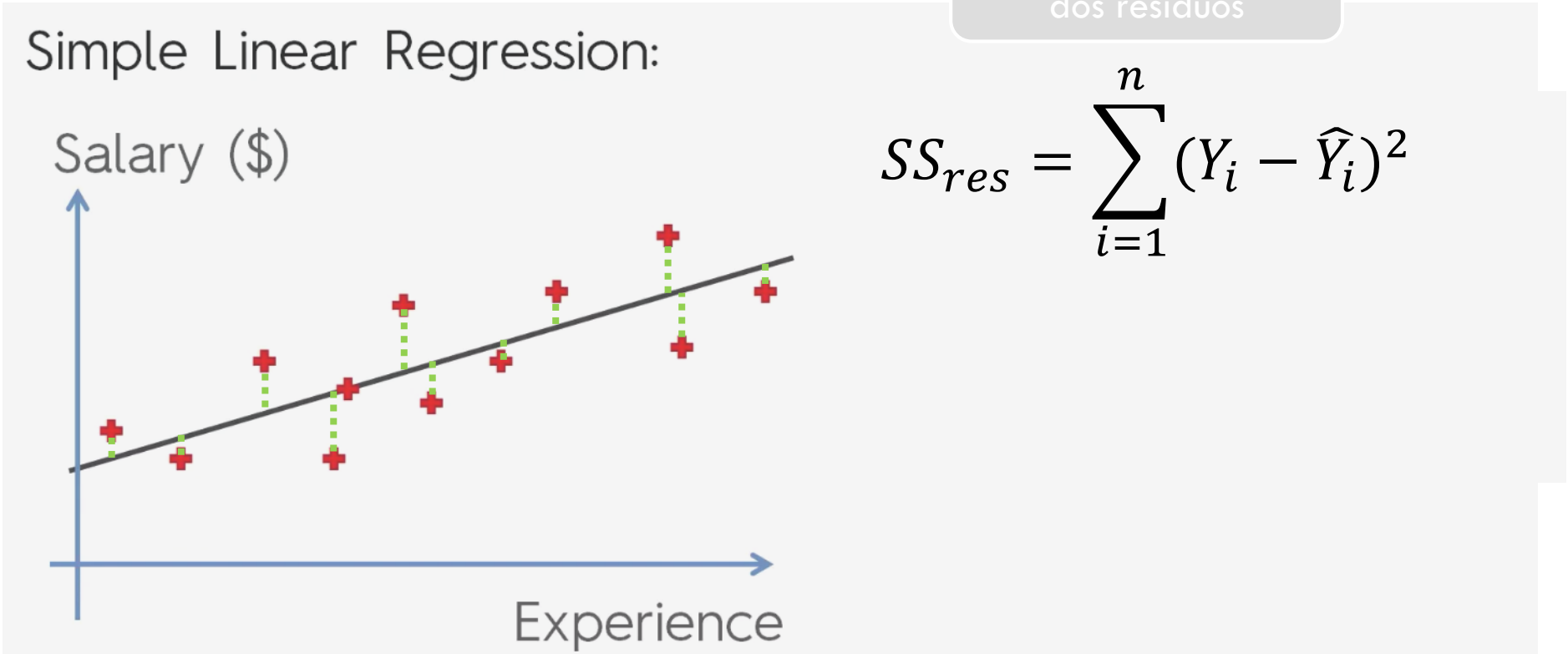


$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Bi Master

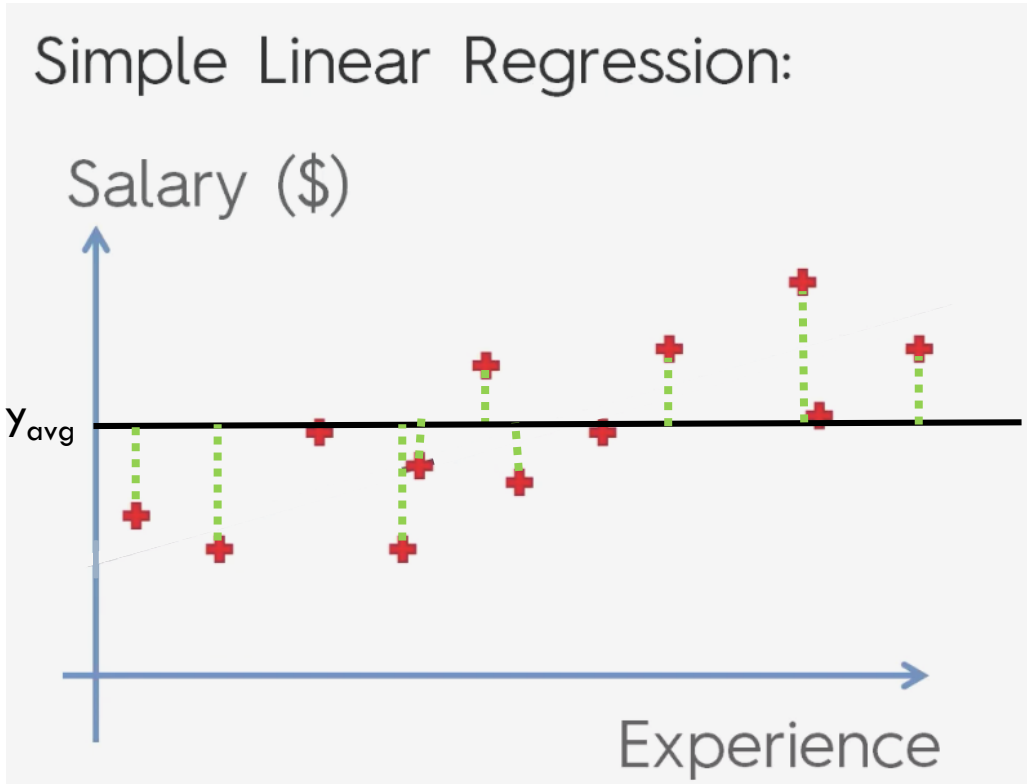
R^2

Soma dos quadrados dos resíduos



$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

R^2

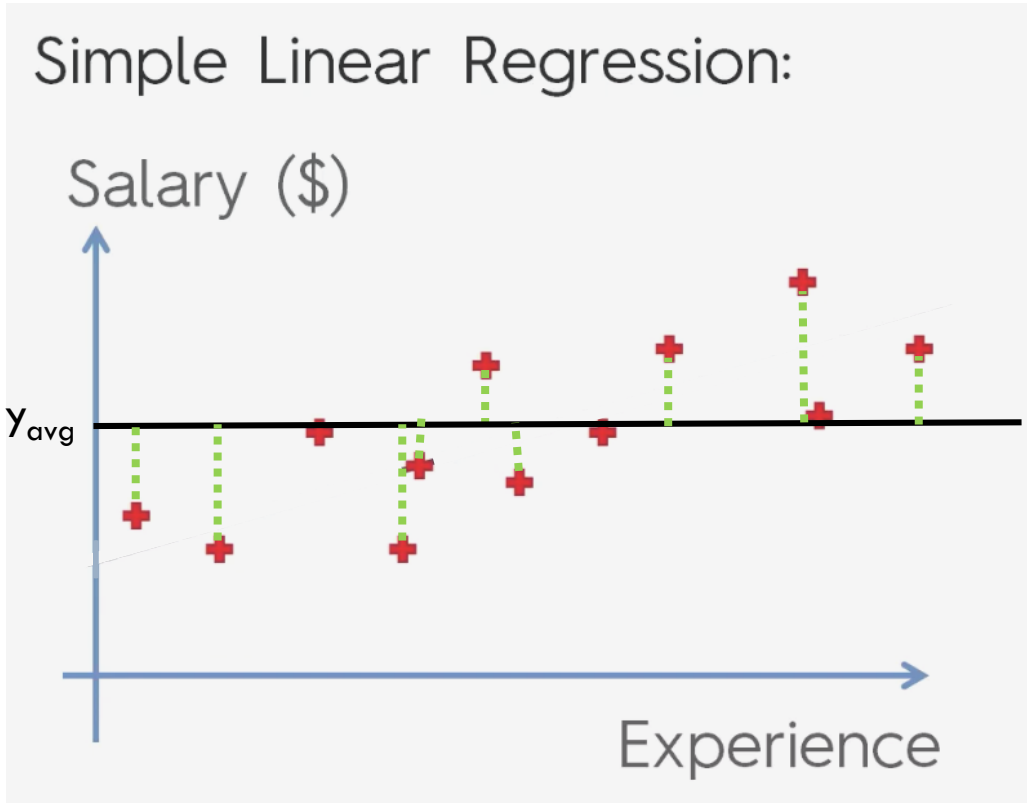


$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$SS_{tot} = \sum_{i=1}^{n} (Y_i - Y_{avg})^2$$

Soma dos quadrados total

R^2



$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$SS_{tot} = \sum_{i=1}^{n} (Y_i - Y_{avg})^2$$

$$R^2 = 1 - \frac{SSres}{SStot}$$

R^2 AJUSTADO

$$R^2 = 1 - \frac{SSres}{SStot}$$

R^2 - Quão bom é o modelo

(quanto maior, melhor)

$y = b_0 + b_1 x_1$

$y = b_0 + b_1 x_1 + b_2 x_2$

Problema:

$+b_3 x_3$

SS_{res} => Min

R^2 nunca diminuirá

Master

R^2 AJUSTADO

$$R^2 = 1 - \frac{SSres}{SStot}$$

$$R^2{}_{aj} = 1 - (1 - R2)\frac{n-1}{n-p-1}$$

n = número de dados

p = número de regressores

Regressão Linear Múltipla

REGRESSÃO LINEAR MÚLTIPLA

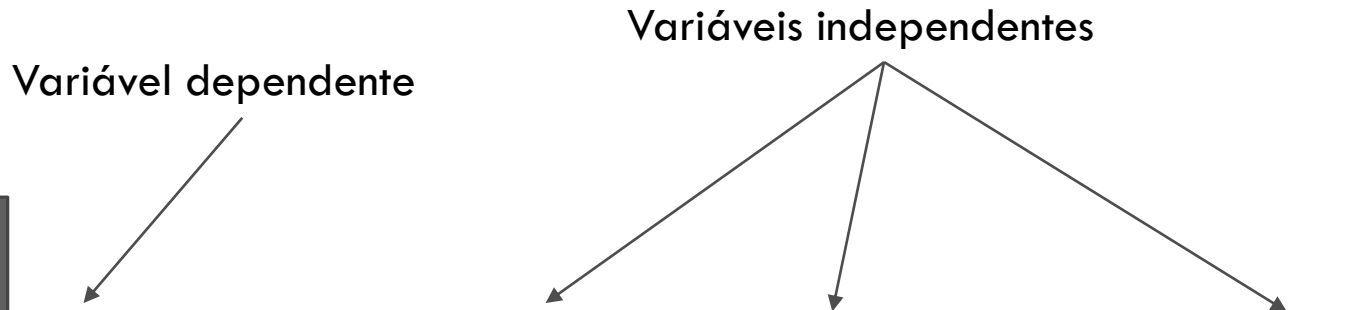
A regressão linear múltipla examina as relações lineares entre uma resposta contínua e dois ou mais preditores.

REGRESSÃO

Regressão Linear Simples

$$y = b_0 + b_1 x_1$$

Regressão Linear Múltipla



ESTUDO DE CASO

Startups

50 Startups:

- 4 variáveis independentes (explicativas):
  - Gastos com P&D;
  - Gastos administrativos;
  - Gastos com marketing;
  - Estado.
- Variável dependente (resposta):
  - Lucro

Regressão com objetivo de criar um modelo para investidores.

