# A Jornada do Dado:
## Da Preparação à Previsão

Machine Learning

# **Objetivos** da aula

- A Jornada do Dado
- Prevendo o Futuro com Regressão
- Mão na Massa com Regressão
- Conclusão e Dúvidas

Machine Learning

# **Da Matéria-Prima** ao Ingrediente Perfeito

- **Limpeza de Dados**

  Tratar Valores Ausentes, Outliers (pontos muito fora da curva) e Dados Duplicados.

- **Pré-processamento**
  - Normalização: Colocar todas as características na mesma escala (ex: de 0 a 1).
  - Codificação: Transformar texto (ex: "vermelho") em números que o modelo entenda.

- **Feature Engineering**

  A "arte" de criar novas características a partir das existentes para melhorar o modelo.

Machine Learning

# Dividir para Conquistar
## e os Vilões do ML

### Divisão dos Dados

- **Treino:** Onde o modelo aprende.
- **Validação/Teste:** Onde o modelo é avaliado com dados que nunca viu, para garantir que ele aprendeu de verdade e não apenas "decorou" as respostas.

### Dados Desbalanceados

- Ocorre quando uma classe tem muito mais exemplos que outra.
- **Solução:** Balancear com técnicas com Oversampling (SMOTE).

### Underfitting vs. Overfitting

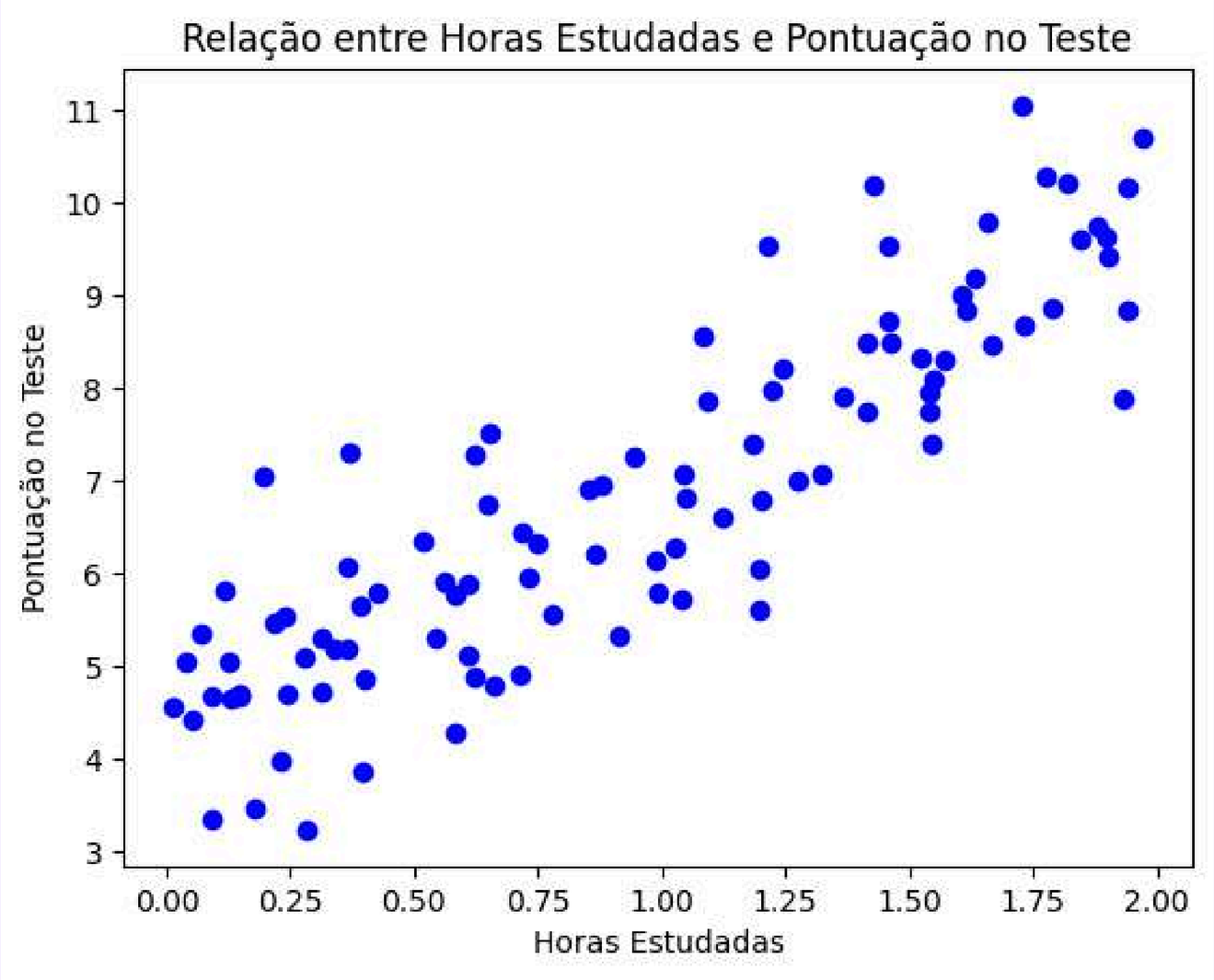
- **Underfitting:** O modelo é simples demais e não aprende o padrão.
- **Overfitting:** O modelo é complexo demais, decora os dados de treino e não generaliza para novos dados.

