Universidade Federal do Pará
Programa de Pós-graduação em Engenharia de Processos (PPGEP)
Disciplina: Tópicos Especiais em Engenharia de Processos
Prof. Dr. Alan de Souza

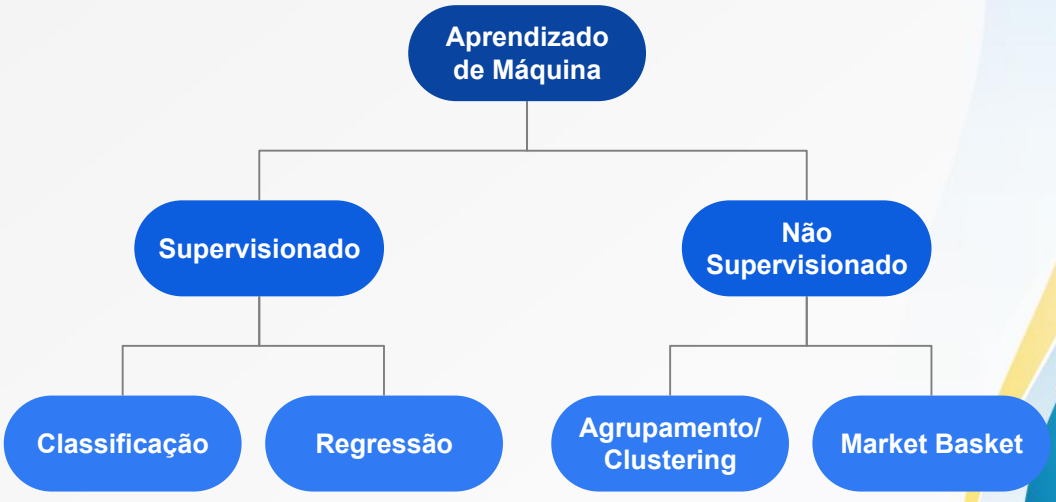
Aula 2

Setembro/2025

Na aula anterior...

1. Apresentação do professor, da disciplina e dos alunos;
2. Dados: importância e possibilidades;
3. Profissionais de dados;
4. Definição de machine learning;
5. Árvore de decisão para classificação;
6. Projetos práticos de classificação de frutas e de cerveja;
7. Exercício com os dados de cerveja.

Tipos de Aprendizado



Problemas de regressão

- São voltados à estimativa alvo (*target*) ou saída (y), sendo este um número/valor. Por exemplo:
  - Quantidade de vendas
  - Receita presumida
  - Valor de crédito
  - Precificação de imóvel
  - Volume de chuva

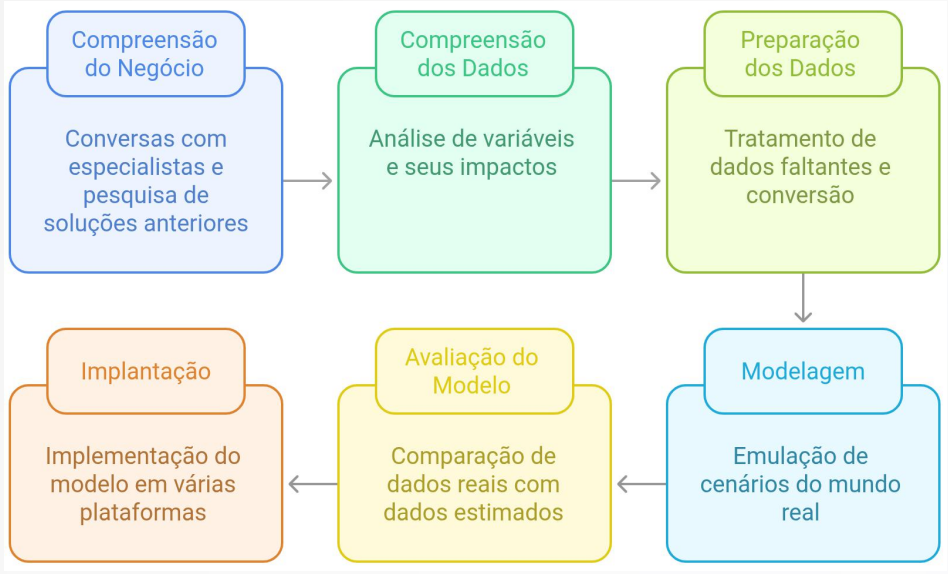
Problemas de classificação

- São voltados à estimativa alvo (*target*) ou saída (y), sendo este um rótulo ou classe, por exemplo:
  - Compradores ou Não compradores
  - Inadimplente ou Adimplente
  - Equipamento operando ou com defeito
  - Vitória, derrota ou empate
  - Reconhecimento facial
  - Satisfação do cliente

