

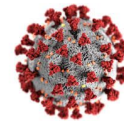
INSTITUTO FEDERAL
Goiás

Regressão Linear

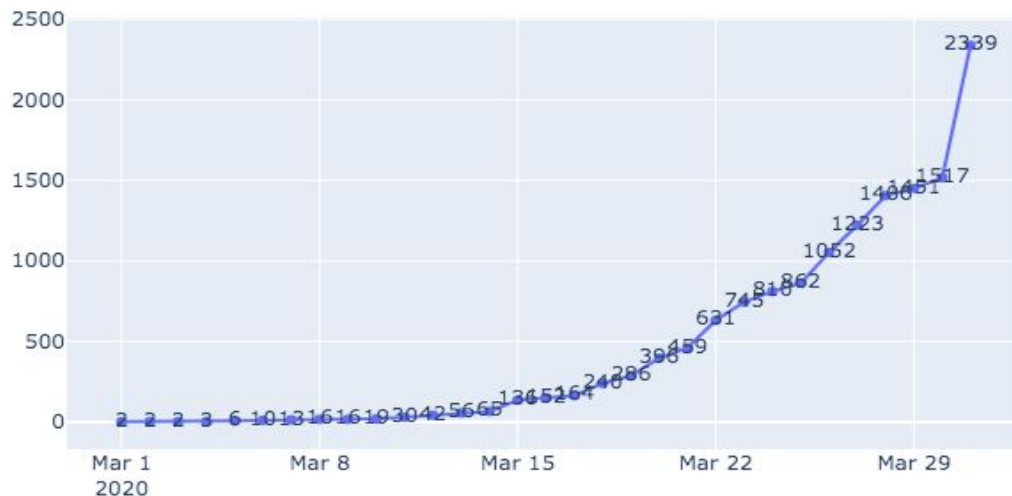
Prof. Dr. Eduardo Noronha
Instituto Federal de Goiás (IFG)

eduardonaf@gmail.com

Regressão Linear (COVID-19)



Casos Confirmados em SP



$$y = w_1 * x + w_0$$

Qual é a melhor reta?

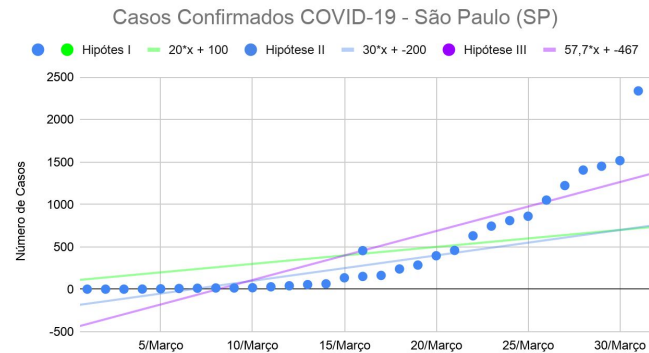
Qual a Hipótese com maior capacidade de generalizar?

Como encontrá-la?

Fonte: Ministério da Saúde (<https://covid.saude.gov.br/>)

Escolhendo as Melhores Hipóteses (Least Squares Method)

