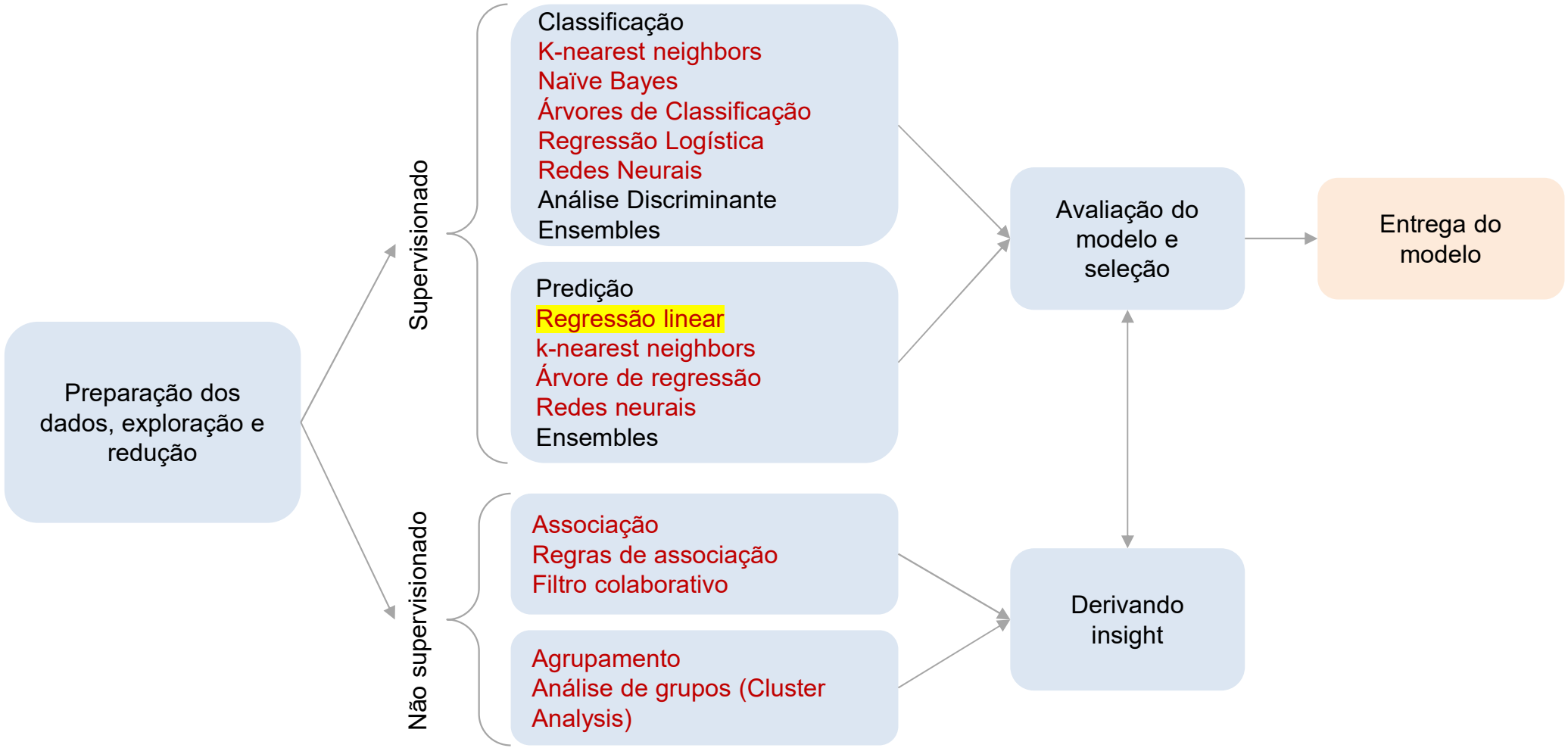
Agenda

UNIDADE 3: Regressão Linear

- Previsões simples (Regressão linear)
- Previsões complexas (Regressão linear múltipla)