Metodologia CRISP-DM

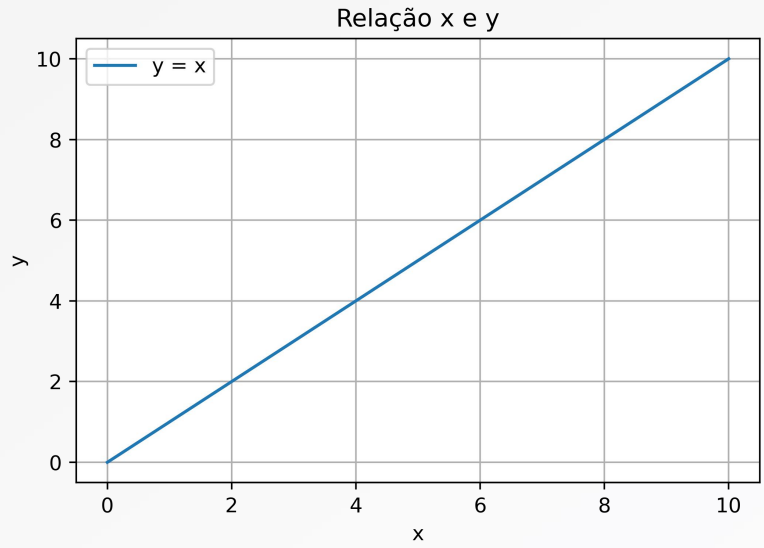
- *CRoss-Industry Standard Process for Data Mining*
- Etapas:
  1. **Compreensão do negócio**: conversas com o especialista, pesquisa de solução anteriores;
  2. **Compreensão dos dados**: o que cada variável representa? como cada variável de entrada impacta na variável de saída (correlação)?
  3. **Preparação dos dados**: dados faltantes, outliers, desbalanceamento dos dados, conversão de dados, dados de treino e teste, normalização de dados;
  4. **Modelagem**: emular a realidade;
  5. **Avaliação do modelo**: comparar dados reais *vs* dados estimados;
  6. **Implantação**: site, aplicativo de celular, programa de computador...

Metodologia CRISP-DM

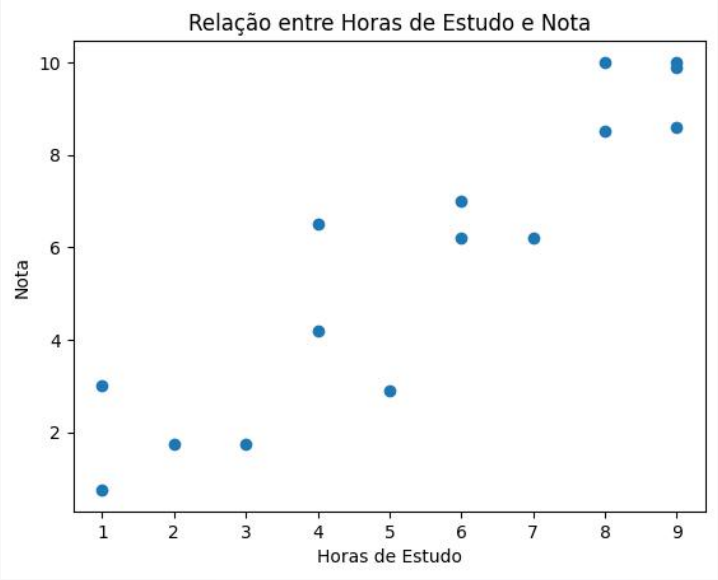


DÚVIDAS?

