Dilema Bias-Variância

1. Geração dos Dados Sintéticos

Nessa etapa foi feita a geração dos dados através de uma função, levando em consideração o ruído que queremos para termos uma relação não-linear nos dados.

Abaixo se encontra nossos dados plotados:

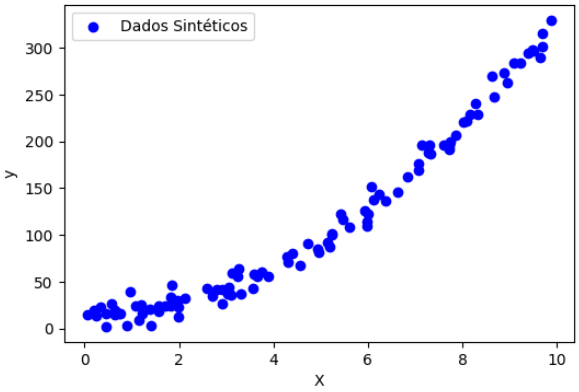

Figura 1.

Também foi feita a divisão dos dados em treino e teste, com a porcentagem de teste = 20% do total dos dados.

2. Modelos de Diferentes Complexidade

Nesse momento, treinaremos três tipos diferentes de modelo para sabermos qual se ajusta melhor aos nossos dados.

Então teremos:

- Modelo linear simples (apenas o termo linear, y=a+bx).
- Modelo polinomial de grau 2 (captura a estrutura correta, y=a+bx+cx2).
- Modelo polinomial de grau 10 (modelo muito complexo).

Para cada modelo calcularemos as métricas RSE e R^2 para avalia-los,

É esperado que, o modelo linear simples apresente um underfitting em relação aos nossos dados, não conseguindo capturar de forma otimizada sua complexidade, além de que o modelo muito complexo (de grau 10) apresente um overfitting em relação aos dados, ou seja, não consiga generalizar para novos dados que a entrada de nosso modelo possa enfrentar.

3. Avaliação

Após o treinamento dos três modelos, nessa etapa faremos sua avaliação.

Então, temos:

Polinômio de grau 1:

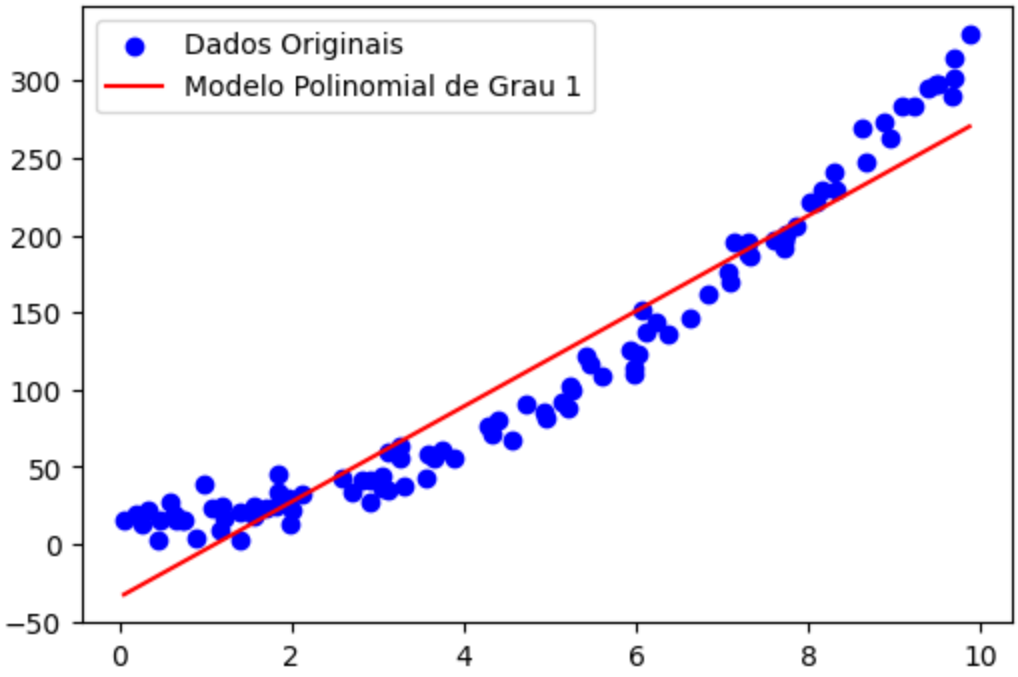

Figura 2.