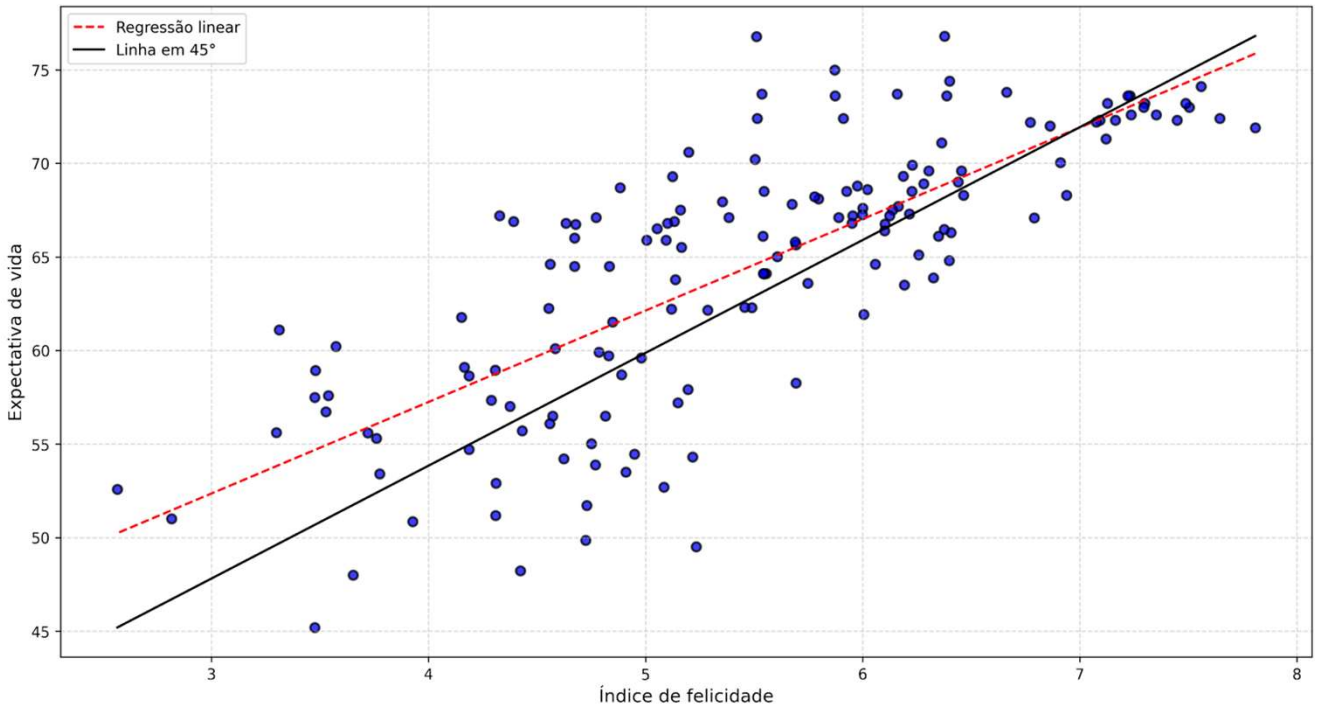


Fonte: Data Mining for Business Analytics Concepts, Techniques and Applications with XLMiner. Wiley, 2016

Relação entre variáveis

- Cenário: determinar a associação entre duas (ou mais) informações:
  - Relação entre **índice de felicidade** e **expectativa de vida**
  - Relação entre **número de processos pendentes** e **quantidade de juízes**

- Variáveis relacionadas são ditas correlacionadas.
- Exemplo de variáveis que aparentemente não são relacionadas:
  - Relação entre a **altura de uma criança e a de seus pais**
  - Relação entre **número de absolvições e a cidade de nascimento dos réus**



Fonte: https://dataat.github.io/introducao-ao-machine-learning/regressão.html

É preciso estabelecer uma relação matemática para entender como uma variável influência outra.

Para que serve descobrir a relação entre variáveis?

Regressão x Classificação

- Regresão:
  - Verificação se duas ou mais variáveis estão relacionadas, como se influenciam.
  - A regressão linear produz uma previsão numérica, como base em valores conhecidos.

- Exemplo de **regressão:**
  - Predição de Montantes de Danos em Casos de Indenização
  - Estimativa de Custos Processuais
  - Projeção de Casos de Litígios Trabalhistas

- Exemplo de **classificação**
  - Classificação de Tipos de Casos Jurídicos (Criminal, Civil, Trabalhista, Familiar, etc.)
  - Classificação de Risco de Reincidência Criminal (Alto risco, médio risco, baixo risco)
  - Classificação de Documentos Jurídicos (contratos, petições, sentenças, etc.)

