Exercício Dilema Bias-Variância

Objetivo:

Os alunos irão:

- Treinar modelos de regressão linear em diferentes cenários de ajuste.
- Observar o comportamento de bias e variância em conjuntos de treino e teste.
- Interpretar o impacto no erro.

Descrição do Exercício:

1. Geração de Dados Sintéticos

Os alunos criarão um conjunto de dados simples com uma relação não-linear entre as variáveis, por exemplo:

$y=10+2x+3x2+\epsilon$

onde ϵ é ruído gaussiano (N(0,1)).

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import mean_squared_error

# 1. Gerando os dados sintéticos
np.random.seed(42)
X = np.random.rand(100, 1) * 10 # Valores entre 0 e 10
y = 10 + 2 * X + 3 * X**2 + np.random.normal(0, 10, size=(100, 1)) #
Relação não-linear com ruído
```

2. Modelos de Diferentes Complexidades

Os alunos treinarão:

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).

```
# Criando as features polinomiais
poly = PolynomialFeatures(degree=grau)
X_poly_train = poly.fit_transform(X_train)
X_poly_test = poly.transform(X_test)
# Ajustando o modelo
modelo = LinearRegression()
```

```
modelo.fit(X_poly_train, y_train)
# Fazendo previsões
y_pred_train = modelo.predict(X_poly_train)
y_pred_test = modelo.predict(X_poly_test)
```

3. Avaliação

Os alunos devem:

- Dividir os dados em treino e teste.
- Ajustar os modelos aos dados de treino.
- Avaliar os modelos nos conjuntos de treino e teste usando o **MSE** e o **R-quadrado**.

Discussão:

1. **Responder às Perguntas:**
    - Qual modelo apresentou **erro alto em ambos os conjuntos** (alto bias)?
    - Qual modelo apresentou **baixo erro no treino e alto erro no teste** (alta variância)?
    - Qual modelo tem o **melhor desempenho geral**?
    - Como o aumento da complexidade do modelo impactou bias e variância?
    - Por que o modelo de grau 10 teve um desempenho ruim no teste, mesmo que o treino fosse bom?
    - Aumentar o ruído gere alguma alteração nos resultados observados para os três modelos? Porque?
    - Triplique o número de dados. O que isso causa nos resultados observados dos modelos?

O relatório em PDF deve conter tudo o que é solicitado acima.