Mean Squared Error (MSE): 549.97
R^2 Score: 0.94

Polinômio de grau 2:

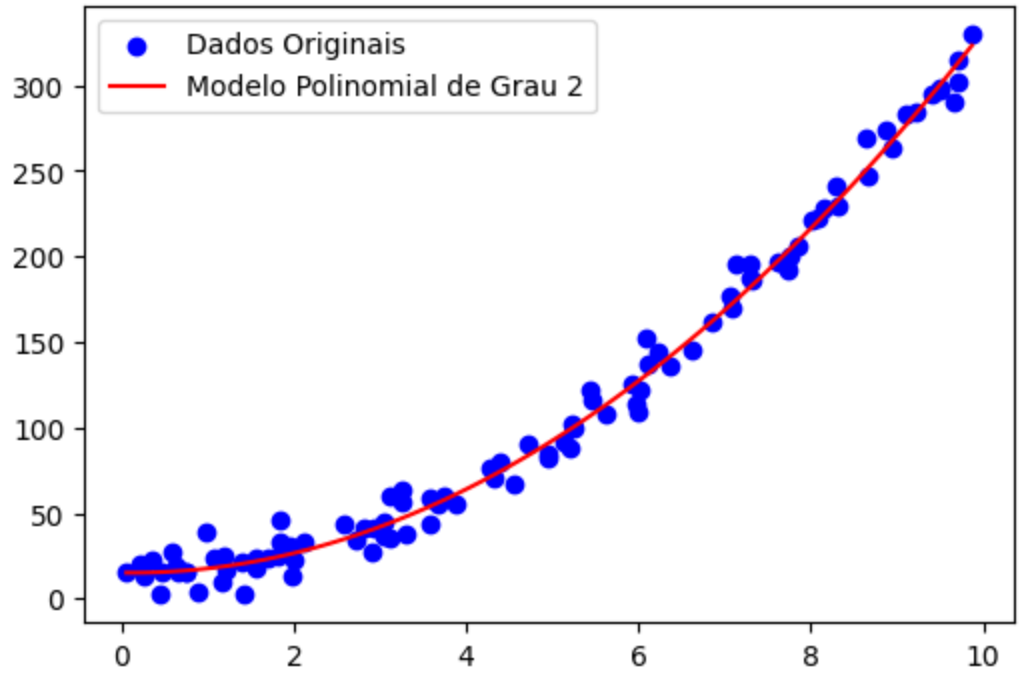

Figura 3.

Mean Squared Error (MSE): 63.58
R^2 Score: 0.99

Polinômio de grau 10:

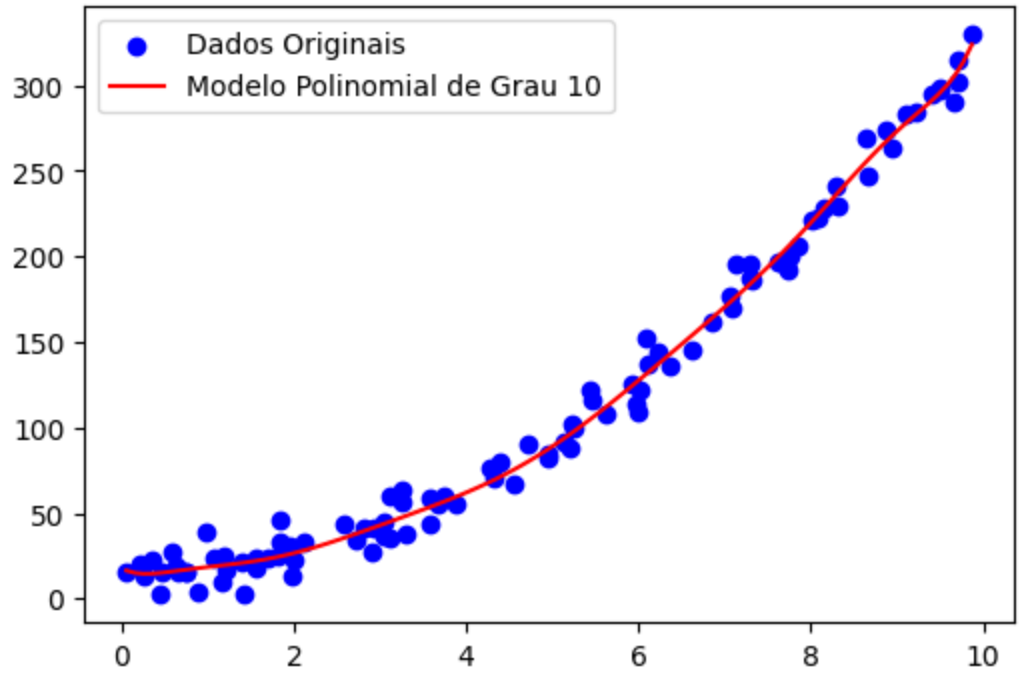

Figura 4.

Mean Squared Error (MSE): 66.12
R² Score: 0.99

Vemos que o MSE reduziu drasticamente seu valor do modelo simples para o modelo quadrático, e o valor do R^2 aumentou, o que mostra logo de início que o modelo quadrático teve uma boa performance. Também é possível analisar isso qualitativamente comparando os gráficos da Figura 2 e 3.

Em relação ao modelo mais complexo (de grau 10), podemos ver que o MSE chegou a até ser maior em comparação ao modelo de grau 2, enquanto o R^2 se manteve parecido entre os dois.

Isso mostra que não necessariamente um modelo mais complexo será melhor para o seu conjunto de dados, dependendo do problema que você deseja resolver e das variáveis do seu dataset também. Além de que, é muito provável que o modelo de grau 10 esteja em um estado de overfitting, em que não conseguirá generalizar para possíveis novos dados que o modelo tenha em sua entrada.

Discussão:

1. Qual modelo apresentou erro alto em ambos os conjuntos (alto bias)?

O modelo linear (de grau 1). Por ser um modelo muito simples, ele não conseguiu capturar de forma adequada a estrutura da função, ficando em um estado de underfitting em relação aos dados.

2. Qual modelo apresentou baixo erro no treino e alto erro no teste (alta variância)?

O modelo de grau 10. Esse modelo, por estar muito complexo, provavelmente está em overfitting em relação aos dados, apresentando um baixo erro no treino e alto erro no teste, não conseguindo generalizar para novos dados. (Alta variância, e baixo viés).

3. Qual modelo tem o melhor desempenho geral?

O modelo de grau 2. O modelo de grau 1 está em underfitting (alto viés e baixa variância), e o modelo de grau 10 está em overfitting (baixo viés e alta variância), sendo que o modelo de grau 2 é o que melhor se comporta em relação a função de entrada.

4. Como o aumento da complexidade do modelo impactou bias e variância?

No geral, temos:

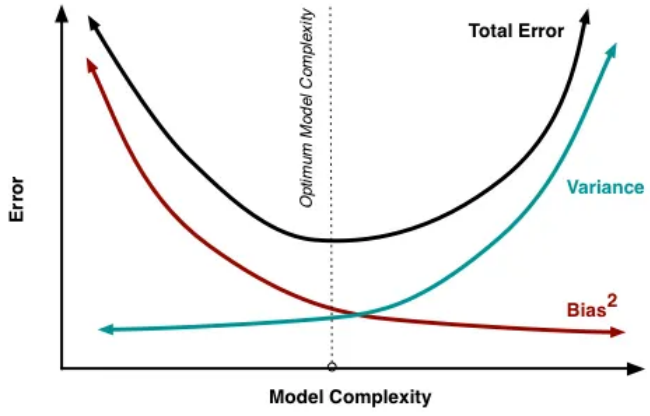

Figura 5.