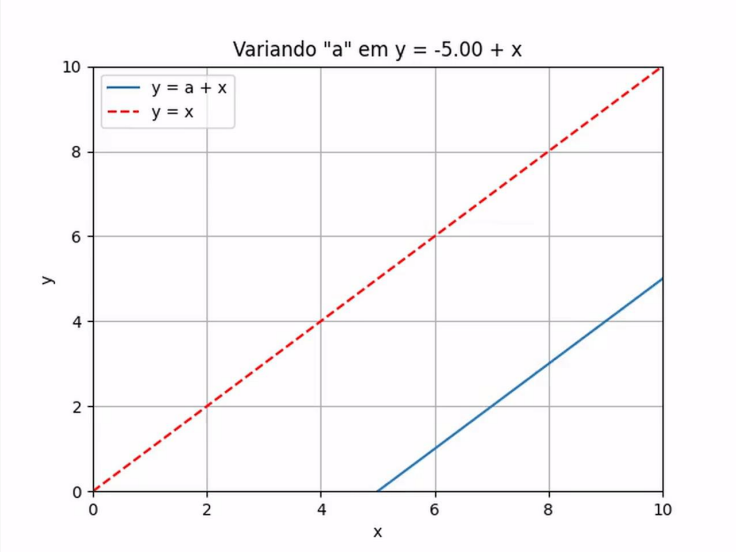
Regressão linear - gráfico genérico



Regressão linear - exemplo



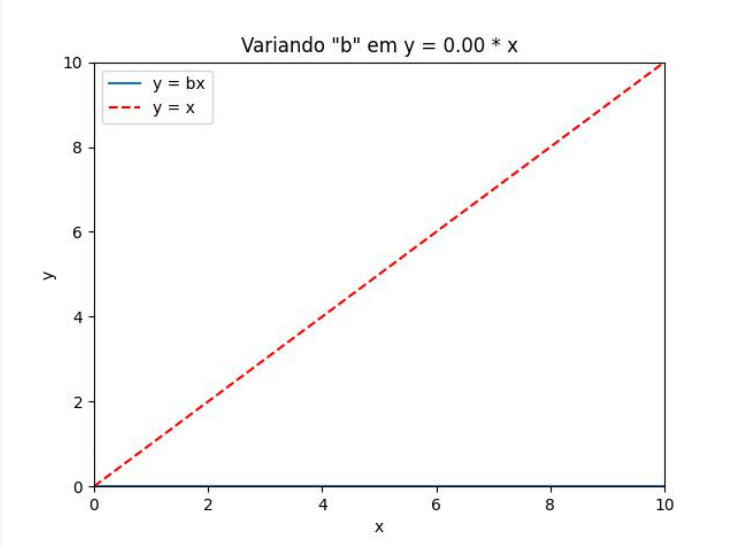
Regressão linear - gráfico animado variando “a”



$$y = a + x$$

Coeficiente linear

Regressão linear - gráfico animado variando “b”

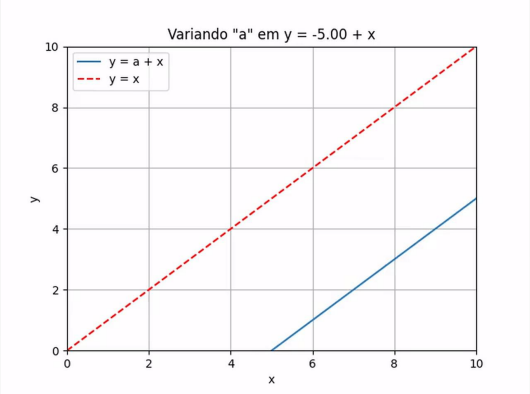


$$y = b.x$$

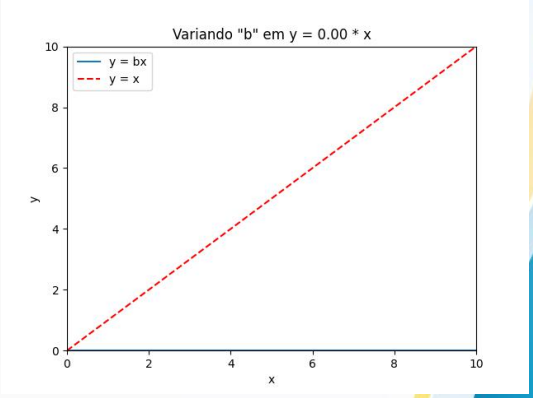
Coeficiente angular

Regressão linear - variando “a” e “b”

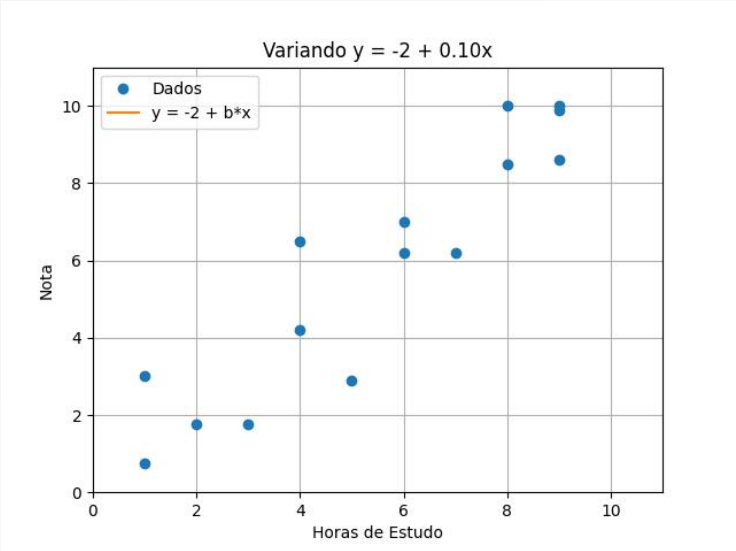
$$y = a + x$$



$$y = b.x$$



Regressão linear - exemplo nota vs horas de estudo



DÚVIDAS?

