AULA 9: APRENDIZADO DE MÁQUINAS
MODELOS DE REGRESSÃO

INTRODUÇÃO A CIÊNCIA DE DADOS NA ENGENHARIA DE PETRÓLEO

Maria Clara Duque (clara_duque@poli.ufrj.br)

Calendário

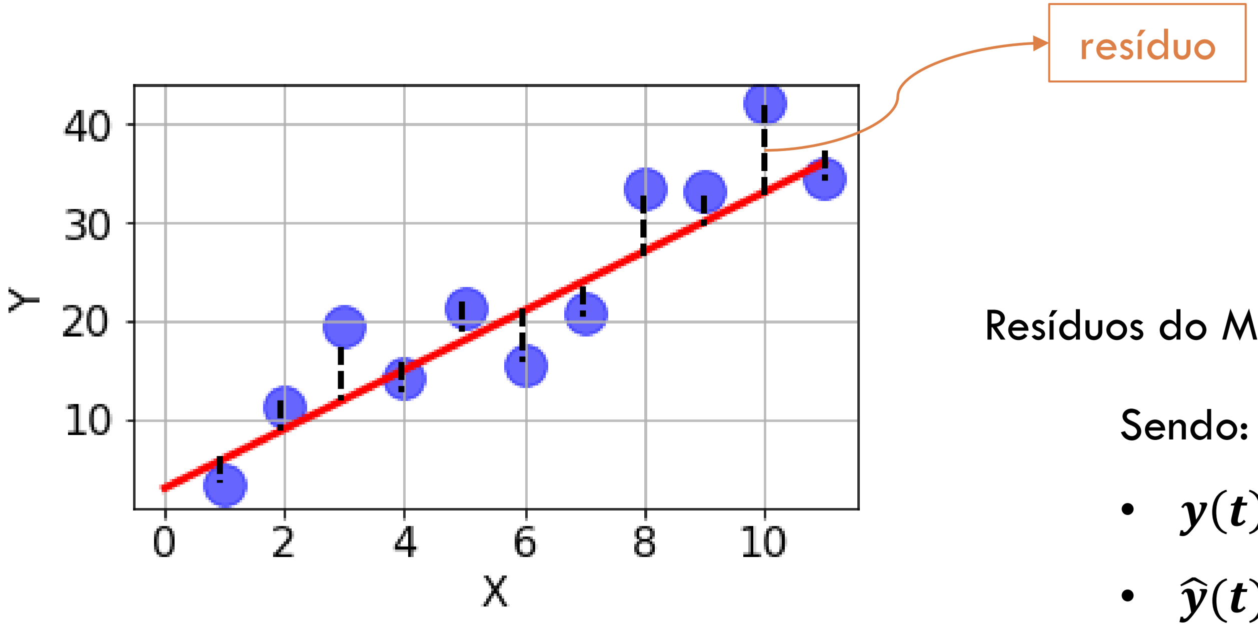
DATA	ATIVIDADE
26/08	Introdução
02/09	Tipos de dados/ Pré-processamento
09/09	Aula Prática 1
16/09	Aula Prática 2
23/09	Aula Prática 3
30/09	Introdução ML
07/10	ML Classificação
14/10	Aula Prática 4
21/10	**ML Regressão**
28/10	Feriado
04/11	ML Agrupamento/Aula Prática 5
11/11	Entrega dos Trabalhos

Tópicos

- Modelos de Regressão;
- Métricas de Avaliação;
- Regressão Linear Simples;
- Regressão Linear Multivariada;
- Regressões de Suporte de Vetores (Support Vector Regression – SVR);

Modelos de Regressão

- Modelo capaz de realizar estimativa do valor da variável de saída contínua a partir das variáveis de entrada.
- Avaliação da predição:
  - Função do erro de predição (Valor Real - Valor Predito).



Resíduos do Modelo: $\boldsymbol{e(t) = y(t) - \hat{y}(t)}$

Sendo:

- $\boldsymbol{y(t)}$: y real
- $\boldsymbol{\hat{y}(t)}$: y ajustado

Métricas de Avaliação da Regressão

- Métricas geralmente usadas para avaliar modelos de regressão:

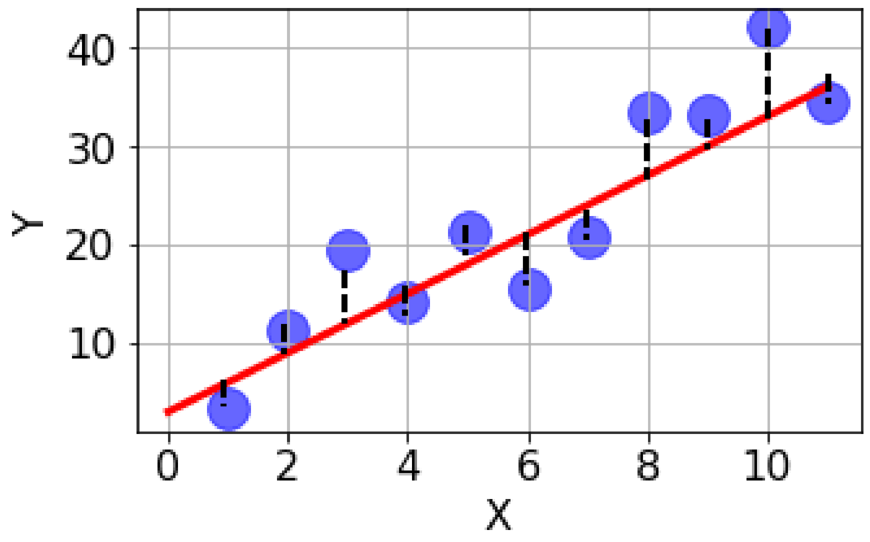
Nome da Métrica	Equação
Erro Quadrático Médio	$EQM = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
Raiz quadrada do Erro Quadrático Médio (RMS)	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Erro absoluto médio percentual (MAPE)	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$
Coeficiente de Determinação (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