Machine Learning

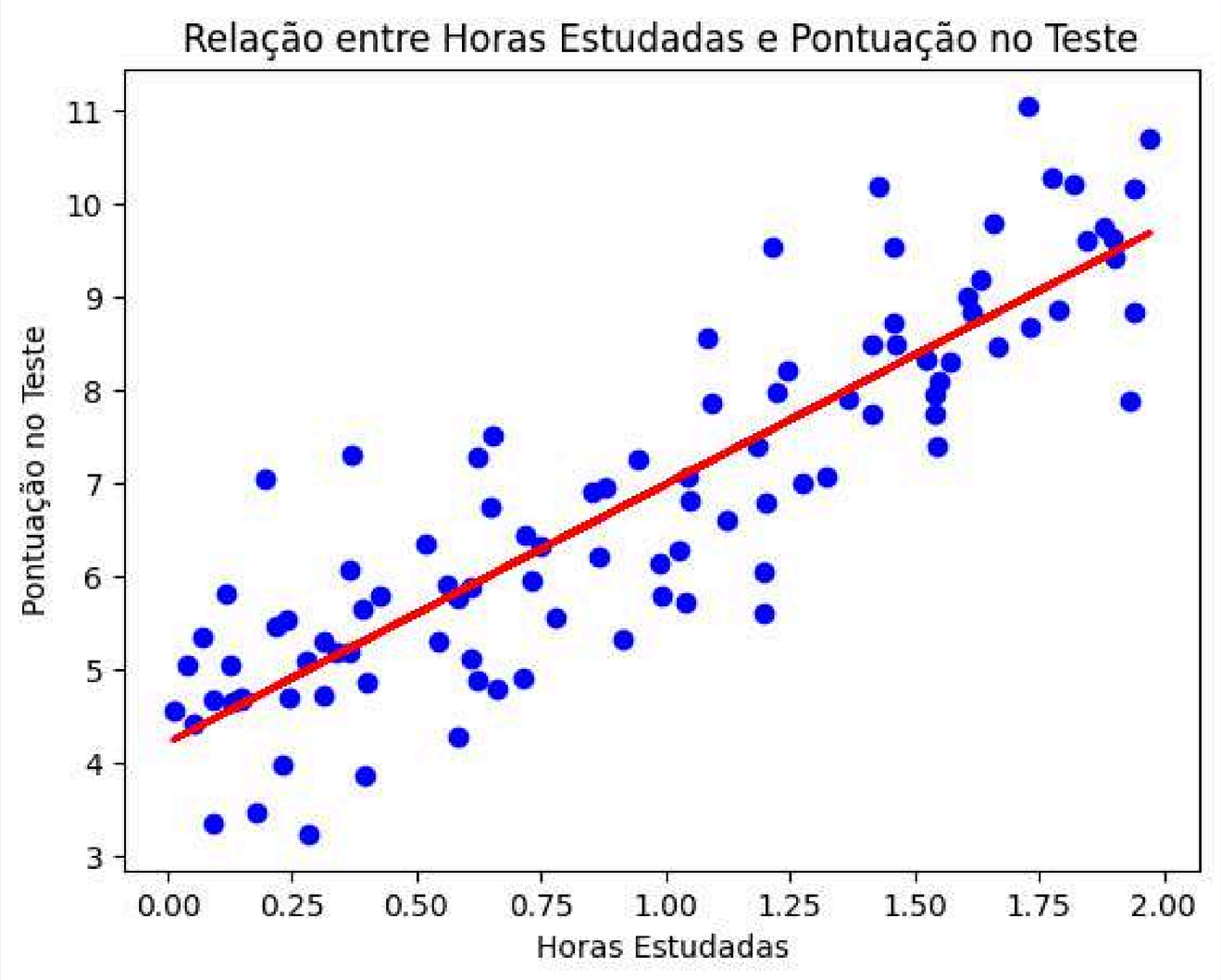

# O que é Regressão?

- **Definição**
  - Uma técnica para prever um valor contínuo (numérico)
- **Diferença Crucial**
  - Classificação prevê "O quê?" (Ex: Este e-mail é spam ou não spam?) .
  - Regressão prevê "Quanto?" (Ex: Qual será o preço desta casa?) .