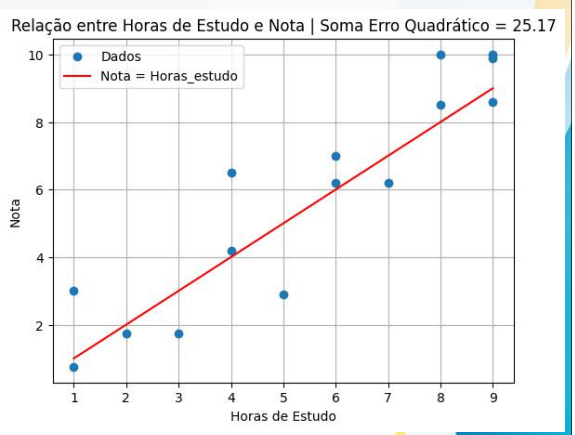
Regressão linear - notação matemática

$$\hat{y} = \hat{a} + \hat{b}x_i \Leftrightarrow \widehat{Nota}_i = \hat{a} + \hat{b}.Horas\_estudo_i$$

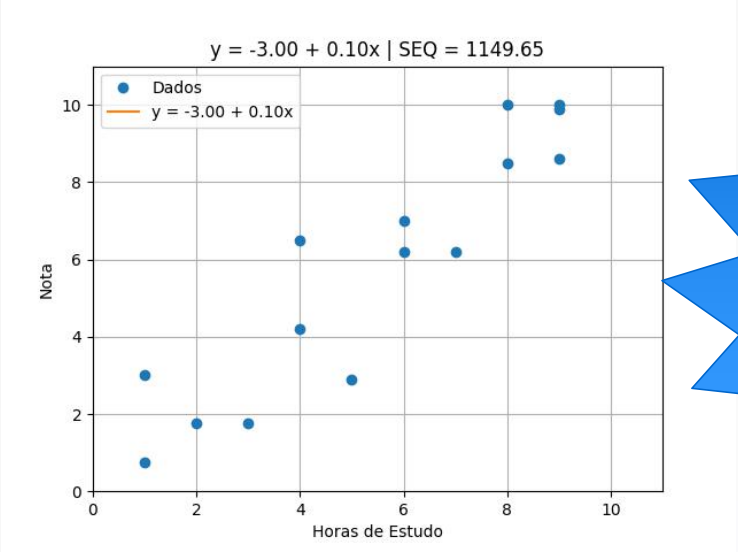
$$\text{Erro} = y_i - \hat{y}_i$$

$$\text{Erro Quadrático} = (y_i - \hat{y}_i)^2$$

$$\text{Soma do Erro Quadrático} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



Regressão linear - notação matemática



Como se calcula os melhores coef. "a^" e "b^" para minimizar o erro quadrático?

Regressão linear - notação matemática

$$\widehat{y} = \hat{a} + \widehat{b}x_i \Leftrightarrow \widehat{Nota}_i = \hat{a} + \widehat{b}.Horas\_estudo_i$$

$$\text{Erro} = y_i - \hat{y}_i$$

$$\text{Erro Quadrático} = (y_i - \hat{y}_i)^2$$

$$\text{Soma do Erro Quadrático} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

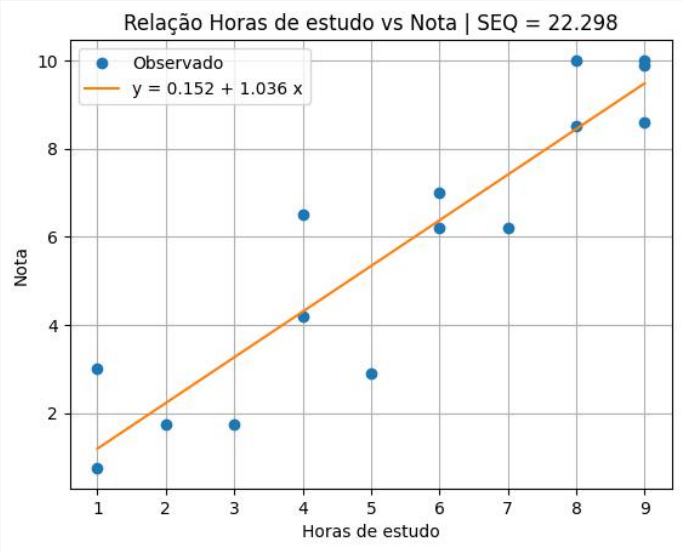
$$\text{Soma do Erro Quadrático} = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

Regressão linear - notação matemática

$$\frac{\partial}{\partial \hat{a}} \sum_{i=1}^{n} (y_i - (\hat{a} + \hat{b}x_i))^2 = 0$$

$$\frac{\partial}{\partial \hat{b}} \sum_{i=1}^{n} (y_i - (\hat{a} + \hat{b}x_i))^2 = 0$$

Regressão linear - melhores “a^”e “b^” para o exemplo



DÚVIDAS?