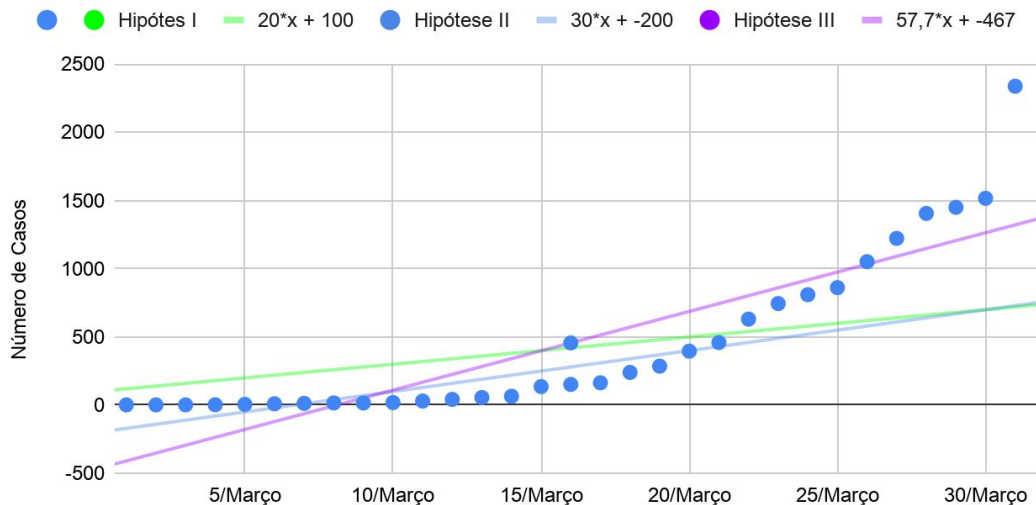
Data	Dias (X)	Casos Acumulados (y)	x-x_med	y-y_med	Hipótese I (h1=w1.x+w0)			Hipótese II (h2=w1.x+w0)			Hipótese III (h3=w1.x+w0)		
					Hipótese I	Erro (y - h1)	Erro ² (y - h1) ²	Hipótese II	Erro (y - h2)	Erro ² (y - h2) ²	Hipótese III	Erro (y - h3)	Erro ² (y - h3) ²
01/03/202	1	2	-15,0	-454,48	120,00	-118,00	13.924,00	-170,00	172,00	29.584,00	-409,29	411,29	169.163,05
02/03/202	2	2	-14,0	-454,48	140,00	-138,00	19.044,00	-140,00	142,00	20.164,00	-351,58	353,58	125.015,85
03/03/202	3	2	-13,0	-454,48	160,00	-158,00	24.964,00	-110,00	112,00	12.544,00	-293,86	295,86	87.531,52
04/03/202	4	3	-12,0	-453,48	180,00	-177,00	31.329,00	-80,00	83,00	6.889,00	-236,14	239,14	57.187,32
05/03/202	5	6	-11,0	-450,48	200,00	-194,00	37.636,00	-50,00	56,00	3.136,00	-178,42	184,42	34.010,80
06/03/202	6	10	-10,0	-446,48	220,00	-210,00	44.100,00	-20,00	30,00	900,00	-120,70	130,70	17.082,91
07/03/202	7	13	-9,0	-443,48	240,00	-227,00	51.529,00	10,00	3,00	9,00	-62,98	75,98	5.773,43
08/03/202	8	16	-8,0	-440,48	260,00	-244,00	59.536,00	40,00	-24,00	576,00	-5,26	21,26	452,18
09/03/202	9	16	-7,0	-440,48	280,00	-264,00	69.696,00	70,00	-54,00	2.916,00	52,45	-36,45	1.328,90
10/03/202	10	19	-6,0	-437,48	300,00	-281,00	78.961,00	100,00	-81,00	6.561,00	110,17	-91,17	8.312,44
11/03/202	11	30	-5,0	-426,48	320,00	-290,00	84.100,00	130,00	-100,00	10.000,00	167,89	-137,89	19.013,96
12/03/202	12	42	-4,0	-414,48	340,00	-298,00	88.804,00	160,00	-118,00	13.924,00	225,61	-183,61	33.712,51
13/03/202	13	56	-3,0	-400,48	360,00	-304,00	92.416,00	190,00	-134,00	17.956,00	283,33	-227,33	51.678,12
14/03/202	14	65	-2,0	-391,48	380,00	-315,00	99.225,00	220,00	-155,00	24.025,00	341,05	-276,05	76.201,82
15/03/202	15	136	-1,0	-320,48	400,00	-264,00	69.696,00	250,00	-114,00	12.996,00	398,77	-262,77	69.045,61