Machine Learning



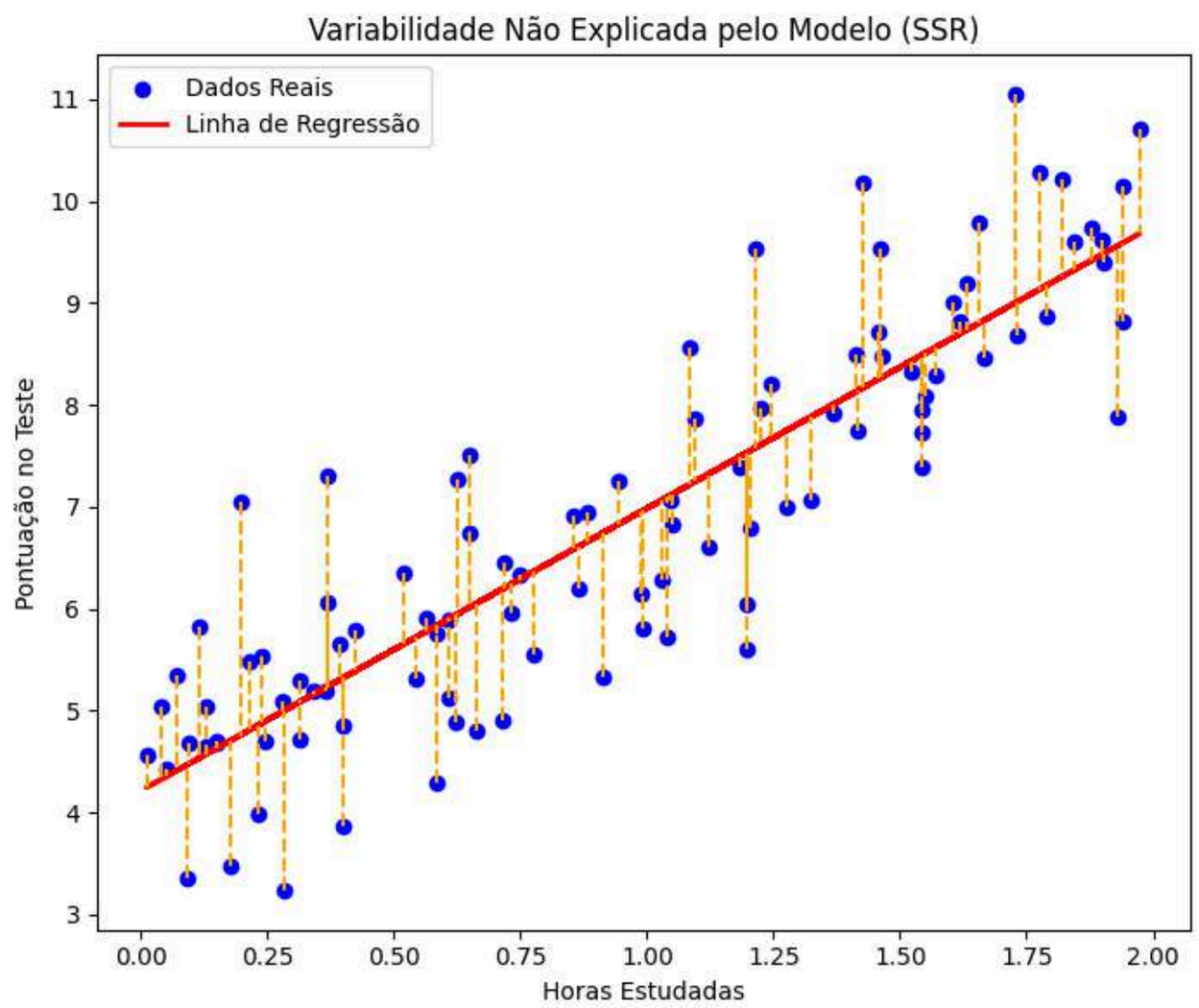
Machine Learning

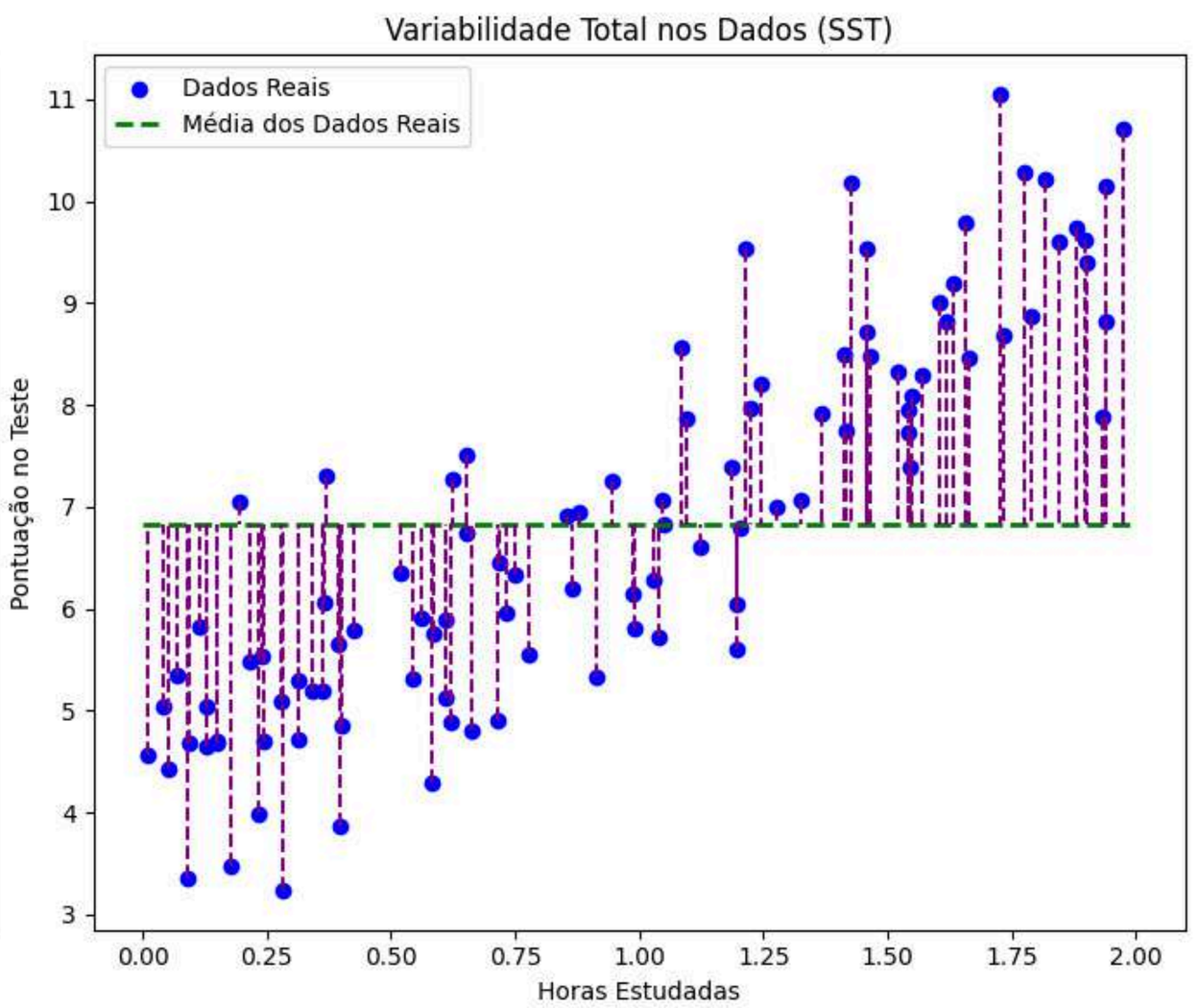


Machine Learning

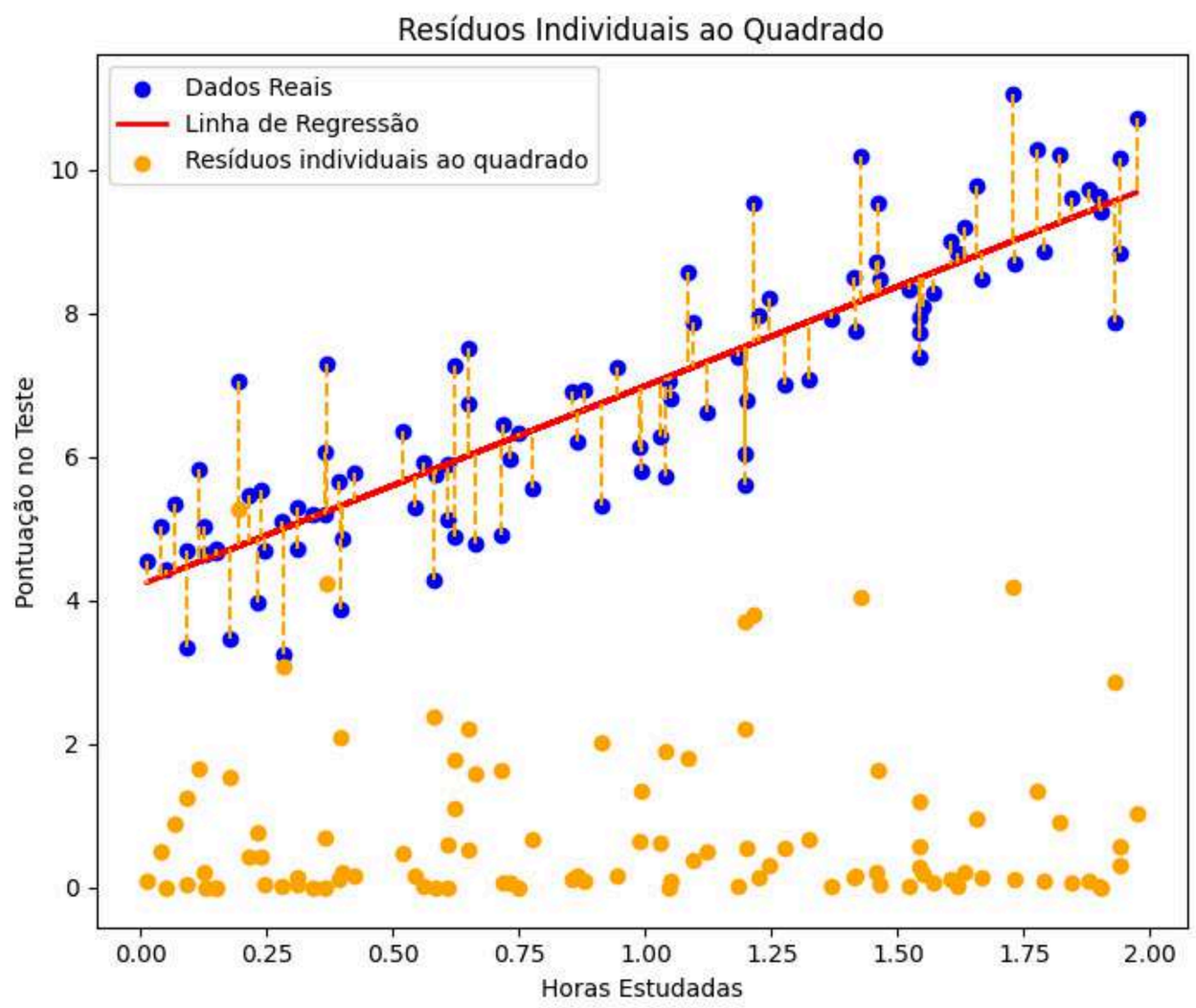

# O que é Regressão?

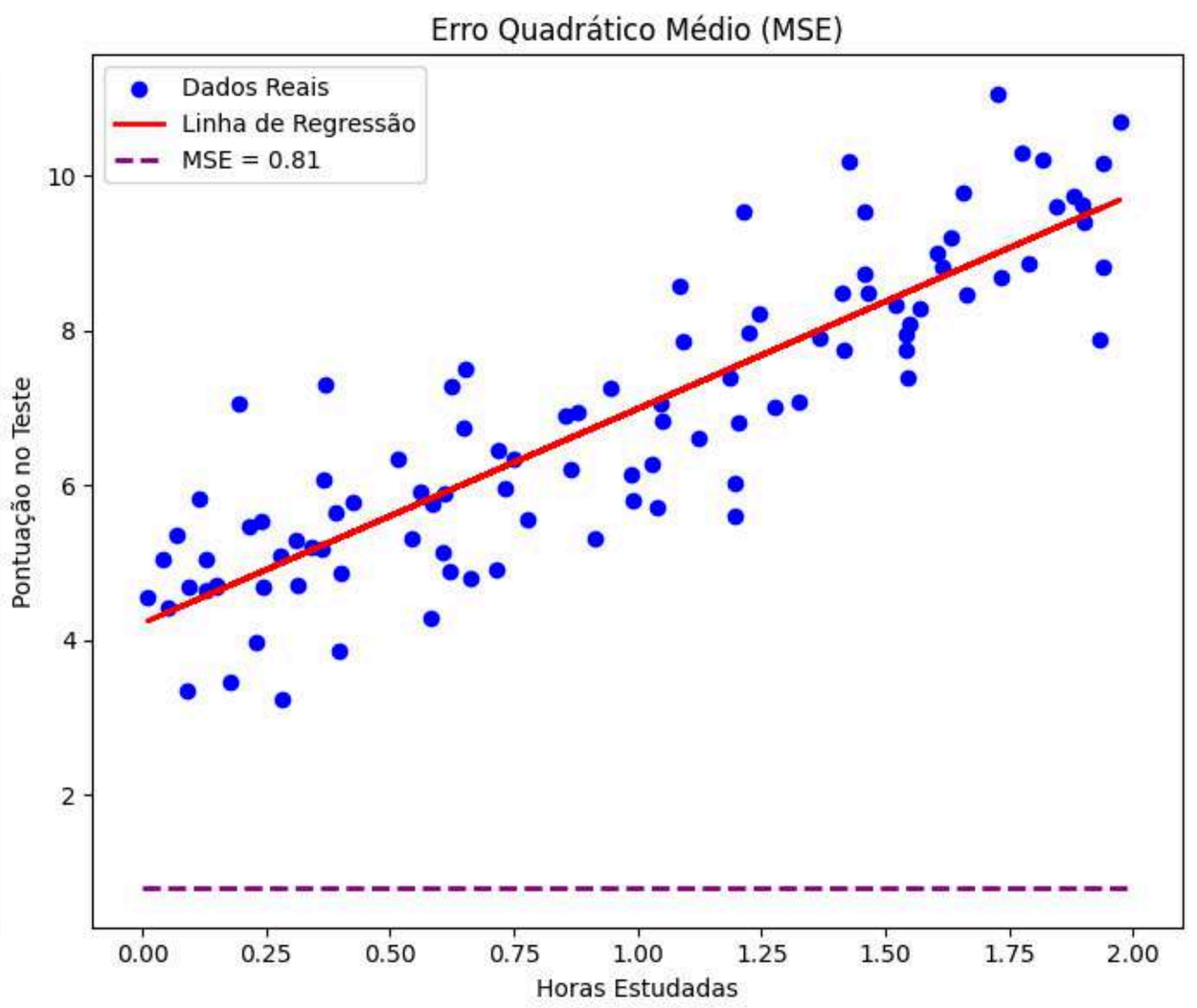
- **Como Medimos o Sucesso?**
  - Com métricas que calculam o erro (a distância entre a previsão e o valor real).