Umidade relativa média do ar

Atributo	Valor
Temperatura	26,4ºC
Chuva acumulada	0 mm
Velocidade do vento	2,5 m/s
Radiação solar	2064 kj/m^2
Sensação térmica	24,7ºC
Umidade	**54%**

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots$$

Predição de Cargas de Trabalho para Juízes

Descrição: Prever o volume de trabalho em diferentes tribunais com base em fatores como tamanho da população, tipos de casos comuns na região, etc.

- Tamanho da população da região onde o tribunal está localizado (valores real entre 30,000 e 500,000)
- Número de juízes ativos no tribunal (valores inteiros entre 1 e 20)
- Número de casos recebidos mensalmente (valores inteiros entre 150 e 800)
- Distribuição dos tipos de casos comuns na região (valores: criminal, civil, trabalhista)
- Taxa de crescimento da população na região (valores reais entre 0.0 e 1.0)
- Taxa de criminalidade na região (valores reais entre 0.0 e 1.0)
- Média de tempo gasto em cada tipo de caso em meses (número inteiro)
- Nível de automação e eficiência do tribunal (valores inteiros entre 0 e 10)
- Nível de congestionamento do sistema judicial na região (baixo, médio, alto)
- Número de advogados atuando na região (valores inteiros entre 1 e 100)
- Número de prédios judiciais na região (valores inteiros entre 1 e 10)
- Nível de urbanização da região (baixo, médio, alto)
- Nível de educação da população na região (valores reais entre 0.0 e 1.0)
- Número de habitantes por juiz (valor fruto da divisão de tamanho da população da região pelo número de juízes ativos no tribunal)
- Índice de Desenvolvimento Humano (valores reais entre 0.0 e 1.0)
- Média de idade dos es (entre 35 e 60)
- **Número de casos pendentes atualmente no tribunal (valores inteiros, coluna que não pode ser nula)**

Fonte: **Provost,Foster e Fawcett,Tom.** Data Science Para Negócios. Alta Books, 2016.

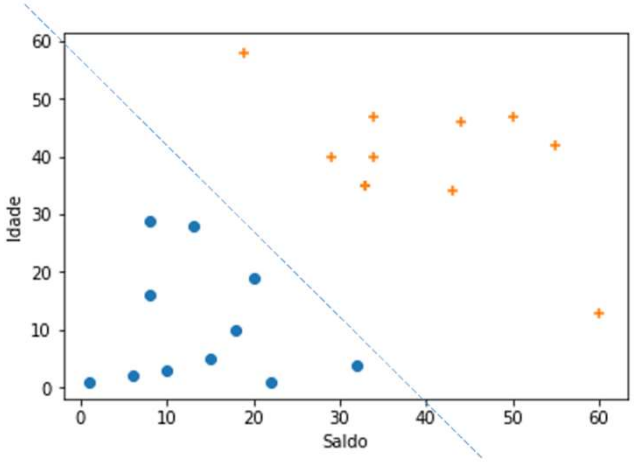
Modelo linear geral

Ponderação

Característica

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots$$

- Vetor de características $x = (x_1, x_2, \ldots x_n)$
- As ponderações da função linear (w_i) são os parâmetros
- A ponderação w_0 é a intersecção



+ Se 60 - 1.0 x Idade - 1.5 x Saldo ≤ 0

● Se 60 - 1.0 x Idade - 1.5 x Saldo > 0

Ajustar o modelo significa encontrar um bom conjunto de ponderações das características.

Quanto maior a magnitude da ponderação mais importante é a característica para a classificação.

Variáveis Dependentes e Independentes

- Variável dependente é o valor que estamos prevendo
- Variável independente é a variável que estamos usando para prever uma variável dependente.

$$y = a + bx$$

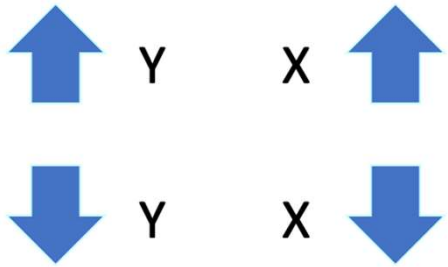
- **Uma variável independente X, explica a variação em outra variável, que é chamada variável dependente Y.**