DEVEMOS SABER

DUMMY VARIABLES

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

DUMMY VARIABLES

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$y =$

DUMMY VARIABLES

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 +$$

DUMMY VARIABLES

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1$$

DUMMY VARIABLES

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1 + b_2 * x_2$$

DUMMY VARIABLES

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3$$

DUMMY VARIABLES

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

DUMMY VARIABLES

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York		
191,792.06	162,597.70	151,377.59	443,898.53	California		
191,050.39	153,441.51	101,145.55	407,934.54	California		
182,901.99	144,372.41	118,671.85	383,199.62	New York		
166,187.94	142,107.34	91,391.77	366,168.42	California		

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

DUMMY VARIABLES

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	
191,792.06	162,597.70	151,377.59	443,898.53	California	0	
191,050.39	153,441.51	101,145.55	407,934.54	California	0	
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	
166,187.94	142,107.34	91,391.77	366,168.42	California	0	

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

DUMMY VARIABLES

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

DUMMY VARIABLES

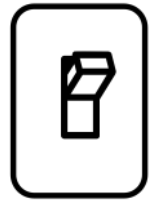
Profit	R&D Spend	Admin	Marketing	State	Dummy Variables: New York	Dummy Variables: California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1$$

DUMMY VARIABLES

Profit	R&D Spend	Admin	Marketing	State	Dummy Variables: New York	Dummy Variables: California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1$$



DUMMY VARIABLE TRAP

Profit	R&D Spend	Admin	Marketing	State	Dummy Variables: New York	Dummy Variables: California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1 + b_5*D_2$$

DUMMY VARIABLE TRAP

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,3			
191,792.06	162,5			
191,050.39	153,4			
182,901.99	144,3			
166,187.94	142,107.34			California

$$D_2 = 1 - D_1$$

Multicolinearidade

Dummy Variables

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1 + b_5*D_2$$

O modelo não funcionará de forma apropriada

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33128  -4865      5   6098  18065

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.965e+04  7.637e+03   6.501 1.94e-07 ***
R.D.Spend        7.986e-01  5.604e-02  14.251 6.70e-16 ***
Administration  -2.942e-02  5.828e-02  -0.505    0.617
Marketing.Spend  3.268e-02  2.127e-02   1.537    0.134
State2           1.213e+02  3.751e+03   0.032    0.974
State3           2.376e+02  4.127e+03   0.058    0.954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9908 on 34 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9425
F-statistic:   129 on 5 and 34 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33128  -4865      5   6098  18065

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.965e+04  7.637e+03   6.501 1.94e-07 ***
R.D.Spend        7.986e-01  5.604e-02  14.251 6.70e-16 ***
Administration  -2.942e-02  5.828e-02  -0.505    0.617
Marketing.Spend  3.268e-02  2.127e-02   1.537    0.134
State2           1.213e+02  3.751e+03   0.032    0.974
State3           2.376e+02  4.127e+03   0.058    0.954
---
F-statistic:   129 on 5 and 34 DF,  p-value: < 2.2e-16
```

Coeficiente: o modelo estima um aumento esperado de 0.79 no lucro para cada 1 unidade (no caso, 1 dólar) de aumento de gasto com pesquisa e desenvolvimento (quando as outras variáveis são mantidas constantes).

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33128  -4865      5   6098  18065

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.965e+04  7.637e+03   6.501 1.94e-07 ***
R.D.Spend        7.986e-01  5.604e-02  14.251 6.70e-16 ***
Administration  -2.942e-02  5.828e-02  -0.505    0.617
Marketing.Spend  3.268e-02  2.127e-02   1.537    0.134
State2           1.213e+02  3.751e+03   0.032    0.974
State3           2.376e+02  4.127e+03   0.058    0.954
---
```

Desvio padrão: quão precisamente o modelo estimou o coeficiente da variável em questão. Quanto menor, mais precisa é a estimativa.

```
F-statistic:   129 on 5 and 34 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33128  -4865      5   6098  18065

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.965e+04  7.637e+03   6.501 1.94e-07 ***
R.D.Spend        7.986e-01  5.604e-02  14.251 6.70e-16 ***
Administration  -2.942e-02  5.828e-02  -0.505    0.617
Marketing.Spend  3.268e-02  2.127e-02   1.537    0.134
State2           1.213e+02  3.751e+03   0.032    0.974
State3           2.376e+02  4.127e+03   0.058    0.954
---
```

T value: Estimate/Standard Error
Quantos desvios padrões do zero o coeficiente estimado está.

```
F-statistic:   129 on 5 and 34 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33128  -4865      5   6098  18065

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.965e+04  7.637e+03   6.501 1.94e-07 ***
R.D.Spend        7.986e-01  5.604e-02  14.251 6.70e-16 ***
Administration  -2.942e-02  5.828e-02  -0.505    0.617
Marketing.Spend  3.268e-02  2.127e-02   1.537    0.134
State2           1.213e+02  3.751e+03   0.032    0.974
State3           2.376e+02  4.127e+03   0.058    0.954
---
```

P-value: testa a hipótese nula onde o coeficiente é igual a zero (sem efeito). Um p-value < 0.05 indica que você pode rejeitar a hipótese nula. Em outras palavras, um preditor que tem um p-value pequeno é provavelmente uma boa adição ao seu modelo (estatisticamente significante).

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33128  -4865      5   6098  18065

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.965e+04  7.637e+03   6.501 1.94e-07 ***
R.D.Spend        7.986e-01  5.604e-02  14.251 6.70e-16 ***
Administration  -2.942e-02  5.828e-02  -0.505    0.617
Marketing.Spend  3.268e-02  2.127e-02   1.537    0.134
State2           1.213e+02  3.751e+03   0.032    0.974
State3           2.376e+02  4.127e+03   0.058    0.954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9908 on 34 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9425
F-statistic:   129 on 5 and 34 DF,  p-value: < 2.2e-16
```

R SCRIPT

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33128  -4865      5   6098  18065

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.965e+04  7.637e+03   6.501 1.94e-07 ***
R.D.Spend        7.986e-01  5.604e-02  14.251 6.70e-16 ***
Administration  -2.942e-02  5.828e-02  -0.505    0.617
Marketing.Spend  3.268e-02  2.127e-02   1.537    0.134
State2           1.213e+02  3.751e+03   0.032    0.974
State3           2.376e+02  4.127e+03   0.058    0.954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9908 on 34 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9425
F-statistic:   129 on 5 and 34 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33117  -4858    -36   6020  17957

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.970e+04  7.120e+03   6.980 3.48e-08 ***
R.D.Spend        7.983e-01  5.356e-02  14.905  < 2e-16 ***
Administration  -2.895e-02  5.603e-02  -0.517    0.609
Marketing.Spend  3.283e-02  1.987e-02   1.652    0.107
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9629 on 36 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9457
F-statistic: 227.6 on 3 and 36 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33294  -4763   -354   6351  17693

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.638e+04  3.019e+03  15.364   <2e-16 ***
R.D.Spend        7.879e-01  4.916e-02  16.026   <2e-16 ***
Marketing.Spend 3.538e-02  1.905e-02   1.857   0.0713 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9533 on 37 degrees of freedom
Multiple R-squared:  0.9495,    Adjusted R-squared:  0.9468
F-statistic: 348.1 on 2 and 37 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-34334  -4894   -340   6752  17147

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.902e+04  2.748e+03   17.84   <2e-16 ***
R.D.Spend   8.563e-01  3.357e-02   25.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9836 on 38 degrees of freedom
Multiple R-squared:  0.9448,    Adjusted R-squared:  0.9434
F-statistic: 650.8 on 1 and 38 DF,  p-value: < 2.2e-16
```

Bi Master

R SCRIPT

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33128  -4865      5   6098  18065

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.965e+04  7.637e+03   6.501 1.94e-07 ***
R.D.Spend        7.986e-01  5.604e-02  14.251 6.70e-16 ***
Administration  -2.942e-02  5.828e-02  -0.505    0.617
Marketing.Spend  3.268e-02  2.127e-02   1.537    0.134
State2           1.213e+02  3.751e+03   0.032    0.974
State3           2.376e+02  4.127e+03   0.058    0.954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9908 on 34 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9425
```

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33117  -4858    -36   6020  17957

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.970e+04  7.120e+03   6.980 3.48e-08 ***
R.D.Spend        7.983e-01  5.356e-02  14.905  < 2e-16 ***
Administration  -2.895e-02  5.603e-02  -0.517    0.609
Marketing.Spend  3.283e-02  1.987e-02   1.652    0.107
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9629 on 36 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9457
```

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33294  -4763   -354   6351  17693

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.638e+04  3.019e+03  15.364   <2e-16 ***
R.D.Spend        7.879e-01  4.916e-02  16.026   <2e-16 ***
Marketing.Spend  3.538e-02  1.905e-02   1.857   0.0713 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9508 on 37 degrees of freedom
Multiple R-squared:  0.9495,    Adjusted R-squared:  0.9468
```

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-34334  -4894   -340   6752  17147

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.902e+04  2.748e+03   17.84   <2e-16 ***
R.D.Spend      8.563e-01  3.357e-02   25.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9836 on 38 degrees of freedom
Multiple R-squared:  0.9448,    Adjusted R-squared:  0.9434
```

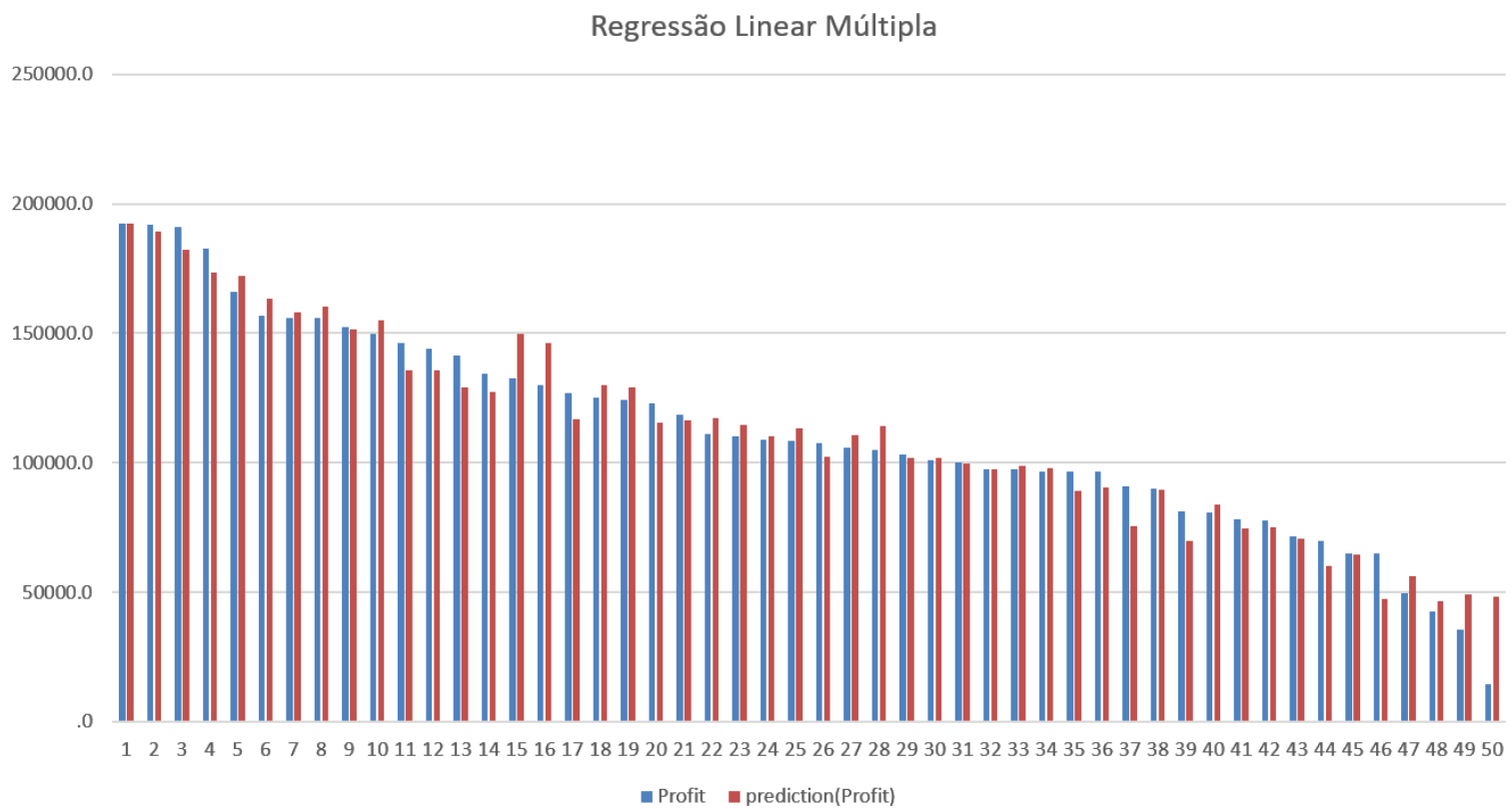
R SCRIPT

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33128  -4865      5   6098  18065