  - **MSE:** Penaliza muito os erros grandes.
  - **MAE**: Menos sensível a outliers.
  - **$R^2$:** De 0 a 1, mede o quão bem o modelo "explica" os dados.
  - **RMSE:** traz a medida de erro para a mesma escala da variável dependente


Variabilidade Não Explicada pelo Modelo (SSR)
Dados Reais
Linha de Regressão
Pontuação no Teste
Horas Estudadas
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
3
4
5
6
7
8
9
10
11

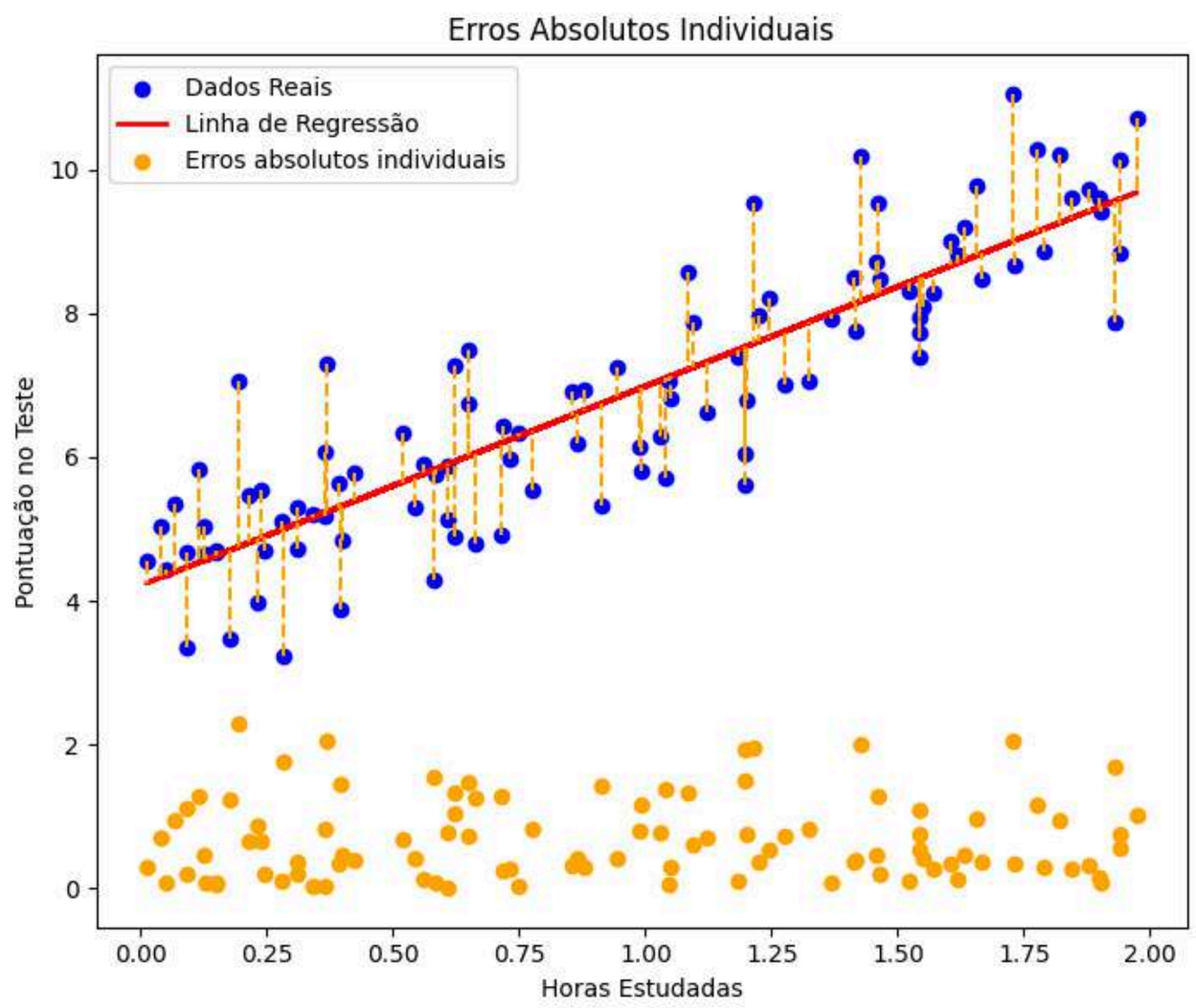

Variabilidade Total nos Dados (SST)
Dados Reais
Média dos Dados Reais
Pontuação no Teste
Horas Estudadas
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
3
4
5
6
7
8
9
10
11