Sendo:

- y **: Valor Real**
- $\hat{y}$ **: Valor Ajustado**
- $\bar{y}$ **: Valor Médio**
- **N : Número de Amostras**

Métricas de Avaliação da Regressão

- Erro Quadrático Médio:

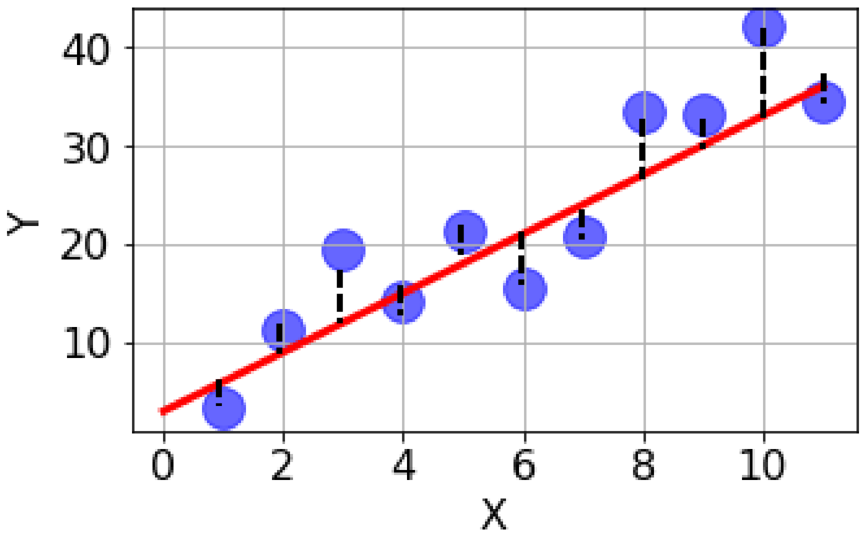


Nome da Métrica	Equação
Erro Quadrático Médio	$EQM = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

Atentar que não está na mesma escala que a variável resposta.

Métricas de Avaliação da Regressão

- Raiz Quadrada do Erro Quadrático Médio:

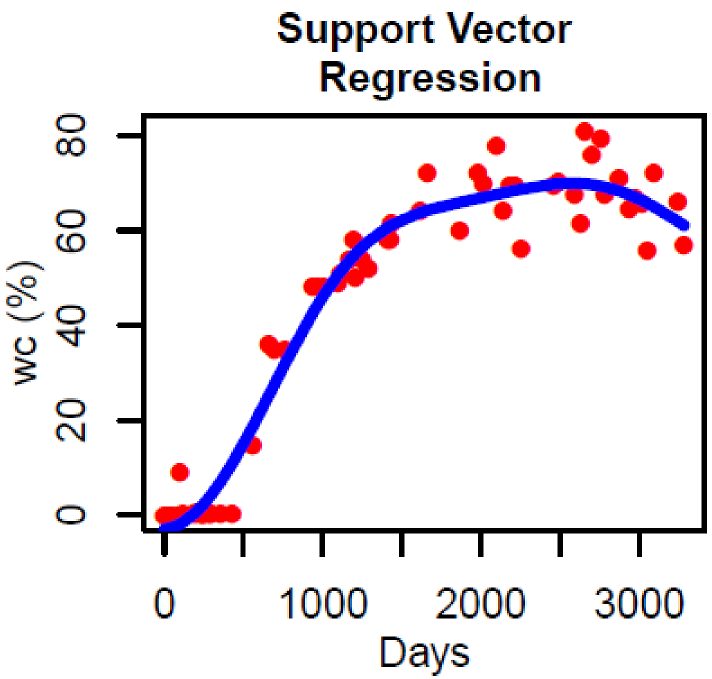


Nome da Métrica	Equação
Raiz quadrada do Erro Quadrático Médio (RMS)	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

Mesma escala da variável de saída e pode ser analisado no contexto do problema.

Métricas de Avaliação da Regressão

- Erro Absoluto Médio Percentual:



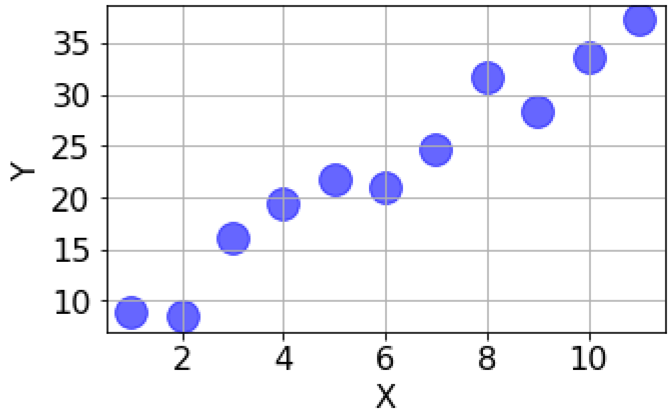
Nome da Métrica	Equação
Erro absoluto médio percentual (MAPE)	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left\|\frac{y_i - \hat{y}_i}{y_i}\right\|$

Tem que ser utilizado com cautela.
Evitar usar quando existem valores reais que são 0 ou muito próximos de 0.

Métricas de Avaliação da Regressão

- Coeficiente de Determinação (R^2):

Equação
$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$



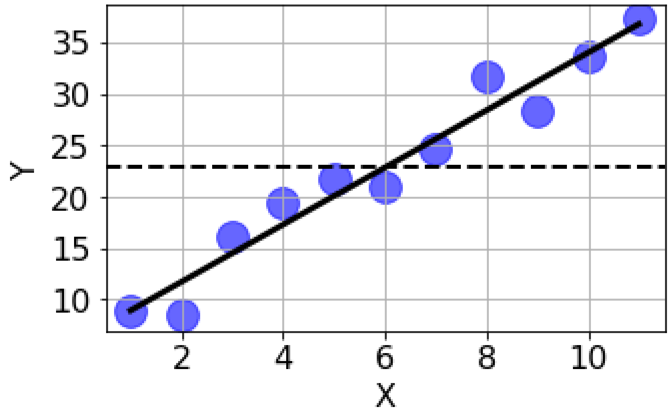
A média de y ($\bar{y}$) é melhor forma de prever o y?