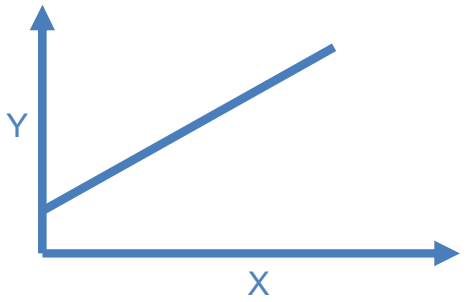
- Esse relacionamento existe em apenas uma direção

  Variável independente (x) → variável dependente (y)

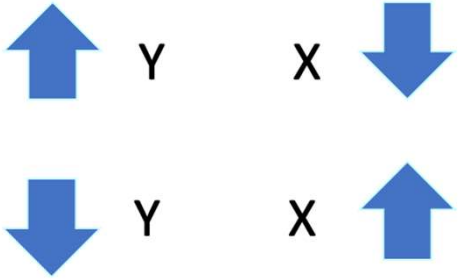
Relação entre variáveis Dependentes e Independentes

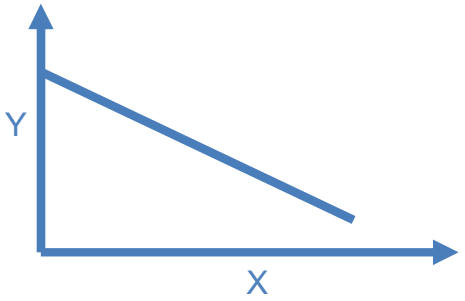
Direta (ou positiva) quando os valores de Y aumentam em decorrência do aumento dos valores de X .





Inversa (ou negativa) quando os valores de Y variam inversamente em relação aos de X.





Correlação

- **Correlação**
  - Valor **entre -1 e 1**
  - Quanto mais **próximo de 1 ou -1, mais forte** é a relação
  - Quanto mais **próximo de zero, mais fraca** é a relação
  - O "**valor**" significa a "**força**" da relação
  - O "**sinal**" significa o "**sentido**" da relação
- **1 ou -1 é uma relação perfeita**
- **0 é uma relação inexistente**

Correlação x Causalidade

- **Correlação**
  - Valor **entre -1 e 1**
  - Quanto mais **próximo de 1 ou -1, mais forte** é a relação
  - Quanto mais **próximo de zero, mais fraca** é a relação
  - O "**valor**" significa a "**força**" da relação
  - O "**sinal**" significa o "**sentido**" da relação
- **1 ou -1 é uma relação perfeita**
- **0 é uma relação inexistente**

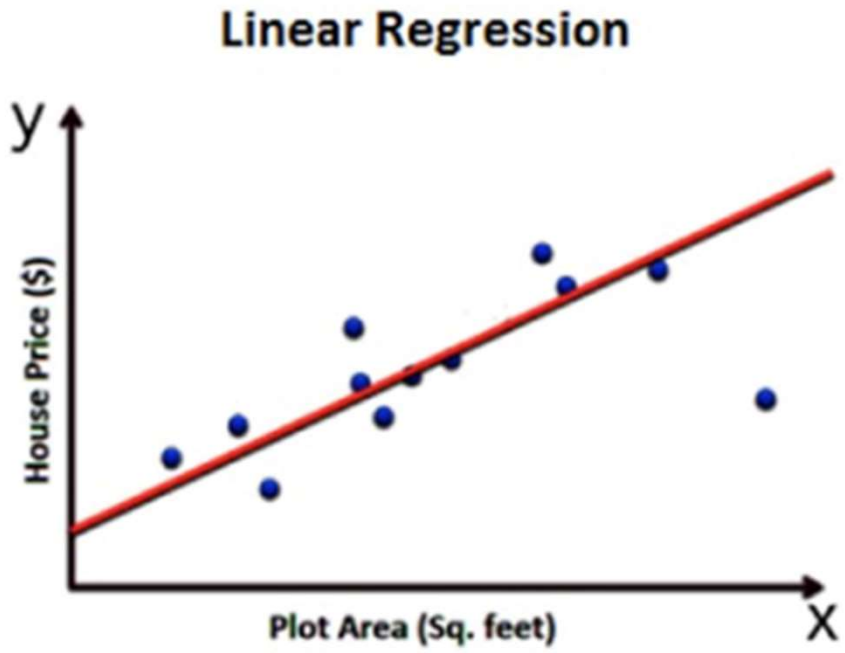
Só porque existe uma correlação entre duas variáveis, isso não significa que exista uma relação causal entre elas.