Regressão linear - mais notação matemática

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Regressão linear - mais notação matemática

- E se for preciso utilizar mais variáveis?

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_p x_{ip}$$

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$$

DÚVIDAS?

Projeto prático - Nota x Horas de estudo

```
# Importação de dados
import pandas as pd
df = pd.read_excel('dados_estudo_nota.xlsx')
df
```

Projeto prático - Nota x Horas de estudo

```
# Separação de dados de entrada e de saída:
X = df[['horas_estudo']] # Isso é uma matriz (dataframe)
y = df['nota']           # Isso é um vetor (series)
```

Projeto prático - Nota x Horas de estudo

```
# Criação e treino do modelo:
from sklearn import linear_model
regressor = linear_model.LinearRegression(fit_intercept=True)
regressor.fit(X, y)
# Cálculo do “a^” e do “b^”:
a, b = regressor.intercept_, regressor.coef_[0]
```

Projeto prático - Nota x Horas de estudo

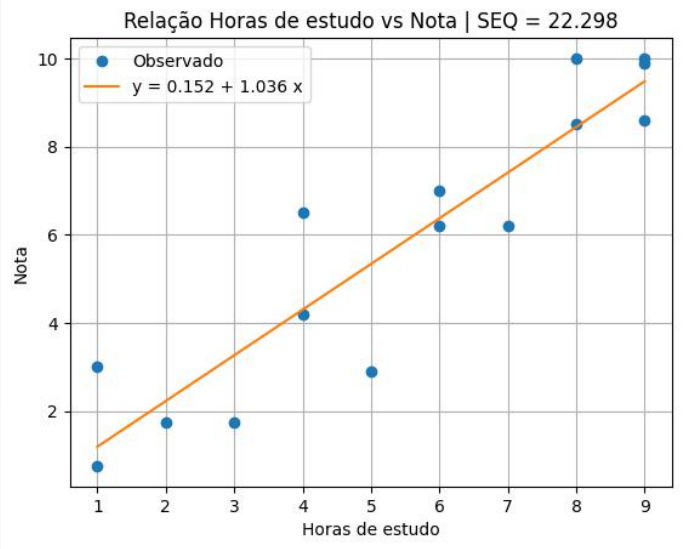
```
# Uso do modelo para fazer previsões:
previsao_regressor = regressor.predict(X)

# Cálculo da soma dos erros quadráticos (SEQ):
seq = np.sum((df['nota'] - previsao_regressor)**2)
```

Projeto prático - Nota x Horas de estudo

```
# Mostrando o gráfico de análise:
import matplotlib.pyplot as plt
plt.plot(X['horas_estudo'], y, 'o')
plt.grid(True)
plt.title(f"Relação Horas de estudo vs Nota | SEQ = {seq:.3f}")
plt.xlabel("Horas de estudo")
plt.ylabel("Nota")
plt.plot(X['horas_estudo'], previsao_regressor)
plt.legend(['Observado',
        f'y = {a:.3f} + {b:.3f} x'
        ])
```

Projeto prático - Nota x Horas de estudo - resultado



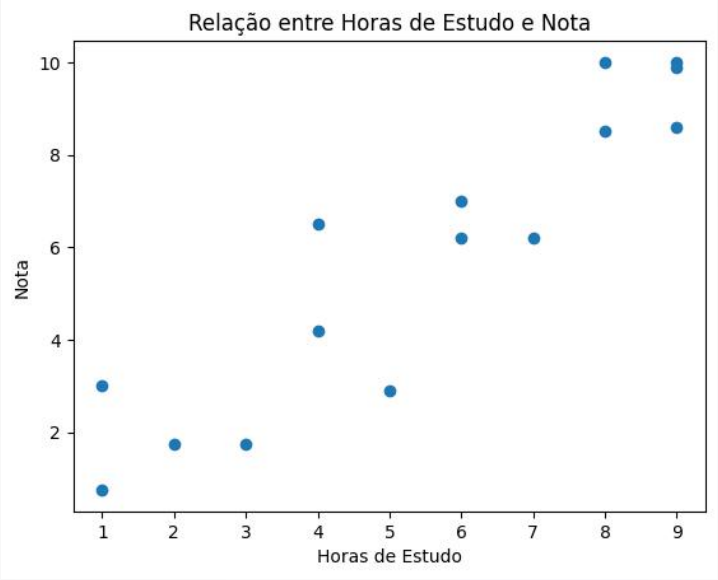
DÚVIDAS?