16/03/202	16	152	0,0	-304,48	420,00	-268,00	71.824,00	280,00	-128,00	16.384,00	456,48	-304,48	92.710,43
17/03/202	17	164	1,0	-292,48	440,00	-276,00	76.176,00	310,00	-146,00	21.316,00	514,20	-350,20	122.641,73
18/03/202	18	240	2,0	-216,48	460,00	-220,00	48.400,00	340,00	-100,00	10.000,00	571,92	-331,92	110.171,53
19/03/202	19	286	3,0	-170,48	480,00	-194,00	37.636,00	370,00	-84,00	7.056,00	629,64	-343,64	118.088,12
20/03/202	20	396	4,0	-60,48	500,00	-104,00	10.816,00	400,00	-4,00	16,00	687,36	-291,36	84.889,52
21/03/202	21	459	5,0	2,52	520,00	-61,00	3.721,00	430,00	29,00	841,00	745,08	-286,08	81.839,83
22/03/202	22	631	6,0	174,52	540,00	91,00	8.281,00	460,00	171,00	29.241,00	802,80	-171,80	29.513,58
23/03/202	23	745	7,0	288,52	560,00	185,00	34.225,00	490,00	255,00	65.025,00	860,51	-115,51	13.343,42
24/03/202	24	810	8,0	353,52	580,00	230,00	52.900,00	520,00	290,00	84.100,00	918,23	-108,23	11.714,22
25/03/202	25	862	9,0	405,52	600,00	262,00	68.644,00	550,00	312,00	97.344,00	975,95	-113,95	12.984,79
26/03/202	26	1052	10,0	595,52	620,00	432,00	186.624,00	580,00	472,00	222.784,00	1.033,67	18,33	336,01
27/03/202	27	1223	11,0	766,52	640,00	583,00	339.889,00	610,00	613,00	375.769,00	1.091,39	131,61	17.321,74
28/03/202	28	1406	12,0	949,52	660,00	746,00	556.516,00	640,00	766,00	586.756,00	1.149,11	256,89	65.994,30
29/03/202	29	1451	13,0	994,52	680,00	771,00	594.441,00	670,00	781,00	609.961,00	1.206,83	244,18	59.621,43
30/03/202	30	1517	14,0	1060,52	700,00	817,00	667.489,00	700,00	817,00	667.489,00	1.264,54	252,46	63.734,26
31/03/202	31	2339	15,0	1882,52	720,00	1.619,00	2.621.161,00	730,00	1.609,00	2.588.881,00	1.322,26	1.016,74	1.033.755,96
Média	16,00	456,48				MSE	201.410	MSE	178.876	MSE	86.264		