O fato de as pessoas usarem guarda-chuvas quando chove não significa que os guarda-chuvas façam a chuva cair.

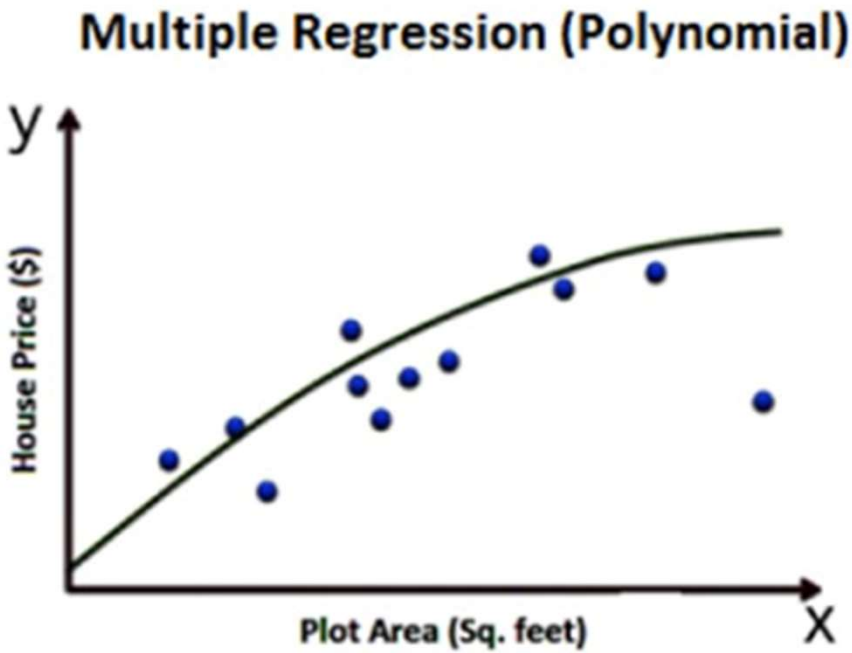
A correlação nos mostra a força de relacionamento entre duas variáveis mas não mostra "como".
Se o relacionamento é forte (± 1) ou não (0)

Uma **análise de regressão** nos permite **começar a ver como**.

Modelos de Regressão



$$y = \mathrm{a} + bx$$



$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots$$

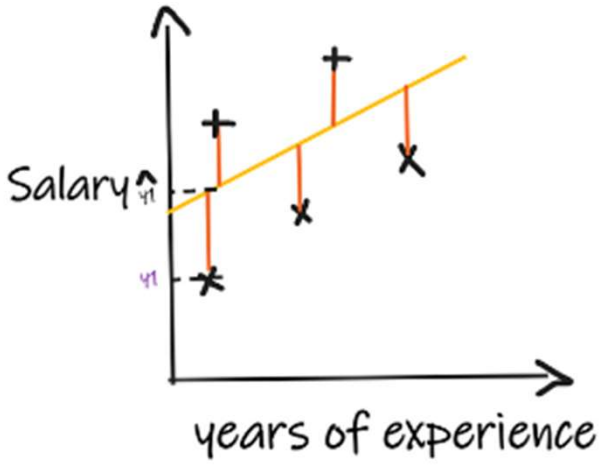

TEORIA + PRÁTICA

Como avaliar um modelo

- Como avaliar a qualidade do modelo?
  - Os erros têm distribuição normal?
  - Existem “outliers” no conjunto de dados?
  - O modelo gerado é adequado?

Métricas de avaliação: regressão

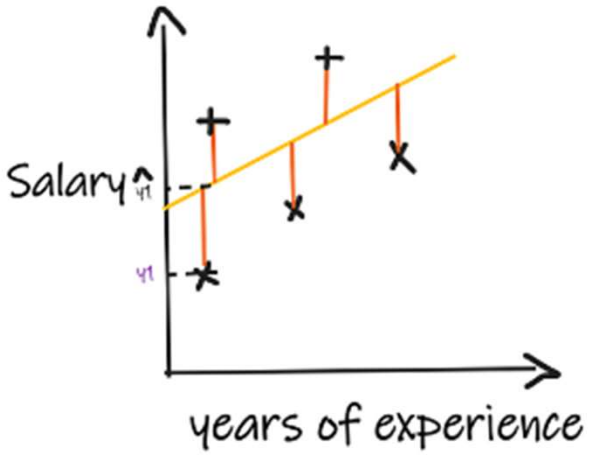
- **Mean Absolute Error** (MAE): Erro Absoluto Médio é a **média do valor absoluto dos erros (diferenças absolutas entre os valores preditos e os valores observados)**.