Métricas de Avaliação da Regressão

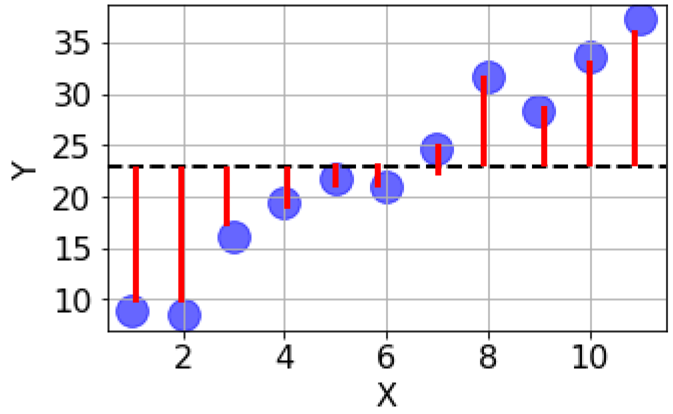
- Coeficiente de Determinação (R^2):

A média de y ($\bar{y}$) é melhor forma de prever o y?

Equação
$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$



Ajuste do Modelo



$\bar{y} = 22{,}86$

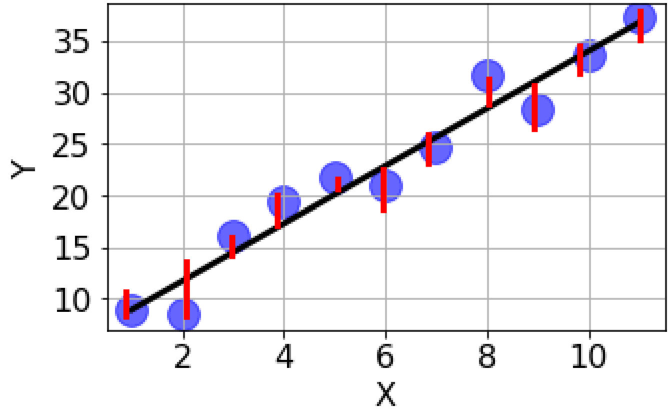
$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 896$

Métricas de Avaliação da Regressão

- Coeficiente de Determinação (R^2):

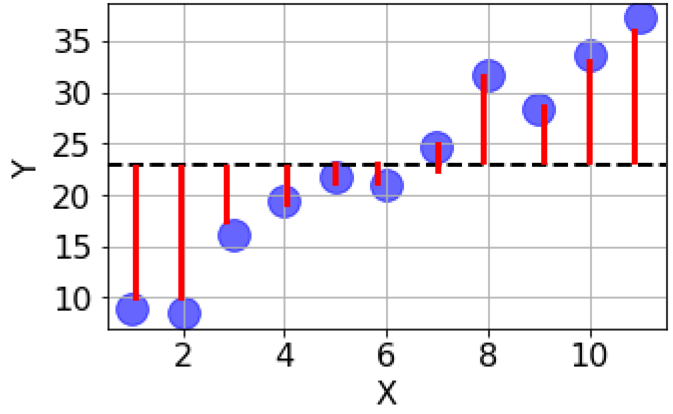
A média de y $(\bar{y})$ é melhor forma de prever o y?

Equação
$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$



Ajuste do Modelo

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 43.95$

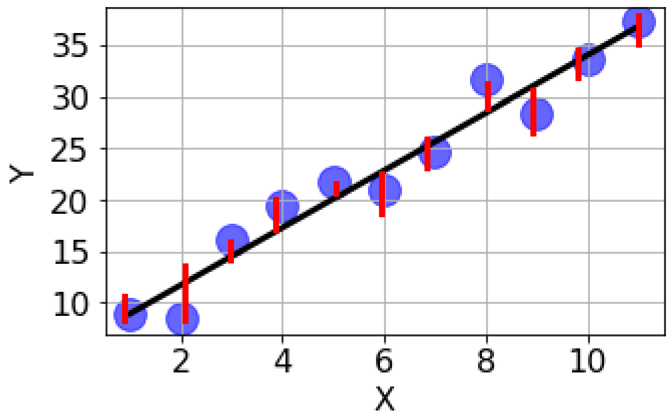


$\bar{y} = 22{,}86$

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 896$

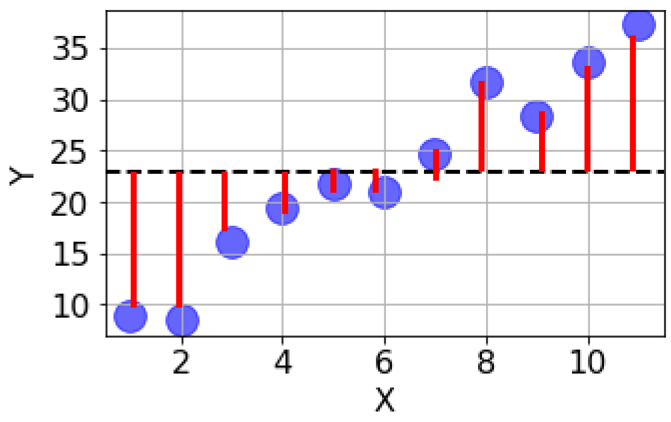
Métricas de Avaliação da Regressão

- Coeficiente de Determinação (R^2):



- Ajuste da Reta:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 43.95$$



- Média do y:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 896$$

Equação
$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

$$R^2 = \frac{\text{variação(média)} - \text{variação(ajuste)}}{\text{variação(media)}}$$

$$R^2 = \frac{896 - 43{,}95}{896} = 0{,}95$$

- Ou seja, a relação entre x e y explica 95% da variação dos dados.
- Os dados tem 95% menos variação em relação ao ajuste do que em relação a média.