Exercício:

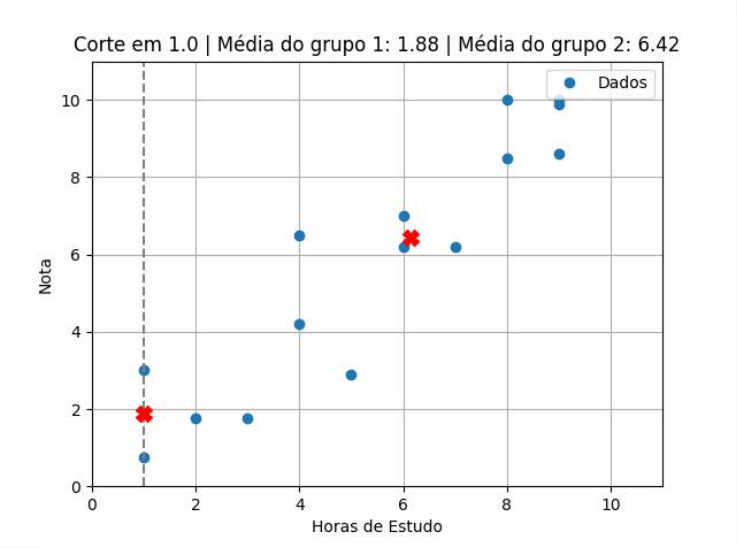
1) Baixe os dados através do link abaixo e crie um modelo, usando a técnica de regressão linear, para estimar o nível de automação de uma fábrica dado a quantidade de operadores que trabalham nela. Construa o gráfico que mostra os pontos observados e a linha de previsão do modelo linear. Quais são as conclusões a respeito da análise?

   Link: https://github.com/amarcel/ML-PPGEP-set2025

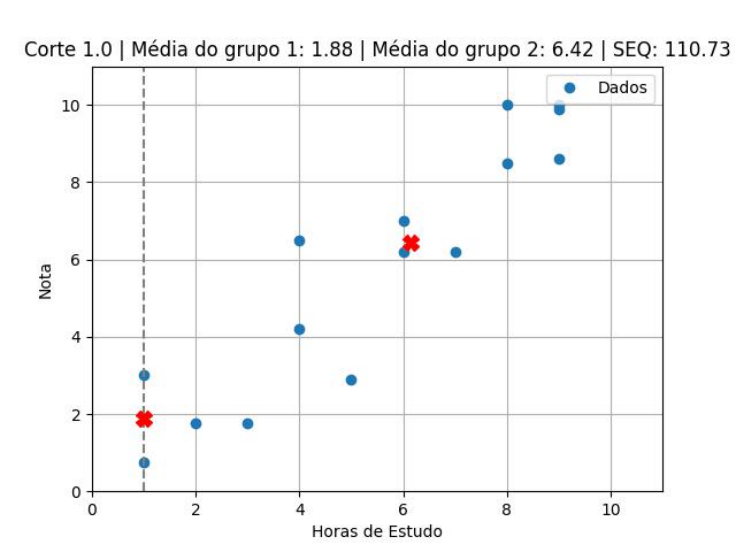
Árvore de decisão como estimadora (regressão)



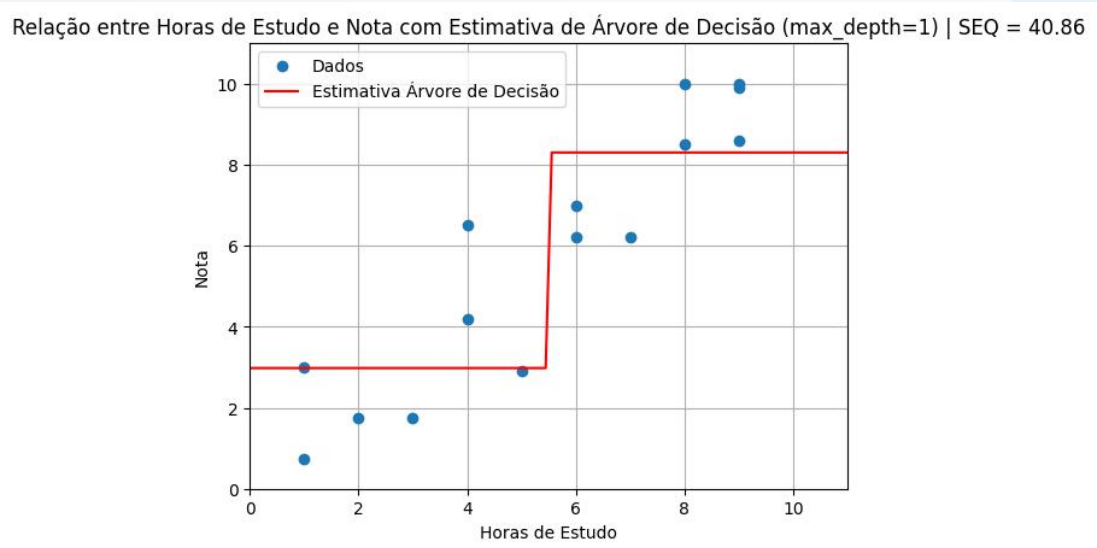
Árvore de decisão como estimadora (regressão)