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.965e+04  7.637e+03   6.501 1.94e-07 ***
R.D.Spend        7.986e-01  5.604e-02  14.251 6.70e-16 ***
Administration  -2.942e-02  5.828e-02  -0.505    0.617
Marketing.Spend  3.268e-02  2.127e-02   1.537    0.134
State2           1.213e+02  3.751e+03   0.032    0.974
State3           2.376e+02  4.127e+03   0.058    0.954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9908 on 34 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9425
F-statistic:   129 on 5 and 34 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33117  -4858    -36   6020  17957

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.970e+04  7.120e+03   6.980 3.48e-08 ***
R.D.Spend        7.983e-01  5.356e-02  14.905  < 2e-16 ***
Administration  -2.895e-02  5.603e-02  -0.517    0.609
Marketing.Spend  3.283e-02  1.987e-02   1.652    0.107
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9629 on 36 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9457
F-statistic: 227.6 on 3 and 36 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-33294  -4763   -354   6351  17693

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.638e+04  3.019e+03  15.364   <2e-16 ***
R.D.Spend        7.879e-01  4.916e-02  16.026   <2e-16 ***
Marketing.Spend 3.538e-02  1.905e-02   1.857   0.0713 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9533 on 37 degrees of freedom
Multiple R-squared:  0.9495,    Adjusted R-squared:  0.9468
F-statistic: 348.1 on 2 and 37 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ ., data = training_set)