Hipótese	w1	w0	Erro
I	20	100	201.410
II	30	-200	178.876
III	57,72	-467	86.264

Escolhendo as Melhores Hipóteses (Least Squares Method)

Casos Confirmados COVID-19 - São Paulo (SP)



Hipótese	w1	w0	Erro
I	20	100	201.410
II	30	-200	178.876
III	57,72	-467	86.264

Otimizando as Variáveis de Decisão (Least Squares Method)

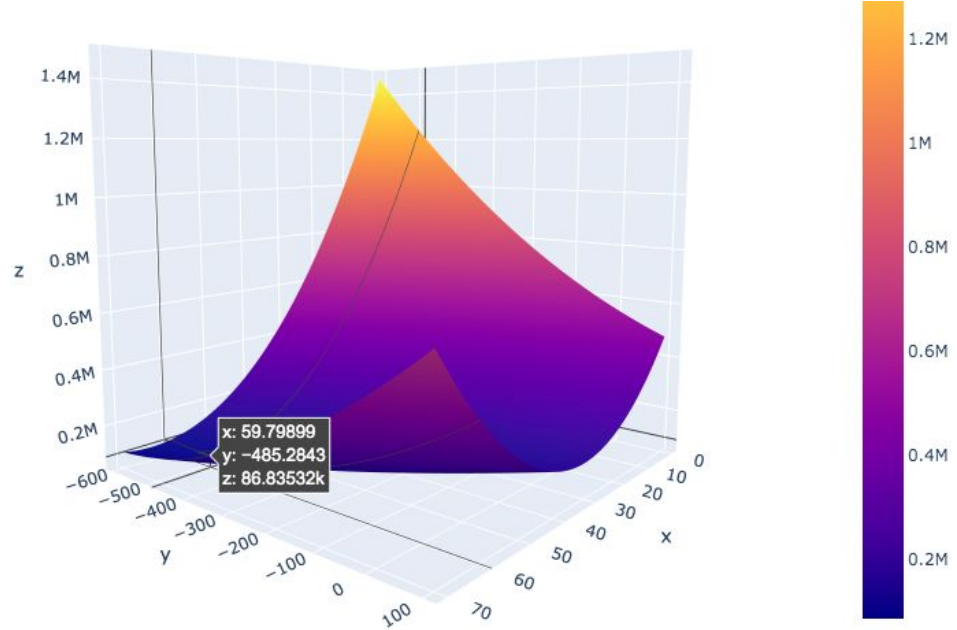
$$E = \frac{1}{n} \sum_{i=1}^n (y_i - h_i)^2$$

$$E = \frac{1}{n} \sum_{i=1}^n (y_i - (w_1 * x_i + w_0))^2$$

Para minimizar o Erro:

$$w_1 = \frac{\partial E}{\partial w_1} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0$$

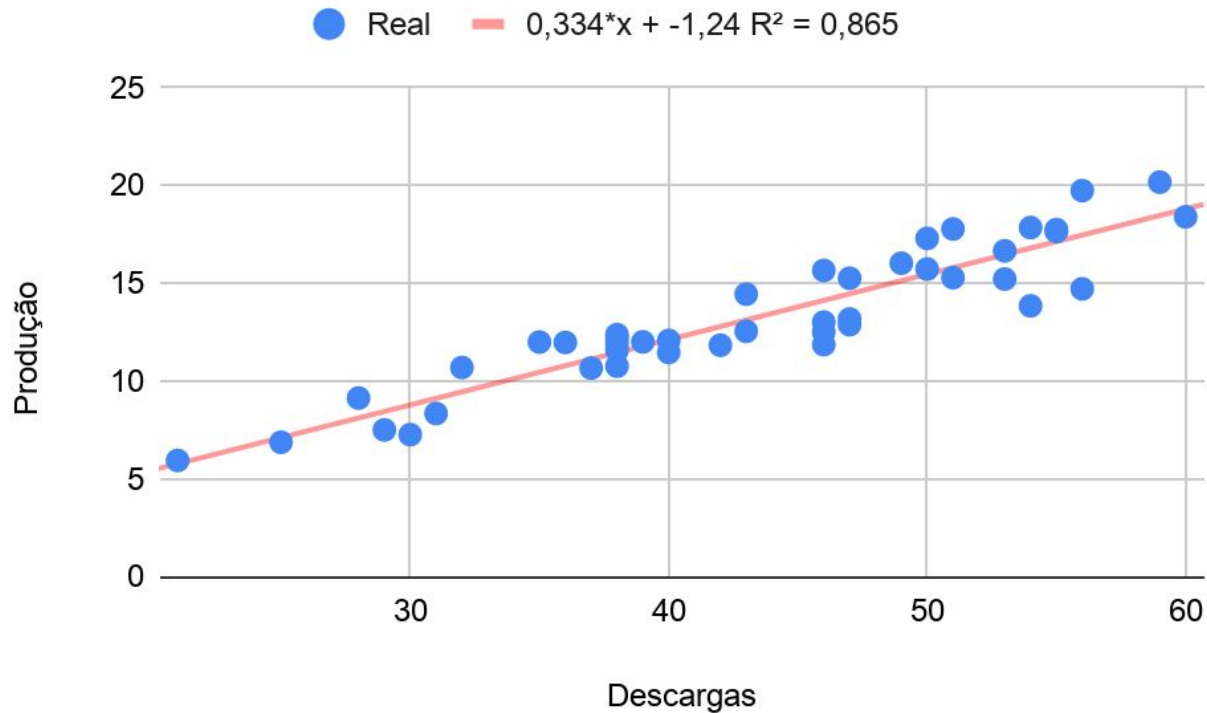
$$w_0 = \frac{\partial E}{\partial w_0} = \bar{y} - w_1 * \bar{x} = 0$$