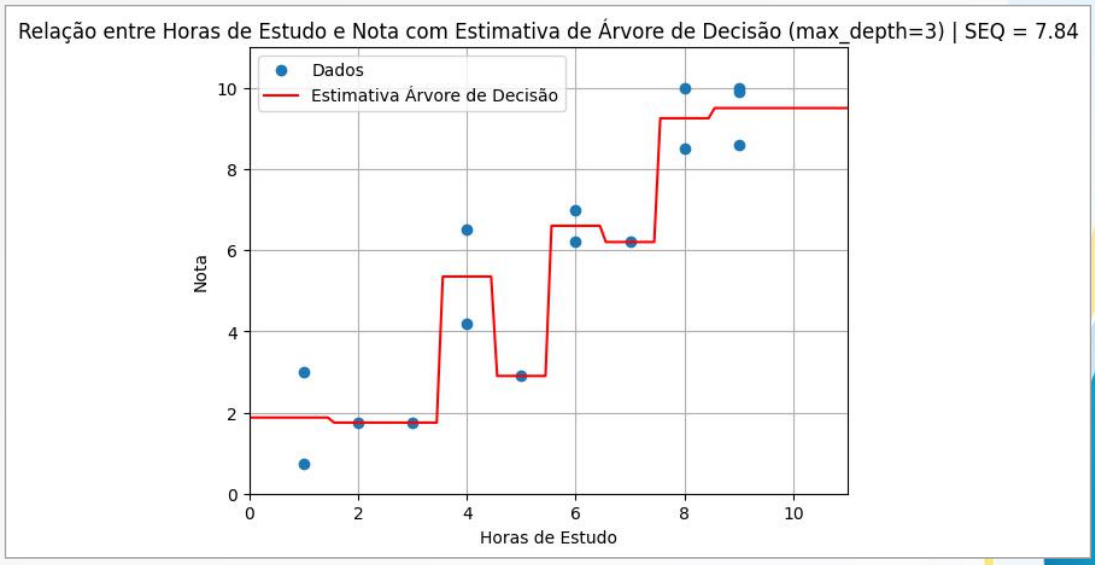
Árvore de decisão como estimadora (Erro)



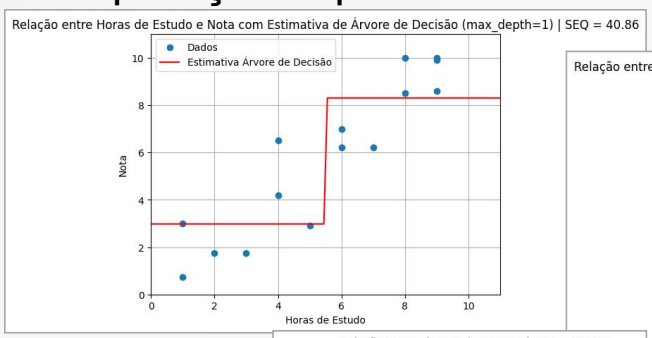
Árvore de decisão como estimadora (1 nível)

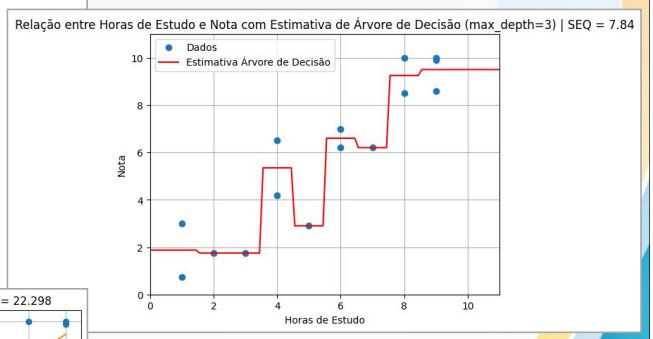


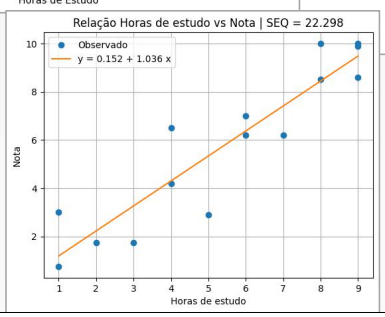
Árvore de decisão como estimadora (3 níveis)



Comparação - qual é o melhor? por quê?