Residuals:
   Min     1Q Median     3Q    Max
-34334  -4894   -340   6752  17147

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.902e+04  2.748e+03   17.84   <2e-16 ***
R.D.Spend   8.563e-01  3.357e-02   25.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9836 on 38 degrees of freedom
Multiple R-squared:  0.9448,    Adjusted R-squared:  0.9434
F-statistic: 650.8 on 1 and 38 DF,  p-value: < 2.2e-16
```

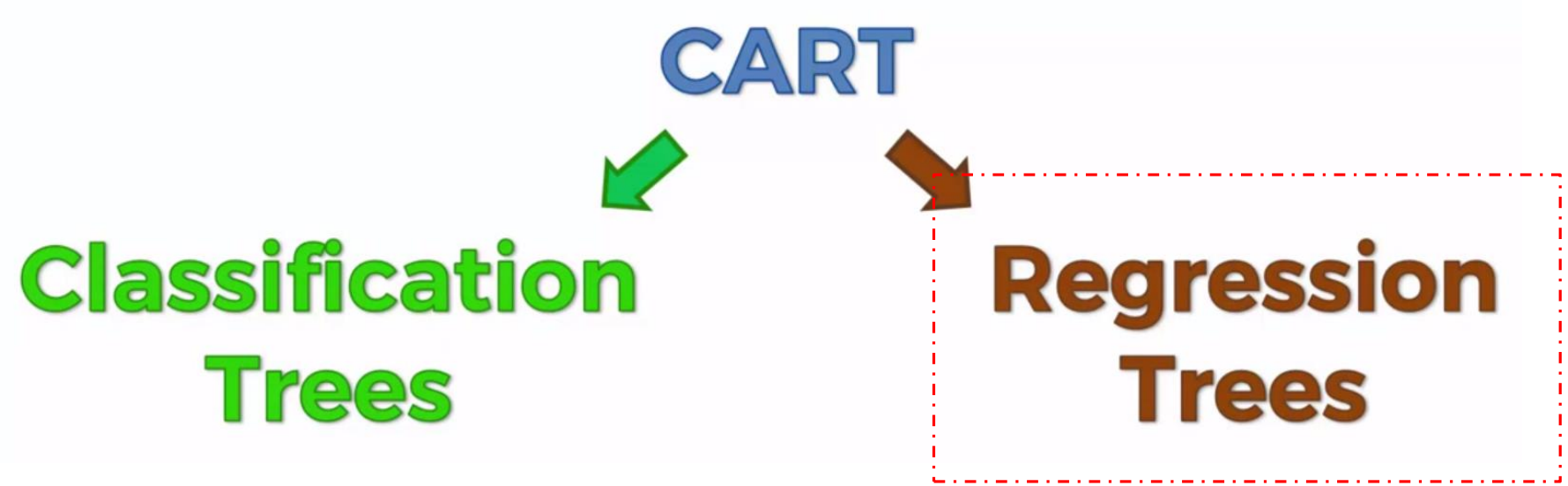

Regressão Linear Múltipla
250000.0
200000.0
150000.0
100000.0
50000.0
.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Profit
prediction(Profit)

Árvores de Regressão

ÁRVORES DE REGRESSÃO

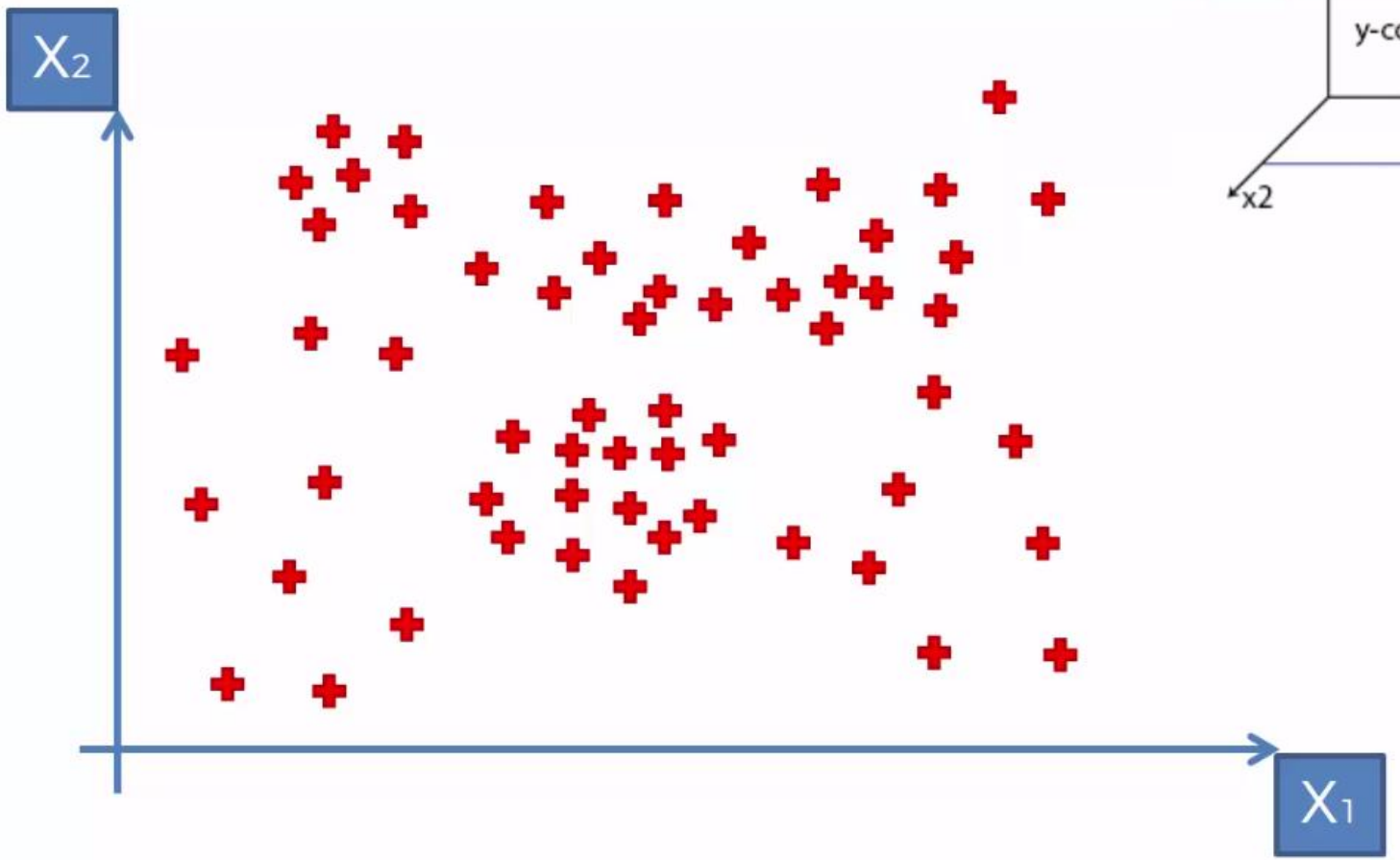
CART

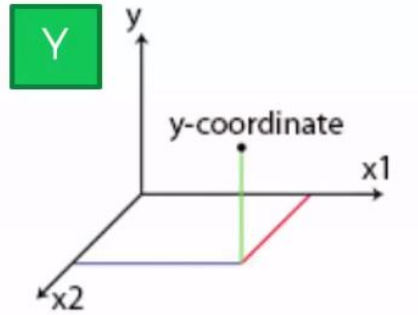
ÁRVORES DE REGRESSÃO



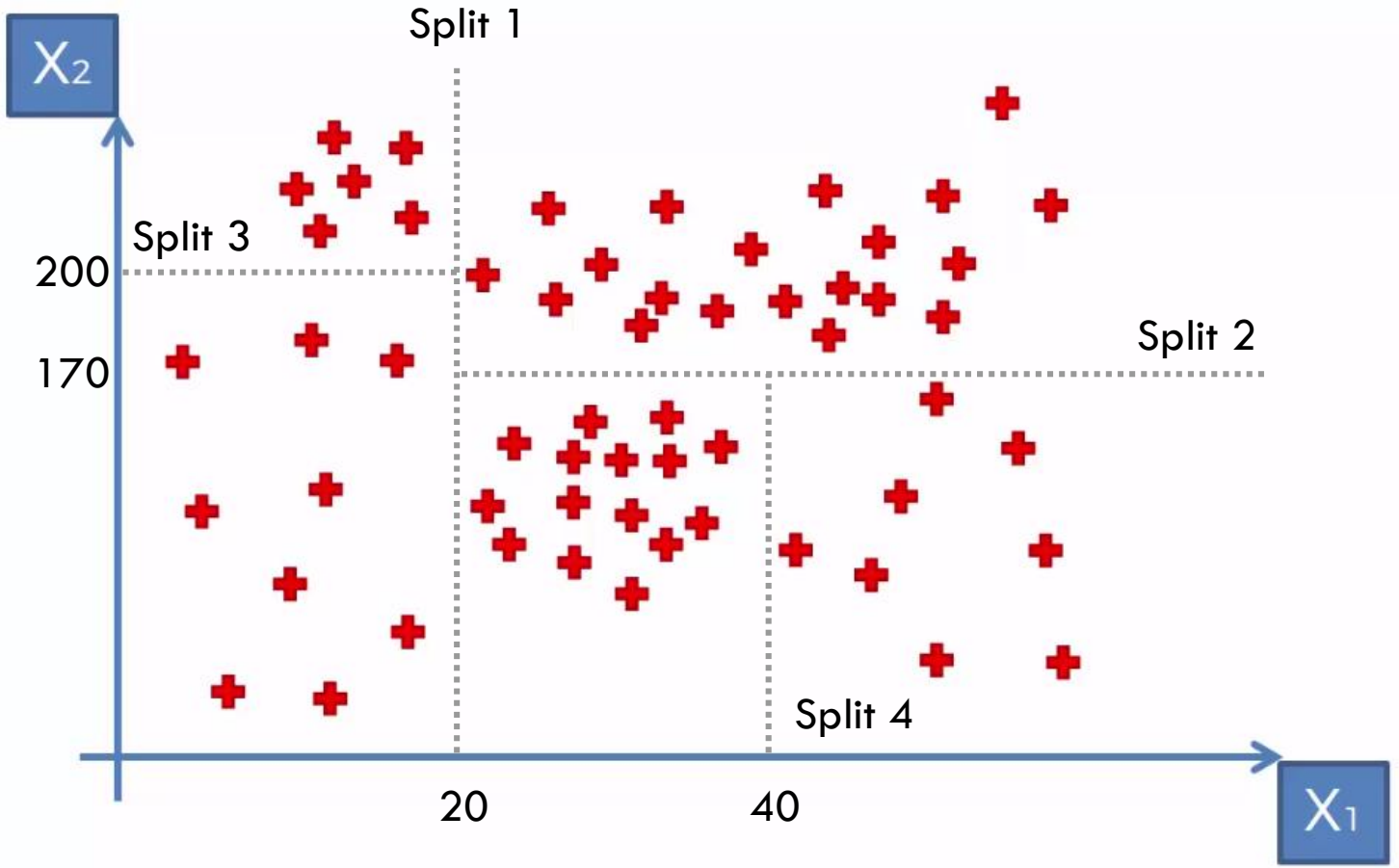