Métricas de Avaliação da Regressão

- Coeficiente de Determinação (R^2):

Equação
$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

- Coeficiente de Determinação (R^2) pode ser negativo?
  - Sim! Em casos que o modelo é pior que a média da variável de saída.

- R^2 compara o erro do modelo estimado pelo erro da média da variável de saída.
  - $\boldsymbol{R^2 \approx 0}$: Modelo não consegue explicar as observações melhor que a média da variável de saída.
  - $\boldsymbol{R^2 < 0}$: Indica que o modelo é pior que a média da variável de saída.
  - $\boldsymbol{R^2 \approx 1}$: Maior parte da variabilidade total é explicada pelo modelo.

Regressão

- Regressão Linear Simples
- Regressão Linear Múltipla
- Regressão Polinomial
- Regressão de Suporte de Vetores (Support Vector Regression – SVR)

Exemplo de Dataset

X y

	Well	Por	Perm	AI	Brittle	TOC	VR	Prod
0	1	12.08	2.92	2.80	81.40	1.16	2.31	4165.196191
1	2	12.38	3.53	3.22	46.17	0.89	1.88	3561.146205
2	3	14.02	2.59	4.01	72.80	0.89	2.72	4284.348574
3	4	17.67	6.75	2.63	39.81	1.08	1.88	5098.680869
4	5	17.52	4.57	3.18	10.94	1.51	1.90	3406.132832
5	6	14.53	4.81	2.69	53.60	0.94	1.67	4395.763259
6	7	13.49	3.60	2.93	63.71	0.80	1.85	4104.400989
7	8	11.58	3.03	3.25	53.00	0.69	1.93	3496.742701

https://aegis4048.github.io/mutiple_linear_regression_and_visualization_in_python

Parte do Dataframe com 200 registros e 8 categorias

Variável	Descrição
Well	Well Index
Por	Well Average Porosity (%)
AI	Accoustic Impedance (kg/m^{2}s*10^6)
Brittle	Brittleness ratio (%)
TOC	Total Organic Carbon (%)
VR	Vitrinite Reflectance (%)
Prod	Gas production per day (MCFD)

Perguntas:

- Variável “Well” é importante para análise?
- Se fosse utiliza-la, deveria ser feita alguma transformação?

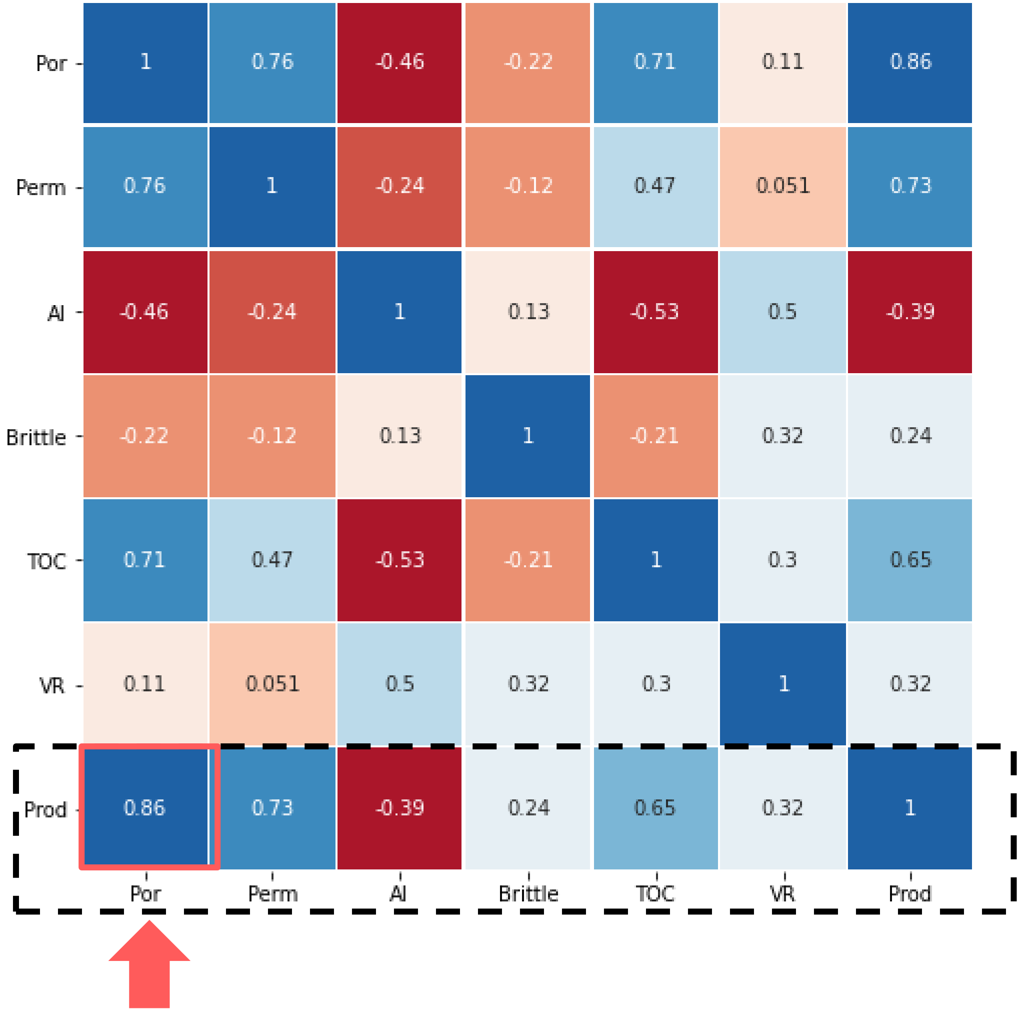
Regressão Linear Simples

X y

Por	Perm	AI	Brittle	TOC	VR	Prod
12.08	2.92	2.80	81.40	1.16	2.31	4165.196191
12.38	3.53	3.22	46.17	0.89	1.88	3561.146205
14.02	2.59	4.01	72.80	0.89	2.72	4284.348574
17.67	6.75	2.63	39.81	1.08	1.88	5098.680869
17.52	4.57	3.18	10.94	1.51	1.90	3406.132832
14.53	4.81	2.69	53.60	0.94	1.67	4395.763259
13.49	3.60	2.93	63.71	0.80	1.85	4104.400989
11.58	3.03	3.25	53.00	0.69	1.93	3496.742701