$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

Métricas de avaliação: regressão

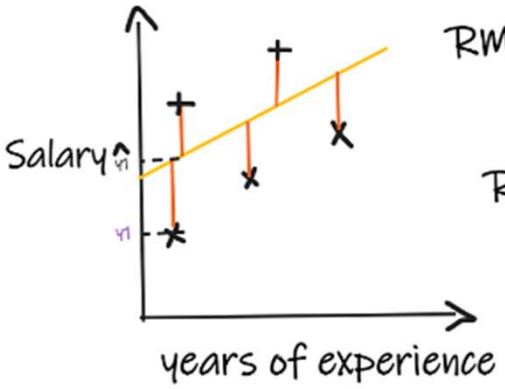
- **Mean Squared Error** (MSE): Erro Médio Quadrático é a **média dos erros quadrados, pune erros maiores**. MSE é preferido quando se deseja **penalizar mais fortemente os erros maiores**, o que é comum em modelos onde grandes desvios são particularmente indesejáveis.



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

Métricas de avaliação: regressão

- **Root Mean Square Error** (RMSE): Raiz do Erro Quadrático Médio é a **raiz quadrada da média dos erros quadrados**. RMSE é frequentemente usado em contextos onde é importante **manter as unidades do erro comparáveis com os dados originais e ao mesmo tempo penalizar erros maiores**. O RMSE, em particular, **nos dá uma ideia do erro médio em relação às unidades dos dados originais**.

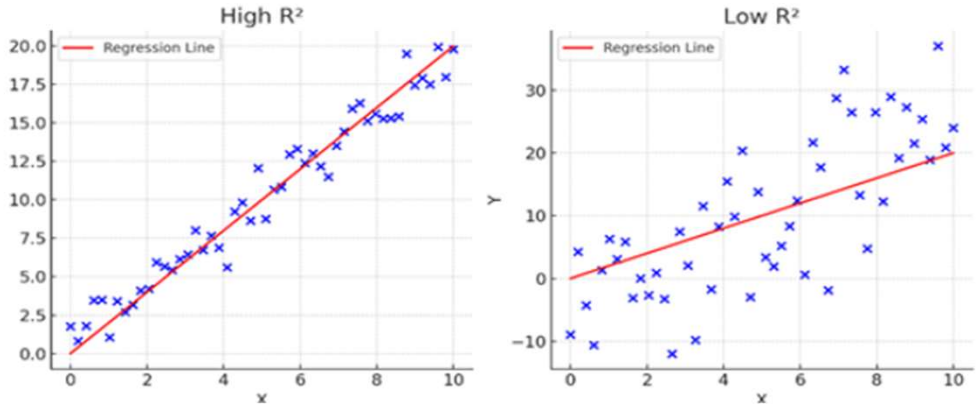


$$RMSE = \sqrt{MSE}$$

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

Métricas de avaliação: regressão

- **Coeficiente de Determinação** (R^2): Ele é uma medida de **quão bem os valores preditos se ajustam aos valores reais**. O R^2 varia de 0 a 1. Um valor de R^2 de 1 indica que o modelo explica 100% da variância nos dados; um valor de 0.75 que explica 75% dos dados, um valor de 0 indica que o modelo não explica nenhuma variância em relação aos valores reais.