Breiman, Leo, et al. Classification and regression trees. CRC press, 1984

ÁRVORES DE REGRESSÃO

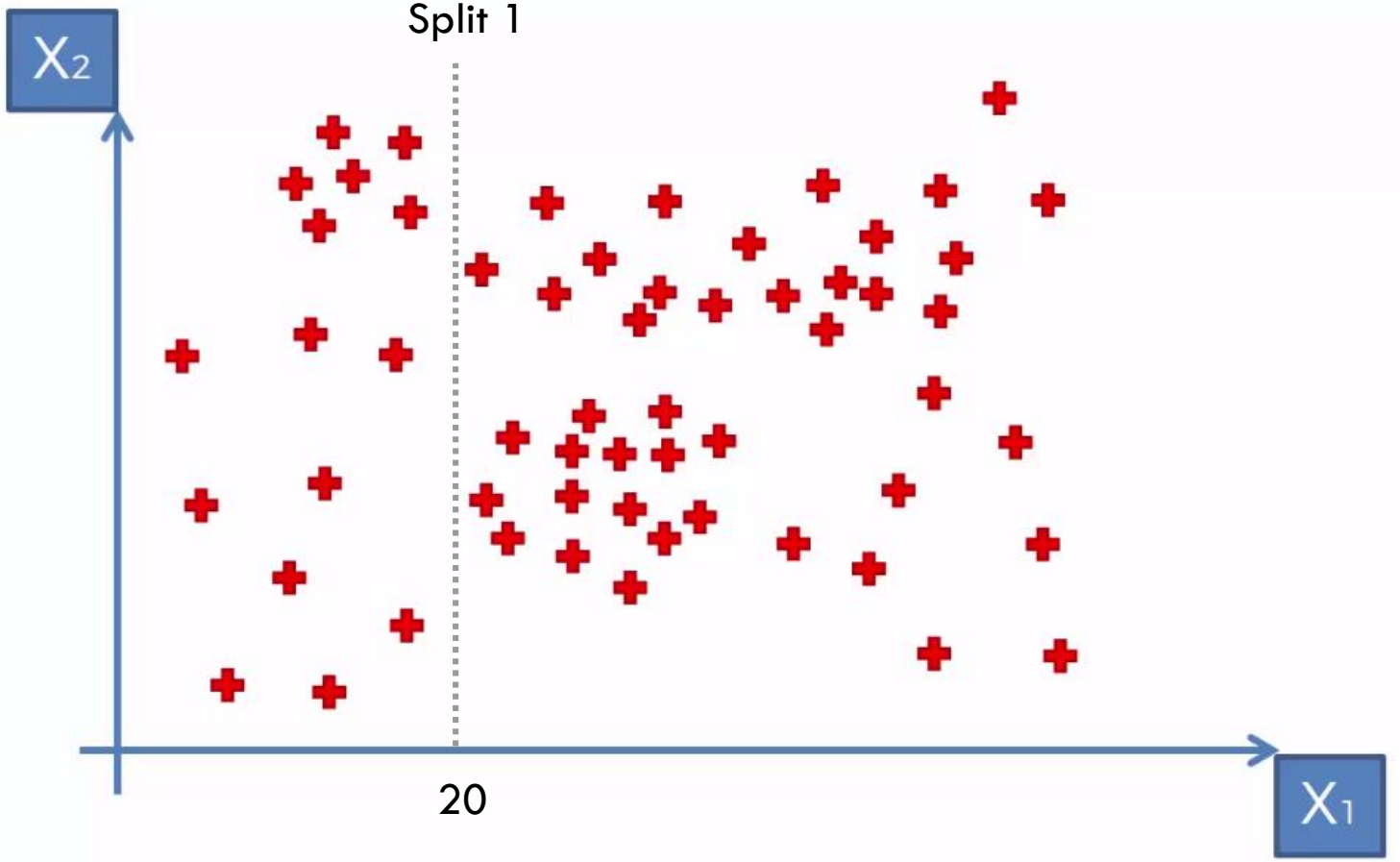




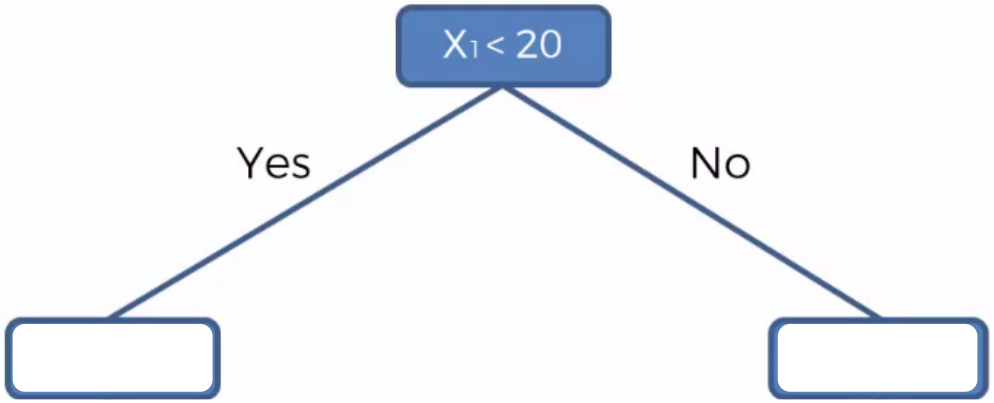
ÁRVORES DE REGRESSÃO



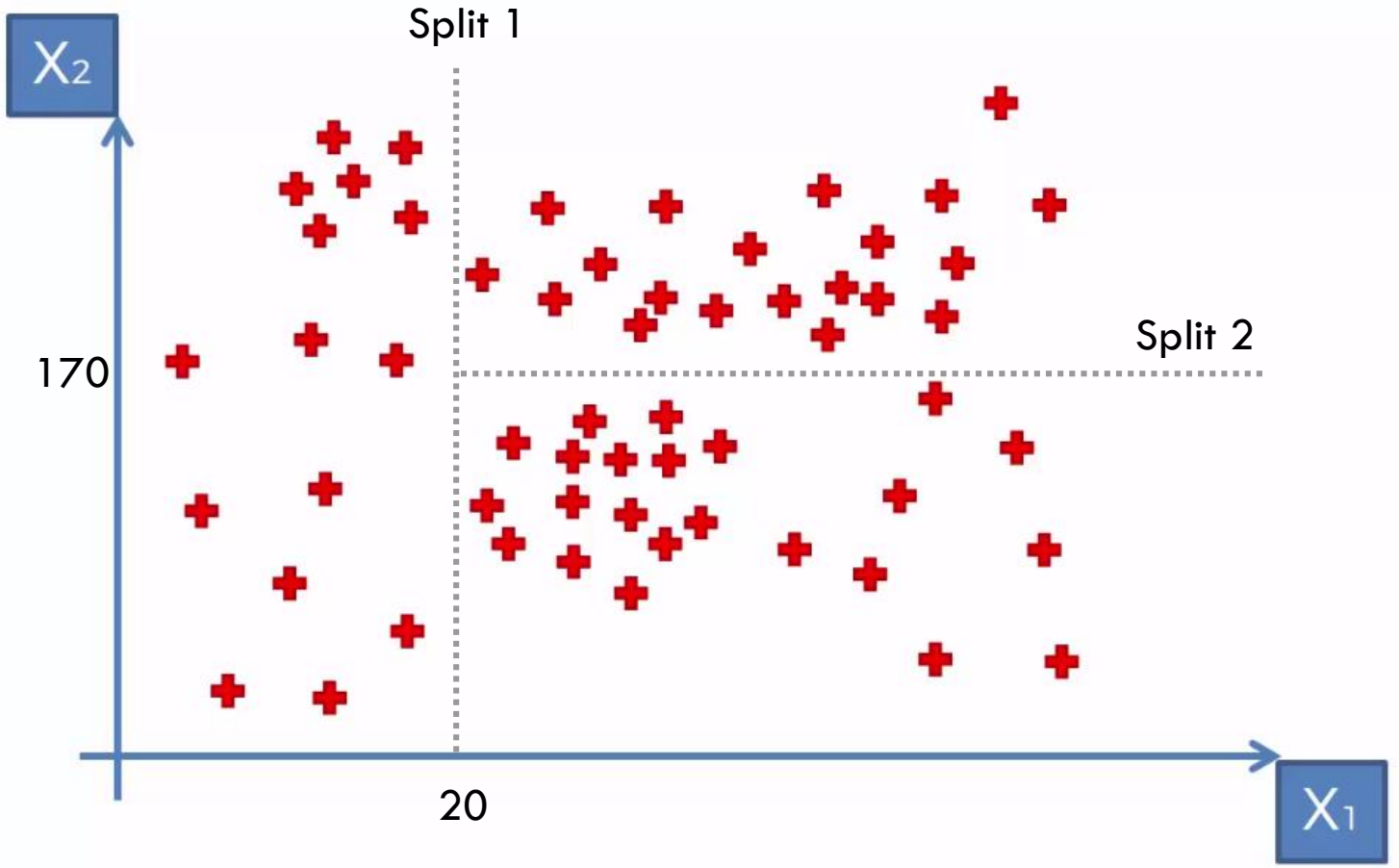
ÁRVORES DE REGRESSÃO



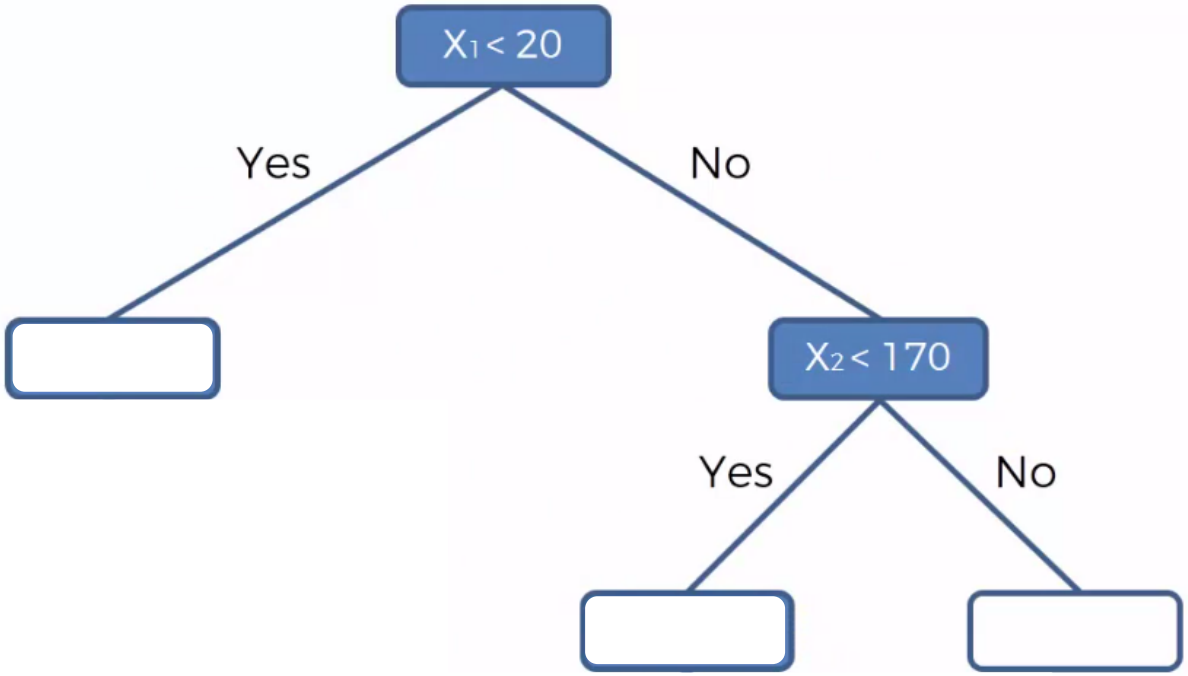
ÁRVORES DE REGRESSÃO



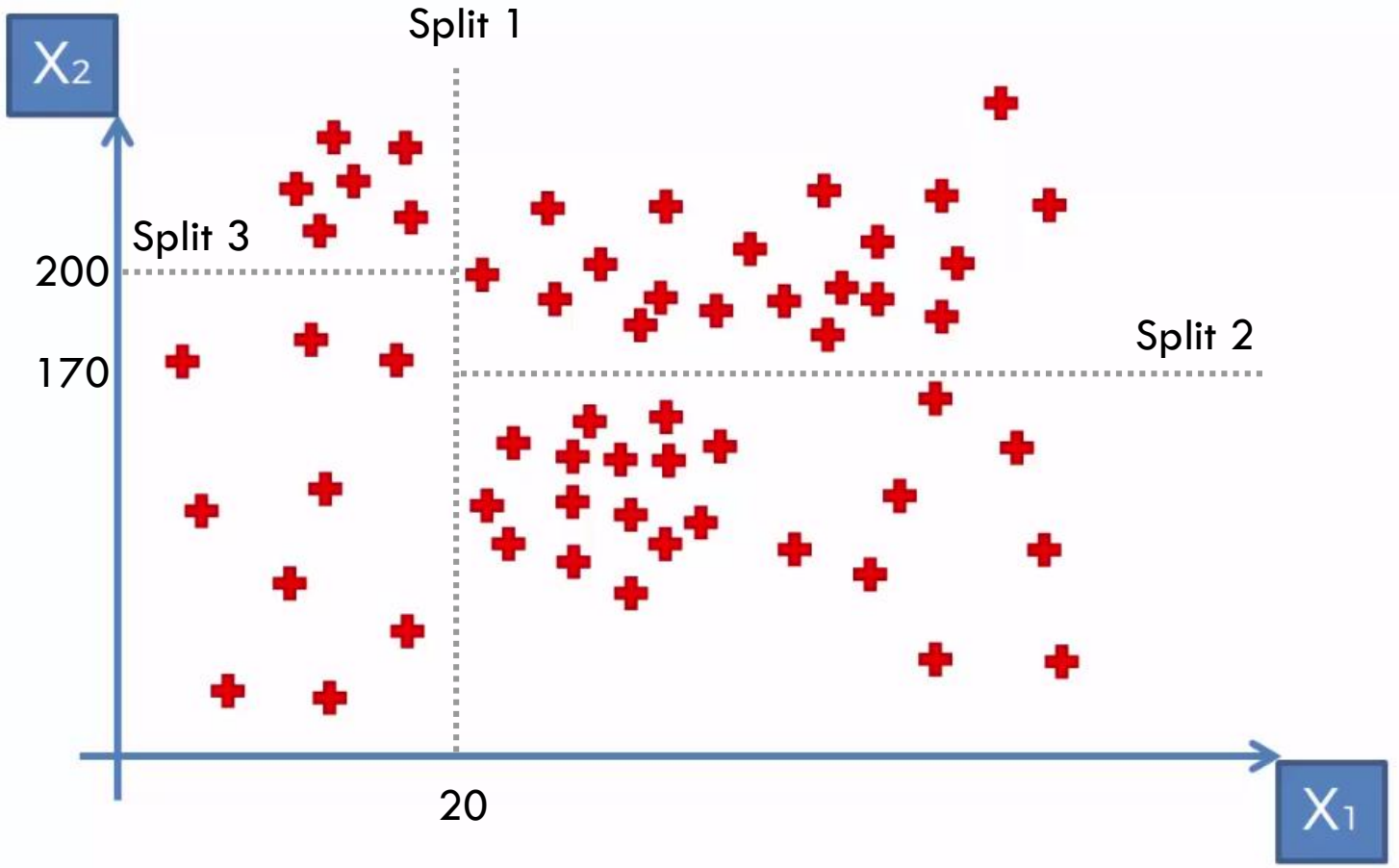
ÁRVORES DE REGRESSÃO



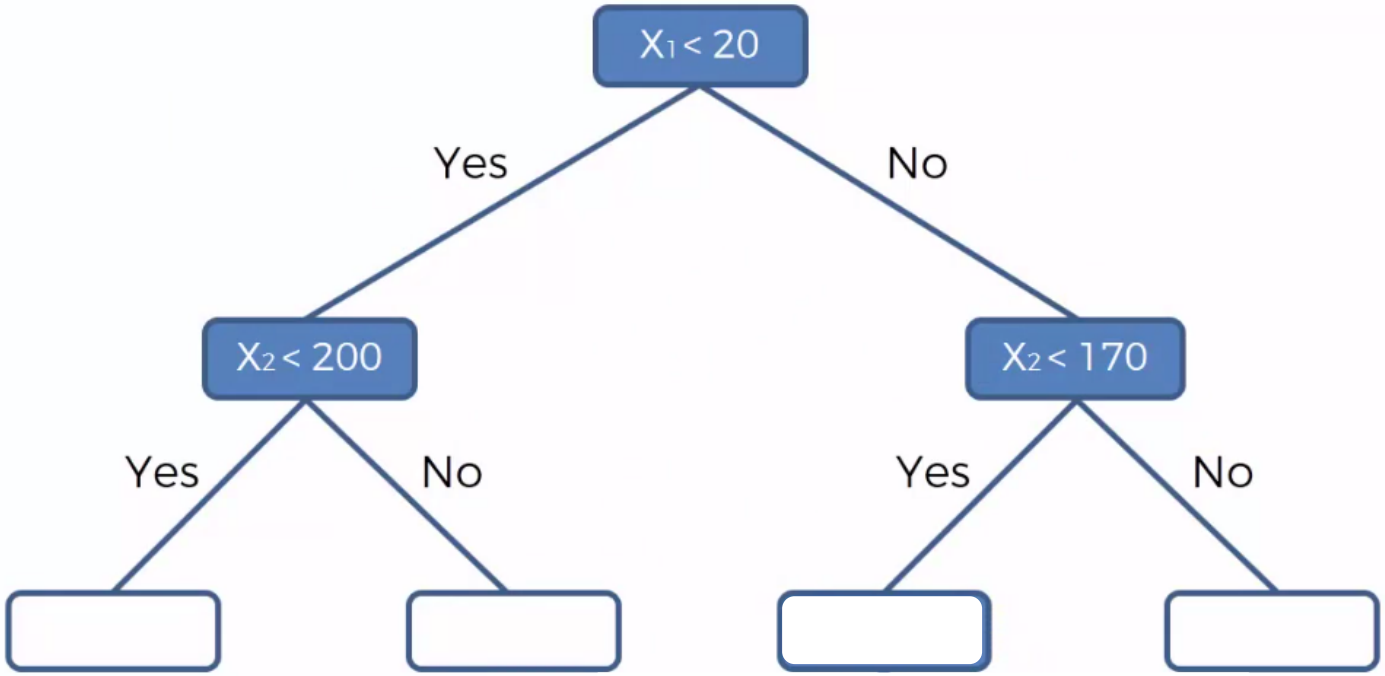
ÁRVORES DE REGRESSÃO



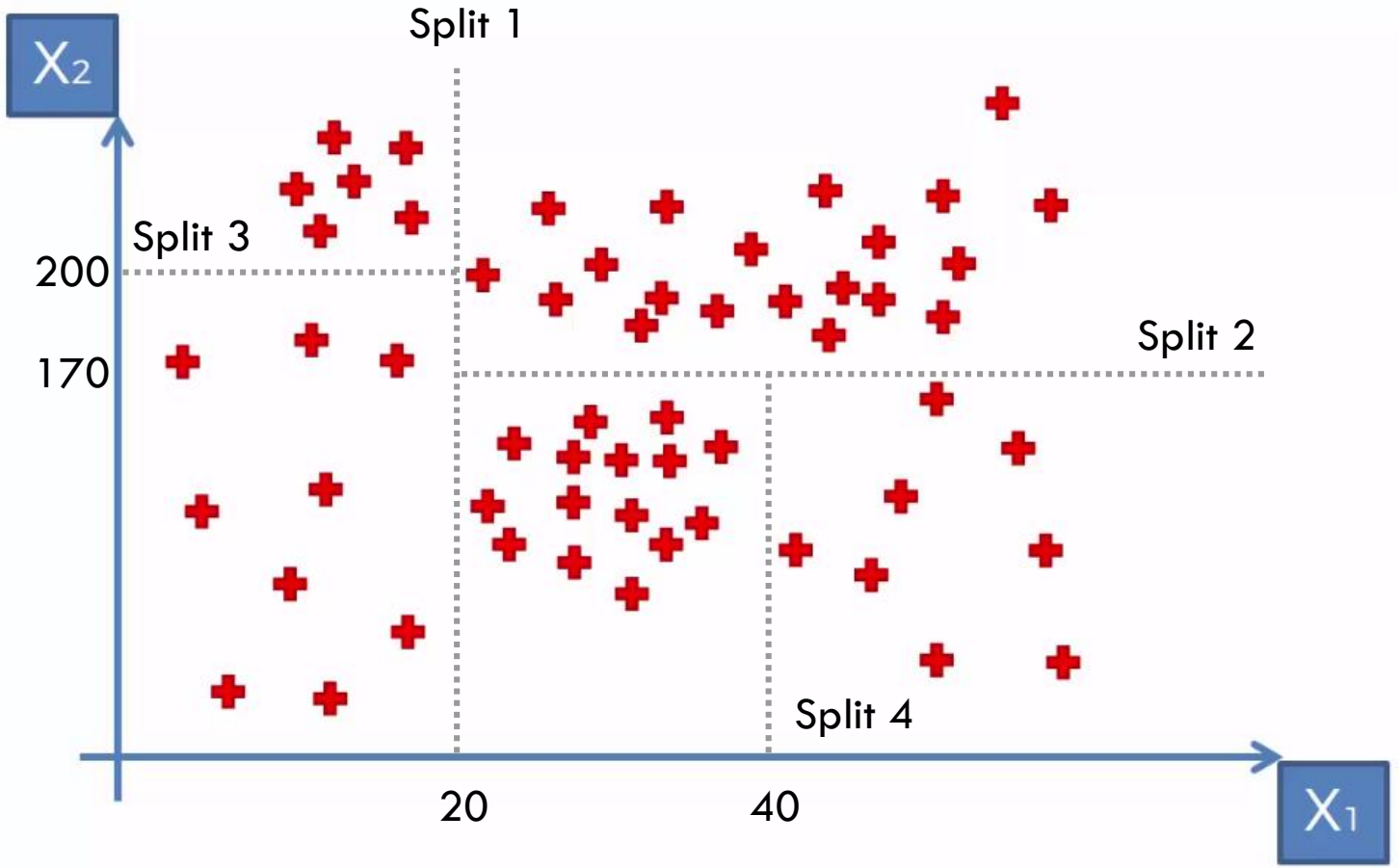
ÁRVORES DE REGRESSÃO



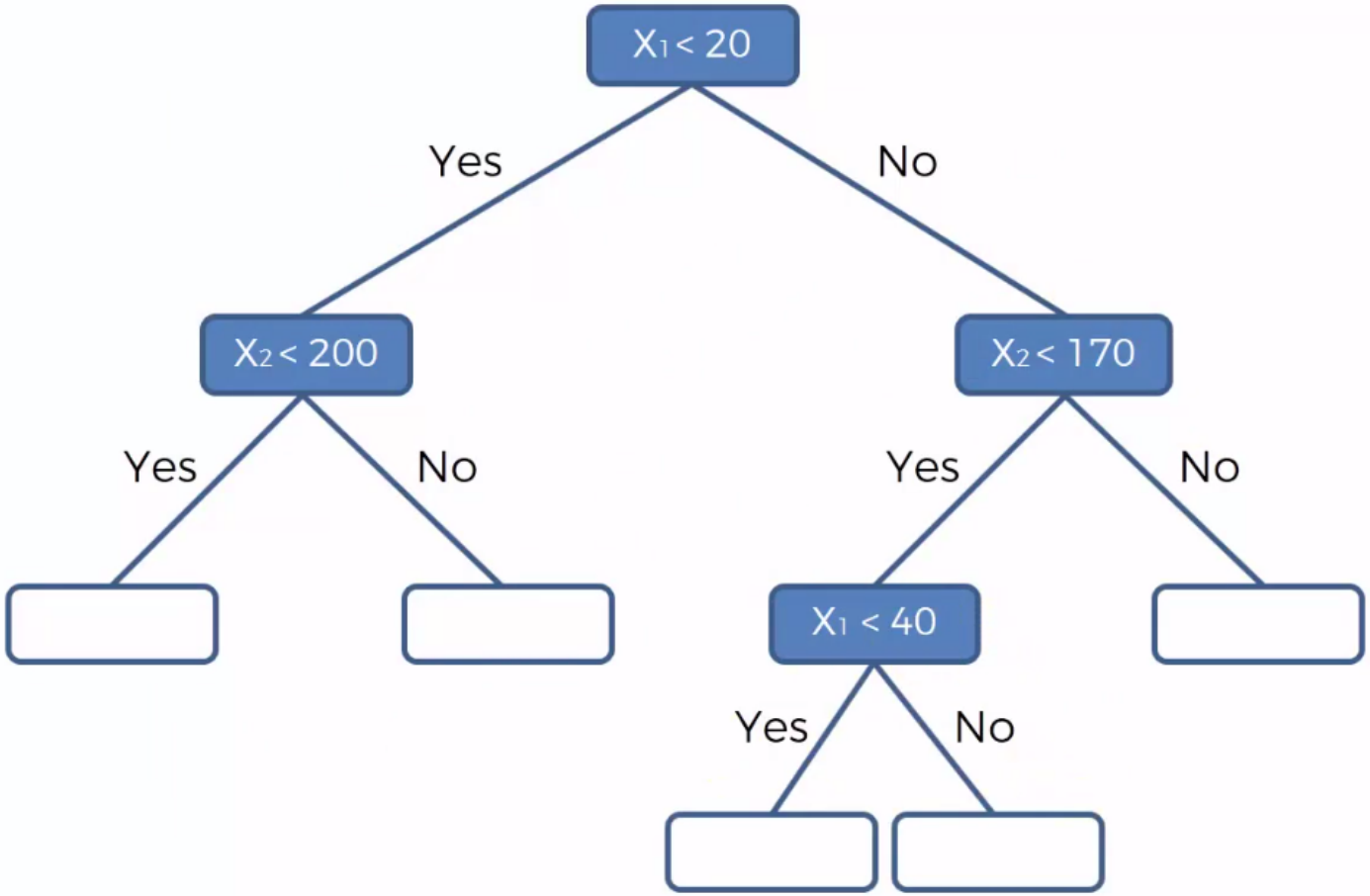
ÁRVORES DE REGRESSÃO



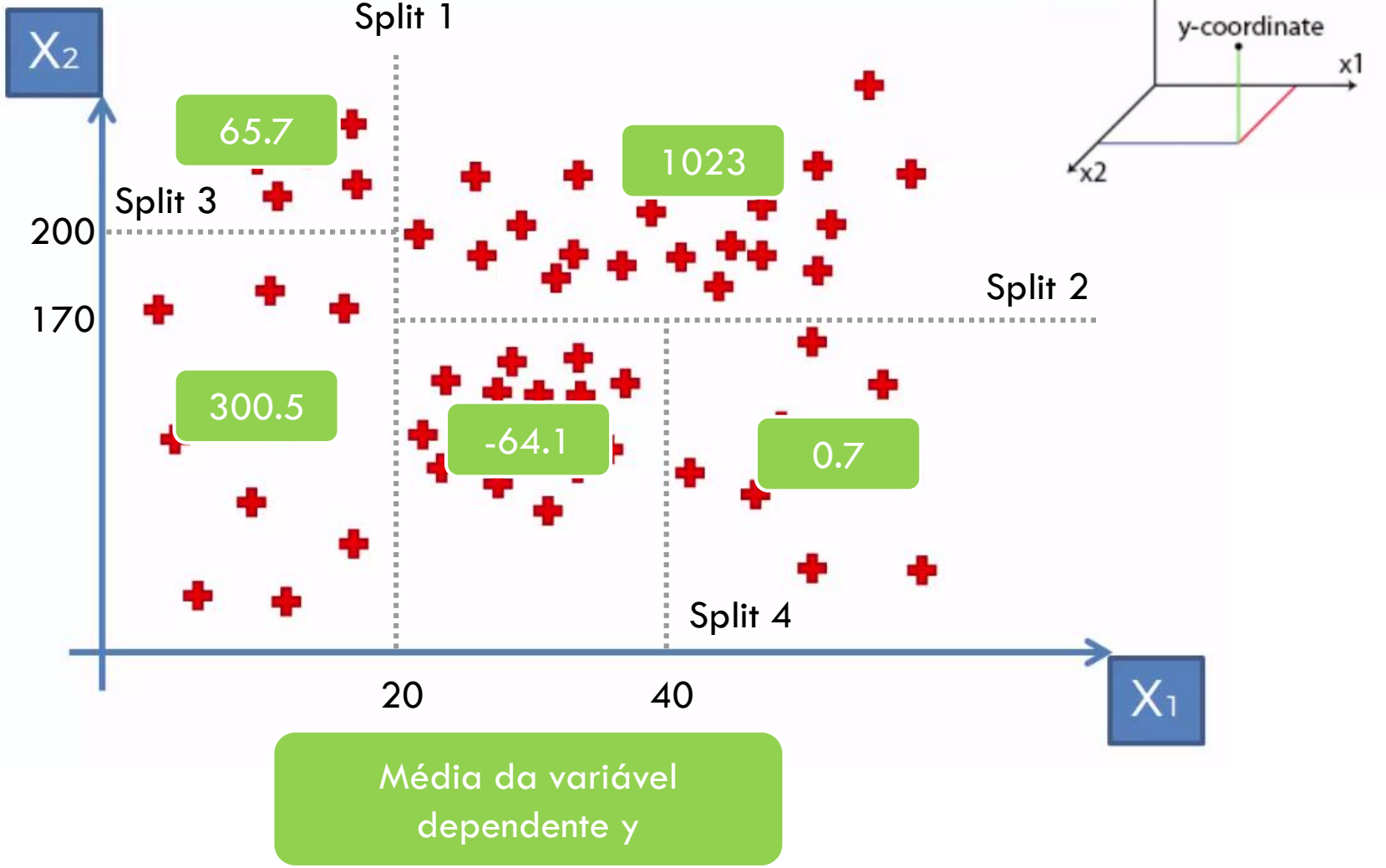
ÁRVORES DE REGRESSÃO



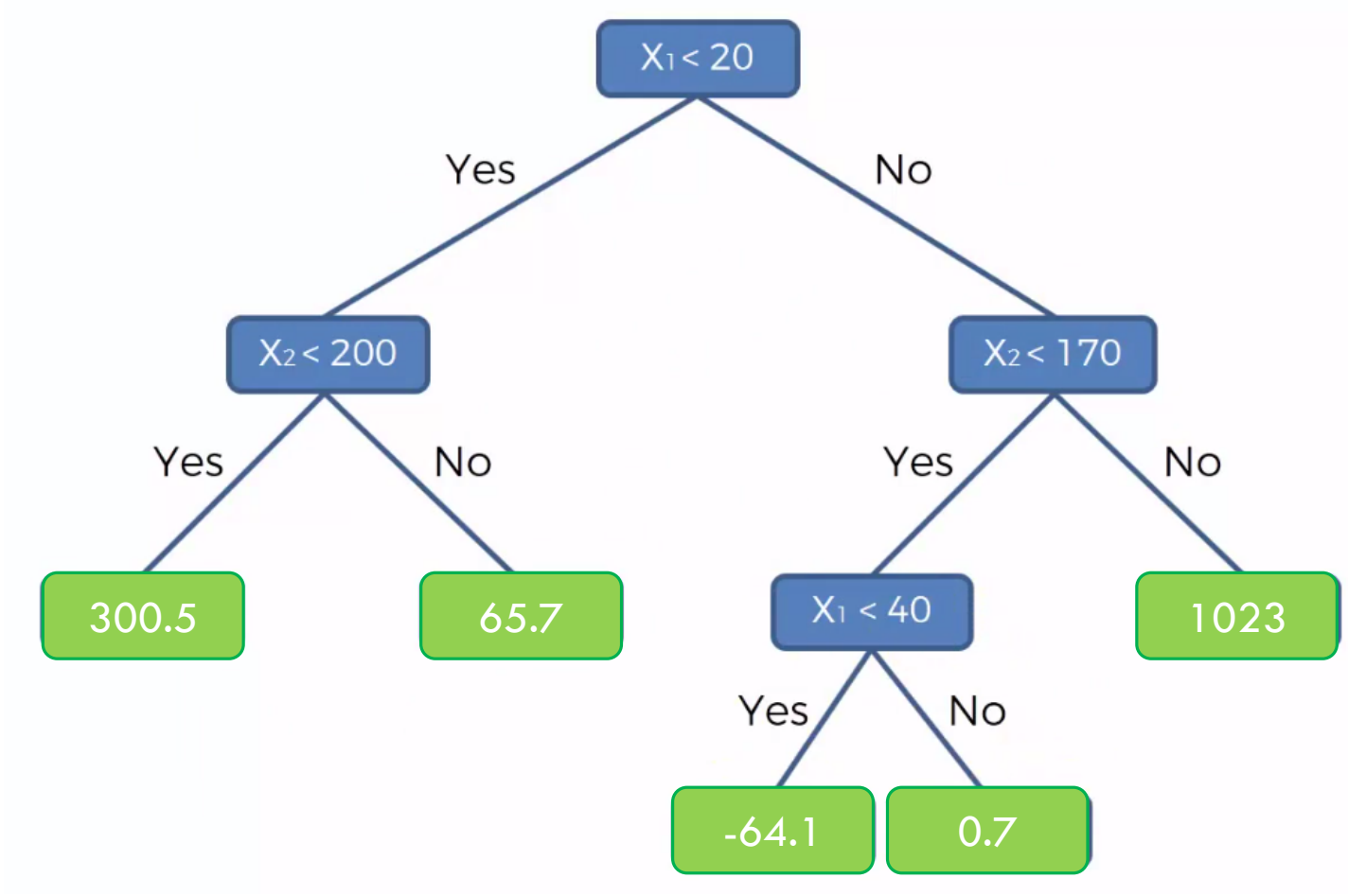
ÁRVORES DE REGRESSÃO



ÁRVORES DE REGRESSÃO

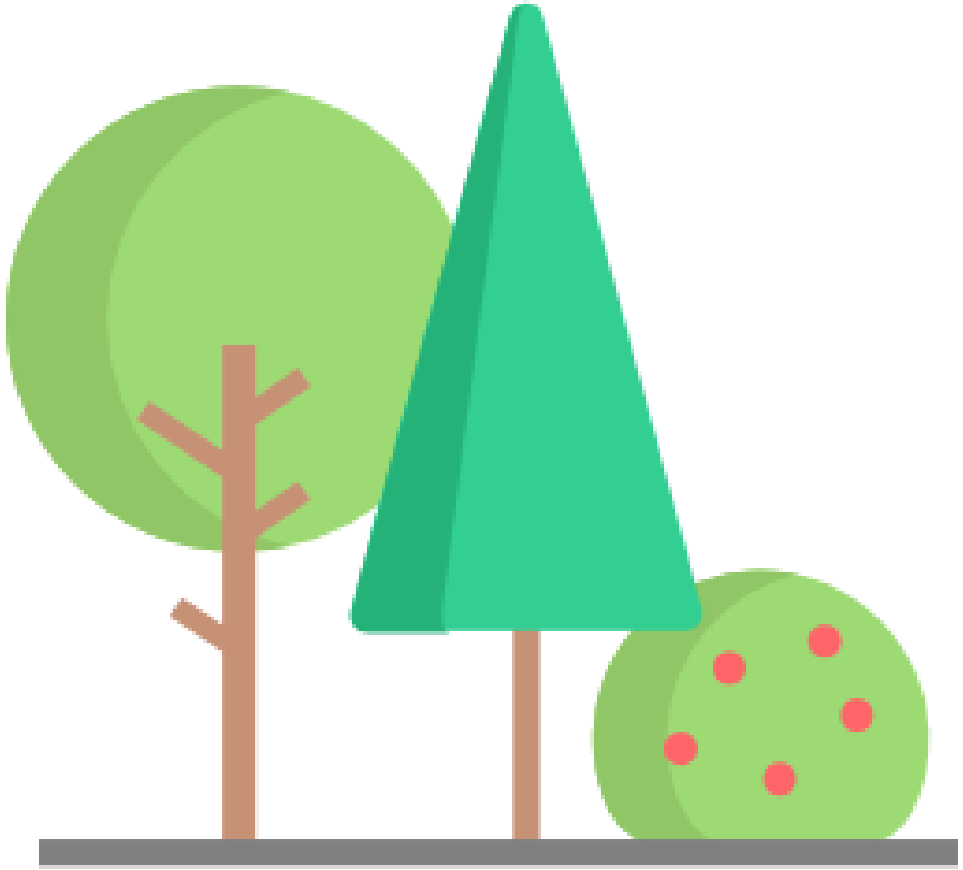


ÁRVORES DE REGRESSÃO



Random Forest