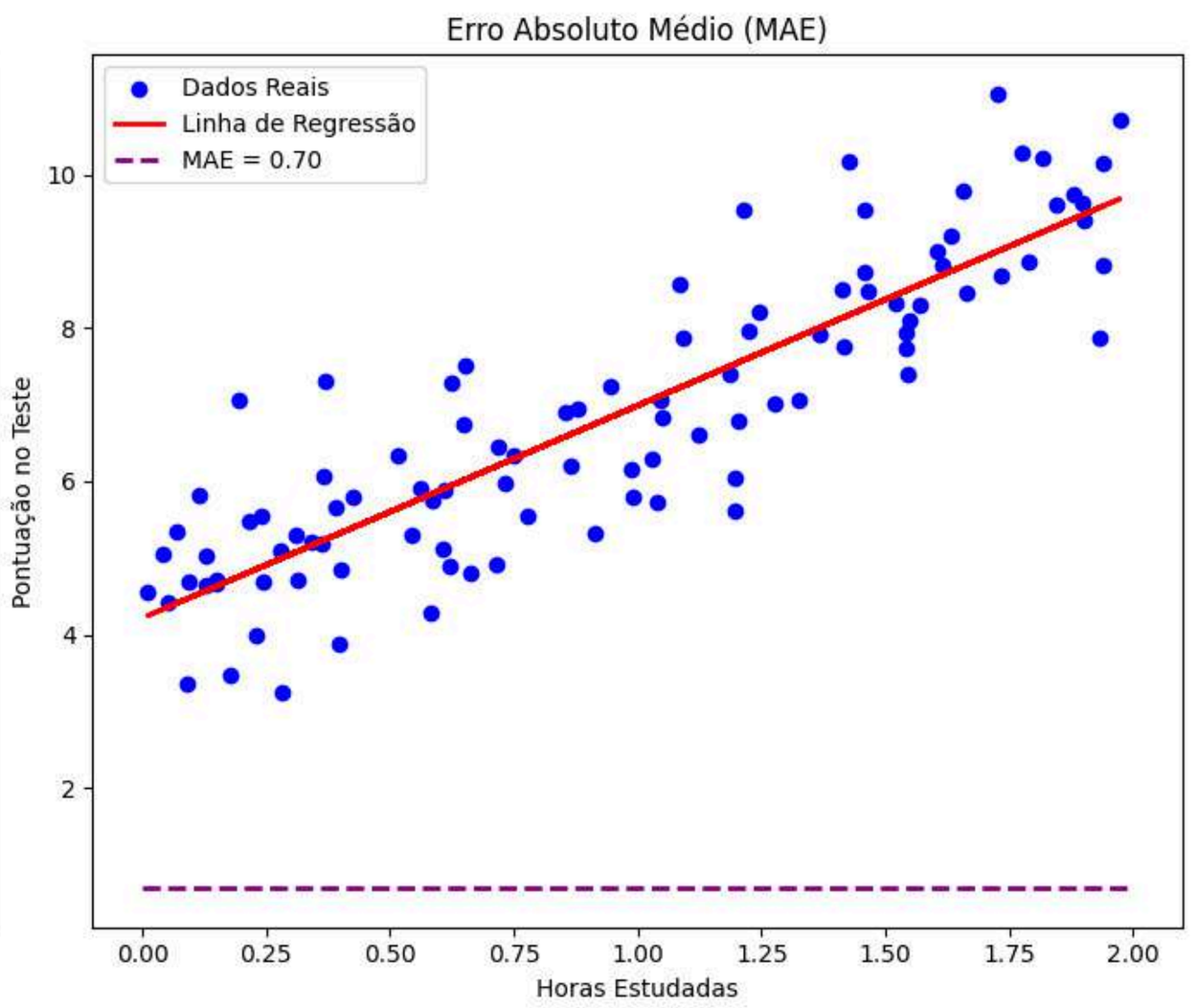
Machine Learning

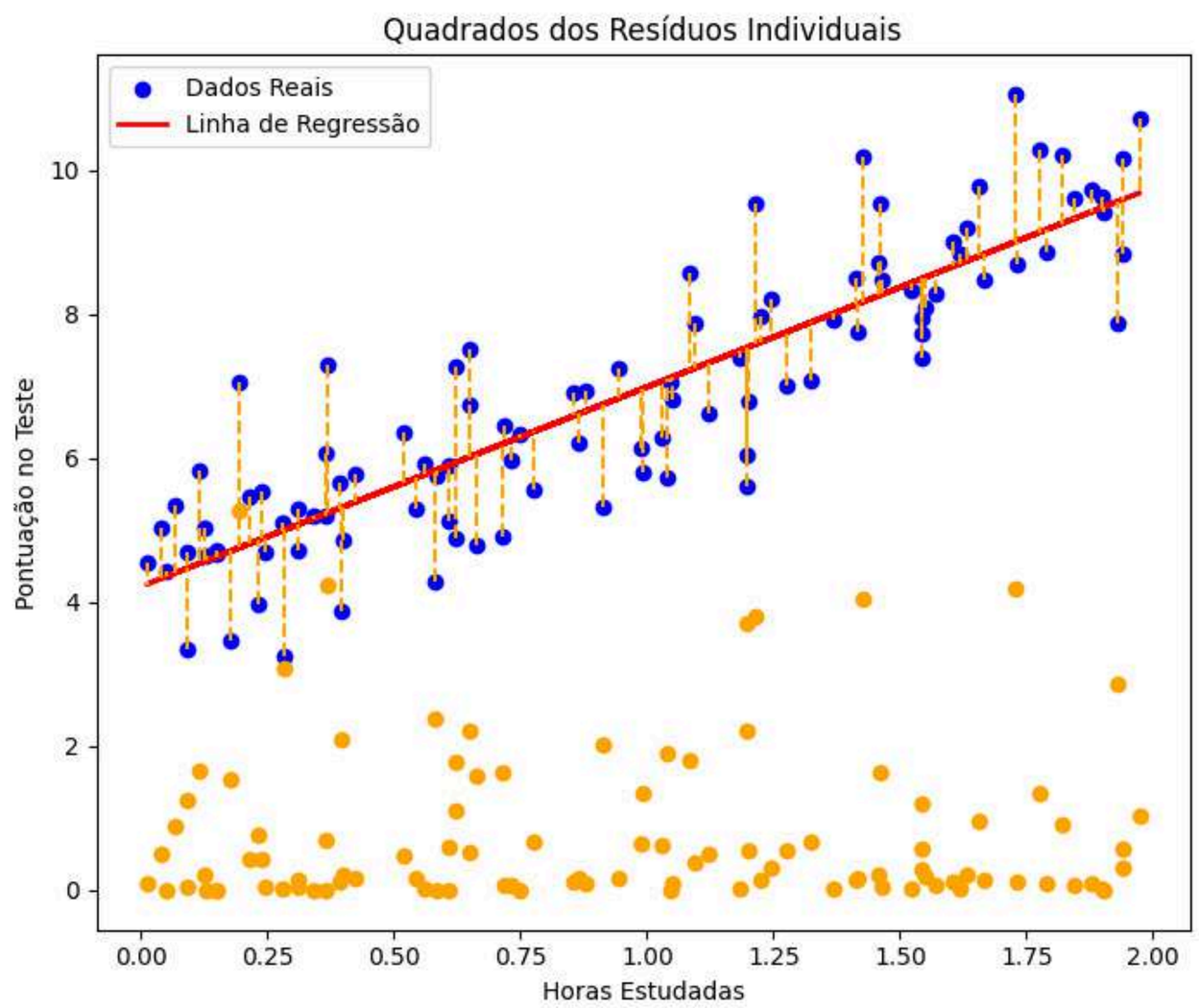


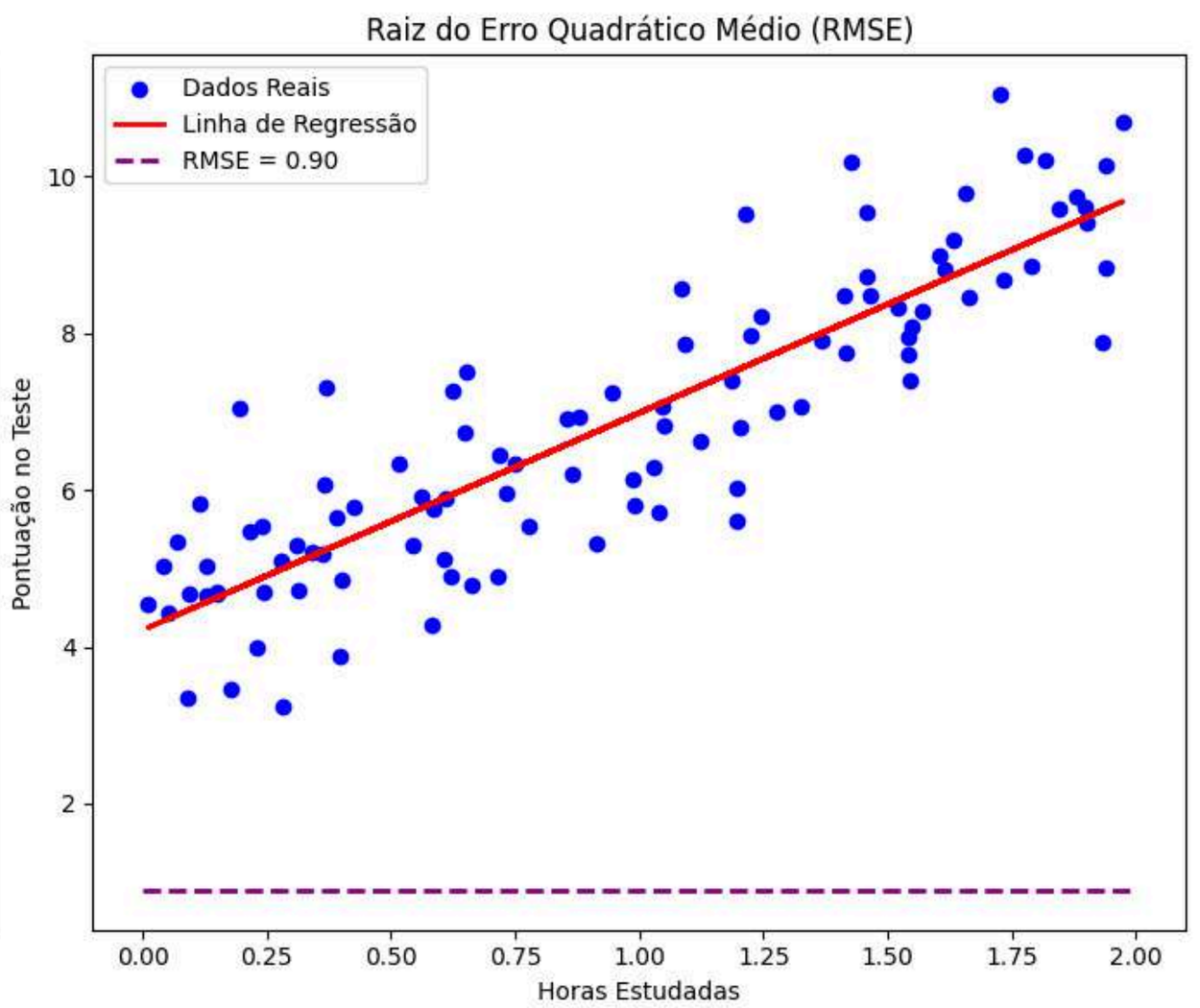



Machine Learning

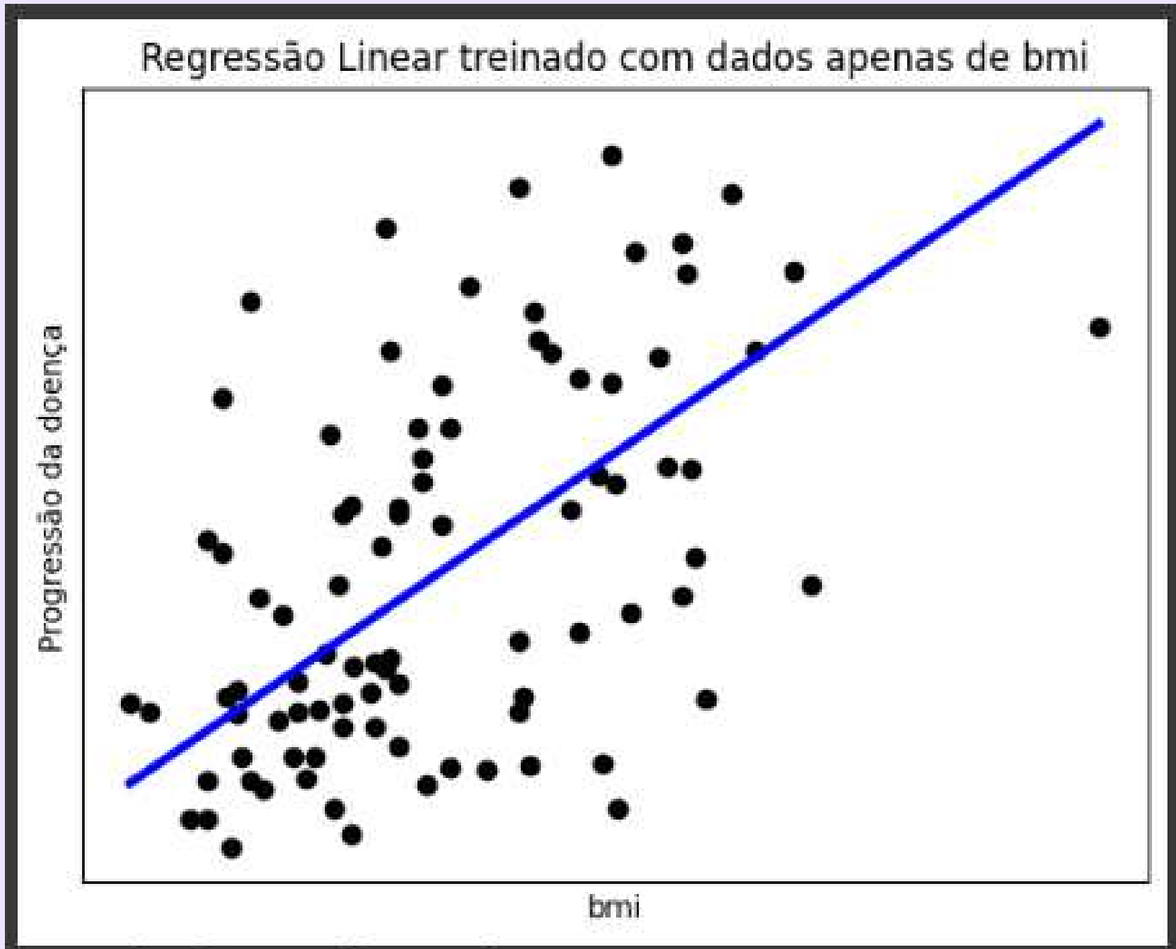


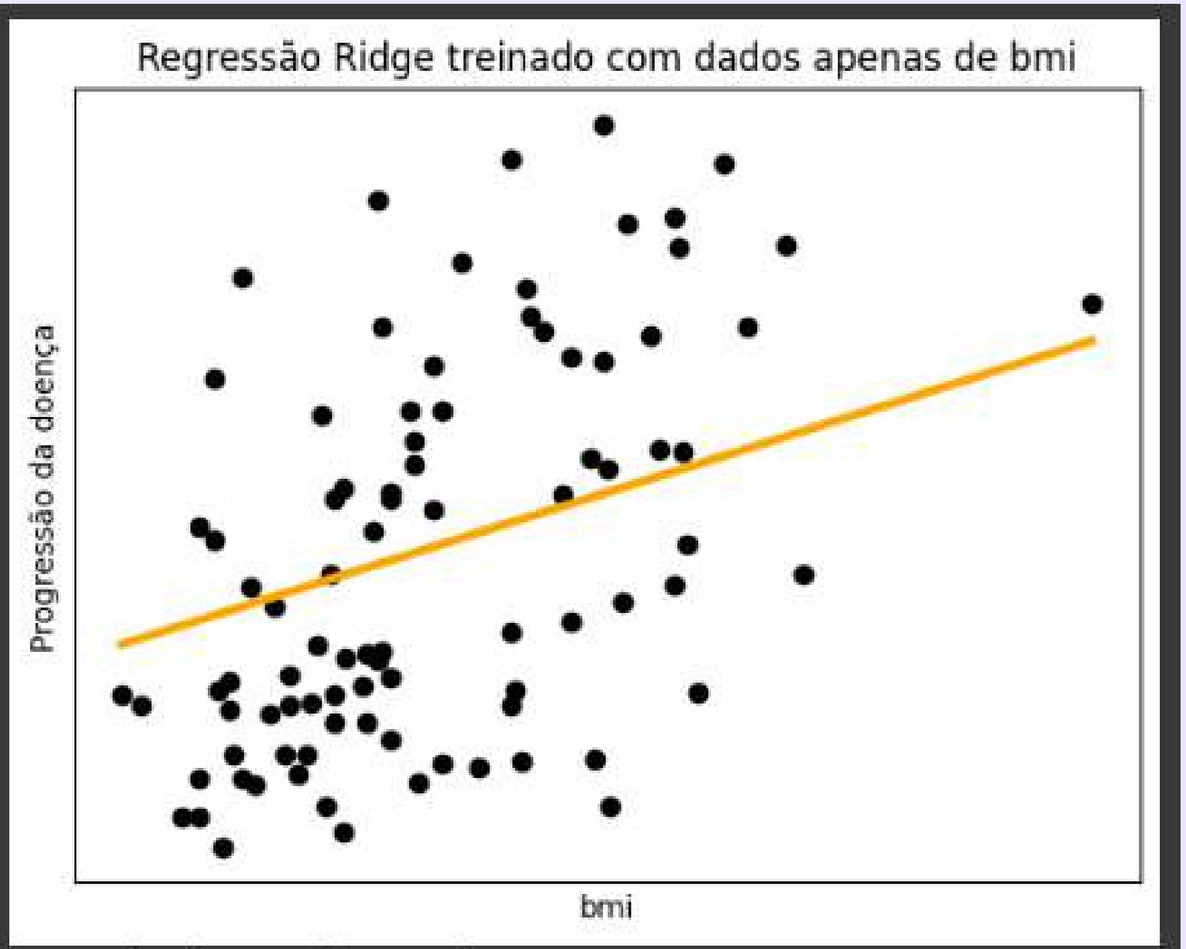




Quadrados dos Resíduos Individuais
Dados Reais
Linha de Regressão
Pontuação no Teste
Horas Estudadas
0
2
4
6
8
10
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


Raiz do Erro Quadrático Médio (RMSE)
Dados Reais
Linha de Regressão
RMSE = 0.90
Pontuação no Teste
Horas Estudadas
2
4
6
8
10
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