Exemplo - Regressão Linear Simples

Real	Estimado	Vazao_regular	Vazao_pico	Descargas	Erro
9,13	8,4	4,29	10,87	28	-0,08
7,5	8,7	6,03	19,74	29	0,16
11,83	12,6	5,45	18,84	42	0,07
12,5	13,8	5,47	19,17	46	0,1
15,21	15,9	4,91	12,03	53	0,05
19,73	16,8	3,95	17,32	56	-0,15
11,97	10,8	4,18	20,77	36	-0,1
11,57	11,4	4,45	16,47	38	-0,01
10,66	11,1	5,06	22,83	37	0,04
11,89	11,4	4,25	15,8	38	-0,04
13,84	16,2	6,09	16,05	54	0,17
8,34	9,3	5,41	19,99	31	0,12
12,54	12,9	4,56	15,91	43	0,03
17,84	16,2	4,12	10,79	54	-0,09
11,99	10,5	3,9	20,36	35	-0,12
10,69	9,6	4,1	16,79	32	-0,1
11,86	13,8	5,83	15,23	46	0,16

Exemplo de Regressão Linear



Bibliotecas em Python

```
import pandas as pd
from datetime import datetime
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.linear_model import LinearRegression
```


Código em Python - Regressão Linear

```
df = pd.read_csv(DIR+'/dados.csv', sep=';', header=0)

X = df.iloc[:, 4:5] # Pegar somente descargas (variáveis independentes)
y = df.iloc[:, 0:1] # Pegar somente Producao Real (variável dependente)
print(X)           # Imprimindo as variáveis independentes

regressor = LinearRegression()
regressor.fit(X, y) # Fitting Multiple Linear Regression to the Training set
y_pred = regressor.predict(X) # Predicting the Test set results
score=r2_score(y,y_pred)
print('Score: '+str(score))
print('Coefficients: \n', regressor.coef_)
print('intercept_: \n', regressor.intercept_)
print('Mean squared error: %.2f' % mean_squared_error(y, y_pred))
print('Coefficient of determination: %.2f' % r2_score(y, y_pred))
```

Exercícios

- 1) Realizar em Python a Regressão Linear dos Dados de COVID-19 no estado de São Paulo no mês de março de 2020.

Variável independente: dias (1, 2, ...31)