Então, ao **aumentarmos a complexidade** de nosso modelo, a tendência é ele cada vez possuir uma precisão e acurácia melhor em relação aos dados de treino,

mas não conseguir generalizar para dados de teste e novos dados que a entrada possa enfrentar (**overfitting**, com alta variância e baixo viês).

5. Por que o modelo de grau 10 teve um desempenho ruim no teste, mesmo que o treino
   fosse bom?

Devido ao fato de ele ser **muito complexo** para o problema que queremos resolver, ou seja, para nossa função de entrada do problema. Com isso, o modelo está em **overfitting,** não conseguindo uma boa perfoarmance nos dados de teste ou novos dados que o modelo possa enfrentar.

6. Aumentar o ruído gera alguma alteração nos resultados observados para os três
   modelos? Porque?

O resultado é alterado. É notável que, ao aumentarmos o ruído, o desvio padrão das amostras do conjunto de dados fica maior também, o que acaba sendo mais perceptível o overfitting que acontece caso nosso modelo for complexo demais.

Temos como exemplo:

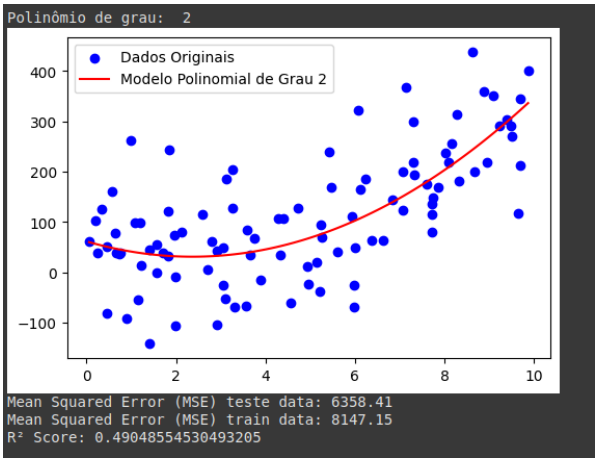

Figura 6.

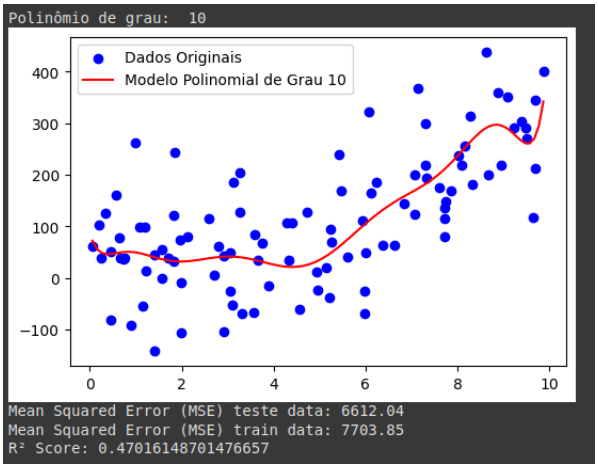

Figura 7.

É notável que o modelo mais complexo tentou se ajustar completamente aos dados de treino, e provavelmente não conseguirá generalizar para novos dados que o modelo possa enfrentar. Esse é o caso de alta variância e baixo viés.

7. Triplique o número de dados. O que isso causa nos resultados observados dos
   modelos?

Ao triplicarmos o número de dados, a tendência é que nosso modelo mais complexo tente se ajustar mais ainda aos dados de treino, piorando seu ajuste. Isso é mostrado abaixo: