UNIVERSIDADE FEDERAL
DE MINAS GERAIS

EXERCÍCIO
DILEMA BIAS-VARIÂNCIA

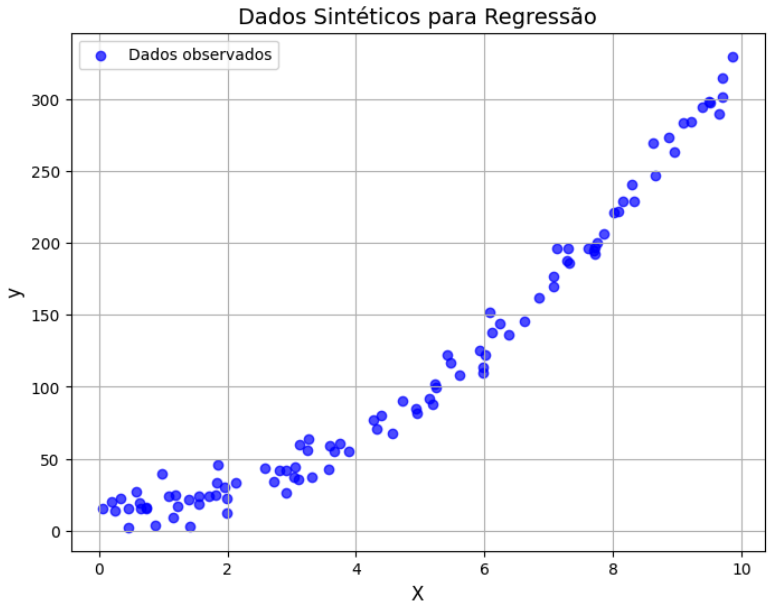
ALUNO: FELIPPE VELOSO MARINHO
MATRÍCULA: 2021072260
DISCIPLINA: APRENDIZADO DE MÁQUINA

Objetivo:

O objetivo do exercício é analisar um conjunto de dados aplicando a regressão linear em diferentes cenários de ajuste. A partir disso, será observado o comportamento de bias e variância em conjuntos de treino e teste.

Gerando os Dados:

O conjunto de dados criado possui uma relação não-linear entre as variáveis e se utiliza de um ruído gaussiano.

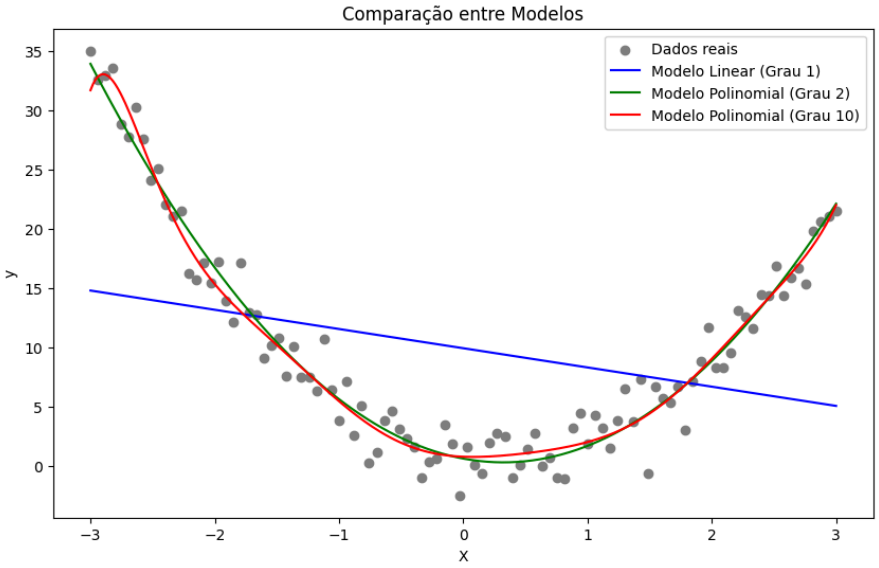


A partir disso importamos 3 modelos diferentes com as seguintes características:

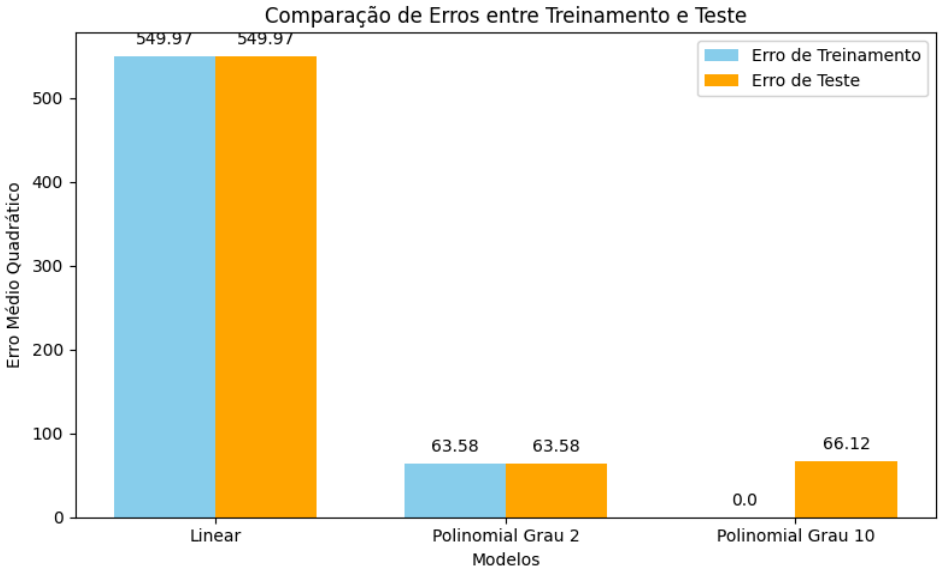
1. Modelo linear simples (apenas o termo linear, y=a+bx).

2. Modelo polinomial de grau 2 (captura a estrutura correta, y=a+bx+cx2).

3. Modelo polinomial de grau 10 (modelo muito complexo).



A comparação entre os modelos nos mostra uma maior adaptabilidade nos modelos polinomiais com um possível overfitting no modelo polinomial de grau 10. Abaixo estão as comparações entre os erros de treinamento e teste.



O **modelo polinomial de grau 10** apresenta erro quase nulo no treinamento (overfitting) e um erro levemente maior no teste. Isso sugere que o modelo ajusta demais os dados de treinamento, prejudicando a generalização.
Bias-Variância: Baixo bias, alta variância.

O **modelo de grau 2** tem o menor erro em ambos os conjuntos, mostrando que captura adequadamente a relação sem superajustar os dados.
Bias-Variância: Baixo bias, baixa variância.

O **modelo linear** apresenta erro elevado tanto em treinamento quanto em teste. Isso sugere **alto bias**, indicando que o modelo não captura bem a relação subjacente nos dados.
Bias-Variância: Alto bias, baixa variância.

Os dados de erro quadrático médio e R-quadrado confirmam as observações com o modelo linear refletindo uma limitação notável na captura da relação entre as variáveis podendo representar um alto viés.

O modelo polinomial de grau 2 se mostra como o melhor ajustado e equilibra bem o bias e a variância.

O modelo de grau 10 possui um MSE um pouco maior que o de grau 2, o que sugere novamente um overfitting nos dados de treinamento.

```
Erro médio quadrático (Linear): 549.97
r2 Score (Linear): 0.94

Erro médio quadrático (Polinomial Grau 2): 63.58
r2 Score (Polinomial Grau 2): 0.99

Erro médio quadrático (Polinomial Grau 10): 66.12
r2 Score (Polinomial Grau 10): 0.99
```

Por fim, o modelo linear apresentou erro alto em ambos os conjuntos, caracterizando **alto bias**, pois não conseguiu capturar a relação subjacente nos dados. Já o modelo polinomial de grau 10 apresentou **baixa variância no treinamento**, mas alta no teste, indicando overfitting, ou seja, ele se ajustou excessivamente aos dados de treino. O modelo de grau 2 teve o **melhor desempenho geral**, equilibrando bias e variância com baixo erro nos dois conjuntos. À medida que a complexidade aumenta, o bias tende a diminuir, enquanto a variância aumenta, o que explica o desempenho ruim do modelo de grau 10 no teste, mesmo com ótimo ajuste nos dados de treino.

```
Modelo Linear:
Erro médio quadrático: 28.79
R²: 0.78
Modelo Polinomial Grau 2:
Erro médio quadrático: 19.21
R²: 0.85
Modelo Polinomial Grau 10:
Erro médio quadrático: 19.48
R²: 0.85
```

O aumento do ruído nos dados tende a impactar o desempenho dos modelos, pois os erros nas previsões aumentam, especialmente para modelos de

alta variância como o polinomial de grau 10. O modelo linear será menos afetado pelo ruído devido ao seu **alto bias** e simplicidade, enquanto o modelo de grau 2 continuará relativamente robusto. Já o modelo de grau 10 pode piorar significativamente, pois ele tenta ajustar até mesmo as flutuações aleatórias causadas pelo ruído.

Com o aumento dos dados, a generalização dos modelos melhorou, e os valores de erro médio quadrático diminuíram, especialmente para o modelo polinomial de grau 2, que manteve o melhor desempenho geral. O modelo de grau 10 apresenta desempenho próximo ao de grau 2, mas o aumento da complexidade não trouxe vantagens significativas, mostrando estabilidade na variação com os dados ampliados.