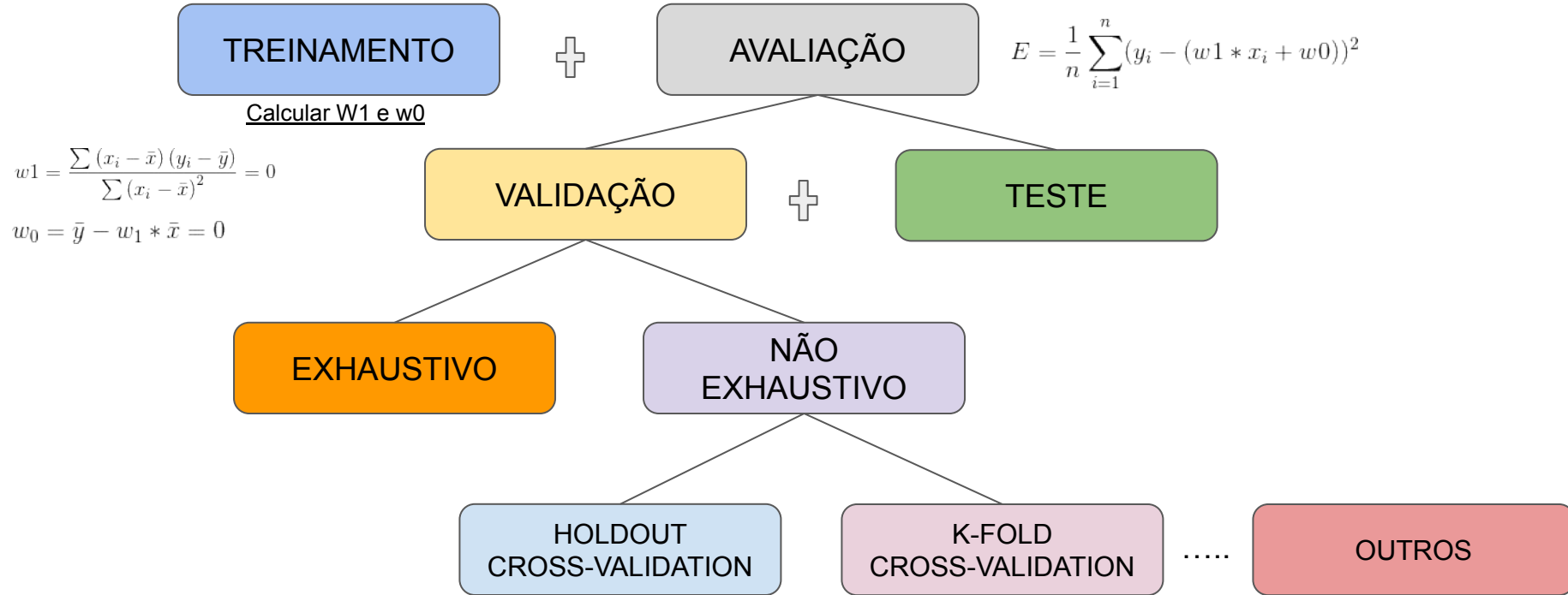
Variável dependente: Número de casos confirmados acumulados

Fonte: <https://covid.saude.gov.br/>

- 2) Gerar Regressão Linear por países

<https://c3.ai/covid-19-api-documentation>

Processo de Avaliação



Exemplo de Cross Validation

```
import pandas as pd
from datetime import datetime
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score

DIR = '/Users/noronha/Documents/InteligenciaArtificial/LinearRegression'
df = pd.read_csv(DIR+'/dados.csv', sep=';', header=0)
X = df.iloc[:, 4:5] # Pegar somente descargas (variáveis independentes)
y = df.iloc[:, 0:1] # Pegar somente Producao Real (variável dependente)

regressor = LinearRegression() #
scores = cross_val_score(regressor, X, y, cv=5)
print(scores)
print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))
```