Maior correlação com Prod: Por

Matriz de Correlação



Regressão Linear Simples

	X	y
	Por	Prod
0	12.08	4165.196191
1	12.38	3561.146205
2	14.02	4284.348574
3	17.67	5098.680869
4	17.52	3406.132832
5	14.53	4395.763259
6	13.49	4104.400989
7	11.58	3496.742701
8	12.52	4025.851153
9	13.25	4285.026122

- Ajuste do Modelo de Regressão Linear Simples:

$$\boxed{\mathbf{y} = \boldsymbol{\theta}_0 + \boldsymbol{\theta}_{Por} \boldsymbol{x}_{Por}}$$

- Ajuste da série de Treino:

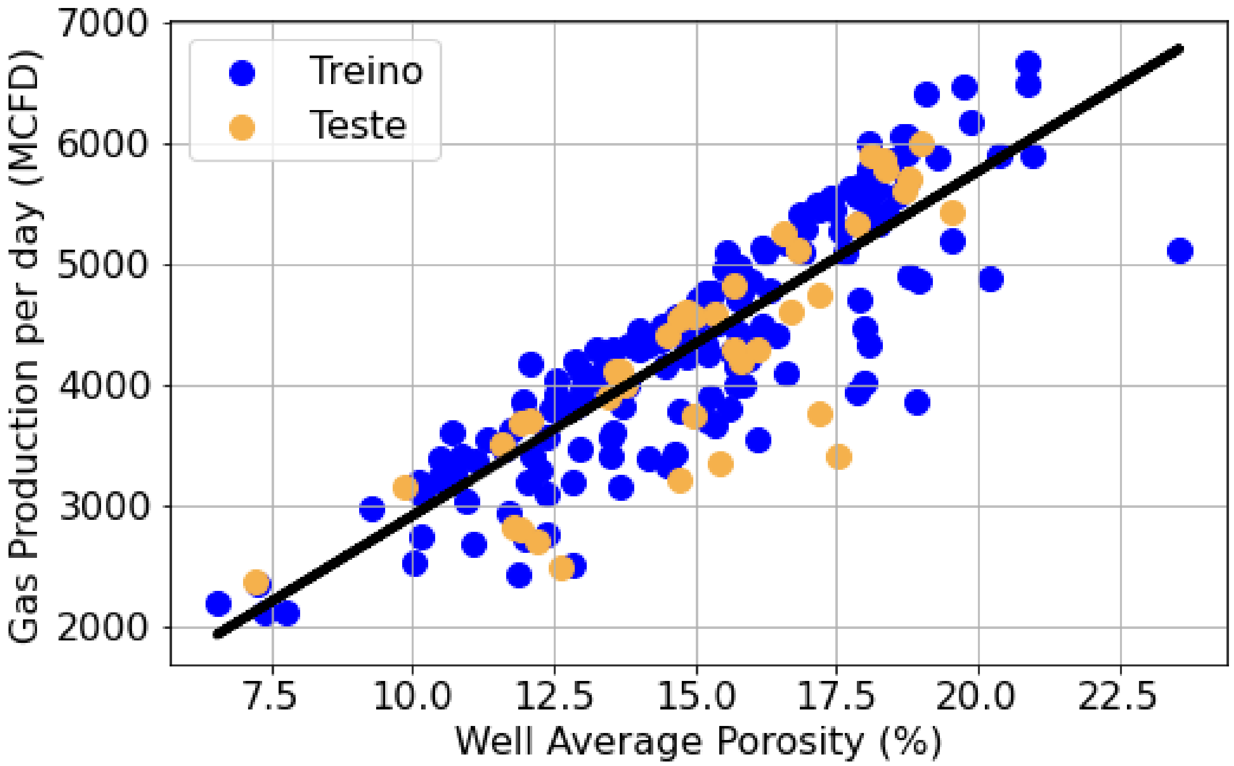
$$\boxed{\mathbf{y} = 60.36 + 284.82\ \boldsymbol{x}_{Por}} \quad \left\{ \begin{array}{l} R^2_{treino} = 0.758 \\ RMSE = 486.64 \end{array} \right.$$

- Avaliação na Série de Teste:

$$\boxed{\boldsymbol{R^2_{test}} = \mathbf{0.678}} \quad \boxed{\boldsymbol{EQM} = \mathbf{313735}} \quad \boxed{\boldsymbol{RMSE} = \mathbf{560}}$$

Regressão Linear Simples

	Por	Prod
0	12.08	4165.196191
1	12.38	3561.146205
2	14.02	4284.348574
3	17.67	5098.680869
4	17.52	3406.132832
5	14.53	4395.763259
6	13.49	4104.400989
7	11.58	3496.742701
8	12.52	4025.851153
9	13.25	4285.026122



$R^2_{test} = 0.678$

Regressão Linear Múltipla

		X				y
Por	Perm	AI	Brittle	TOC	VR	Prod
12.08	2.92	2.80	81.40	1.16	2.31	4165.196191
12.38	3.53	3.22	46.17	0.89	1.88	3561.146205
14.02	2.59	4.01	72.80	0.89	2.72	4284.348574
17.67	6.75	2.63	39.81	1.08	1.88	5098.680869
17.52	4.57	3.18	10.94	1.51	1.90	3406.132832
14.53	4.81	2.69	53.60	0.94	1.67	4395.763259
13.49	3.60	2.93	63.71	0.80	1.85	4104.400989
11.58	3.03	3.25	53.00	0.69	1.93	3496.742701