Sinais de Overfitting

- Treino: BAIXO erro (MAE, MSE, RMSE) e ALTO R^2
- Teste: ALTO erro (MAE, MSE, RMSE) e BAIXO R^2

Sinais de Underfitting

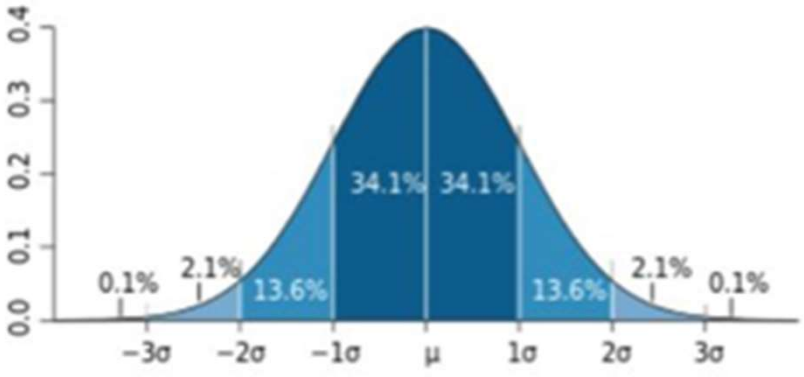
- Treino: ALTO erro (MAE, MSE, RMSE) e BAIXO R^2
- Teste: ALTO erro (MAE, MSE, RMSE) e BAIXO R^2

Diagnóstico de Resíduos (erro)

- Analise a distribuição dos resíduos (**diferença entre valores previstos e reais**) $e_i = y_i - \hat{y}_i$
  - **Overfitting**: PODE apresentar **resíduos sistematicamente distribuídos em uma direção específica ou grandes resíduos em dados de teste**.
  - **Underfitting**: **Resíduos altos e de distribuição uniforme** PODE indicar que o modelo não está capturando bem a variação nos dados.

Quando o erro segue a distribuição normal, podemos garantir que nossos parâmetros beta estimados com mínimos quadrados são iguais a uma estimativa gerada via máxima verossimilhança.

A recíproca não é verdadeira, ou seja, a não normalidade dos resíduos não nos permite afirmar que, com certeza, nossos parâmetros estão errados



https://medium.com/turing-talks/diagnosticando-a-sua-regress%C3%A3o-linear-4c35fcdf3b9d

O histograma de resíduos deve ser semelhante a uma normal

- Se os erros tiverem uma distribuição normal,
- Aproximadamente, 95% dos resíduos estarão no intervalo de um desvio padrão da média
- Caso contrário, deve existir a presença de “outlier”


TEORIA + PRÁTICA

Alguns outros modelos de regressão linear

Lasso

A regressão lasso adiciona uma penalização do tipo L1 ao termo de erro da regressão linear, que é a soma das magnitudes dos coeficientes (elimina variáveis irrelevantes)

$$\mathrm{RSS}_{\mathrm{lasso}} = \sum_{i=1}^{n} \left[y_i - (\mathbf{w} \cdot \mathbf{x}_i + b) \right]^2, \boxed{+\alpha \sum_{j=1}^{p} |w_j|}$$

regularização ℓ_1

Soma dos quadrados dos resíduos + penalidade * |inclinação|

Ridge

A regressão ridge adiciona uma penalização do tipo L2 ao termo de erro da regressão linear, que é o quadrado da magnitude dos coeficientes.

$$\mathrm{RSS}_{\mathrm{ridge}} = \sum_{i=1}^{n} \left[y_i - (\mathbf{w} \cdot \mathbf{x}_i + b) \right]^2 \boxed{+\alpha \sum_{j=1}^{p} w_j^2}$$

regularização ℓ_2

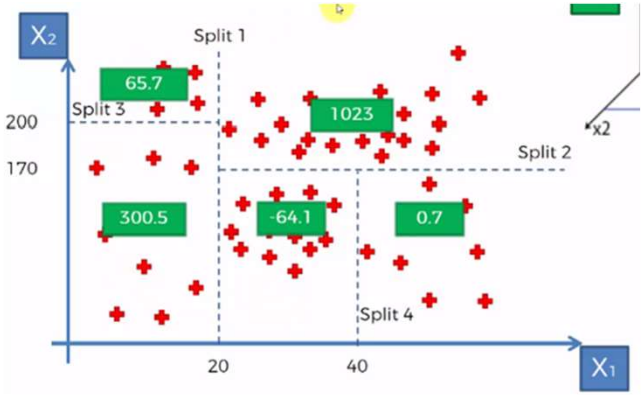
Soma dos quadrados dos resíduos + penalidade * (inclinação)²

Elastic-Net Regression

Combinação das técnicas de regularização L1 (Lasso) e L2 (Ridge).