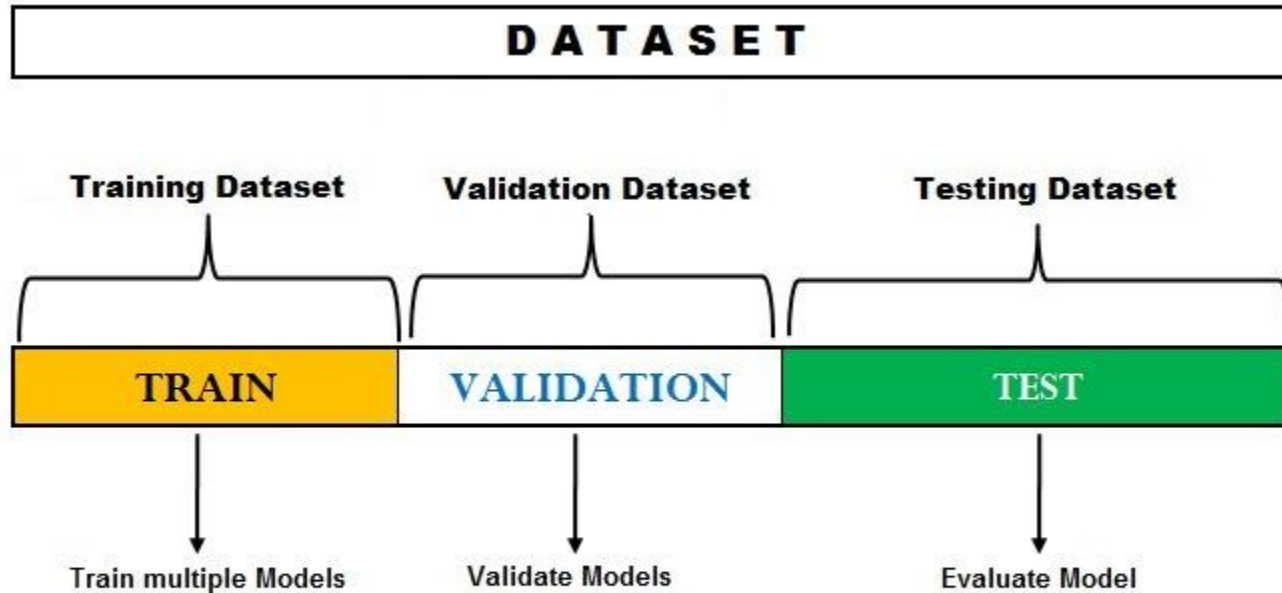
Modelo utilizado

Número de splits

```
[0.86176169 0.7184365 0.84186277 0.89923958 0.547708 ]
Accuracy: 0.77 (+/- 0.26)
```

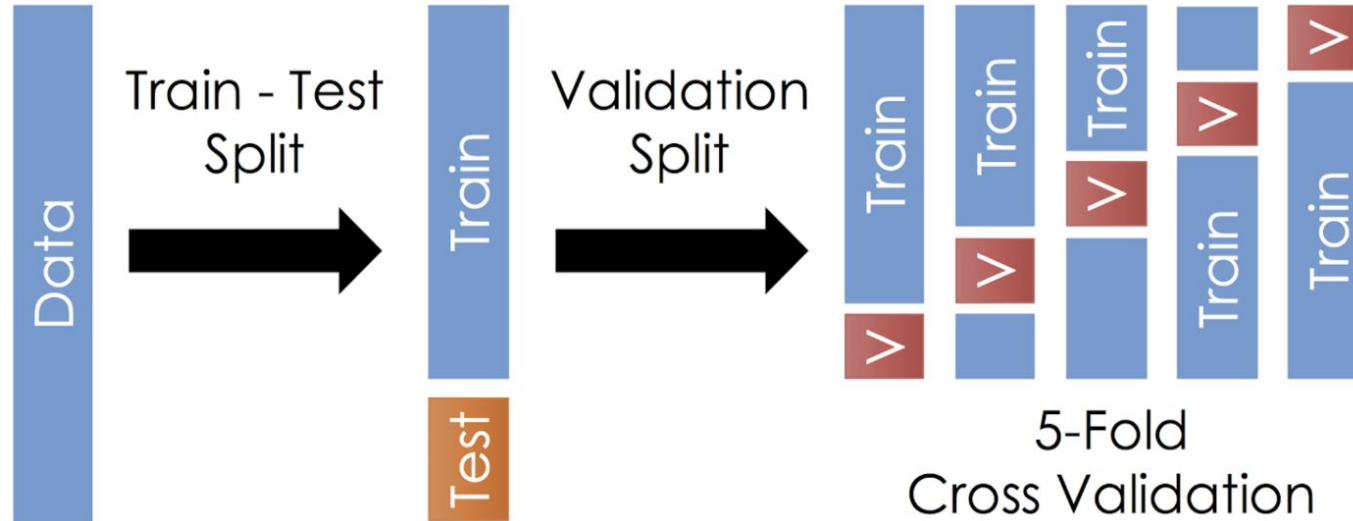
Avaliação da Hipótese

Tarefas: Treinamento, Validação e Teste



Avaliação da Hipótese

K-fold cross-validation



Exercício

- 1) Aplicar técnicas de validação Cross Validation e K-fold ($n=5$) nos exercícios anteriores, e verificar o desempenho dos ajustes realizados (Mean Squared Error - MSE)