Machine Learning

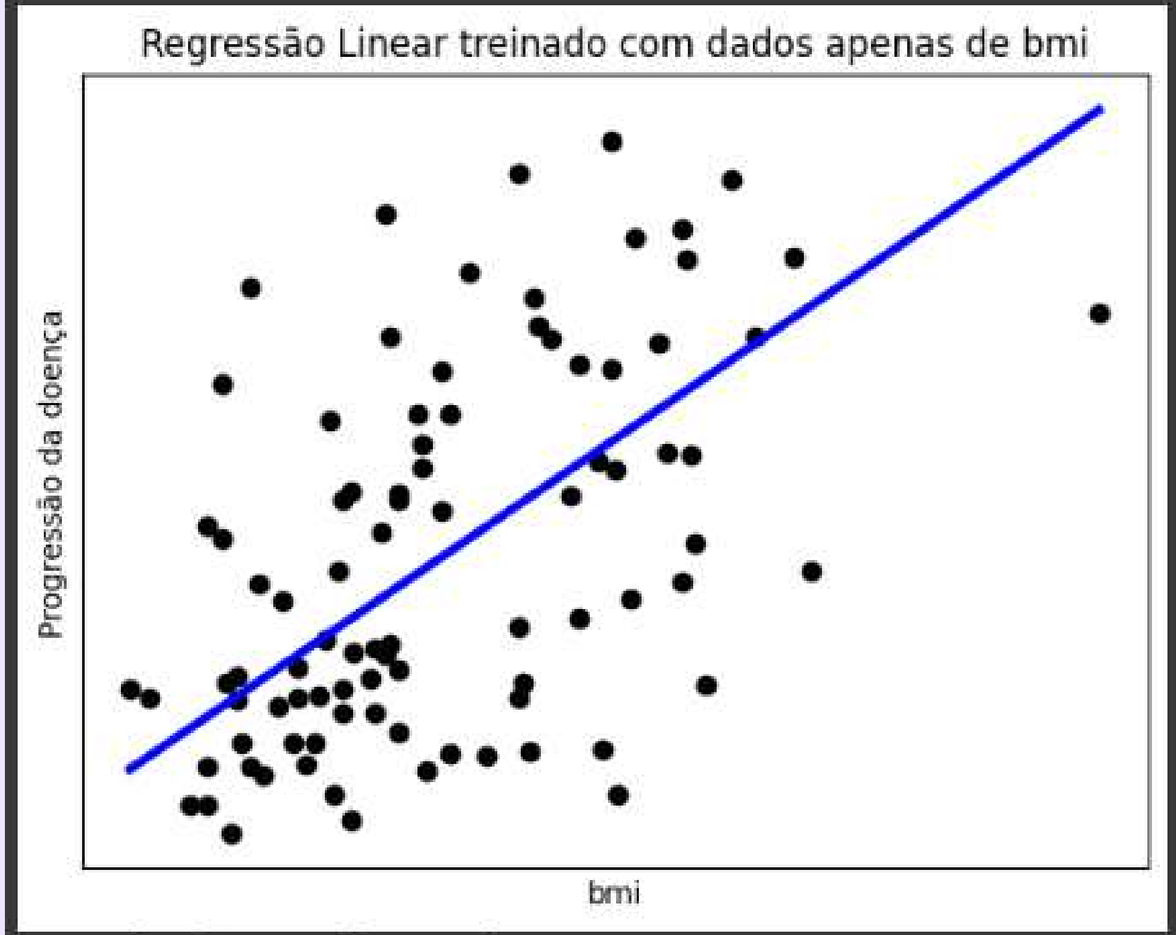


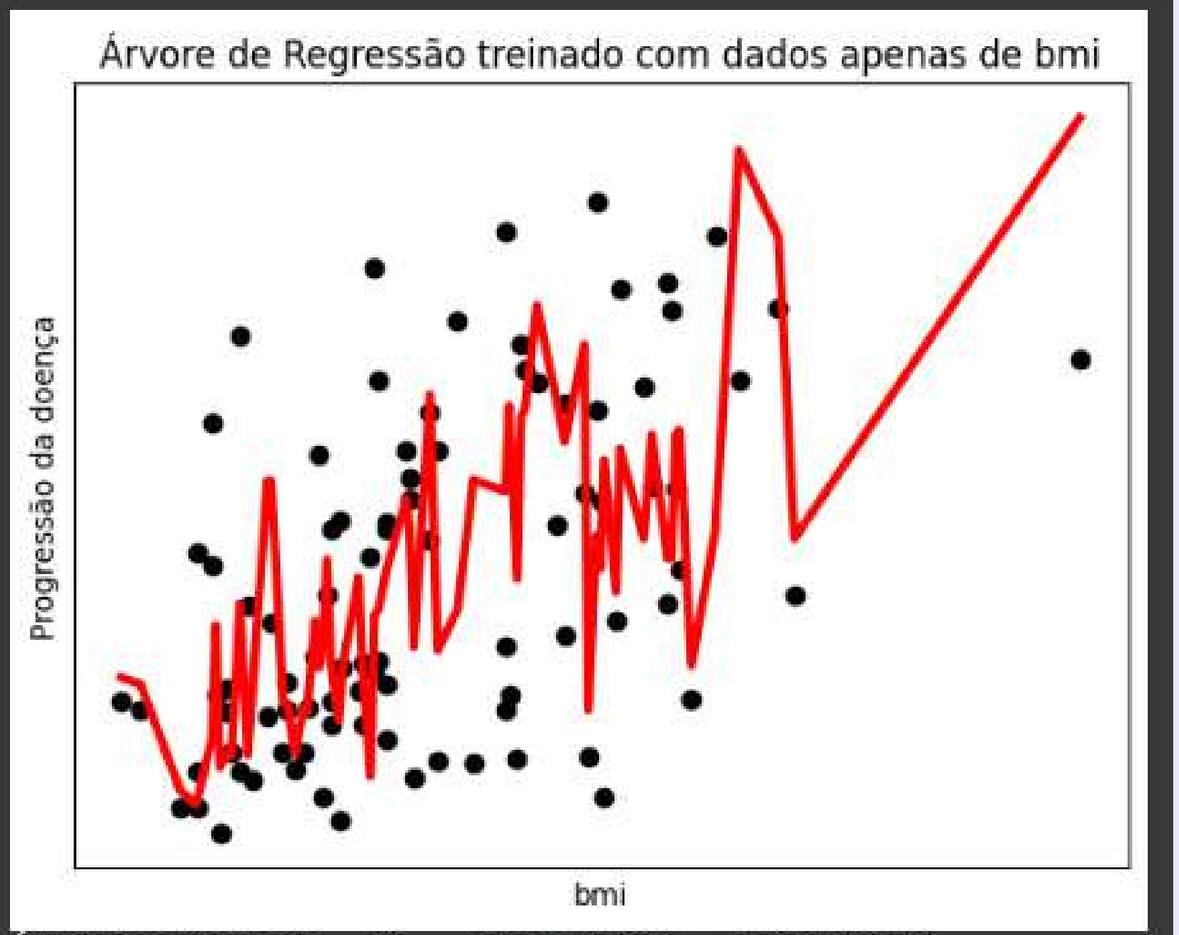




Regressão Linear - $R^2$ - Treino 0.3657 - Teste 0.2334
Regressão Linear - MSE - Treino 3854.1127 - Teste 4061.8259
Regressão Linear - MAE - Treino 51.3797 - Teste 52.2600

Regressão Ridge - $R^2$ - Treino 0.2512 - Teste 0.2079
Regressão Ridge - MSE - Treino 4550.2502 - Teste 4196.6518
Regressão Ridge - MAE - Treino 57.7377 - Teste 55.9570

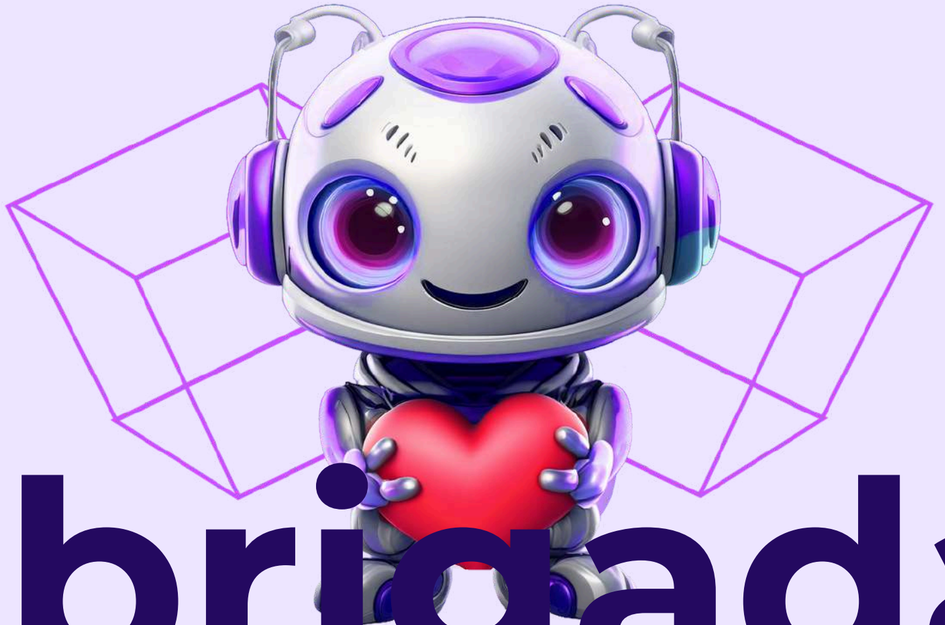
Machine Learning




Regressão Linear - $R^2$ - Treino 0.3657 - Teste 0.2334
Regressão Linear - MSE - Treino 3854.1127 - Teste 4061.8259
Regressão Linear - MAE - Treino 51.3797 - Teste 52.2600

Árvore de Regressão - $R^2$ - Treino 0.6175 - Teste 0.1560
Árvore de Regressão - MSE - Treino 2324.3150 - Teste 4471.5940
Árvore de Regressão - MAE - Treino 36.3725 - Teste 52.3906



# Obrigada!

*Perguntas?*