- Ajuste do Modelo de Regressão Linear Múltipla:

$$\boxed{\mathbf{y} = \boldsymbol{\theta_0 + \theta_{Por} x_{Por} + \theta_{Perm} x_{Perm} + \theta_{AI} x_{AI} + \theta_B x_B + \theta_{TOC} x_{TOC} + \theta_{VR} x_{VR}}}$$

- Ajuste da série de Treino:

$$\boxed{\mathbf{y} = \boldsymbol{-1431 + 235.\, x_{Por} + 108.\, x_{Perm} - 285.\, x_{AI} + 26 x_B + 14.\, x_{TOC} + 685.\, x_{VR}}}$$

$R^2_{treino} = 0.960$ $\quad$ $RMSE = 197$

- Avaliação na Série de Teste:

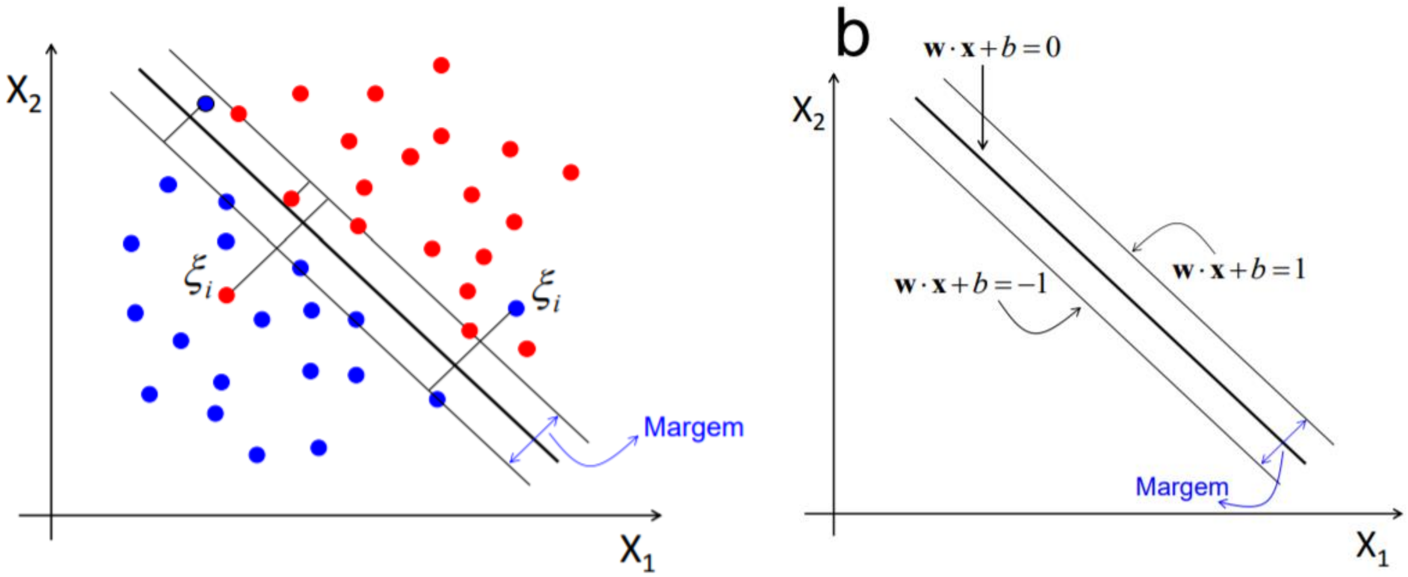
$\boxed{\boldsymbol{R^2_{test} = 0.955}}$ $\quad$ $\boxed{\boldsymbol{EQM = 43732}}$ $\quad$ $\boxed{\boldsymbol{RMSE = 209}}$

SVR – Support Vector Regression

- Recapitulando SVM (Support Vector Machine)...

Objetivo:

- Encontrar um hiperplano que separe as classes sem erros: $$\boldsymbol{f(x) = w.x + b}$$



Filgueiras (2014)

Equação:

$$Minimizar\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i)$$

Restrições:

$$wx_i + b \geq +1 - \xi_i$$

$$w.x + b \leq -1 + \xi_i$$

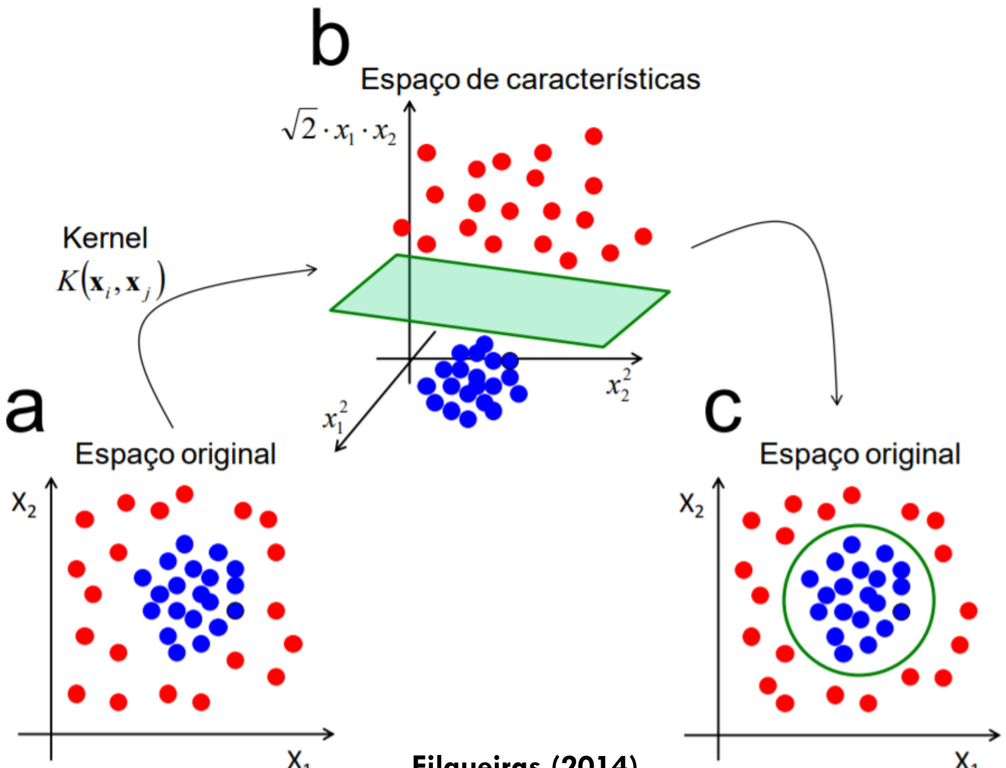
$$\xi_i \geq 0$$

SVR – Support Vector Regression

- Recapitulando SVM...