DÚVIDAS?

Projeto prático - Nota x Horas de estudo

```
# Importação de dados
import pandas as pd
df = pd.read_excel('dados_estudo_nota.xlsx')
df
```

Projeto prático - Nota x Horas de estudo

```
# Separação de dados de entrada e de saída:
from sklearn.tree import DecisionTreeRegressor
X = df[['horas_estudo']] # Isso é uma matriz (dataframe)
y = df['nota']           # Isso é um vetor (series)
```

Projeto prático - Nota x Horas de estudo

```
# Criação e treino do modelo:
tree_reg = DecisionTreeRegressor(random_state=42,
                                 max_depth=3)
tree_reg.fit(X, y)
```

Projeto prático - Nota x Horas de estudo

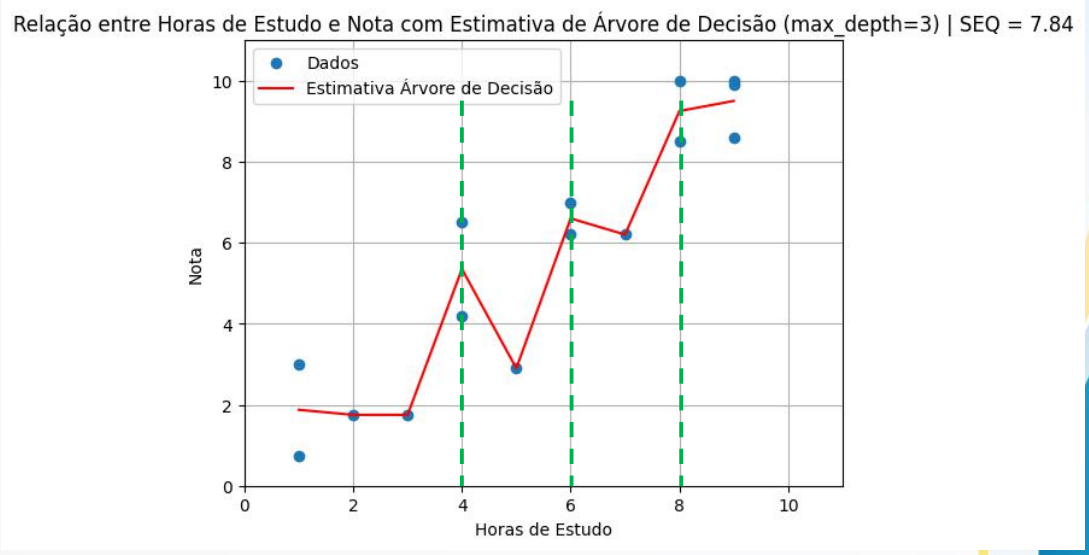
```
# Uso do modelo para fazer previsões:
previsao_tree_reg = tree_reg.predict(X)

# Cálculo da soma dos erros quadráticos (SEQ):
import numpy as np
seq = np.sum((df['nota'] - previsao_tree_reg)**2)
```

Projeto prático - Nota x Horas de estudo

```
# Mostrando o gráfico de análise:
import matplotlib.pyplot as plt
fig, ax = plt.subplots()
ax.set_xlim(0, 11)
ax.set_ylim(0, 11)
ax.set_xlabel('Horas de Estudo')
ax.set_ylabel('Nota')
ax.set_title(f'Relação entre Horas de Estudo e Nota com Estimativa de Árvore de Decisão (random_state=42, max_depth=3) | SEQ = {seq:.2f}')
ax.grid(True)
ax.plot(X, y, 'o', label='Dados')
ax.plot(X, previsao_tree_reg, 'r-', label='Estimativa Árvore de Decisão')
ax.legend()
plt.show()
```

Projeto prático - Nota x Horas de estudo - resultado



DÚVIDAS?

Exercício:

2) Baixe os dados através do link abaixo e crie um modelo, usando a técnica de árvore de decisão, para estimar o nível de automação de uma fábrica sabendo a quantidade de operadores que trabalham nela. Construa o gráfico que mostra os pontos observados e as linhas de previsão de modelos baseados em árvores de 1 até 5 níveis. Quais são as suas conclusões?

Link: https://github.com/amarcel/ML-PPGEP-set2025