Regularização

Regularização é um método que aplicamos em modelos de Machine Learning para, principalmente, reduzir a variância dos modelos. Portanto, o principal ponto é evitar o overfitting e melhorar a generalização do modelo em situações onde há **multicolinearidade** ou um **grande número** de **variáveis**.

Parte 1: Best Subset Selection

1. Carregamento dos Dados

Nessa etapa, foi feito o carregamento dos dados como pedido no exercício, separando as variáveis preditoras e a variável dependende que queremos prever. Foi utilizado o dataset Boston Housing.

2. Implementação do Best Subset Selection

Best Subset Selection é um método para estimar o quanto as variáveis preditoras do nosso conjunto de dados impacta a variável dependente. Ou seja, busca o subconjunto de variáveis independentes que melhor predizem o resultado. Para isso, considera todas as combinações possíveis de variáveis independentes.

Obtivemos o seguinte gráfico:

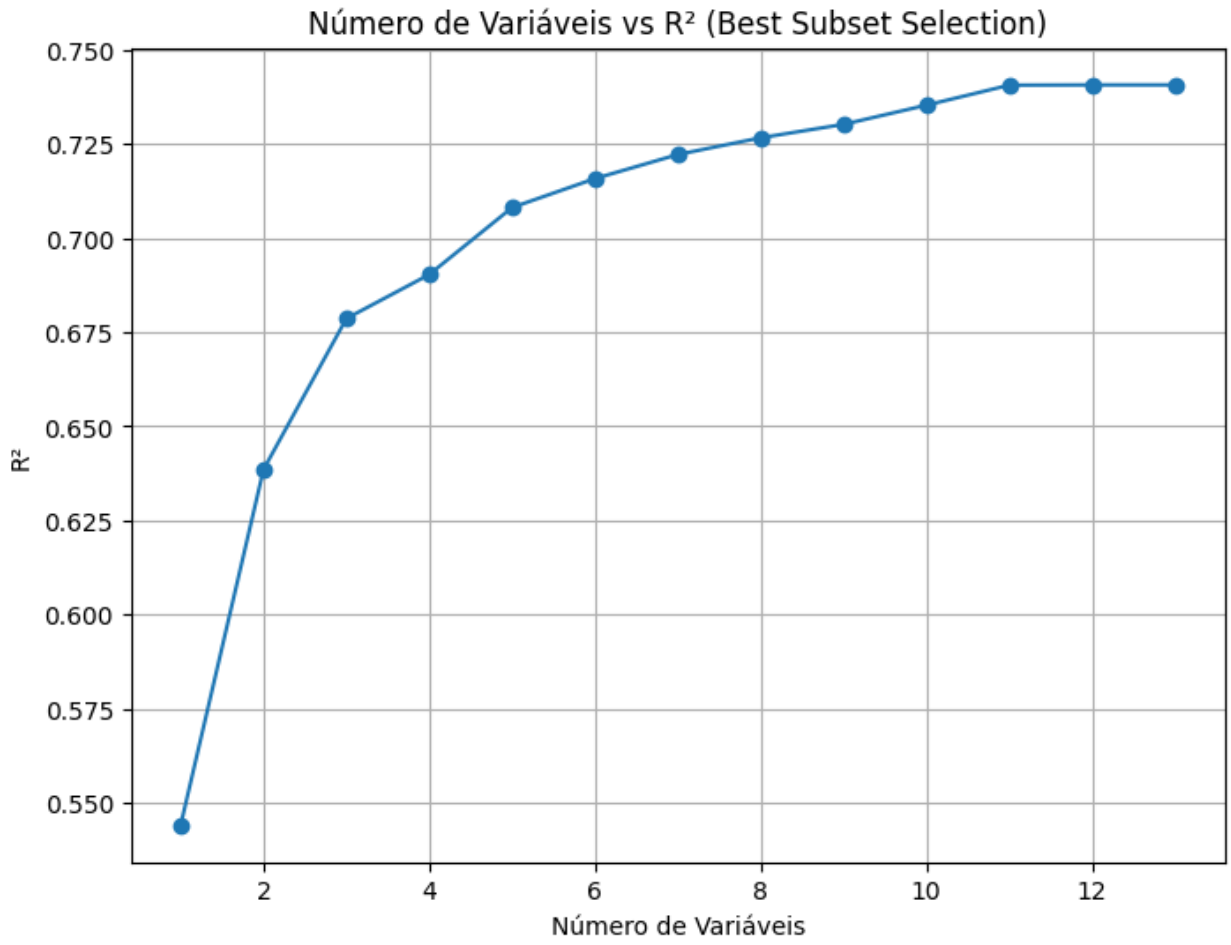

Figura 1. Tempo total de execução: 27 segundos.

3. Validação Cruzada para Subconjuntos Selecionados

A partir disso, foi feita uma validação cruzada para ajudarmos a estimar o número ideal de subconjuntos de variáveis de acordo com o MSE de cada uma (Mean Squared Errror), cujo gráfico está abaixo:

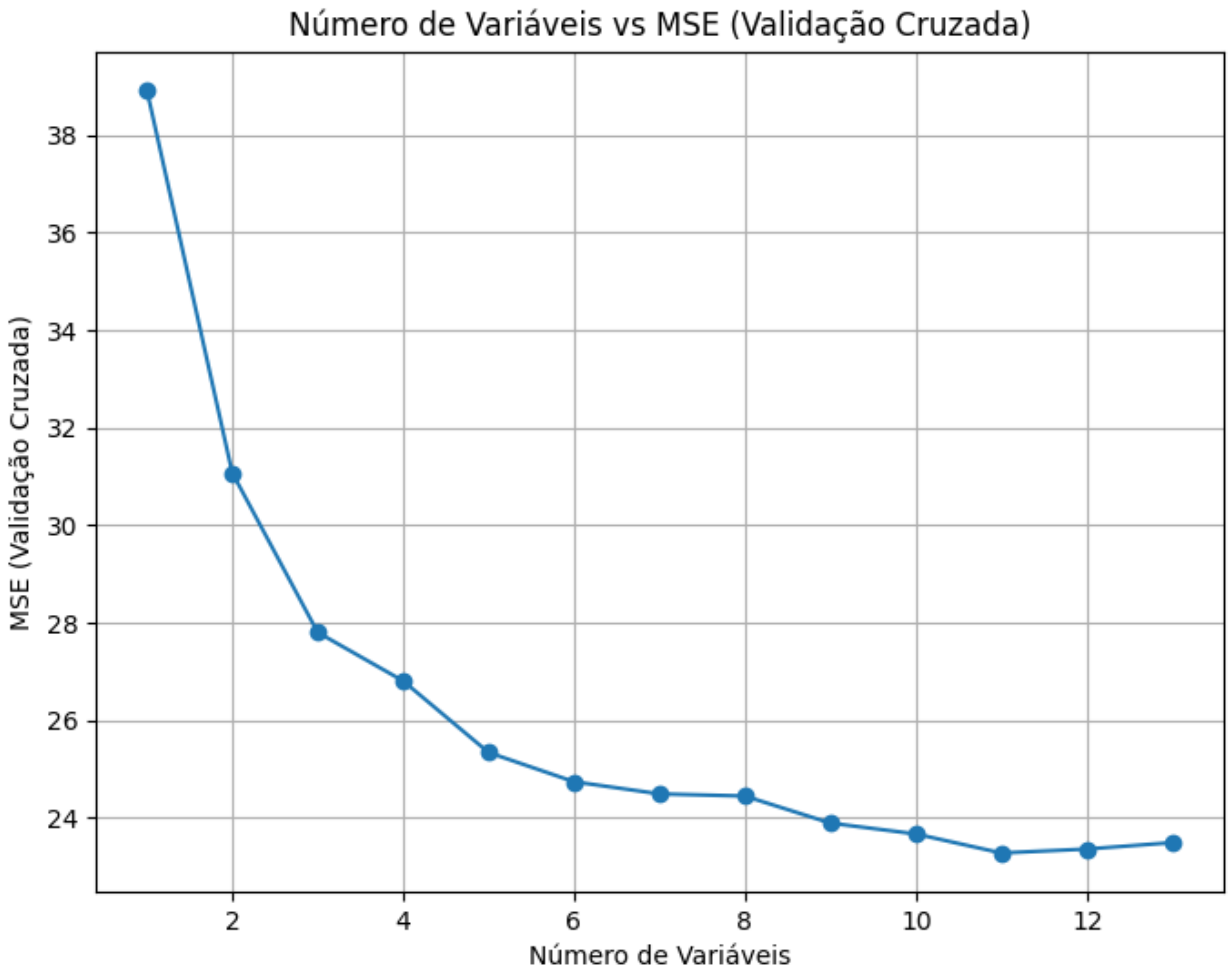

Figura 2.

4. Análise

1. **Levando em conta a simplicidade e eficiência, você escolheria um subconjunto com quantas variáveis? Por que?**

De 6 a 8 variáveis. A partir de 6 variáveis, vemos que o MSE ainda diminui um pouco seu valor, porém tende a se estabilizar, então não é um bom trade-off escolher muito mais que 8 variáveis já que grande parte da redução do MSE já foi realizada.
Além disso, vemos que o
R^2 a partir de 6 variáveis, começa a crescer lentamente e muito pouco, o que

indica que já atingimos uma estabilidade e que nosso modelo já está se ajustando adequadamente a nossas amostras.

2. **Observe o tempo de execução do algoritmo. Em que casos você usaria ele?**

O custo computacional do Best Subset Selection é elevado, já que envolve analisar todos os possíveis subconjuntos de variáveis preditoras para encontrar aquele que maximiza o R^2, nesse caso.

Portanto, o ideal é utiliza-lo para conjunto de dados pequenos.

Parte 2: Regularização em Modelos de Regressão

1. Preparação dos Dados

Nessa etapa, foi utilizado o dataset California Housing. Temos como variável dependente o valor da casa que queremos prever em um distrito da California e diversas variáveis indepentendes referente ao censo realizado na época da análise.

2. Aplicação da Regressão Ridge

Nessa etapa foi feita a aplicação do modelo Ridge Regression no Dataset. Segue a mesma ideia que a regressão linear, porém adiciona uma penalidade para evitar alta variância e overfitting do modelo. A penalização aplicada é L2.

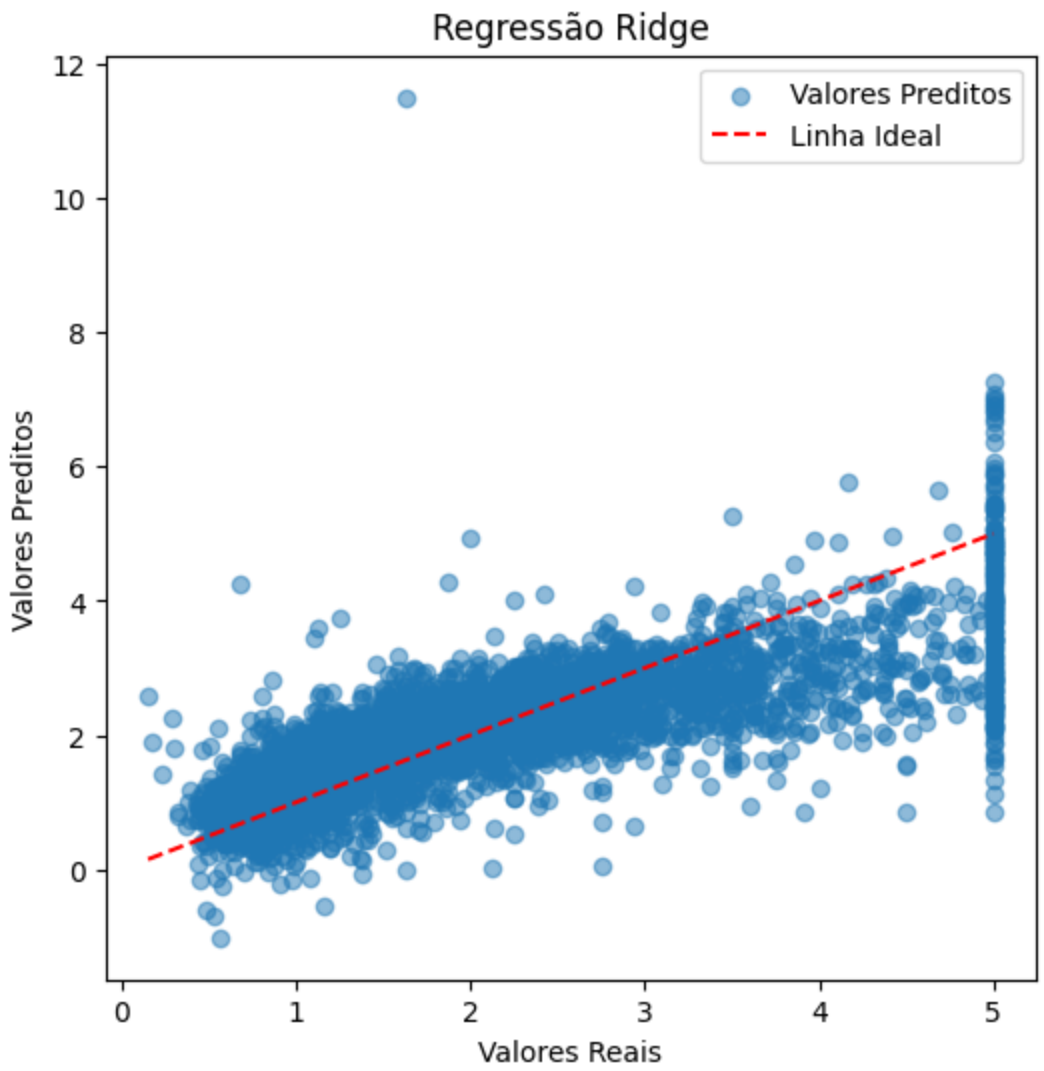

Figura 3. MSE = 0.5559

3. Aplicação da Regressão Lasso

A regressão lasso, diferentemente da regressão ridge, aplica uma penalização L1, mas segue a mesma ideia de evitar overfitting do modelo em comparação com a regressão linear.

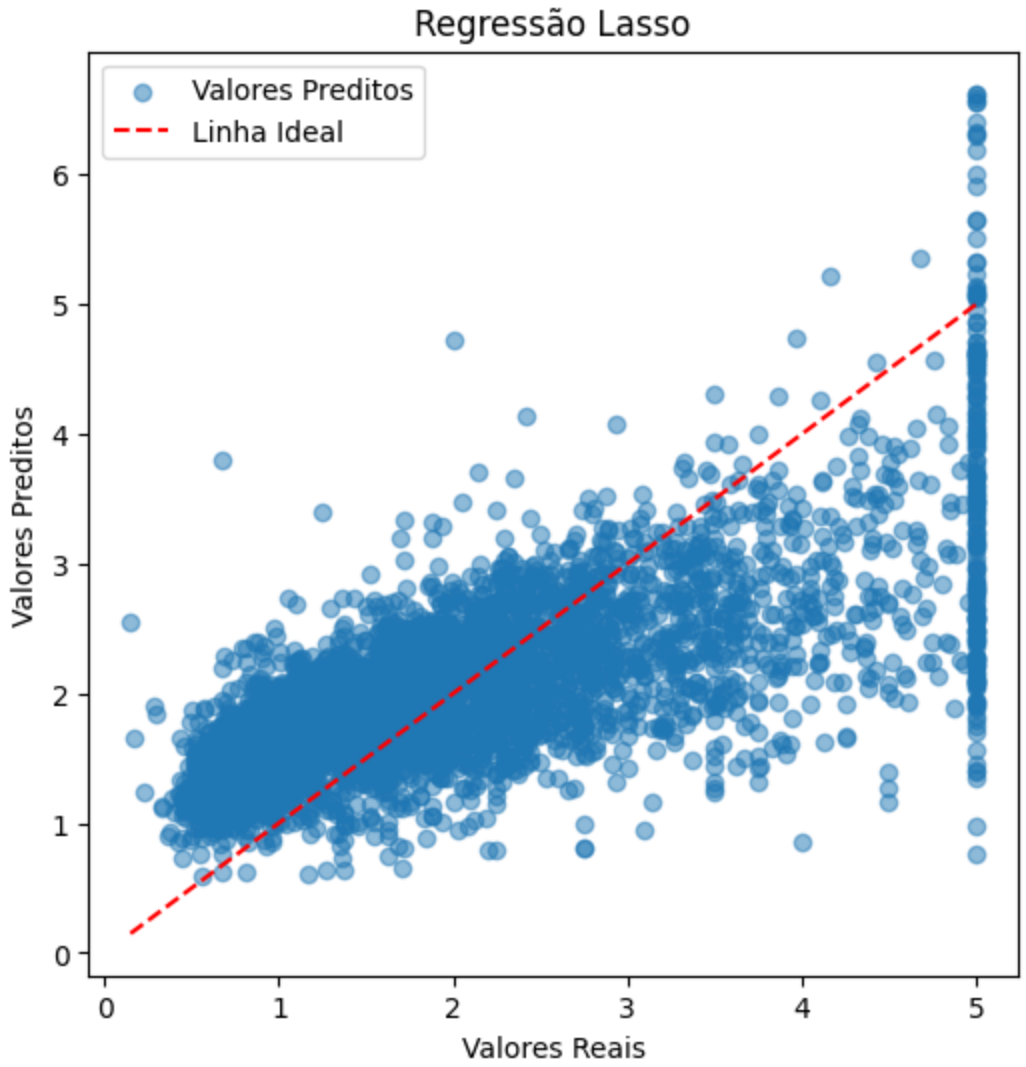

Figura 4. MSE = 0.6657

4. Impacto da Multicolinearidade

Nessa etapa, foi feita a análise da regressão ridge e lasso em relação a diversos valores de correlação entre variáveis de dados sintéticos:

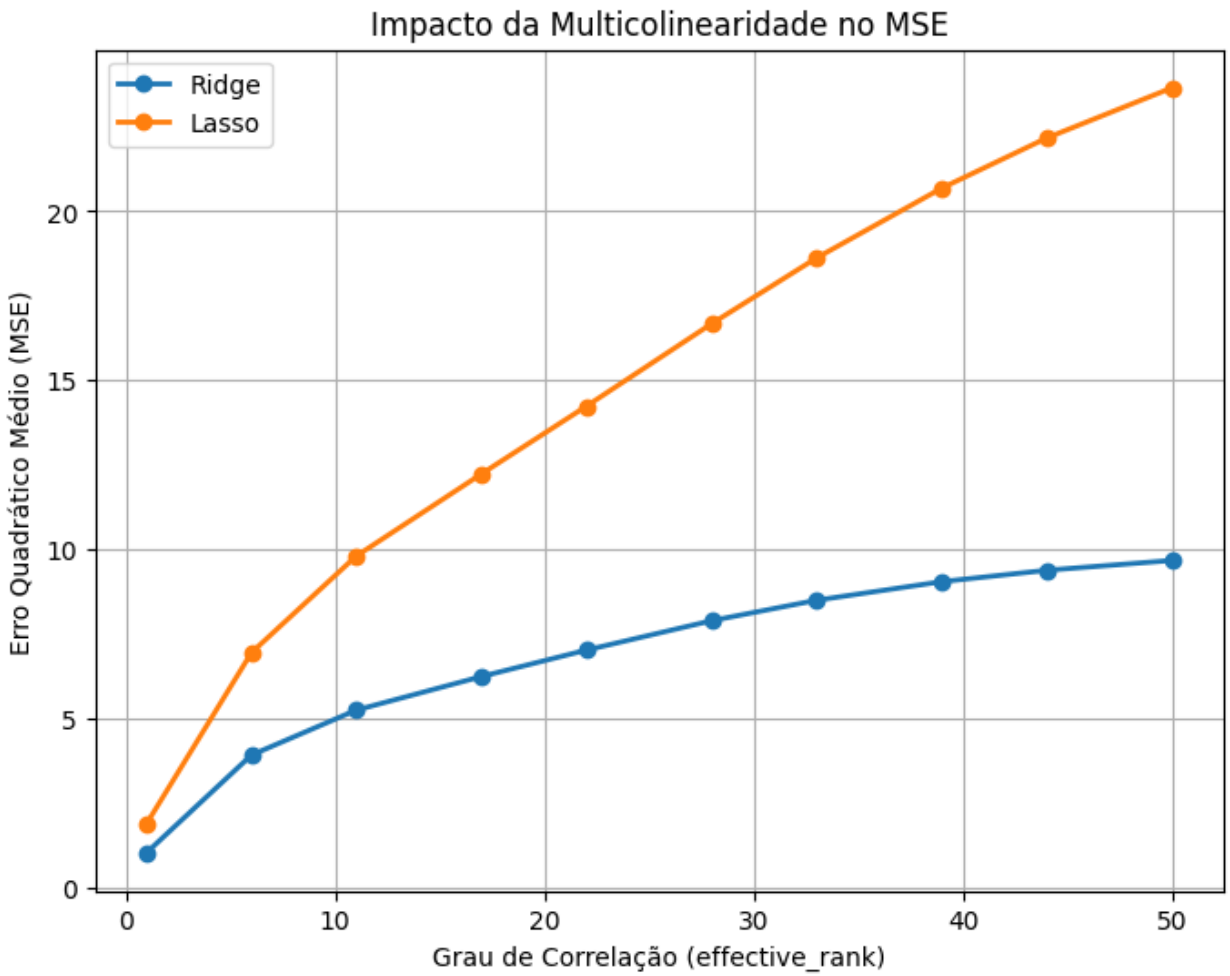

Figura 5.

É perceptível que, a medida que o grau de correlação entre certas variáveis aumenta, o MSE do Lasso, em que a norma de penalização é L1, fica drasticamente maior que a Ridge.

Isso é devido ao fato de o Lasso, ao penalizar certas variáveis do dataset, reduzir drasticamente seus valores a ponto de chegar a 0, fazendo com que certas variáveis sejam eliminadas, é o que chamamos de **seleção de variáveis.** Se as variáveis eliminadas contiverem informações úteis, isso prejudica a explicabilidade do modelo e aumenta o MSE, que foi o nosso caso.

5. Discussão

1. **Qual dos dois modelos se saiu melhor no dataset California Housing?**

A Ridge Regression. Seu MSE foi menor, o que indica que a reta de regressão está levemente melhor ajustada aos dados.

2. **Altere os valores de alpha nos dois modelos para 1 e depois para 10. O que acontece?**

A Regressão Lasso teve uma piora significativa em sua performance, aumetando drasticamente o MSE, enquanto a Ridge permaneceu, majoritariamente, da mesma forma.

Isso é devido ao fato de que, com o aumento de alpha, a Lasso pode reduzir alguns coeficientes a 0, devido a penalização ser L1, o que pode produzir efeito indesejado de underfitting, em que a reta de regressão não explica a variável preditora.

Enquanto isso, a Ridge reduz a magnitude dos coeficientes, mas nunca os reduz exatamente a 0, portanto a penalização é diferente da realizada pela Lasso.

3. **Após analisar o desempenho dos dois métodos com o aumento da colinearidade, a que conclusão você chega?**

Os dois métodos, diferente da Regressão Linear, lidam bem com a colinearidade, adicionando regularização (a penalidade, em cada método) que reduz a magnitude dos coeficientes e, dessa forma, controla o impacto da colinearidade entre variáveis preditorias.

Porém, a forma que cada método lida com a colinearidade é diferente:

Ridge: Com o aumento da colinearidade, o desempenho do modelo de Ridge Regression tende a ser mais estável (já que reduz a magnitude dos coeficientes, mas não zera), com uma melhora na capacidade de generalização (menor erro de validação) em comparação com a Regressão Linear simples.

Lasso: À medida que a colinearidade aumenta, Lasso tende a eliminar algumas variáveis redundantes, o que pode melhorar a interpretabilidade do modelo, mas, em alguns casos, pode prejudicar a capacidade preditiva se as variáveis eliminadas contêm informações úteis, que provavelmente foi o que aconteceu em nosso caso, quando aumentamos o valor de alpha.