Se for não linear:

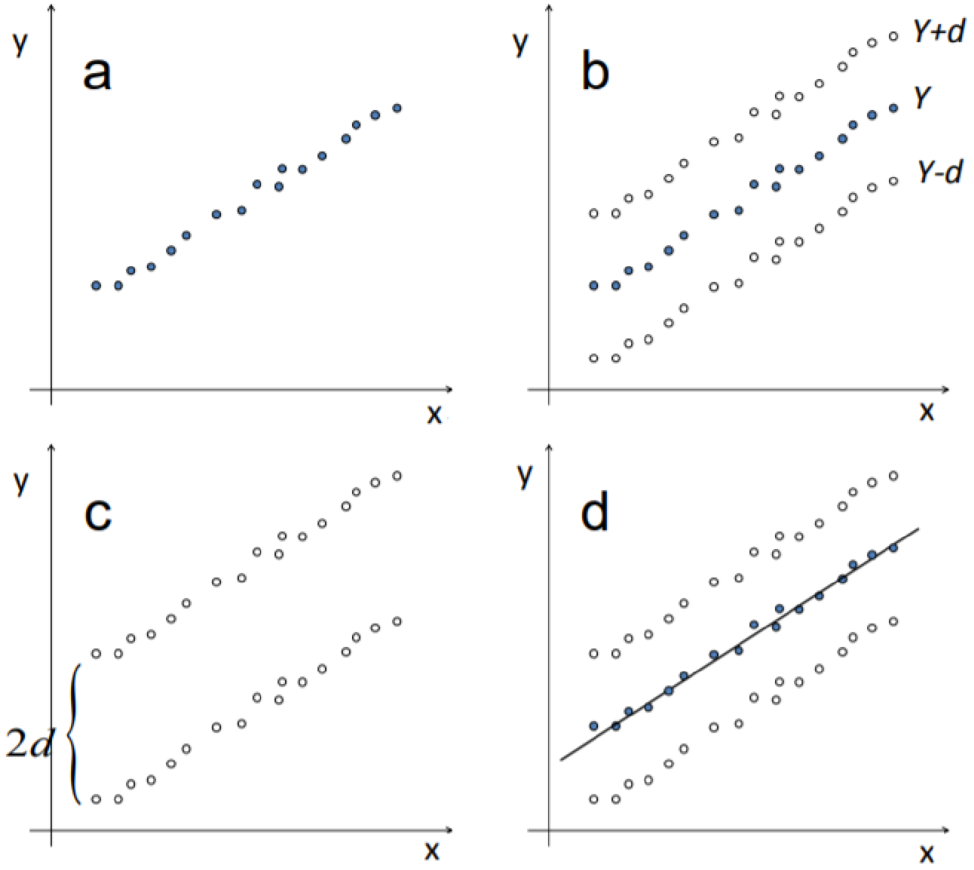
- Utilizar funções de núcleo que aumentam a dimensionalidade dos dados para eles se tornarem linearmente separáveis.



Filgueiras (2014)

SVR – Support Vector Regression

- E o SVR?
  - Problemas de regressão podem ser resolvidos pelo método de classificação binária.
  - Para cada amostra x_i da regressão, um número positivo d é adicionado e subtraído do correspondente valor y_i.



Filgueiras (2014)

SVR – Support Vector Regression

- Equação:

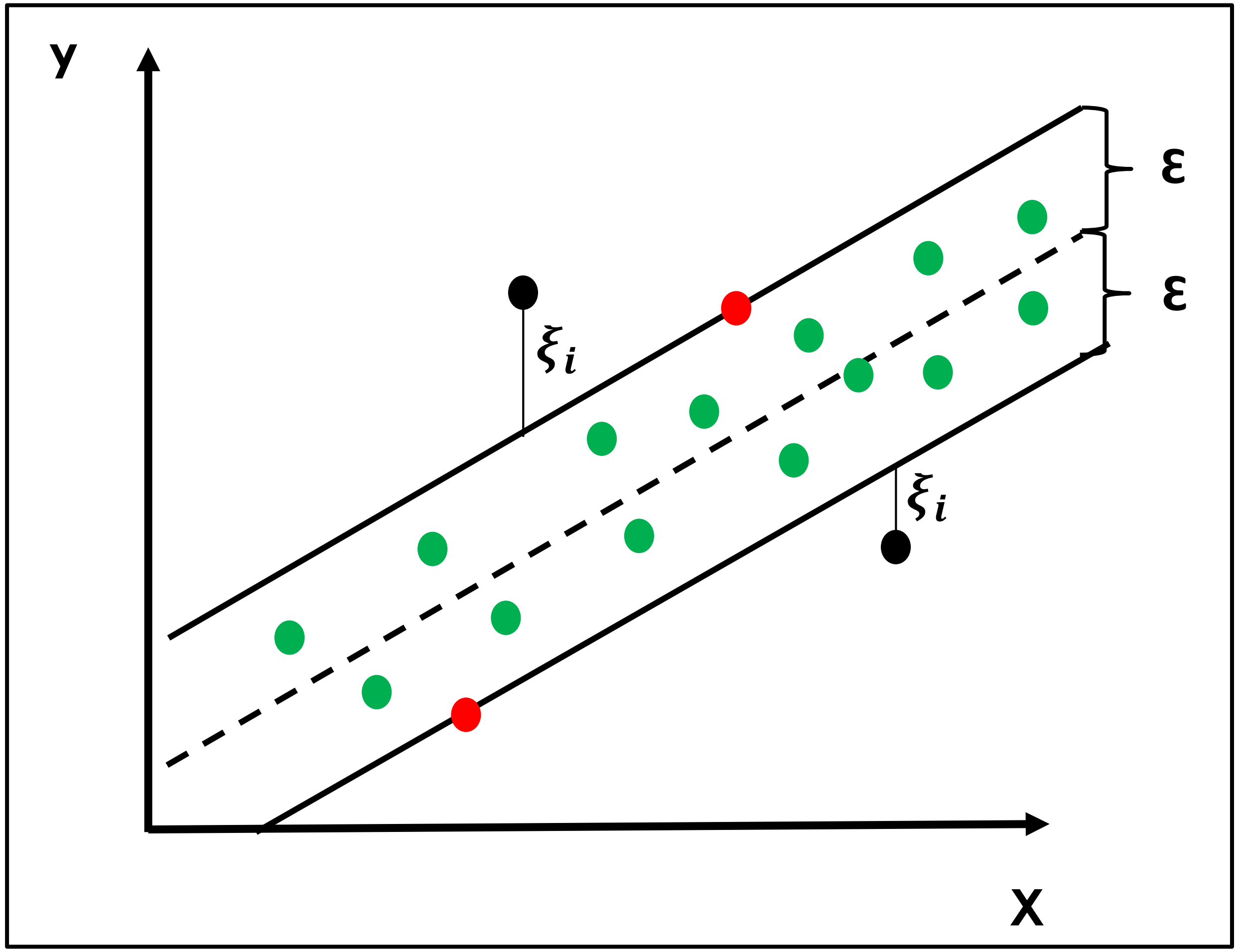
$$Minimizar\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

Restrições:

$$y_i - wx_i - b \leq \epsilon + \xi_i$$

$$wx_i + b - y_i \leq \epsilon + \xi_i^*$$

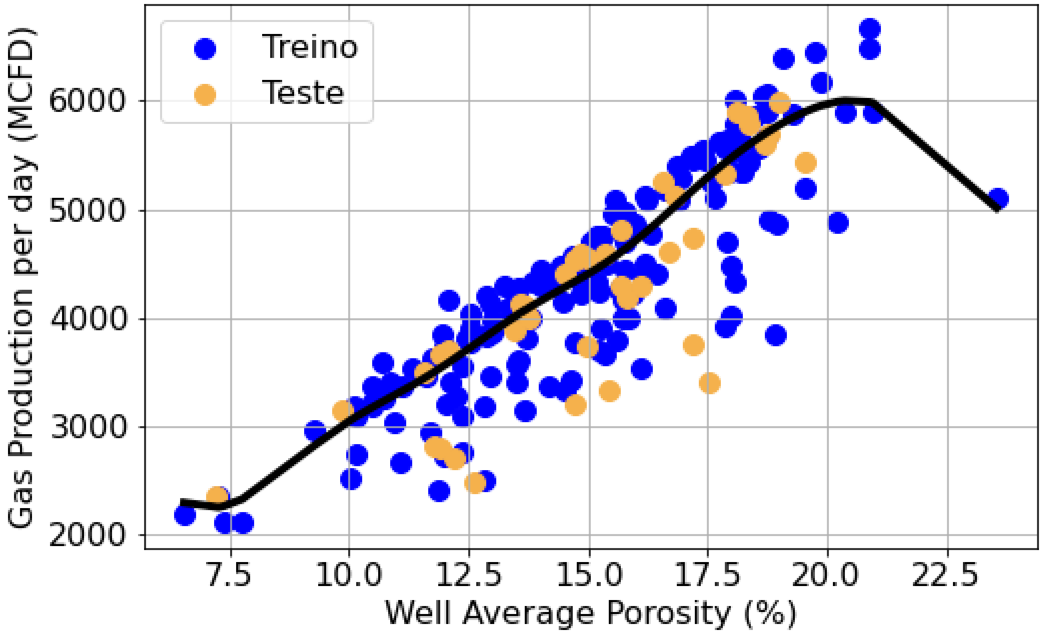
$$\xi_i, \xi_i^* \geq 0$$



y = wx + b

SVR – Support Vector Regression

	X Por	y Prod
0	12.08	4165.196191
1	12.38	3561.146205
2	14.02	4284.348574
3	17.67	5098.680869
4	17.52	3406.132832
5	14.53	4395.763259
6	13.49	4104.400989
7	11.58	3496.742701
8	12.52	4025.851153
9	13.25	4285.026122



Base Radial e parâmetros *default*

- Ajuste da série de Treino:

$$R^2_{treino} = 0.760$$

$$RMSE = 993$$

- Avaliação na Série de Teste:

$$\boldsymbol{R^2_{test} = 0.644}$$

$$\boldsymbol{RMSE = 589}$$

$$\boldsymbol{EQM = 347447}$$

Referências Bibliográficas

- Evsukoff, A G. **INTELIGÊNCIA COMPUTACIONAL Fundamentos e aplicações**. 2020.
- Filgueiras, P. R. **REGRESSÃO POR VETORES DE SUPORTE APLICADO NA DETERMINAÇÃO DE PROPRIEDADES FÍSICO-QUÍMICAS DE PETRÓLEO E BIOCOMBUSTÍVEIS**. Tese. 2014.
- Grus, J. **Data Science from Scratch. First Principles with Python**. 2015
- Muller, A and Guido, S. **Introduction to Machine Learning with Python. A guide for Data Scientists**. 2016.
- VanderPlas, J. **Python Data Science Handbook**. 2016.