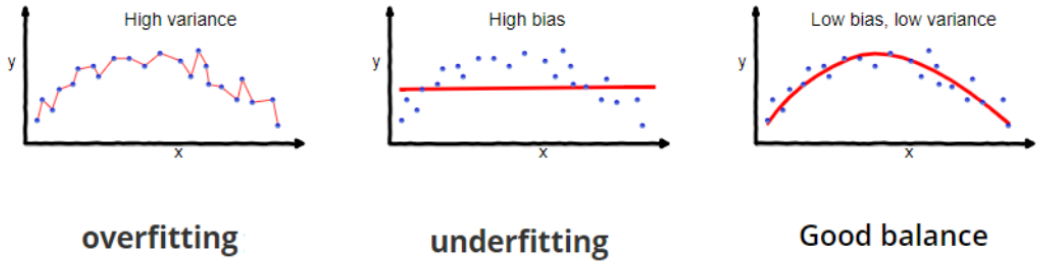
EXERCÍCIO: MODELO DE REGRESSÃO LINEAR E MÉTODO DE VALIDAÇÃO CRUZADA (K-FOLD)

O dataset utilizado para este exercício refere-se à **dados imobiliários de casas à venda.** Os atributos registram informações como localização, idade da casa, dentre outras. O atributo classe desse dataset é o **preço de venda** da casa. Para este exercício, realize o download do arquivo .csv e execute o que se pede.
Dica: utilize o mesmo valor de **random_state** em todos os métodos para garantir a reprodutibilidade dos seus resultados.

Lembrando:
Overfitting: super-ajuste do modelo aos dados (alta variância e baixo bias);
Underfitting: baixo ajuste do modelo as dados, o modelo não é generalizado o suficiente para conseguir prever novos valores fora do conjunto de treino (alto bias, baixa variância).



1. Carregue o arquivo, faça uma análise exploratório e a limpeza dos dados:
    a. Calcule estatísticas básicas usando o **describe**;
    b. Verifique se há dados faltantes. Caso haja, trate-os da forma que achar mais conveniente para o contexto do problema;
    c. Verifique se há dados duplicados. Caso haja, trate-os da forma adequada.
    d. Calcule a correlação dos atributos utilizando a função **corr** do pandas com o método de **pearson**. O que você observa de acordo com a matriz de correlações? Faça um gráfico de dispersão, e procure indícios visuais que confirmem os valores de correlação encontrados na matriz.
    e. Verifique os valores dos dados, e os padronizem na mesma **escala**. Você pode usar a função **MinMaxScaler** do sci-kit learn para isso.
    f. Defina as variáveis **X** (variáveis independentes) e **Y (**variável dependente, atributo **classe** da predição)**.**

2. Crie um modelo de Regressão Linear que, dando como entrada os atributos da casa, retorne o seu preço de venda:
    a. Defina seus conjuntos de **treino** e **teste** (utilize as proporções 70% - 30% ou 80% - 20%);

b. Instancie o modelo de regressão linear usando o **LinearRegression** do sci-kit learn e faça o ajuste (**fit**) dos seus dados;
c. Após o ajuste, realize a predição de novos valores usando o modelo que você treinou.
d. Obtenha as métricas de avaliação do modelo (score, erro, coeficientes). O que eles dizem sobre seu modelo?
e. Plote os resultados em um gráfico de dispersão. O que ele diz sobre o seu modelo?

3. Realize os mesmos passos da questão anterior utilizando a técnica de validação cruzada. Para isso, use o **KFold**, **cross_val_score** e **cross_val_predict** do sci-kit learn. Houve mudanças no seu modelo?

4. Avalie ambos os modelos através das métricas **bias**, **variance** e **mse**. Para isso, utilize a função **bias_variance_decomp** da biblioteca **mlxtend**. De acordo com elas, houve **overfitting** ou **underfitting** do seu modelo?