RANDOM FOREST



RANDOM FOREST

PASSO 1: Escolhe o número de árvores que se deseja construir e repete os passos 2 e 3 para cada uma das árvores.

PASSO 2: Escolhe aleatoriamente K dados do conjunto de treinamento.

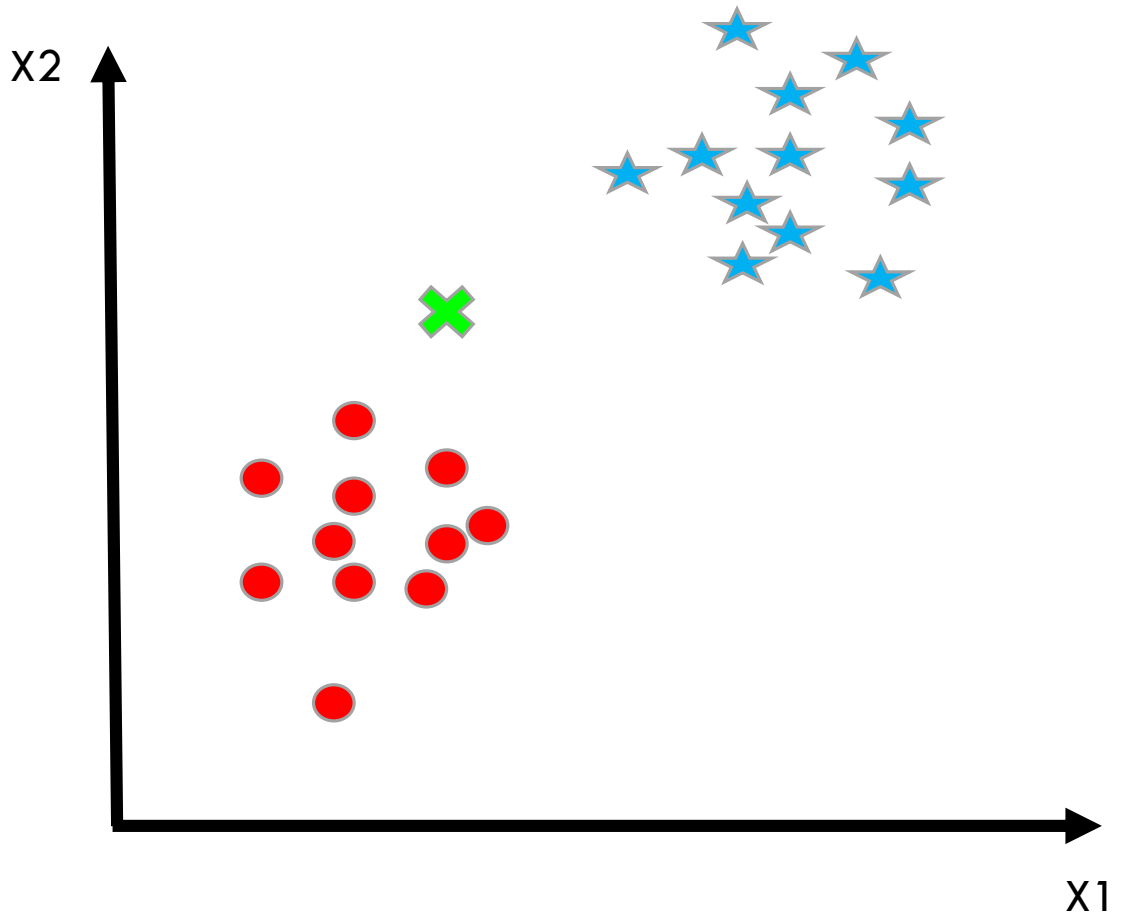
PASSO 3: Constrói a árvore de decisão associada àqueles K dados.

PASSO 4: Para cada dado novo, faça com que cada árvore da sua floresta infira um valor da variável resposta (dependente). A resposta do comitê será, por exemplo, a média aritmética da inferência de todas as árvores.

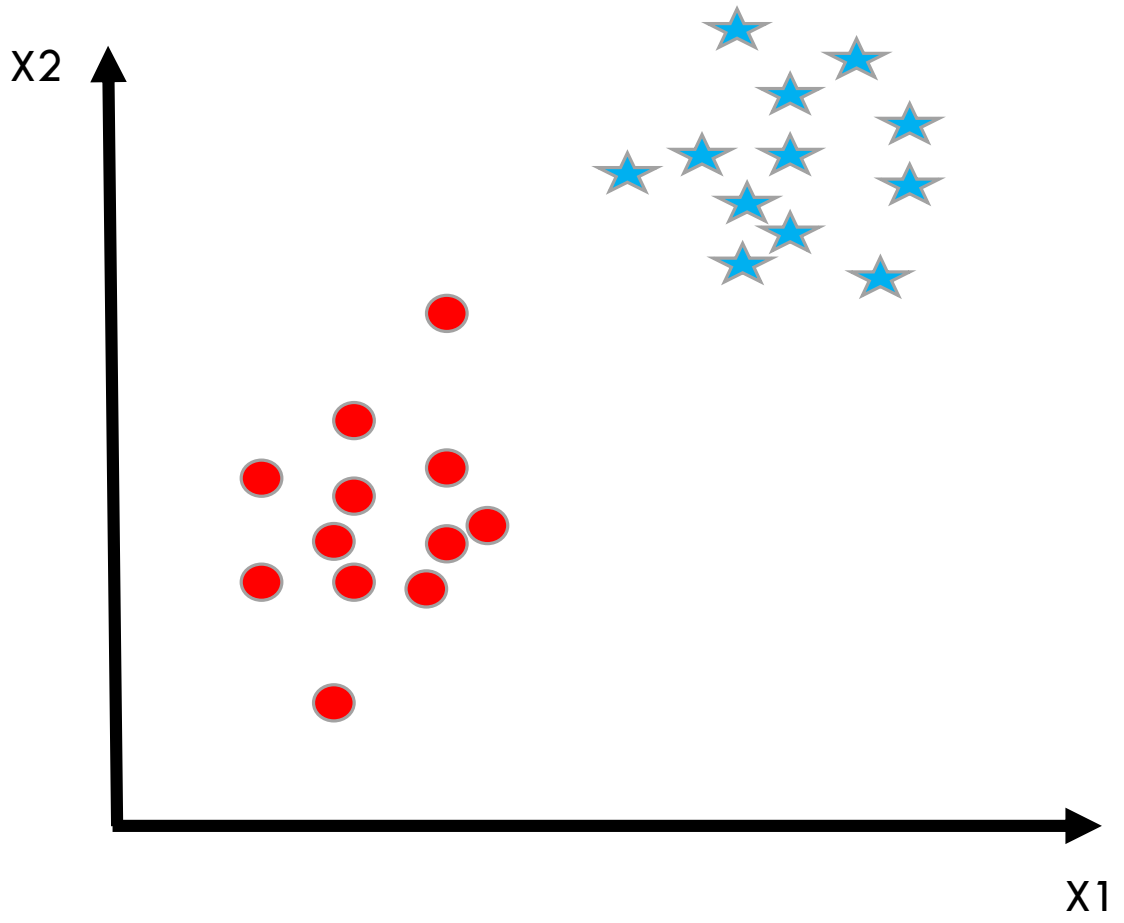
Master

K Nearest Neighbors

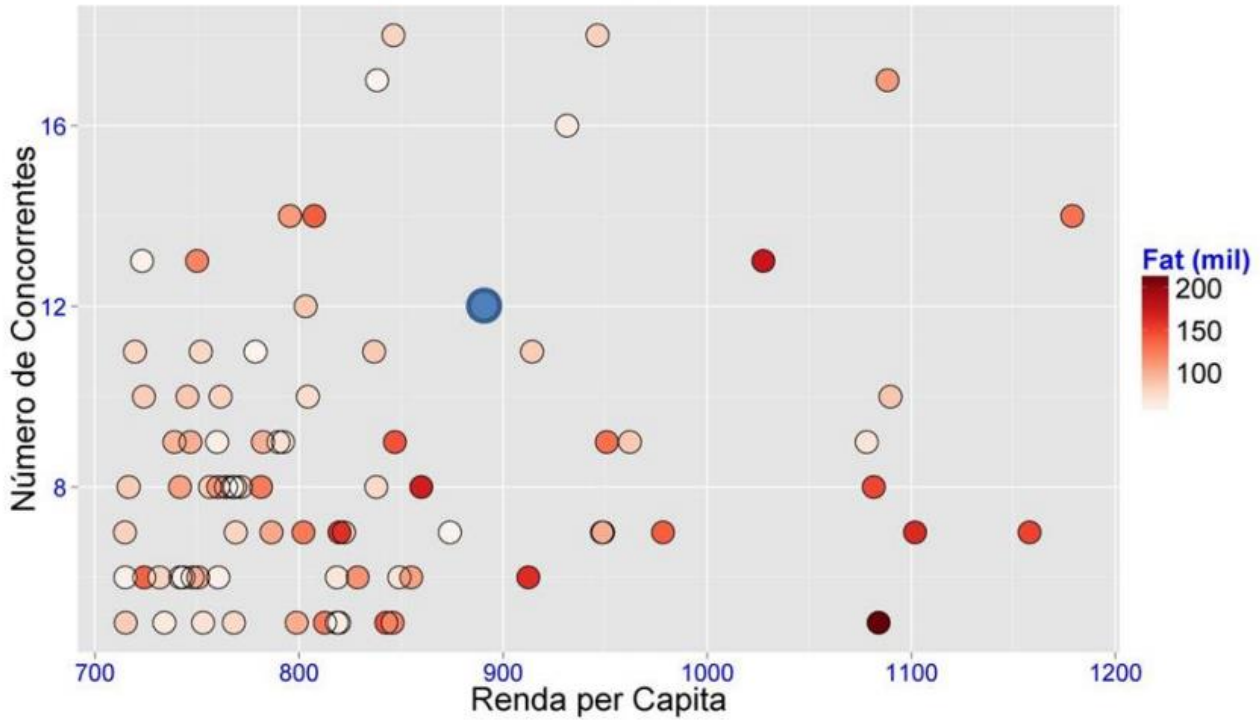
KNN - CLASSIFICAÇÃO



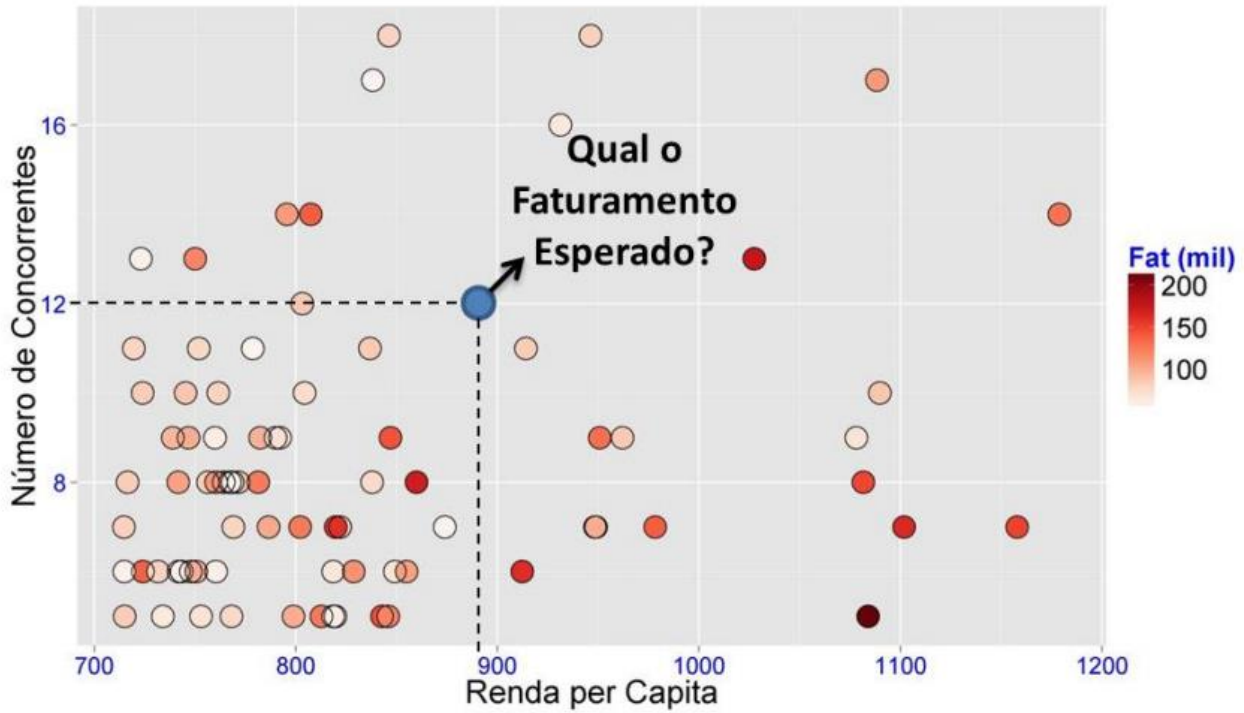
KNN - CLASSIFICAÇÃO



KNN PARA REGRESSÃO

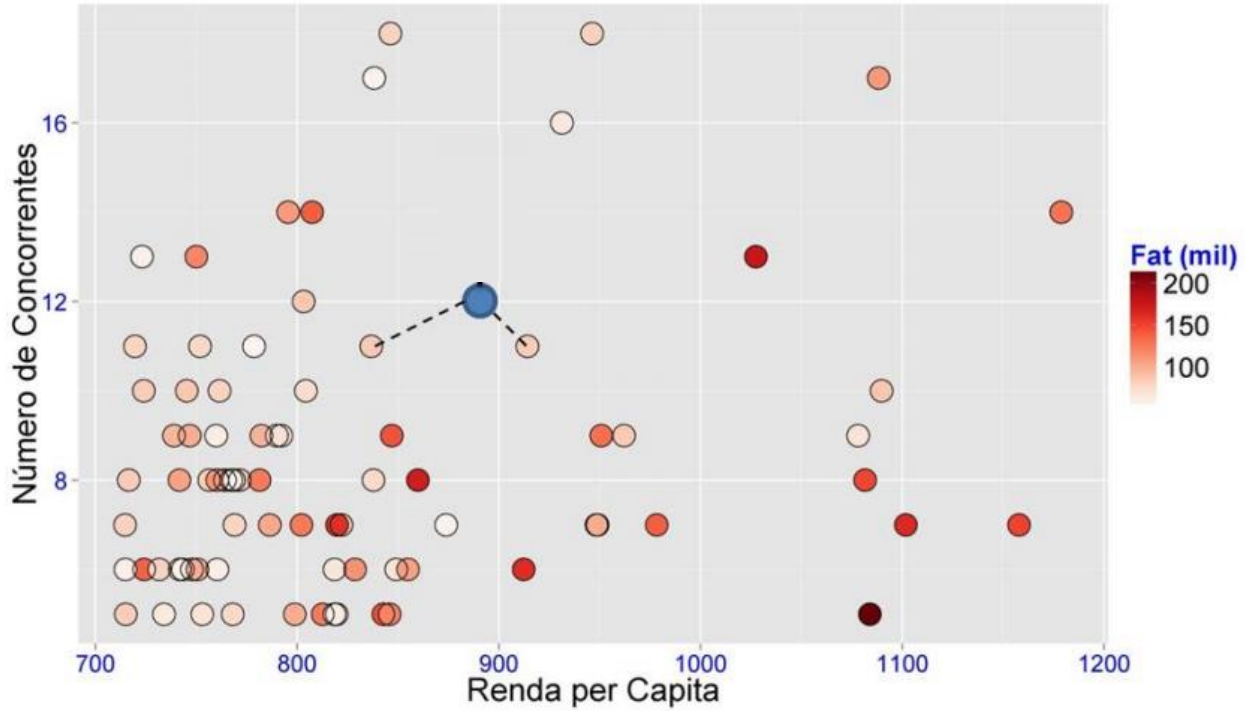


KNN PARA REGRESSÃO



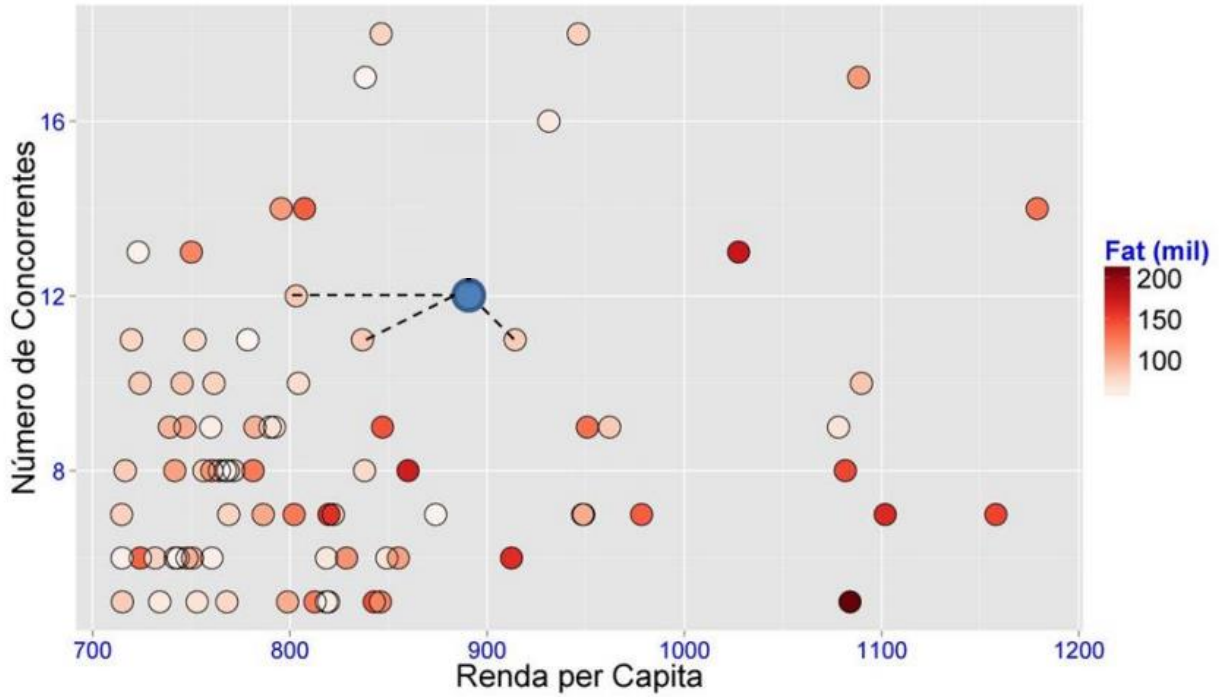
KNN PARA REGRESSÃO

- **2NN para Regressão**: O Faturamento Esperado é a Média Aritmética dos **2** vizinhos mais próximos!



KNN PARA REGRESSÃO

- **3NN para Regressão**: O Faturamento Esperado é a Média Aritmética dos **3** vizinhos mais próximos!



KNN

1. Dados com domínio definido e dados suficientes dentro deste domínio;
2. Não paramétrico;
3. Simples;
4. Local.

ESTUDO DE CASO

Aluguel de Bicicletas (2011-2012)

https://www.capitalbikeshare.com/

- **9 variáveis independentes:**
  - Estação do ano (1:primavera, 2:verão, 3:outono, 4:inverno);
  - Feriado;
  - Dia de semana;
  - Dia de trabalho;
  - Tempo (1:limpo, 2:nublado, 3:neve / chuva);
  - Temperatura;
  - Sensação térmica;
  - Humidade;
  - Velocidade do vento.

Variável resposta: horas de uso de bike.

Bi Master

BOSTON HOUSING

506 registros e 14 atributos:

Crim: per capita crime rate by town.

Zn: proportion of residential land zoned for lots over 25,000 sq.ft.

Indus: proportion of non-retail business acres per town.

Chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

Nox: nitrogen oxides concentration (parts per 10 million).

Rm: average number of rooms per dwelling.

Age: proportion of owner-occupied units built prior to 1940.

Dis: weighted mean of distances to five Boston employment centres.

Rad: index of accessibility to radial highways.

Tax: full-value property-tax rate per $10,000.

Ptratio: pupil-teacher ratio by town.

Black: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

Lstat: lower status of the population (percent).

Medv: median value of owner-occupied homes in $1000s.

Master