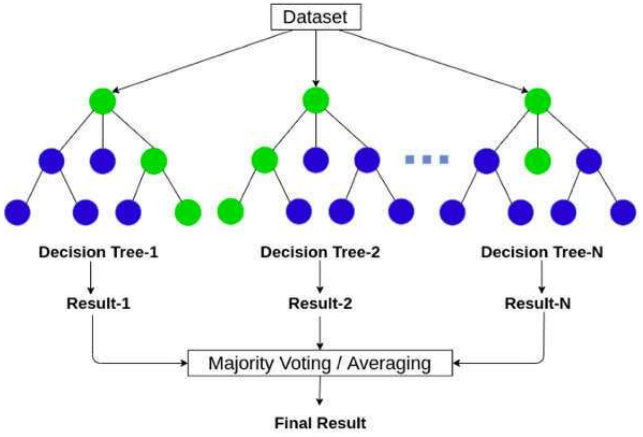
DecisionTreeRegressor

O critério para dividir um nó pode ser baseado em diferentes métricas de erro, como o mean squared error (MSE) ou mean absolute error (MAE)



RandomForestRegressor

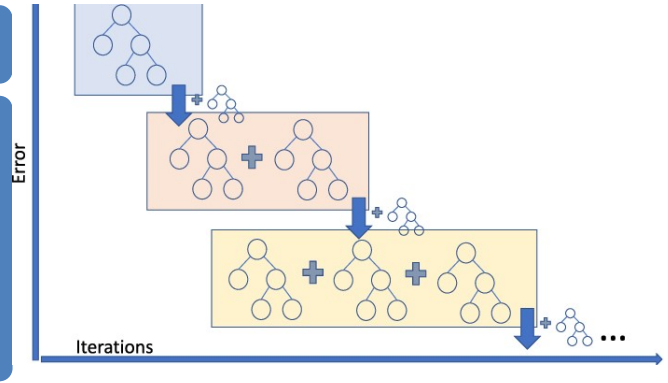
Baseado em florestas aleatórias que são métodos de ensemble que combina múltiplas árvores de decisão fazendo a média das previsões de todas as árvores para produzir uma previsão final.



Alguns outros modelos de regressão linear

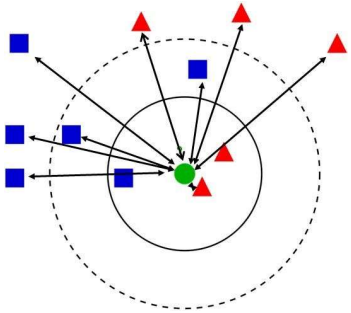
GradientBoostingRegressor

Baseado em gradiente para problemas de regressão, adicionando iterativamente modelos fracos (geralmente árvores de decisão) para corrigir os erros dos modelos anteriores.



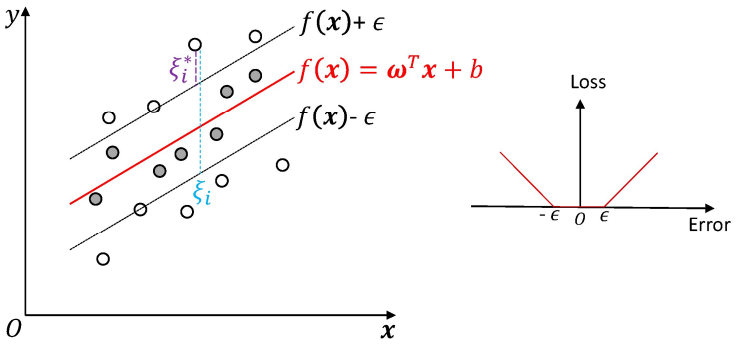
KNeighborsRegressor

Implementação do algoritmo de regressão baseado nos vizinhos mais próximos (k-Nearest Neighbors, k-NN).



Support Vector Regression

Implementação do algoritmo de regressão baseado em SVM, tenta encontrar uma função que tenha no máximo uma margem de erro epsilon para todos os pontos de treinamento.



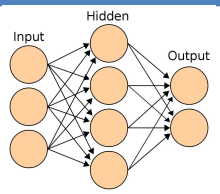
BayesianRidge

Combina princípios de regressão ridge com probabilidade bayesiana para fornecer estimativas dos coeficientes de regressão.

$p(y|\lambda)=N(w|0, \lambda^{-1}Ip)$

MLPRegressor

Implementação de uma rede neural feedforward (perceptron multicamada) para problemas de regressão.




TEORIA + PRÁTICA