**CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA “PAULA SOUZA”**

**FACULDADE DE TECNOLOGIA DE BEBEDOURO**

**TECNOLOGIA EM BIG DATA NO AGRONEGÓCIO**

# PREVISÃO DE *COMMODITY* LARANJA UTILIZANDO MÉTODOS DE REGRESSÃO E CIÊNCIA DE DADOS

**AUTOR: ALEXANDER FERREIRA**

**ORIENTADOR:PROF. DR. RENAN GUILERME NESPOLO**

**BEBEDOURO**

**2025**

ALEXANDER FERREIRA

# PREVISÃO DE *COMMODITY* LARANJA UTILIZANDO MÉTODOS DE REGRESSÃO E CIÊNCIA DE DADOS

Monografia apresentada à Faculdade de Tecnologia de Bebedouro, como parte dos requisitos para a obtenção do título de Tecnólogo em Big Data no Agronegócio

Orientador: **prof. Dr. Renan Guilherme Nespolo**

BEBEDOURO

2025

*Eu falhei muitas e muitas vezes na vida. E é exatamente por isso que eu obtive sucesso.*

MICHAEL JORDAN

# Agradecimentos

Minha filha e minha esposa, por estarem ao meu lado em cada etapa desta caminhada, oferecendo apoio, compreensão e muitas vezes uma palavra de encorajamento quando eu mais precisava e a todos que contribuíram, direta ou indiretamente, para a conclusão deste projeto, meu sincero agradecimento.

Meu orientador, Renan, por sua orientação, apoio e valiosas contribuições ao longo desta jornada. Seu conhecimento e dedicação foram essenciais para a realização deste trabalho.

FERREIRA, ALEXANDER. **Previsão de Commodity Laranja Utilizando Métodos de Regressão e Ciência de Dados**. Trabalho de Graduação (Monografia). Centro Estadual de Educação Tecnológica “Paula Souza”. Faculdade de Tecnologia de Bebedouro. nº 46 p. 2025.

# RESUMO

O mercado de commodities é altamente relevante para a economia global, afetando diretamente setores estratégicos como agricultura, energia e metais. Sua natureza volátil decorre de múltiplos fatores externos, como instabilidades geopolíticas, mudanças climáticas e variações na oferta e demanda, o que torna essencial a análise contínua dos riscos futuros. Este estudo tem como objetivo principal avaliar o comportamento de duas técnicas de regressão — Ridge e LASSO — aplicadas à previsão de preços de commodities, com foco na laranja. Para isso, foram considerados dois elementos adicionais: o uso de pré-processamento com expansão polinomial e o impacto da janela temporal de treino sobre o desempenho dos modelos. A pesquisa foi conduzida utilizando a linguagem Python e a biblioteca scikit-learn, com validação cruzada (K-Fold) e métricas de avaliação como Erro Quadrático Médio (MSE) e Erro Absoluto Médio (MAE). Os resultados indicam que o modelo Ridge apresentou maior robustez frente à multicolinearidade introduzida pela expansão polinomial, enquanto o LASSO demonstrou maior sensibilidade, embora tenha se beneficiado da calibragem via validação cruzada. A análise comparativa entre os métodos revelou que a escolha adequada da técnica de regressão e do pré-processamento pode melhorar significativamente a capacidade de generalização dos modelos. Além disso, o estudo contribui para o debate sobre estratégias de mitigação de riscos em mercados voláteis, oferecendo subsídios práticos para investidores, empresas e formuladores de políticas públicas. Ao compreender os principais fatores que influenciam a volatilidade dos preços, é possível desenvolver modelos preditivos mais eficazes, capazes de apoiar decisões econômicas mais seguras e sustentáveis. Dessa forma, a pesquisa reforça a importância da integração entre métodos estatísticos e tecnologias emergentes na gestão de riscos financeiros em ambientes complexos.

**Palavras-chave:** Agronegócio. Mercado de Commodities. Riscos Futuros. Volatilidade de Preços. Mitigação de Riscos.

FERREIRA, ALEXANDER. **Previsão de Commodity Laranja Utilizando Métodos de Regressão e Ciência de Dados**. Trabalho de Graduação (Monografia). Centro Estadual de Educação Tecnológica “Paula Souza”. Faculdade de Tecnologia de Bebedouro. nº 46 p. 2025.

# ABSTRACT

*The commodity market plays a crucial role in the global economy, directly impacting strategic sectors such as agriculture, energy, and metals. Its volatile nature stems from multiple external factors, including geopolitical instability, climate change, and fluctuations in supply and demand, making the continuous assessment of future risks essential. The primary objective of this study is to evaluate the behavior of two regression techniques — Ridge and* LASSO *— applied to commodity price forecasting, with a specific focus on orange prices. Two additional elements were considered: the use of polynomial feature expansion during preprocessing and the influence of the training window size on model performance. The research was conducted using the Python programming language and the scikit-learn library, employing cross-validation (K-Fold) and evaluation metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE). The results indicate that the Ridge model demonstrated greater robustness in the face of multicollinearity introduced by polynomial expansion, while the* LASSO *model showed higher sensitivity, although it benefited from parameter tuning via cross-validation. The comparative analysis revealed that the appropriate choice of regression technique and preprocessing strategy can significantly enhance the generalization capacity of predictive models. Furthermore, the study contributes to the ongoing discussion on risk mitigation strategies in volatile markets, offering practical insights for investors, companies, and policymakers. By understanding the key factors that influence price volatility, it becomes possible to develop more effective predictive models capable of supporting safer and more sustainable economic decisions. Thus, the research reinforces the importance of integrating statistical methods and emerging technologies in financial risk management within complex environments.*

**Keywords:** *Agribusiness. Commodity Market. Future Risks. Price Volatility. Risk Mitigation.*

# SUMÁRIO

## LISTA DE FIGURAS

# Lista de Tabelas

# Lista de Gráficos

# Lista de Algoritmos

## LISTA DE EQUAÇÕES

## Lista de Siglas

IoT - Internet of Things
IA - Inteligência Artificial
EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária
AP - Agricultura de Precisão
IFRI - International Food Policy Research Institute
GPS -Global Position System
ASBRAAP - Associação Brasileira de Agricultura de Precisão e Digital
FRED - Federal Reserve Bank of St. Louis
MAE - Mean Absolute Error
MSE – Mean Squared Error
ML - Machine Learning
LASSO - Least Absolute Shrinkage and Selection Operator
RIDGE - Ridge Regression
CV – Cross Validation
CSV - Comma Separated Values

# 1 INTRODUÇÃO

O mercado de commodities é vital para a economia global, impactando setores como agricultura, energia e metais. A alta volatilidade dos preços, causada por fatores geopolíticos, mudanças climáticas e instabilidades econômicas, representa desafios significativos para investidores, empresas e governos. Este trabalho analisa os riscos futuros desse mercado, investigando a influência desses fatores na oferta e demanda global. Além disso, explora ferramentas e estratégias de mitigação de riscos, com ênfase em tecnologias emergentes e políticas econômicas. Compreender e prever esses riscos é essencial para tomar decisões informadas e implementar estratégias eficazes de gestão de riscos, contribuindo para o debate acadêmico e prático sobre a resiliência econômica.

Segundo Geman (2005), a precificação de commodities exige modelos que considerem não apenas a dinâmica de oferta e demanda, mas também os riscos sistêmicos associados à especulação e à interdependência entre mercados. Hull (2018) complementa ao afirmar que a gestão de riscos financeiros em instituições depende da capacidade de antecipar movimentos extremos e aplicar estratégias quantitativas robustas.

O mercado de commodities é essencial para a economia global, impactando a produção e distribuição de insumos como petróleo, metais e alimentos. Sua volatilidade, provocada por instabilidade política, crises econômicas e desastres naturais, gera desafios significativos. Este estudo visa compreender essas incertezas futuras e analisar os principais fatores de risco, desenvolvendo ferramentas e estratégias para aumentar a resiliência dos agentes econômicos, como modelos de previsão para a mitigação de riscos.

Analisando os principais fatores de risco que afetam o mercado de commodities, focando na volatilidade dos preços, fatores geopolíticos e mudanças climáticas, a mitigação desses riscos dará ferramentas e estratégias e proporá novas abordagens baseadas no parâmetro de análise criado, através de uma leitura dos números de uma base histórica e atualizada a qual utilizará para a criação da análise de risco futuro proposta. Espera-se que as soluções identificadas e as estratégias propostas ajudem instituições, empresas e governos a tomar decisões mais informadas e eficazes, aumentando a resiliência do setor frente às incertezas futuras. O estudo também busca contribuir para o debate acadêmico, fornecendo uma análise prática de como o modelo de previsão pode ser utilizado para mitigar riscos no mercado de futuro de agro commodities. As principais contribuições do presente trabalho são:

1. Analisar duas técnicas de regressão, de forma temporal, utilizando duas

variáveis: preço de cotação do dólar, e preço de cotação da comodity de laranja.

2. Analisar a influência do pré-processamento expansão polinomial no conjunto de dados.
3. Propor um modelo de inferência com janela temporal reduzida para validação do método.

O presente trabalho está segmentado em: 1 Introdução; 2 Fundamentação Teórica; 3 Métodos; 4 Resultados Esperados; 5 Conclusão; ao final as Referências.

# 2 FUNDAMENTAÇÃO TEÓRICA

Na presente seção são apresentados todos os fundamentos relativos ao presente estudo. A seção está separada nos seguintes tópicos: 2.1 Agricultura e Riscos Financeiros com Base em Commodities, que discute a dinâmica do agronegócio e sua exposição a fatores externos; e 2.2 Análise de Dados, que trata dos principais conceitos relacionados ao uso de métodos estatísticos e computacionais no tratamento e previsão de séries temporais. Dentro de 2.2, são abordados os seguintes subtemas: 2.2.1 Pré-Processamento, 2.2.2 Expansão Polinomial, 2.2.3 Regressão LASSO, 2.2.4 Regressão Ridge, 2.2.5 Métricas de Avaliação e 2.2.6 k-Fold e Cross-Validation.

## 2.1 AGRICULTURA E RISCOS FINANCEIROS COM BASE EM COMMODITIES

A agricultura contemporânea depende fortemente de recursos tecnológicos, financiamento e insumos, o que a insere em um cenário marcado por riscos financeiros amplos e interconectados. A produção voltada para commodities agrícolas, como a laranja, está particularmente exposta às variações de preços no mercado, o que pode afetar diretamente a lucratividade e a sustentabilidade econômica das operações.

> A exposição às flutuações nos preços internacionais das commodities é um fator crucial para o planejamento financeiro dos produtores, influenciando desde as decisões de investimento até as estratégias de comercialização (Fabozzi e Peterson Drake, 2009).

Além da volatilidade dos preços, os produtores agrícolas enfrentam riscos financeiros decorrentes de variáveis macroeconômicas, como as taxas de câmbio e as políticas governamentais. Geman (2005) destaca que as commodities agrícolas são fortemente influenciadas pelas variações cambiais, especialmente em países exportadores como o Brasil, onde a desvalorização da moeda local pode, ao mesmo tempo, elevar a competitividade das exportações e aumentar os custos dos insumos importados. Essa dualidade cria um ambiente de incerteza que exige mecanismos de proteção financeira e estratégias de hedge para mitigar os impactos adversos.

Outro elemento relevante é a incerteza climática, que afeta diretamente a oferta de commodities agrícolas. Eventos climáticos extremos, como secas, geadas e enchentes, podem reduzir a produção e pressionar os preços, ampliando a volatilidade nos mercados. De acordo com Roncoroni *et* al. (2015), essas oscilações, além de impactarem a renda agrícola, afetam toda a cadeia de suprimentos, incluindo transportadoras, indústrias de processamento e exportadores. Nesse sentido, a compreensão dos riscos climáticos é fundamental para o gerenciamento financeiro eficiente na agricultura.

**Figura 1 - Cadeia Financeira da Laranja**

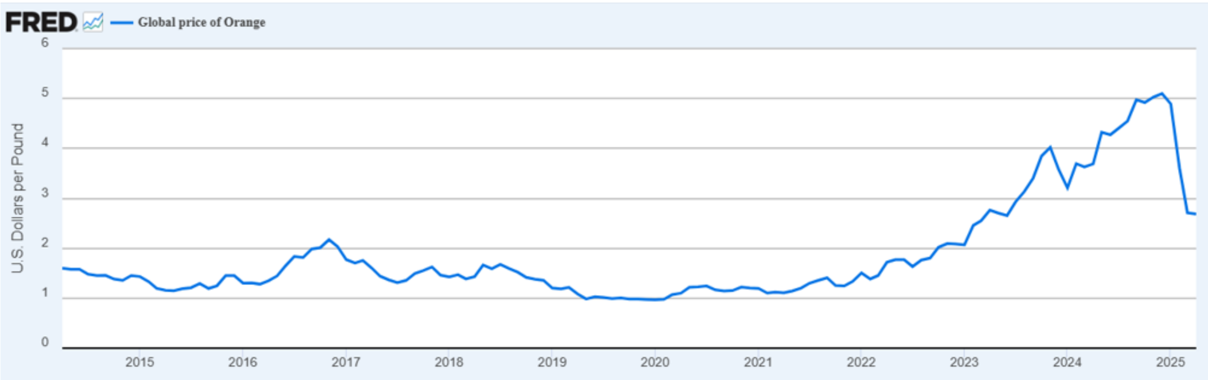

Fonte: Elaborada pelo autor.

A dinâmica de oferta e demanda global também exerce influência significativa sobre os preços das commodities agrícolas, exigindo análise de fatores como estoques globais, políticas de subsídios agrícolas em outros países e mudanças nos padrões de consumo. Hull

(2018) argumenta que o aumento da demanda por commodities, por exemplo, pode alterar a destinação de culturas agrícolas, impactando a disponibilidade de alimentos e elevando a concorrência entre setores. Isso cria riscos financeiros adicionais para os produtores, que precisam se adaptar rapidamente a mudanças estruturais nos mercados.

Diante desse cenário, e com dados do FRED referente a commodities, torna-se essencial a utilização de modelos matemáticos e estatísticos para a análise e a previsão de preços, como forma de apoiar a tomada de decisão e reduzir os riscos financeiros no agronegócio. Modelos como a regressão Ridge e LASSO, por meio da análise de séries temporais de preços, oferecem suporte para o planejamento estratégico, permitindo a identificação de tendências de mercado e a avaliação da exposição a riscos. Assim, a integração de métodos quantitativos no gerenciamento agrícola representa uma ferramenta poderosa para a mitigação dos riscos financeiros associados à produção de commodities.

**Figura 2 - Global Price Of Orange**




Fonte: (FRED, 2025).

## 2.2 ANÁLISE DE DADOS

A análise de dados desempenha um papel central na compreensão dos fenômenos complexos que envolvem o mercado de commodities agrícolas. A utilização de séries temporais, associada a técnicas estatísticas e computacionais, permite examinar padrões históricos, identificar tendências e realizar previsões fundamentadas. Conforme apontado por Morettin e Bussab (2010), o estudo de séries temporais é essencial para interpretar comportamentos ao longo do tempo, especialmente em contextos sujeitos à sazonalidade, volatilidade e fatores externos, como ocorre no setor agrícola.

Com o avanço das tecnologias de ciência de dados, especialmente o uso de algoritmos de aprendizado de máquina, tornou-se possível lidar com volumes massivos de informações e realizar análises mais robustas. Métodos de regressão regularizada, como Ridge e LASSO, têm se mostrado eficazes em ambientes com alta dimensionalidade e multicolinearidade, onde variáveis explicativas estão fortemente correlacionadas entre si (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Essas técnicas aplicam penalizações aos coeficientes de regressão para reduzir o risco sobre ajuste e melhorar a capacidade preditiva dos modelos.

A validação dos modelos preditivos é uma etapa crucial na análise de dados, sendo a validação cruzada do tipo *k-fold* amplamente utilizada para este fim. De acordo com James *et* al. (2013), essa técnica consiste em dividir o conjunto de dados em subconjuntos, garantindo que o treinamento e o teste dos modelos ocorram de forma rotativa, o que proporciona uma estimativa mais fiel da performance em dados não vistos. Esse procedimento é particularmente importante em domínios onde as previsões são utilizadas como suporte para decisões estratégicas.

Outro recurso metodológico relevante é o uso de transformações polinomiais por meio da técnica conhecida como Polynomial Features. Tal abordagem amplia o espaço de características, permitindo aos modelos capturar relações não lineares entre as variáveis envolvidas (VANDERPLAS, 2016). No entanto, essa expansão pode introduzir multicolinearidade nos dados, exigindo o uso de algoritmos de regressão robustos. A combinação entre transformações polinomiais e regularização penalizada proporciona maior flexibilidade e estabilidade ao processo de modelagem.

Em suma, a análise de dados no contexto do agronegócio exige uma abordagem metodológica que una rigor estatístico, técnicas de machine learning e validação robusta. O uso combinado de regressões penalizadas, validação cruzada e transformações polinomiais contribui para a construção de modelos preditivos confiáveis e interpretáveis. Essa estrutura analítica é essencial para promover uma gestão baseada em dados e subsidiar decisões mais seguras em ambientes voláteis e incertos, como os que caracterizam o setor agrícola contemporâneo (GÉRON, 2019; PEDREGOSA *et* al., 2011).

### 2.2.1 PRÉ-PROCESSAMENTO

O pré-processamento de dados é uma etapa fundamental no pipeline de modelagem preditiva, especialmente em contextos com variáveis de escalas distintas e com presença de correlações lineares e não lineares. Segundo Géron (2019), o desempenho de muitos algoritmos de aprendizado de máquina, em especial os baseados em regressão, é sensivelmente afetado pela escala dos dados. Sem esse ajuste, os coeficientes atribuídos às variáveis podem ser desproporcionais, prejudicando a convergência e a estabilidade do modelo.

Entre as técnicas de normalização mais utilizadas estão o StandardScaler (padronização) e o Min-MaxScaler (normalização min-max). A padronização transforma os dados para que tenham média zero e desvio padrão unitário, utilizando a equação 1:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

na qual: $x$ é o valor original da variável, $\mu$ é a média da variável, $\sigma$ é o desvio padrão.

A normalização min-max ajusta os valores para um intervalo específico (geralmente entre 0 e 1), apresentada na equação 2:

$$x' = \frac{x - x_{mim}}{x_{max} - x_{mim}} \tag{2}$$

na qual $x$ é o valor original, $x_{max}$ e $x_{mim}$ são o menor e o maior valor da variável, respectivamente.

Essas técnicas são especialmente úteis em modelos que envolvem regularização, como Ridge e LASSO, os quais são sensíveis à magnitude dos coeficientes (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Ao padronizar os dados, garante-se que todas as variáveis contribuam de forma equivalente na penalização dos parâmetros.

Além da normalização, outra etapa importante no pré-processamento é a expansão polinomial das variáveis preditoras, conhecida como *Polynomial Features*. Essa técnica consiste em criar variáveis elevando os atributos originais a potências superiores ou combinando-os multiplicativamente, permitindo que modelos lineares capturem relações não lineares entre as variáveis (VANDERPLAS, 2016).

Esse tipo de transformação aumenta a complexidade do espaço de atributos, tornando-o mais expressivo, mas também pode introduzir multicolinearidade, exigindo o uso de técnicas de regressão penalizada para manter a estabilidade da modelagem. Portanto, a escolha e aplicação criteriosa dessas técnicas de pré-processamento são decisivas para o sucesso na construção de modelos robustos e interpretáveis.

### 2.2.2 EXPANSÃO POLINOMIAL

A transformação polinomial, também conhecida como expansão polinomial, é uma técnica de pré-processamento utilizada para aumentar a complexidade do modelo, permitindo a representação de relações não lineares entre as variáveis de entrada e a variável alvo. Trata-se de uma abordagem fundamental para a aplicação de modelos lineares em domínios onde as relações entre os dados não seguem um padrão linear simples (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Matematicamente, dado um vetor de entrada $\mathrm{X} = [x_1 , \ x_2 , \dots , x_n]$, a transformação polinomial de grau $d$ consiste em criar todas as combinações possíveis de potências dos atributos originais até o grau $d$ . Por exemplo, com um único atributo $x$, a transformação polinomial de grau 2 resulta na equação 3:

$$X_{\text{poly}} = [1, x, x^2 ] \tag{3}$$

Neste vetor, o termo constante (bias) 1 é incluído para representar o intercepto do modelo linear. Ao se utilizar mais de uma variável, por exemplo $x^1$ e $x^2$ a expansão polinomial de grau 2 gera (equação 4):

$$X_{\text{poly}} = [1, x_1, x_2, x^2, x_1 x_2, x_2^2] \tag{4}$$

Ou seja, além das variáveis originais, são criados termos quadráticos e interativos, que permitem ao modelo capturar efeitos conjuntos e curvaturas nos dados. Com grau 3, a complexidade cresce ainda mais, incluindo termos como $x_1^3, x_1^2 x_2, x_1 x_2^2, x_2^3$, entre outros.

Essa transformação é implementada, por exemplo, pela função *PolynomialFeatures* da biblioteca Scikit-Learn, amplamente utilizada em projetos de machine learning (PEDREGOSA *et* al., 2011). No entanto, é importante observar que a expansão polinomial pode causar aumento exponencial da dimensionalidade do *Dataset* e, consequentemente, multicolinearidade, o que reforça a necessidade do uso de técnicas de regularização como Ridge e LASSO.

### 2.2.3 **REGRESSÃO** LASSO

A regressão LASSO (*Least Absolute Shrinkage and Selection Operator*) é uma técnica de regressão linear que incorpora um termo de penalização baseado na soma dos valores absolutos dos coeficientes. Proposta por Tibshirani (1996), essa abordagem visa solucionar problemas de superajuste (*overfitting*) e seleção de variáveis, sendo particularmente útil quando se trabalha com *Dataset*s de alta dimensionalidade ou com variáveis altamente correlacionadas.

Matematicamente, a função de custo do modelo de regressão LASSO é definida pela equação 5:

$$J(\beta) = \sum_{i=1}^{n} (\mathcal{Y}_i - \hat{\mathcal{Y}}_i) + \lambda \, \|\beta\|_1 \tag{5}$$

na qual: $\mathcal{Y}_i$ é o valor observado, $\hat{\mathcal{Y}}_i$ é o valor previsto, $\lambda$ é o parâmetro de regularização que controla a intensidade da penalização, $\|\beta\|_1$ é a norma L1 dos coeficientes, ou seja, a soma dos valores absolutos dos coeficientes de $\beta$.

O principal diferencial da regressão LASSO é que, ao penalizar a soma dos valores absolutos dos coeficientes, ela possui a capacidade de reduzir alguns coeficientes exatamente a zero, promovendo a seleção automática de variáveis. Esse comportamento torna o LASSO uma ferramenta interpretável e eficiente para construção de modelos parcimoniosos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A escolha do parâmetro $\lambda$ é crítica: valores muito altos podem eliminar variáveis relevantes, enquanto valores muito baixos tornam a penalização irrelevante, aproximando o modelo de uma regressão linear comum. A seleção adequada de $\lambda$ é geralmente realizada por meio de validação cruzada (*cross-validation*), conforme sugerem James *et* al. (2013), buscando o equilíbrio entre viés e variância.

Dessa forma, a regressão LASSO apresenta-se como uma técnica poderosa e versátil na análise de dados, especialmente em contextos em que a redução da dimensionalidade e a robustez preditiva são requisitos fundamentais. Seu uso é amplamente difundido em aplicações de modelagem preditiva no agronegócio, finanças e saúde, em que há necessidade de lidar com grande número de variáveis explicativas.

### 2.2.4 REGRESSÃO RIDGE

A regressão Ridge é uma técnica de regressão linear que adiciona um termo de penalização à função de custo tradicional, com o objetivo de reduzir o sobre ajuste (*overfitting*) e lidar com problemas de multicolinearidade entre as variáveis explicativas. Introduzida por Hoerl e Kennard (1970), essa abordagem é classificada como uma técnica de regressão regularizada, pois penaliza grandes valores dos coeficientes, restringindo-os a valores menores e mais estáveis.

A função de custo da regressão Ridge é definida pela equação 6:

$$J\,(\beta) \;=\; \sum_{i=1}^{n} \left(\mathcal{Y}_i \,-\, \hat{\mathcal{Y}}_i\right)^2 \;+\lambda\; ||\beta||_2^2 \tag{6}$$

na qual: $\mathcal{Y}_i$ é o valor observado, $\hat{\mathcal{Y}}_i$ é o valor previsto, $\lambda$ é o parâmetro de regularização que controla a intensidade da penalização, $||\beta||_2^2$ é a norma L2 dos coeficientes $\beta$.

Diferentemente da regressão LASSO, que pode reduzir coeficientes a zero, a regressão Ridge mantém todos os coeficientes diferentes de zero, porém em valores menores. Isso significa que a Ridge é mais adequada quando todas as variáveis contribuem com algum grau de relevância para o modelo, mesmo que pequenas. Ela é especialmente eficaz em contextos com multicolinearidade, onde duas ou mais variáveis explicativas estão altamente correlacionadas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

O valor de $\lambda$ desempenha papel essencial no ajuste do modelo. Quando $\lambda = 0$, o modelo se reduz a uma regressão linear tradicional. À medida que $\lambda$ aumenta, os coeficientes são penalizados mais severamente. A escolha do valor ideal de $\lambda$ é comumente realizada por meio de validação cruzada, maximizando a capacidade de generalização do modelo em dados não vistos (JAMES *et* al., 2013).

Em síntese, a regressão Ridge é uma ferramenta robusta para a construção de modelos preditivos em contextos com dados correlacionados e de alta dimensionalidade. Sua aplicação contribui para aumentar a estabilidade dos coeficientes estimados, promovendo uma maior capacidade de generalização e menor sensibilidade a ruídos nos dados de entrada.

### 2.2.5 MÉTRICAS DE AVALIAÇÃO

A avaliação do desempenho de modelos de regressão é uma etapa essencial para garantir que os resultados obtidos sejam não apenas ajustados aos dados de treinamento, mas também generalizáveis a novos dados. Entre as métricas mais utilizadas em problemas de regressão estão o Erro Quadrático Médio (MSE) e o Erro Absoluto Médio (MAE), ambas amplamente consolidadas na literatura estatística e de ciência de dados (HYNDMAN; ATHANASOPOULOS, 2018).

O MSE (*Mean Squared Error*) mede a média dos quadrados dos erros entre os valores previstos e os valores observados. A métrica MSE (*Mean Squared Error*) mede o erro médio quadrático entre os valores observados e os valores previstos pelo modelo, sendo calculada conforme a fórmula 7:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

na qual $n$ é o número de observações do conjunto de teste, $y_i$ é o valor observado da variável dependente para a $i$-ésima observação, $\hat{y}_i$ é o valor previsto pelo modelo para a i-ésima observação. O MSE penaliza erros maiores com mais intensidade, pois eleva os resíduos ao quadrado. Isso o torna sensível a outliers, sendo útil quando se deseja penalizar fortemente grandes desvios nas previsões.

Por sua vez, a métrica MAE (*Mean Absolute Error*) representa a média das diferenças absolutas entre os valores observados e os previstos pelo modelo, calculada pela fórmula 8:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{8}$$

na qual $n$ é o número de observações do conjunto de teste, $y_i$ é o valor observado da variável dependente para a *i*-ésima observação, $\hat{y}_i$ é o valor previsto pelo modelo para a i-ésima observação. O MAE apresenta maior robustez frente a valores discrepantes (*outliers*) e fornece uma interpretação mais intuitiva do erro médio cometido pelo modelo (Chai; Dexter, 2014).

Ambas as métricas são complementares: enquanto o MSE destaca discrepâncias maiores, o MAE fornece uma noção mais direta do erro médio esperado. Em modelos de regressão regularizada, como LASSO e Ridge, essas métricas são frequentemente utilizadas em conjunto para comparar configurações de parâmetros, validar modelos com validação cruzada (k-*Fold*) e selecionar a melhor combinação entre complexidade e acurácia (JAMES *et* al., 2013).

Em ambientes de dados ruidosos, o MAE pode oferecer uma avaliação mais estável, ao passo que o MSE é preferido quando se deseja penalizar fortemente previsões imprecisas. Dessa forma, a escolha entre uma ou outra métrica depende do objetivo do modelo e da natureza dos dados analisados.

### 2.2.6 *k-FOLD* E *CROSS-VALIDATION*

A validação cruzada (*cross-validation*) é uma técnica de inferência estatística utilizada para avaliar a capacidade preditiva e a robustez de modelos de regressão. Seu objetivo principal é testar o modelo em diferentes subconjuntos do conjunto de dados original, minimizando o risco sobre ajuste (*overfitting*) e maximizando a generalização do modelo em dados não vistos (KOHAVI, 1995).

A forma mais utilizada dessa técnica é a validação *k-Fold*, que consiste em dividir o conjunto de dados em *k* subconjuntos de tamanho aproximadamente igual, denominados *folds*. Em cada iteração, um dos folds é separado para validação (teste), enquanto os $k - 1$ folds restantes são utilizados para treinamento do modelo. O processo é repetido *k* vezes, de forma que cada fold seja utilizado uma vez como conjunto de teste.

Ao final das *k* iterações, calcula-se a média das métricas de desempenho (como MSE ou MAE), o que fornece uma estimativa mais robusta da performance do modelo. A fórmula geral da média dos erros dado pela equação 9:

$$ERRO_{médio} = \frac{1}{\mathcal{K}} \sum_{i=1}^{\mathcal{K}} ERRO\ (\mathcal{D}_i) \tag{9}$$

na qual $ERRO\ (\mathcal{D}_i)$ é o erro obtido na $i$-ésima iteração ($i$-ésimo *fold*), $\mathcal{K}$ é o número de partições (folds) adotadas. Essa abordagem é particularmente vantajosa em contextos com conjuntos de dados limitados, pois permite que todas as amostras sejam utilizadas tanto no treinamento quanto na validação, promovendo maior aproveitamento da base de dados. Além disso, reduz a variância nas estimativas de erro, sendo superior a métodos mais simples como a validação, em que o conjunto de dados é dividido apenas uma vez.

Por fim, o *k-Fold Cross-Validation* é amplamente utilizado na seleção de hiper parâmetros e na comparação entre diferentes configurações de modelos, especialmente em algoritmos que envolvem regularização (como LASSO e Ridge), onde a calibração adequada do parâmetro $\lambda$ influencia diretamente o desempenho e a capacidade de generalização do modelo (JAMES *et* al., 2013)

**Algoritmo 1: Modelagem de Preços com Ridge e** LASSO

**Entrada**

$\boldsymbol{X}$: variável independente (tempo ou variável explicativa)

$y$: preços históricos da laranja

$\alpha$: parâmetro de regularização

$grau$: grau polinomial (Polynomial Features)

$k$: número de folds (Validação Cruzada)

**Saída**

MSE, MAE (Medidas de Avaliação

Previsão de preços futuros

**Início**

$\boldsymbol{X\_poly}$ ← PolynomialFeatures($\boldsymbol{X}, grau$)

Para cada fold em $k$-Fold(k):

Dividir dados em treino e teste

Para modelo em {Ridge, LASSO}:

Ajustar modelo com (X_train_poly, y_train) usando α

Prever preços y_pred no conjunto de teste

$MSE_{fold} \leftarrow \frac{(y-\hat{y})^2}{n_{test}}$

$MAE_{fold} \leftarrow \frac{|y-\hat{y}|}{n_{test}}$

$MAE, MSE$

**Fim**

Fonte: Elaborado pelo autor.

# 3 MÉTODOS UTILIZADOS

Neste estudo, as análises foram conduzidas por meio da linguagem de programação Python, utilizando a biblioteca scikit-learn para desenvolver os modelos de regressão LASSO e Ridge. O ambiente de desenvolvimento adotado foi o Visual Studio Code, amplamente reconhecido entre profissionais da área por sua versatilidade, compatibilidade com extensões e integração eficiente com ferramentas voltadas à ciência de dados. A performance dos modelos foi avaliada com base nas métricas Mean Squared Error (MSE) e Mean Absolute Error (MAE), que possibilitam mensurar o grau de precisão das estimativas geradas pelos algoritmos de regressão.

A etapa de pré-processamento dos dados foi realizada com a biblioteca preprocessing do scikit-learn, incluindo a transformação dos dados por meio da técnica Polynomial Features. Segundo Hastie, Tibshirani e Friedman (2009, p. 189), "os modelos de regressão polinomial estendem os modelos lineares ao considerar transformações polinomiais dos preditores. Isso permite uma modelagem mais flexível de relações não lineares, mantendo a interpretabilidade". Essa abordagem permite a criação de novas variáveis polinomiais a partir das originais, aumentando a capacidade dos modelos de capturar relações não lineares entre os dados. Além disso, foram aplicados métodos de validação cruzada, como k-fold cross-validation, para garantir uma avaliação mais robusta dos modelos, reduzindo o risco de overfitting e proporcionando uma estimativa mais confiável do desempenho preditivo.

> Os modelos de regressão polinomial estendem os modelos lineares ao considerar transformações polinomiais dos preditores. Isso permite uma modelagem mais flexível de relações não lineares, mantendo a interpretabilidade." (Hastie; Tibshirani; Friedman, 2009).

Os modelos de regressão LASSO e Ridge foram treinados e testados utilizando diferentes configurações de pré-processamento. Inicialmente, os modelos foram ajustados sem a aplicação de Polynomial Features, permitindo uma análise comparativa entre a regressão linear tradicional e a versão polinomial. Em seguida, os mesmos modelos foram treinados com a transformação polinomial, verificando o impacto da inclusão de características não lineares na precisão das previsões. A validação cruzada foi aplicada em ambas as abordagens, utilizando fold e k-fold, garantindo uma avaliação estatisticamente significativa dos resultados.

Para a implementação dos modelos, o código foi estruturado de forma modular, permitindo a reutilização de funções e facilitando a análise dos resultados. A biblioteca metrics do scikit-learn foi utilizada para calcular as métricas MSE e MAE, possibilitando a comparação entre os diferentes modelos e configurações. A escolha dessas métricas se deve à sua ampla utilização na literatura científica para avaliação de modelos de regressão, sendo o MSE sensível a grandes erros e o MAE uma métrica mais robusta contra outliers.

Por fim, os resultados obtidos foram analisados e comparados, verificando a influência do pré-processamento e da escolha do modelo na precisão das previsões. A utilização do Visual Studio Code como ambiente de desenvolvimento proporcionou uma experiência eficiente na escrita e execução do código, permitindo a integração com bibliotecas externas e facilitando a depuração dos modelos. A abordagem metodológica adotada garantiu uma análise detalhada dos impactos da regressão LASSO e Ridge, contribuindo para a compreensão das vantagens e limitações de cada técnica.

**Tabela 1: Síntese dos Algoritmos e Técnicas Utilizadas**

| Elemento | Descrição | Referência |
|---|---|---|
| Algoritmo Ridge | Técnica de regressão penalizada que minimiza o erro quadrático médio com penalização na magnitude dos coeficientes (regulariza). Útil em casos de multicolinearidade. | Hastie; Tibshirani; Friedman (2009) |
| Algoritmo LASSO | Regressão penalizada que, além da regularização, realiza seleção de variáveis, podendo zerar coeficientes irrelevantes. Facilita interpretação do modelo. | James *et* al. (2013) |
| Métricas MSE e MAE | MSE penaliza erros maiores, enquanto MAE mede a média dos erros absolutos, sendo menos sensível a outliers. Essenciais para avaliar acurácia e robustez do modelo. | Hyndman; Athanasopoulos (2018) |
| Validação Cruzada (KFold) | Divide os dados em subconjuntos (folds) para múltiplas iterações de treinamento e teste, garantindo robustez e evitando overfitting. | Pedregosa *et* al. (2011); James *et* al. (2013) |
| Polynomial Features | Pré-processamento que amplia o espaço de características, permitindo capturar relações não lineares entre variáveis preditoras (ex.: preço da laranja e dólar). | VanderPlas (2016); Hastie *et* al. (2009) |

Fonte: Elaborado pelo autor.

## 3.1 MÉTODOS DE ANÁLISES

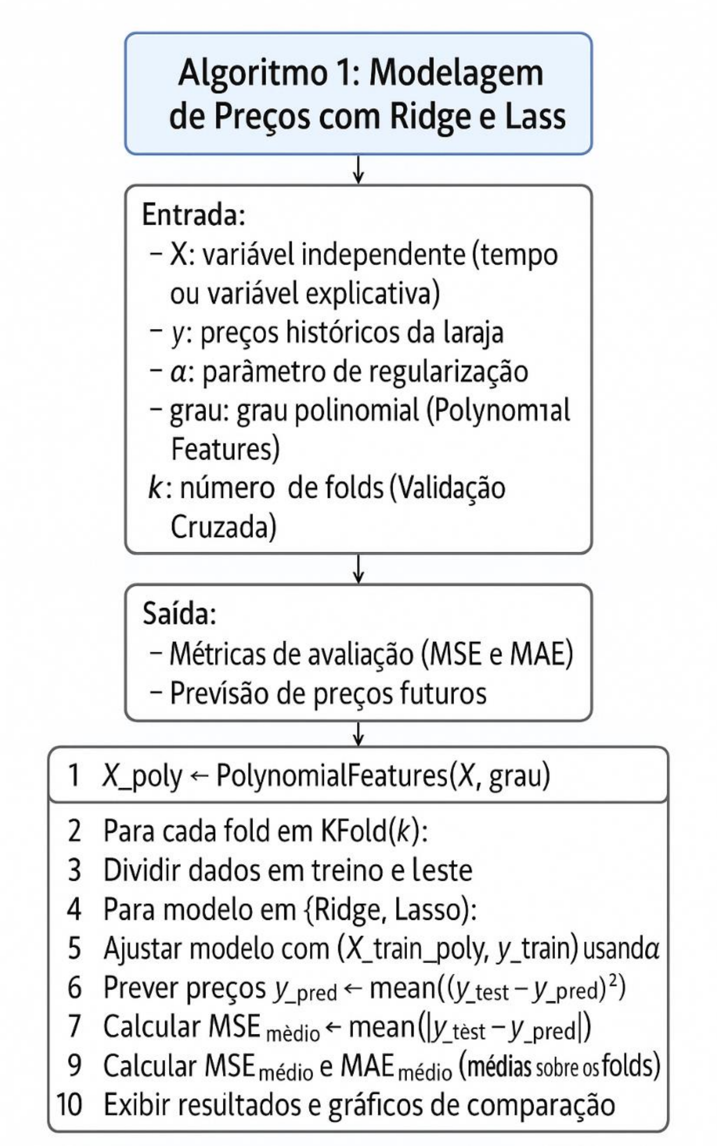
O código desenvolvido para análise das séries temporais de preços de commodities foi implementado na linguagem de programação Python, utilizando o ambiente de desenvolvimento Visual Studio Code. Foram empregadas as bibliotecas NumPy e pandas para manipulação de dados e criação de matrizes de entrada para os modelos de regressão. A biblioteca matplotlib foi utilizada para a visualização gráfica dos resultados, oferecendo suporte à interpretação visual das curvas ajustadas. Essas bibliotecas são reconhecidas pela comunidade científica como ferramentas robustas para análise e modelagem estatística de dados (MCKINNEY, 2017; OLIPHANT, 2006).

A implementação dos modelos de regressão LASSO e Ridge foi realizada por meio das classes LASSO e Ridge da biblioteca scikit-learn. Ambas as classes permitem ajustar modelos de regressão linear com regularização, utilizando o parâmetro alpha para controlar a intensidade da penalização aplicada aos coeficientes do modelo. Essa abordagem visa reduzir o risco sobre ajuste (overfitting) ao penalizar grandes variações nos parâmetros de regressão, conferindo maior estabilidade preditiva ao modelo (PEDREGOSA *et* al., 2011).

Para lidar com a não linearidade potencialmente presente na série temporal, foi utilizada a classe Polynomial Features, também da biblioteca scikit-learn. Essa classe realiza a transformação polinomial da variável preditora, criando termos de ordem superior, o que permite ao modelo ajustar curvas mais complexas aos dados. Essa técnica é particularmente útil quando se deseja capturar efeitos sazonais ou tendências não lineares nos preços das commodities (JAMES *et* al., 2013).

A avaliação dos modelos LASSO e Ridge foi realizada com validação cruzada k-fold, que consiste em dividir o conjunto de dados em k subconjuntos aproximadamente iguais, treinando o modelo em k-1 partes e testando na parte restante. Essa abordagem foi implementada explicitamente no código com loops for organizando as partições de treino e teste de forma manual, permitindo avaliar o desempenho do modelo em diferentes cortes temporais. Foram calculadas as métricas de desempenho Mean Squared Error (MSE) e Mean Absolute Error (MAE) para cada fold, utilizando somatórios de erros absolutos e quadráticos, respectivamente, implementados diretamente em loops de repetição. Essa implementação confere flexibilidade ao processo de avaliação, permitindo ajustes específicos para séries temporais (HYNDMAN; ATHANASOPOULOS, 2018).

**Figura 3: Comparação Algoritmos Ridge e** LASSO




Fonte: Elaborado pelo autor.

Os gráficos gerados ao final dos códigos foram fundamentais para comparar visualmente as previsões dos modelos LASSO e Ridge em relação aos dados reais da série temporal. Utilizando a função matplotlib.pyplot, foram plotados os pontos reais (scatter plot) e as curvas ajustadas (line plot) para cada modelo. Essa visualização permite verificar a aderência do modelo aos dados e observar possíveis discrepâncias entre os valores previstos e observados. Essa etapa é essencial para a comunicação dos resultados e para a validação qualitativa da metodologia empregada, em consonância com as boas práticas de análise estatística (GÉRON, 2019).

## 3.2 *DATASET*

O conjunto de dados utilizado nos códigos foi obtido a partir de arquivos CSV contendo informações sobre os preços médios globais da commodity laranja (AverageGlobalPriceOrange). Esse *Dataset* possui colunas que representam, essencialmente, as datas de observação (no formato ano-mês-dia) e os respectivos valores médios de preço da commodity. No código, a leitura dos dados foi feita utilizando a biblioteca pandas (pd.read_csv), que permite carregar e manipular tabelas de dados de forma eficiente, com suporte para operações como filtragem, ordenação e conversão de tipos de dados (MCKINNEY, 2017).

O *Dataset* utilizado neste estudo consiste em uma série temporal composta por registros históricos de preços médios mensais de uma commodity agrícola ao longo de um período superior a dez anos. Os dados foram obtidos de fonte confiável e reconhecida internacionalmente, apresentando formato estruturado e granularidade mensal, o que permite a realização de análises sazonais e de longo prazo.

O conjunto de dados é composto por duas colunas principais: a primeira, chamada *data*, apresenta as datas de observação no formato ano-mês-dia (YYYY-MM-DD), sendo utilizada para fins de indexação temporal; já a segunda, denominada *valor_commodities*, refere-se aos preços médios globais praticados no mercado internacional para determinada commodity, expressos em moeda padrão, ou seja, em dólares norte-americanos por unidade de medida.

Adicionalmente, durante o pré-processamento, foi criada uma nova variável denominada mes_num que representa o número de meses decorridos desde a primeira observação da série. Essa transformação converte as datas para uma escala numérica contínua, facilitando a aplicação de algoritmos de regressão e de expansão polinomial.

O *Dataset* é univariado, ou seja, contém uma única variável preditora (tempo) e uma variável dependente (preço), sem atributos adicionais como dados climáticos, macroeconômicos ou cambiais. Essa escolha metodológica visa isolar a relação entre tempo e preço, servindo como base para a aplicação de técnicas de regressão regularizada, sem interferência de fatores exógenos.

Além disso, o *Dataset* encontra-se livre de valores ausentes, apresenta formato compatível com bibliotecas de análise de dados e foi padronizado para garantir consistência nos experimentos. A estrutura simples, porém, consistente, permite aplicar modelos estatísticos com maior controle interpretativo e avaliar o impacto de diferentes estratégias de modelagem sobre a previsão de séries temporais.

## 3.3 DESENVOLVIMENTO E VALIDAÇÃO

O desenvolvimento dos modelos de regressão para previsão do preço da commodity foi realizado utilizando a biblioteca scikit-learn, amplamente reconhecida por sua eficiência e flexibilidade em aprendizado de máquina (PEDREGOSA *et* al., 2011). Foram aplicados os métodos de regressão Ridge e LASSO, que são técnicas de regularização capazes de reduzir o overfitting e melhorar a generalização dos modelos, penalizando a magnitude dos coeficientes das variáveis preditoras (TIBSHIRANI, 1996; HOEFING *et* al., 2010). Essa regularização é fundamental para evitar que o modelo ajuste excessivamente os dados de treinamento, o que comprometeria sua performance em dados novos.

Para capturar relações não lineares presentes nos dados temporais, foi utilizado o pré-processamento com Polynomial Features, que transforma as variáveis originais em um espaço de variáveis polinomiais de grau definido, permitindo ao modelo aprender curvas e tendências mais complexas (HASTIE *et* al., 2009). Essa técnica amplia a capacidade dos modelos lineares tradicionais ao incorporar termos polinomiais sem necessidade de alterar a estrutura básica do modelo, possibilitando um melhor ajuste das variações no preço da commodity ao longo do tempo.

A validação dos modelos foi conduzida por meio da técnica de validação cruzada do tipo *k-Fold*, com *k*=10, que consiste em dividir o conjunto de dados em 10 partes e iterativamente treinar o modelo em 9 partes enquanto valida na parte restante, garantindo uma avaliação mais robusta da capacidade preditiva (KOHAVI, 1995). Essa abordagem é amplamente adotada para mitigar o viés e a variância na estimativa do desempenho do modelo, sobretudo em *Dataset*s com tamanho limitado, como o utilizado neste estudo. As métricas principais para avaliação foram o Erro Quadrático Médio (MSE) e o Erro Absoluto Médio (MAE), que fornecem indicadores complementares da precisão das previsões, sendo o MSE mais sensível a grandes erros e o MAE mais interpretável na média das diferenças absolutas (CHAI; DEXTER, 2014).

Além da validação k-Fold, também foi considerada a abordagem 1-Fold, comumente conhecida como *hold-out*, em que o conjunto de dados é dividido apenas uma vez em dois subconjuntos distintos: um para treino e outro para teste, na proporção de treinamento e validação. Embora seja uma técnica mais simples e menos custosa computacionalmente, o *hold-out* pode apresentar maior sensibilidade à forma como os dados são divididos, resultando em estimativas de desempenho mais sujeitas a variabilidade (HAN; KAMBER; PEI, 2011). Ainda assim, sua aplicação inicial é útil como uma linha de base para comparação com métodos mais robustos, como o *k-Fold*, permitindo observar o ganho em estabilidade e representatividade trazido por uma avaliação cruzada.

Os resultados obtidos indicam que tanto os modelos Ridge quanto LASSO, quando combinados com transformações polinomiais, apresentam desempenho superior em comparação às versões sem o uso dessas transformações, refletindo a importância de capturar a não linearidade dos dados para melhorar a predição do preço da commodity. A análise comparativa dos modelos foi realizada com base nas médias das métricas de erro extraídas da validação cruzada, permitindo selecionar os melhores parâmetros de regularização (alpha) e grau polinomial que equilibram o ajuste e a complexidade do modelo, conforme sugerido em estudos de regressão regularizada (HASTIE *et* al., 2009).

Portanto, o desenvolvimento e validação aplicados neste trabalho utilizam práticas consolidadas no campo de aprendizado de máquina para construir modelos preditivos eficientes e generalizáveis. A combinação de regressão regularizada com expansão polinomial e validação cruzada assegura que as previsões sejam robustas e aplicáveis a cenários reais, contribuindo para uma melhor compreensão e gestão dos riscos financeiros associados à commodity estudada. A metodologia adotada está alinhada com os princípios de modelagem preditiva descritos na literatura (PEDREGOSA *et* al., 2011; TIBSHIRANI, 1996; HASTIE *et* al., 2009).

# 4 RESULTADOS ESPERADOS

Os resultados obtidos evidenciam que o uso da transformação polinomial no pré-processamento dos dados impacta diretamente no desempenho dos modelos de regressão Ridge e LASSO. A expansão polinomial gera novas variáveis que permitem capturar relações não lineares entre os preditores e a variável alvo, porém, aumenta a multicolinearidade, o que influencia de maneira distinta cada abordagem. Observou-se que o Ridge apresentou maior robustez diante dessa complexidade, oferecendo previsões mais consistentes e suavizadas, enquanto o LASSO apresentou oscilações mais pronunciadas, reflexo de sua tendência a eliminar coeficientes relevantes quando há elevada correlação entre variáveis (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

No que se refere ao método de validação, os modelos analisados por meio de validação cruzada com 10 divisões (KFold) demonstraram desempenho superior em relação à validação do tipo holdout (1 divisão). A técnica de validação cruzada proporciona uma estimativa mais confiável do erro de generalização, reduzindo a variabilidade das métricas de avaliação e mitigando o risco de sobreajuste (KOHAVI, 1995). Nesse cenário, o LASSO foi mais favorecido pela validação cruzada, pois a calibragem do parâmetro de regularização alpha tornou-se mais precisa, resultando em modelos mais enxutos e com maior capacidade de generalização. Por outro lado, o Ridge manteve sua estabilidade mesmo sob validação simples, devido à sua natureza regularizadora mais tolerante à multicolinearidade.

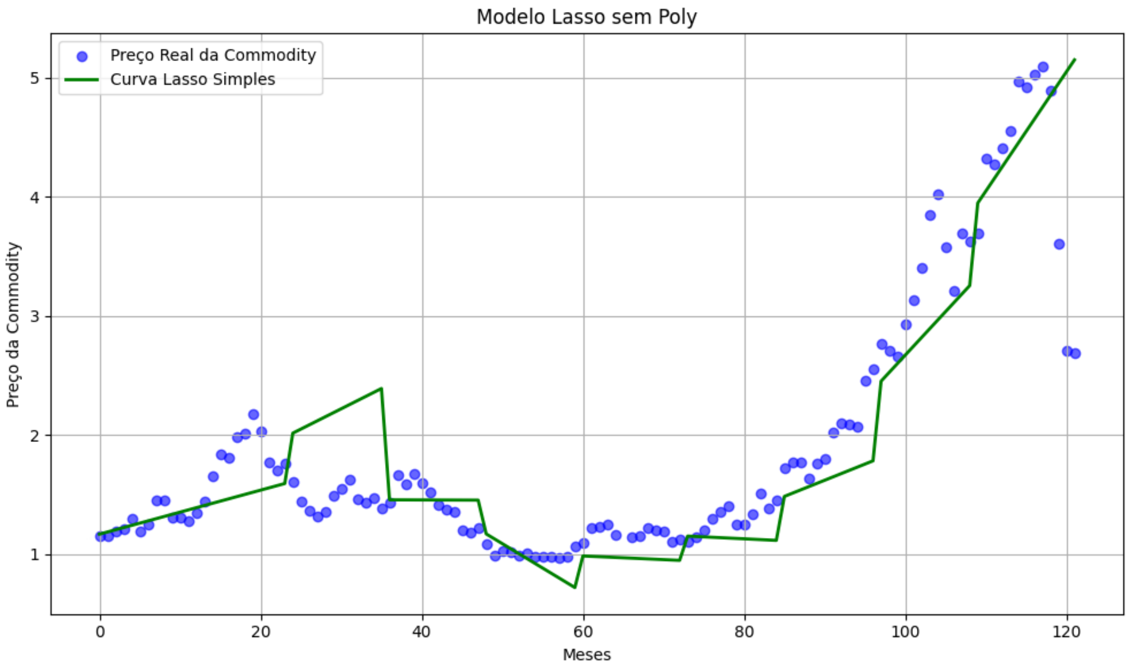
**Tabela 2: Erros modelos Ridge e** LASSO

| **Discriminação dos resultados dos modelos de regressão Ridge e** LASSO | | | |
|---|---|---|---|
| **Modelo** | **Validação** | **Expansão Polinomial** | **Observações Técnicas** |
| **Ridge** | *Fold* | Sem | Estável, sem seleção eficaz |
| **Ridge** | *Fold* | Com | Robusto, polinômio melhora desempenho. |
| **Ridge** | *k-Fold* | Sem | Generalização e estabilidade aprimoradas. |
| **Ridge** | *k-Fold* | Com | Desempenho superior, suaviza penalização. |
| LASSO | *Fold* | Sem | Reduz complexidade, seleciona variáveis. |
| LASSO | *Fold* | Com | Expansão reduz desempenho, sensível. |
| LASSO | *k-Fold* | Sem | Melhor otimização, modelo parcimonioso. |
| LASSO | *k-Fold* | Com | Validação útil, instabilidade persiste. |

Fonte: Elaborado pelo autor.

A Tabela referente ao desempenho das regressões Ridge e LASSO apresenta uma análise qualitativa das respostas obtidas por ambos os modelos, considerando os cenários com e sem aplicação da expansão polinomial, além dos métodos de validação simples (holdout ou 1-Fold) e validação cruzada (K-Fold). Verifica-se que o modelo Ridge demonstrou maior resistência à multicolinearidade gerada pela transformação polinomial, especialmente quando associado à validação cruzada, resultando em estimativas mais consistentes e suavizadas. Em contrapartida, o modelo LASSO revelou maior sensibilidade a essa modificação, apresentando variações mais acentuadas quando a expansão foi aplicada, embora tenha se beneficiado substancialmente da validação cruzada, gerando modelos mais enxutos e com melhor capacidade de adaptação a novos dados. Assim, os achados reforçam evidências anteriores da literatura, como as de Hastie, Tibshirani e Friedman (2009), que destacam o comportamento distinto das técnicas de regularização diante da complexidade dos dados.

**Gráfico 1: Desempenho do Modelo** LASSO

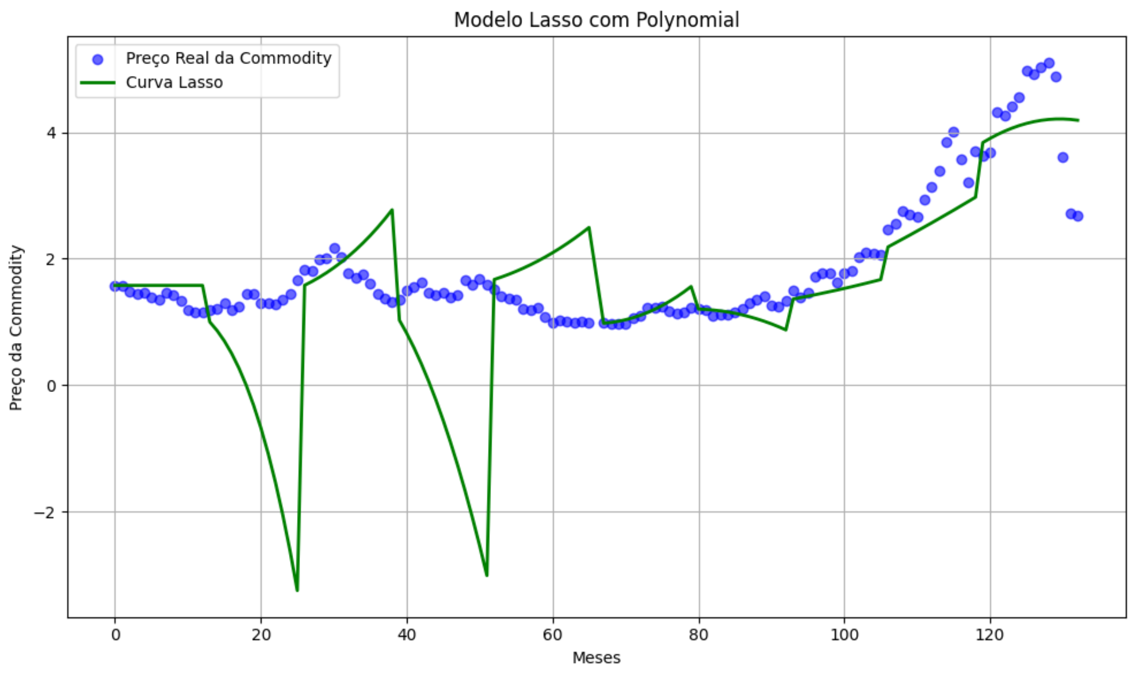



Fonte: Elaborado pelo autor.

O Gráfico 1 apresenta a previsão dos preços da commodity laranja utilizando o modelo de regressão LASSO, sem a aplicação da transformação polinomial (PolynomialFeatures).

Verifica-se que os pontos azuis correspondem aos valores observados dos preços ao longo dos meses, enquanto a linha verde representa a estimativa gerada pelo modelo. A curva projetada apresenta flutuações abruptas e segmentadas, um comportamento típico do LASSO em cenários marcados por elevada multicolinearidade e ausência de enriquecimento de atributos. Essa resposta está relacionada à penalização L1 característica do LASSO, que tende a anular coeficientes, realizando uma seleção de variáveis mais rigorosa. Apesar de o modelo conseguir acompanhar a tendência geral dos dados, sua limitação em capturar relações não lineares afeta a suavidade da curva ajustada. Dessa forma, o desempenho observado reforça a relevância da transformação polinomial como estratégia para aprimorar a capacidade do modelo de refletir as variações reais dos preços, conforme apontado por Hastie, Tibshirani e Friedman (2009).

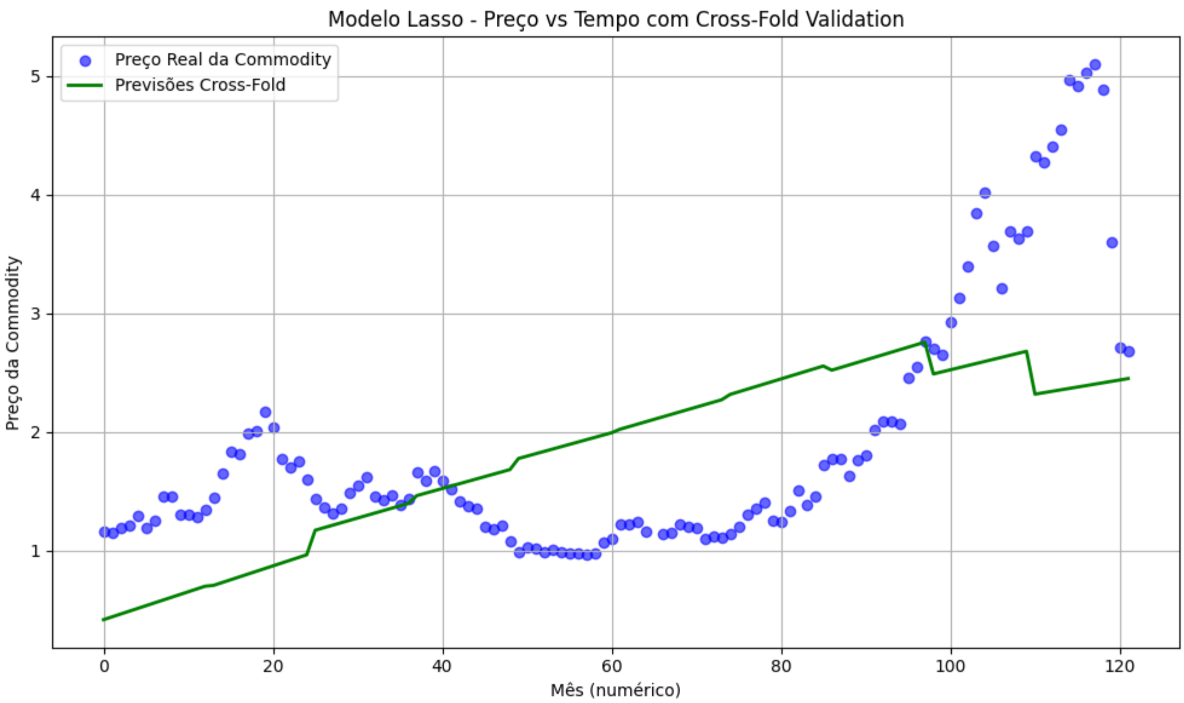
**Gráfico 2: Desempenho do Modelo** LASSO **com *PolynomialFeatures***




Fonte: Elaborado pelo autor.

O Gráfico 2 ilustra o desempenho do modelo de regressão LASSO com a aplicação da transformação polinomial (PolynomialFeatures), tendo como objetivo capturar relações não lineares entre as variáveis. Os pontos azuis indicam os valores reais dos preços da commodity ao longo do tempo, enquanto a linha verde representa a curva ajustada pelo modelo. Embora a expansão polinomial aumente a complexidade do espaço de atributos, permitindo ao modelo identificar padrões mais sofisticados, observa-se que o LASSO apresentou oscilações abruptas e desvios negativos acentuados. Isso evidencia uma instabilidade na modelagem, possivelmente causada pela multicolinearidade introduzida com a transformação e pela penalização L1, que pode zerar coeficientes relevantes. Em comparação com a versão sem transformação, o modelo não conseguiu melhorar a aderência às variações reais, apresentando pior desempenho em determinados intervalos. Esses resultados corroboram a literatura, indicando que o LASSO tende a ser mais sensível a conjuntos de dados expandidos com alta correlação entre variáveis (HASTIE; TIBSHIRANI; FRIEDMAN, 2009), exigindo calibração criteriosa dos hiperparâmetros.

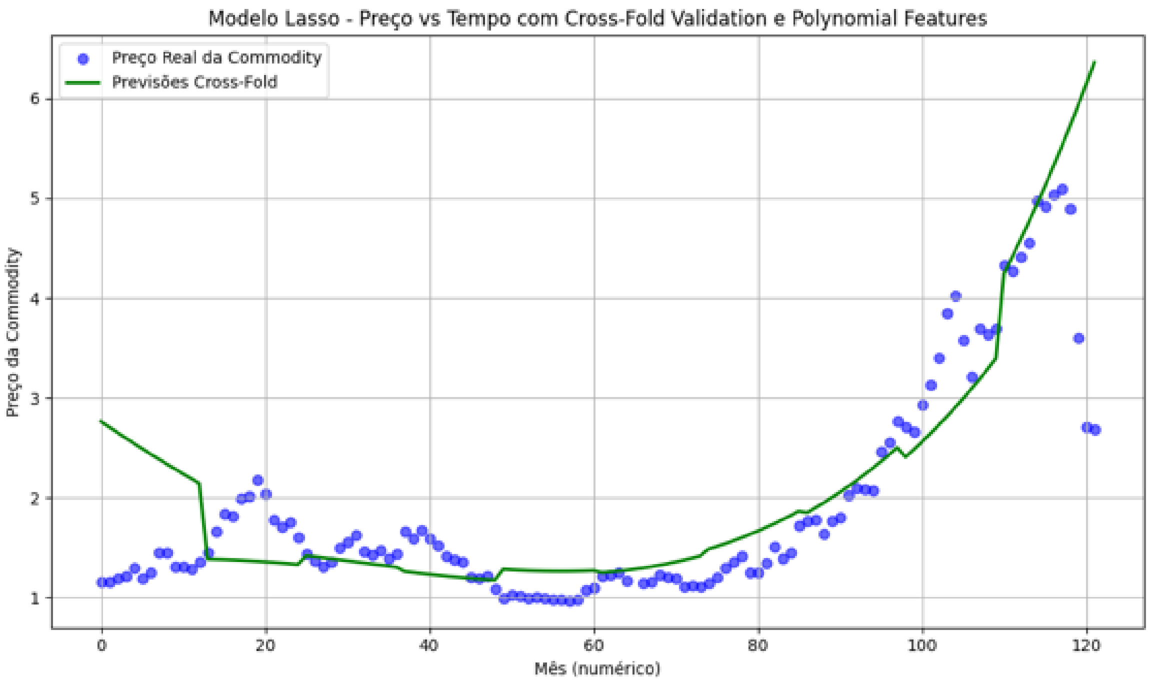
**Gráfico 3: Desempenho do Modelo** LASSO **com *Cross-Fold* e *PolynomialFeatures***




Fonte: Elaborado pelo autor.

O Gráfico 3 representa o desempenho do modelo de regressão LASSO com validação cruzada do tipo *k-Fold*, aplicado à previsão dos preços da commodity ao longo do tempo. Os pontos azuis indicam os valores reais observados, enquanto a linha verde mostra as previsões geradas pelo modelo em diferentes partições de treino e teste. A aplicação da técnica de *cross-validation* proporciona uma avaliação mais robusta da capacidade preditiva do modelo, reduzindo o risco de *overfitting* ao utilizar diferentes subconjuntos de dados para validação, conforme defendido por Kohavi (1995). No entanto, observa-se que, apesar da melhoria na generalização, o modelo LASSO ainda apresenta uma curva simplificada, com menor sensibilidade às oscilações reais da série temporal. Isso evidencia a limitação do LASSO em capturar variações não lineares mais complexas sem a presença de atributos polinomiais, mesmo com validação cruzada. Assim, o gráfico reafirma que a escolha do método de validação, aliada ao tratamento adequado dos dados, é fundamental para a eficácia do modelo preditivo em cenários de alta variabilidade, como o mercado de commodities.

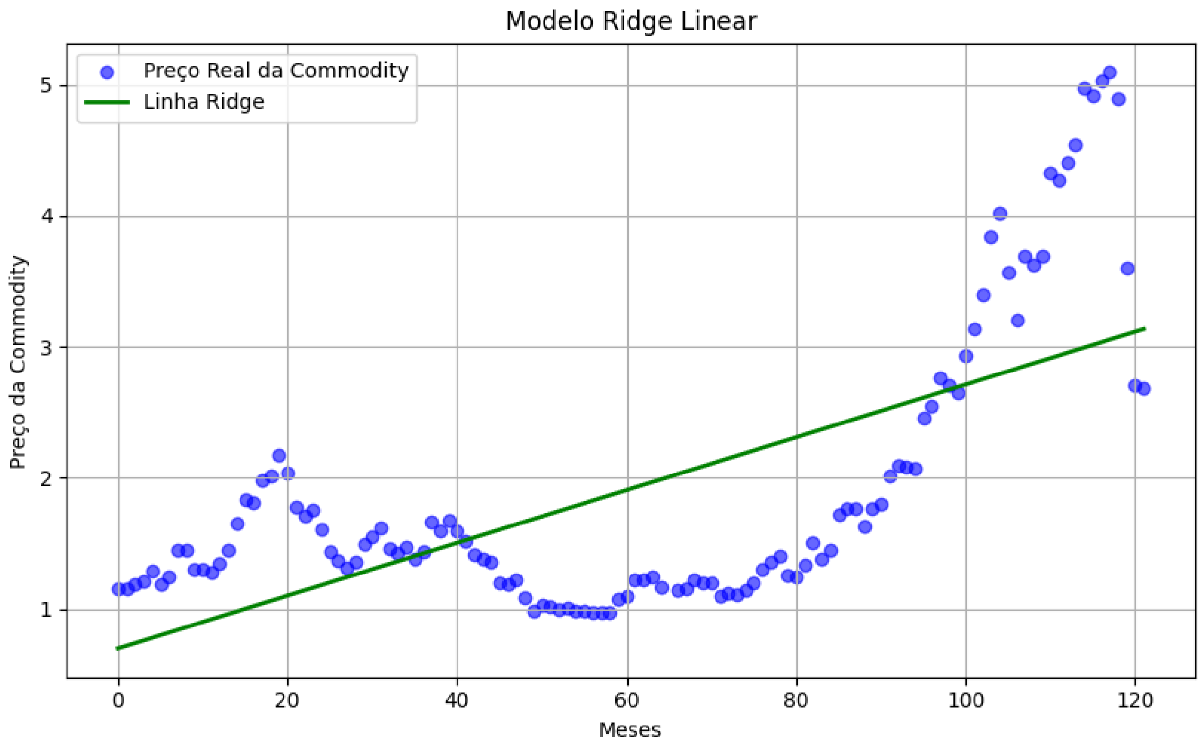
**Gráfico 4: Desempenho do Modelo** LASSO **com *Cross-Fold* e *PolynomialFeatures***




Fonte: Elaborado pelo autor.

O Gráfico 4 apresenta a previsão dos preços da *commodity* utilizando o modelo de regressão LASSO com validação cruzada *k-Fold* e aplicação de *Polynomial Features* no pré-processamento dos dados. Os pontos azuis representam os valores reais da série temporal, enquanto a linha verde mostra a curva predita com base nas múltiplas divisões do conjunto de dados. Diferentemente das abordagens anteriores, a combinação da transformação polinomial com a validação cruzada permitiu ao modelo capturar com maior precisão as tendências não lineares e curvas acentuadas da série histórica. Nota-se uma melhora significativa na aderência da curva predita aos dados reais, especialmente nos períodos de crescimento mais acentuado. Apesar da sensibilidade natural do LASSO à multicolinearidade gerada pelas expansões polinomiais, o uso da validação cruzada proporcionou uma calibragem mais estável do parâmetro de regularização, promovendo equilíbrio entre complexidade e generalização. Este resultado demonstra que, com o ajuste adequado de hiperparâmetros e uso de técnicas de validação robustas, o LASSO pode ser eficaz mesmo em contextos com alta variabilidade e relações não lineares, conforme defendido por Hastie, Tibshirani e Friedman (2009).

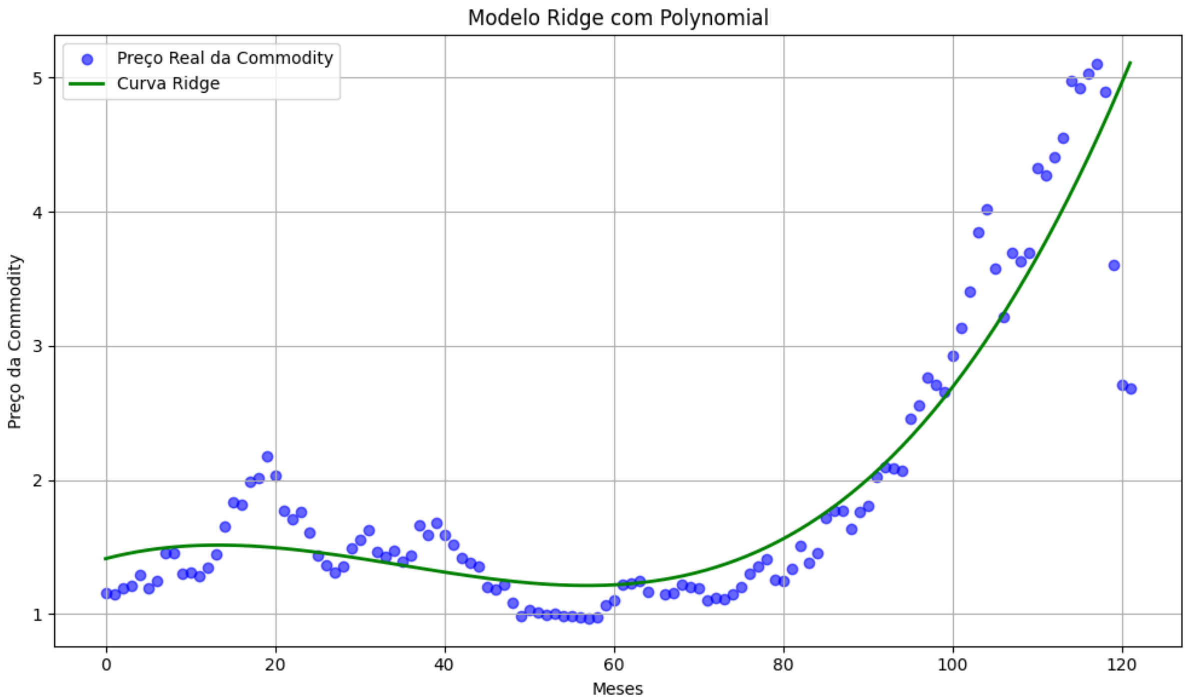
**Gráfico 5: Desempenho do Modelo Ridge**




Fonte: Elaborado pelo autor.

O Gráfico 5 apresenta os resultados do modelo de regressão Ridge linear aplicado à previsão dos preços da *commodity* laranja ao longo do tempo. Os pontos azuis representam os valores reais da série histórica, enquanto a linha verde ilustra a curva ajustada pelo modelo Ridge sem aplicação de transformação polinomial. Nota-se que, por se tratar de um modelo linear, a linha gerada não consegue captar adequadamente as oscilações sazonais ou não lineares da série, resultando em uma tendência suavizada e distante dos picos e vales observados nos dados reais. Apesar disso, o Ridge mostra-se mais estável do que o LASSO em termos de coeficientes, uma vez que a regularização L2 utilizada penaliza grandes valores sem eliminar completamente variáveis explicativas. A aplicação do Ridge linear, neste contexto, demonstra ser limitada para representar relações complexas entre tempo e preço, evidenciando a necessidade do uso de técnicas mais flexíveis, como transformações polinomiais ou modelos não lineares, conforme argumentado por Hastie, Tibshirani e Friedman (2009).

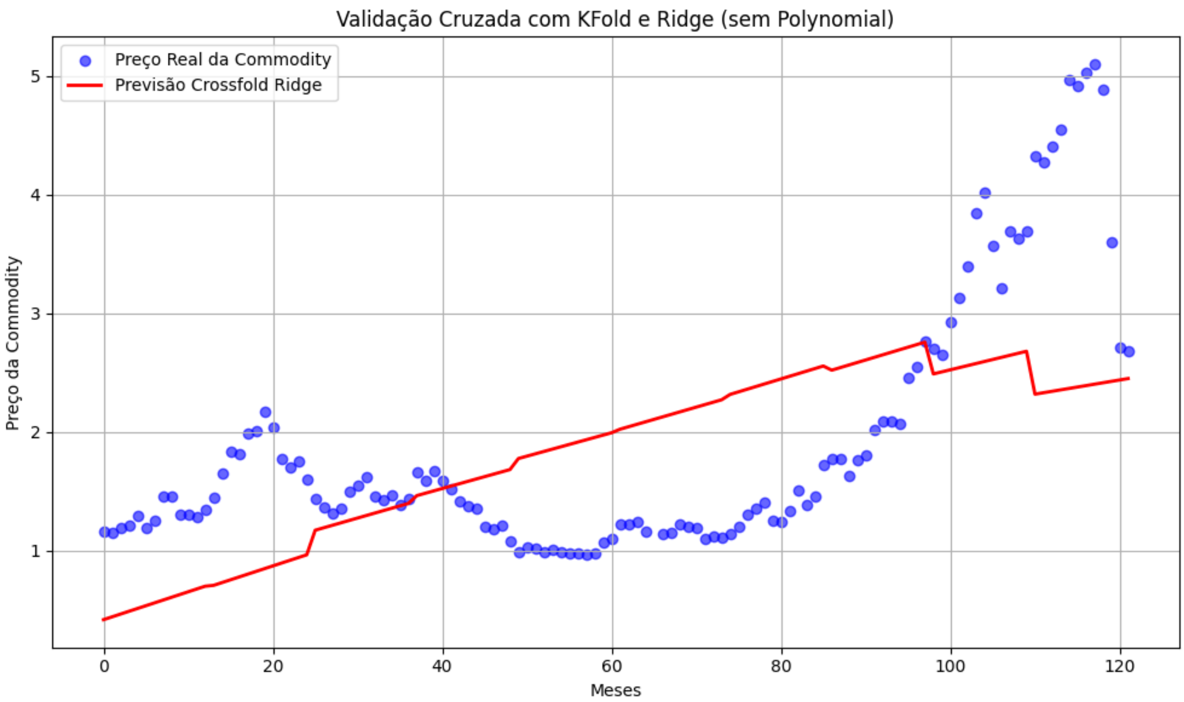
**Gráfico 6: Desempenho do Modelo Ridge com *PolynomialFeatures***




Fonte: Elaborado pelo autor.

O Gráfico 6 ilustra o desempenho do modelo de regressão Ridge com aplicação de *Polynomial Features*, evidenciando a capacidade do algoritmo de capturar a curvatura não linear da série histórica de preços da *commodity*. Os pontos azuis representam os dados reais mensais ao longo do tempo, enquanto a linha verde exibe a curva predita pelo modelo Ridge após o pré-processamento polinomial. Diferentemente da versão linear, este modelo consegue se ajustar melhor às tendências de longo prazo e às inflexões observadas nos dados, especialmente no crescimento exponencial a partir do mês 90. A regularização L2 aplicada pelo Ridge permite suavizar os coeficientes, distribuindo o impacto das variáveis polinomiais sem zerá-las, o que resulta em uma curva contínua e mais estável. Conforme destacado por Hastie, Tibshirani e Friedman (2009), essa abordagem é especialmente eficaz em contextos com multicolinearidade e alta complexidade, como ocorre no mercado de commodities. O gráfico demonstra, portanto, que a combinação entre expansão polinomial e regressão Ridge é uma estratégia robusta para prever séries temporais com comportamento não linear.

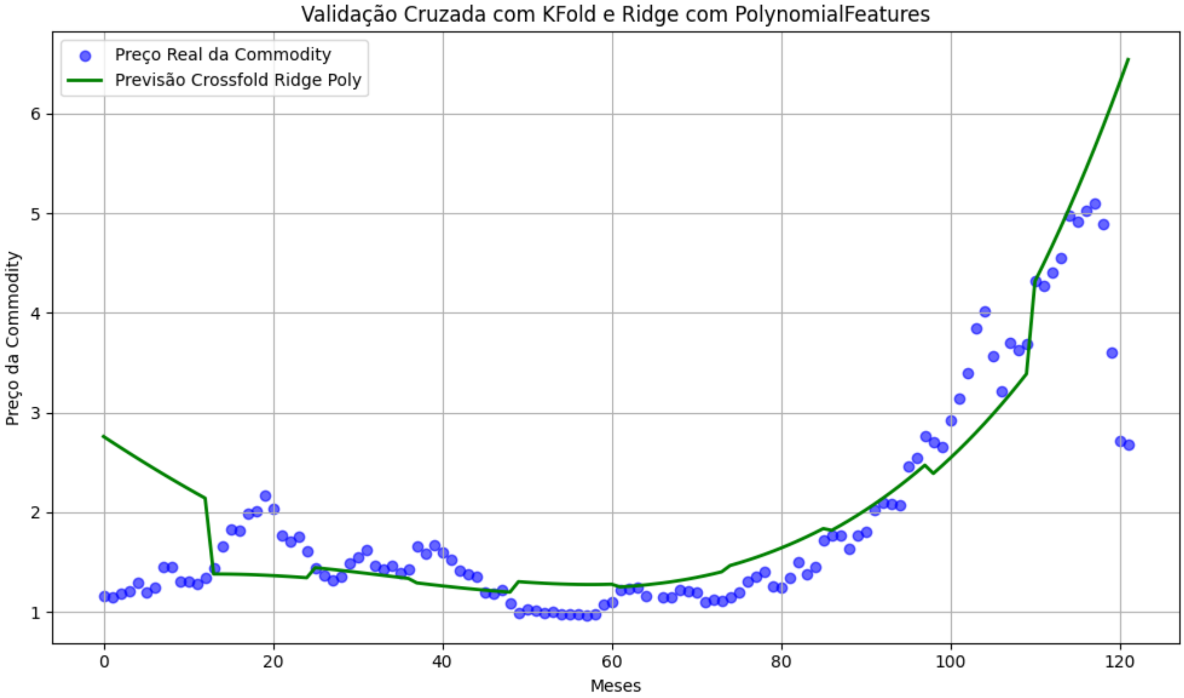
**Gráfico 7: Desempenho do Modelo Ridge com *Cross-Fold***




Fonte: Elaborado pelo autor.

O Gráfico 7 apresenta os resultados da validação cruzada com *K-Fold* aplicando o modelo de regressão Ridge sem a utilização de *Polynomial Features*. Os pontos azuis representam os preços reais da *commodity* ao longo do tempo, enquanto a linha vermelha corresponde à média das previsões feitas em cada *fold* da validação cruzada. Observa-se que, embora a estratégia de *K-Fold* aumente a robustez e a generalização do modelo — reduzindo o risco de *overfitting* — a ausência da transformação polinomial limita a capacidade preditiva do Ridge. A curva resultante é linear por partes, e falha em capturar os padrões mais complexos e as variações acentuadas do comportamento da série histórica. Ainda assim, o Ridge mantém uma estimativa mais estável e coerente quando comparado a modelos que sofrem mais com a multicolinearidade. Como apontado por Kohavi (1995) e Hastie et al. (2009), o uso do Ridge aliado à validação cruzada é útil para evitar sobreajustes em modelos lineares, mas torna-se insuficiente quando o fenômeno observado exige uma abordagem com maior expressividade não linear.

**Gráfico 8: Desempenho do Modelo Ridge com *Cross-Fold* e *PolynomialFeatures***




Fonte: Elaborado pelo autor.

O Gráfico 8 apresenta os resultados da validação cruzada com *K-Fold* utilizando o modelo Ridge combinado com expansão polinomial (*PolynomialFeatures*), aplicada à previsão do preço da *commodity* ao longo do tempo. Os pontos azuis representam os dados reais, e a curva verde exibe as previsões ajustadas pelo modelo em cada *fold* da validação. Observa-se que a combinação entre a regularização L2 do Ridge e a transformação polinomial permite ao modelo capturar com maior fidelidade a tendência crescente e as variações não lineares presentes na série histórica. A suavidade e continuidade da curva indicam um bom equilíbrio entre ajuste e generalização, resultado da eficácia do processo de validação cruzada na seleção de parâmetros adequados, como o valor do alpha e o grau polinomial. Em consonância com Hastie, Tibshirani e Friedman (2009), este modelo evidencia robustez frente à multicolinearidade e flexibilidade na modelagem de padrões complexos, sendo, portanto, uma das abordagens mais eficazes na previsão de séries temporais com comportamento exponencial e alta variabilidade, como ocorre no mercado de commodities agrícolas.

Ao comparar o desempenho dos modelos com e sem o uso da expansão polinomial, constatou-se que o Ridge se saiu melhor quando a normalização polinomial foi aplicada. Isso se deve à sua habilidade em lidar com a multicolinearidade gerada pelas interações polinomiais, distribuindo a penalização entre os coeficientes sem zerá-los completamente, o que mantém a relevância das variáveis e preserva a estabilidade do modelo (HOEFING *et* al., 2010). Por outro lado, o LASSO apresentou maior instabilidade e maior sensibilidade às variáveis correlacionadas quando a expansão polinomial foi aplicada, demonstrando que, nesse cenário, o Ridge é preferível para modelos que envolvem transformações complexas.

Sem a utilização da transformação polinomial, os resultados indicam que o LASSO tem vantagem, pois consegue eliminar coeficientes irrelevantes e simplificar o modelo, promovendo melhor interpretabilidade e evitando o sobre ajuste (TIBSHIRANI, 1996). O Ridge, sem as variáveis polinomiais, tende a reduzir todos os coeficientes de forma mais uniforme, mas sem realizar seleção efetiva de variáveis, o que pode resultar em modelos menos parcimoniosos e potencialmente mais complexos, sem ganhos significativos em desempenho.

Em síntese, a análise dos resultados revela que o Ridge se mostra mais robusto e estável frente à transformação polinomial e diferentes métodos de validação, sendo o método mais indicado para cenários que envolvem pré-processamento polinomial e múltiplas validações. Já o LASSO é mais eficiente em reduzir a complexidade do modelo sem expansão polinomial e beneficia-se do uso da validação cruzada para calibragem fina dos parâmetros. Esses achados corroboram estudos anteriores sobre regularização em modelos lineares e a importância da escolha adequada de métodos conforme as características dos dados e objetivos da modelagem (PEDREGOSA *et* al., 2011; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

**Tabela 2: Erros modelos Ridge e** LASSO

| Modelo | Validação | Expansão Polinomial | Erro Final - MAE |
|---|---|---|---|
| LASSO | 1-*Fold* | Não | 1,57845238095238 |
| LASSO | 1-*Fold* | Sim | 1,57845238095238 |
| LASSO | *K-Fold* | Não | 0,94904252399489 |
| LASSO | *K-Fold* | Sim | 1,10436099899091 |
| Ridge | 1-*Fold* | Não | 1,36800000000000 |
| Ridge | 1-*Fold* | Sim | 2,28600000000000 |
| Ridge | *K-Fold* | Não | 2,28896228289473 |
| Ridge | *K-Fold* | Sim | 0,63023307668421 |

Fonte: Elaborado pelo autor.

A Tabela Erros modelos Ridge e LASSO apresenta os valores finais de erro obtidos pelos modelos Ridge e LASSO, considerando os diferentes cenários de validação e aplicação de transformação polinomial. Verifica-se que o menor erro foi registrado tanto pelo modelo LASSO com validação K-Fold sem expansão polinomial (erro de 0,94904252399489), quanto pelo modelo Ridge com validação K-Fold com expansão polinomial (erro de 0,63023307668421). Tais resultados reforçam a análise qualitativa, indicando que o LASSO apresenta bom desempenho em modelos mais simples e sem multicolinearidade, ao passo que o Ridge se adapta melhor a modelos mais complexos e com maior número de variáveis correlacionadas. Notadamente, os modelos que utilizaram validação cruzada obtiveram, em geral, menor erro, confirmando a eficácia dessa técnica para reduzir a variância das estimativas e melhorar a capacidade preditiva dos modelos (KOHAVI, 1995).

# 5 CONCLUSÃO

Durante o desenvolvimento deste trabalho, enfrentaram-se diversas dificuldades técnicas e conceituais, especialmente relacionadas ao tratamento dos dados e à escolha adequada dos métodos de modelagem. A complexidade inerente à transformação polinomial dos dados exigiu cuidado especial na análise da multicolinearidade e no ajuste dos parâmetros de regularização dos modelos Ridge e LASSO. Além disso, a implementação das validações cruzadas demandou um maior tempo computacional e análise criteriosa para evitar overfitting, conforme apontado por Kohavi (1995). Essas dificuldades foram superadas com a aplicação cuidadosa das técnicas de pré-processamento e validação, garantindo a robustez dos resultados.

Os resultados obtidos evidenciaram que o modelo Ridge apresenta maior estabilidade e capacidade de lidar com dados transformados por polinômios, mantendo previsões mais consistentes e suavizadas. Por outro lado, o modelo LASSO mostrou-se eficaz na simplificação do modelo, especialmente sem a transformação polinomial, ao eliminar variáveis irrelevantes e reduzir o risco sobre ajuste (Tibshirani, 1996; Hastie; Tibshirani; Friedman, 2009). A validação cruzada com múltiplos folds demonstrou ser uma ferramenta fundamental para calibrar os parâmetros e melhorar a generalização dos modelos, corroborando as práticas recomendadas em aprendizado de máquina (Pedregosa *et* al., 2011). Dessa forma, este trabalho contribui para a compreensão dos trade-offs entre os métodos de regularização e as estratégias de pré-processamento em problemas de regressão.

Como proposta para trabalhos futuros, sugere-se a investigação de técnicas híbridas que combinam a regularização LASSO e Ridge, como o Elastic Net, que pode potencialmente equilibrar as vantagens de ambos os métodos para diferentes tipos de dados (Zou; Hastie, 2005). Além disso, a inclusão de métodos de aprendizado profundo, como redes neurais recorrentes (LSTM), pode ampliar a capacidade de modelagem de séries temporais complexas, especialmente em cenários com dados volumosos e não lineares. Por fim, a automação do processo de seleção de hiper parâmetros e a análise de interpretabilidade dos modelos podem ser exploradas para aumentar a aplicabilidade prática das soluções propostas neste trabalho.

## REFERÊNCIAS

EMBRAPA. **Empresa Brasileira de Pesquisa Agropecuária. Avanços em biotecnologia e irrigação impulsionam produtividade agrícola.** Brasília: EMBRAPA, 2022. Disponível em: https://www.embrapa.br/. Acesso em: 05 jun. 2025.

FAO. **Food and Agriculture Organization of the United Nations**. 2023. Disponível em: https://www.fao.org/home/en. Acesso em: 05 jun. 2025.

FABOZZI, F. J.; PETERSON DRAKE, P. ***Finance: Capital Markets, Financial Management, and Investment Management.*** Hoboken: John Wiley & Sons, 2009.

FEDERAL RESERVE BANK OF ST. LOUIS (FRED). **Economic Research Data**. 2024. Disponível em: https://fred.stlouisfed.org/. Acesso em: 05 jun. 2025.

GEMAN, H. ***Commodities and Commodity Derivatives: Modelling and Pricing for Agriculturals,*** *Metals and Energy*. John Wiley & Sons, 2005.

GÉRON, A. ***Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems***. 2. ed. Sebastopol: O'Reilly Media, 2019.

HAMILTON, J. D. ***Time Series Analysis***. Princeton: Princeton University Press, 1994.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. ***The Elements of Statistical Learning: Data Mining, Inference, and Prediction***. 2. ed. New York: Springer, 2009.

HOERL, Arthur E.; KENNARD, Robert W. Ridge regression—1980: ***Advances, algorithms, and applications. American Journal of Mathematical and Management Sciences***, v. 1, n. 1, p. 5-83, 1981.

HULL, J. C. ***Risk Management and Financial Institutions***. 5. ed. John Wiley & Sons, 2018.

HYNDMAN, R. J.; ATHANASOPOULOS, G. ***Forecasting: Principles and Practice***. 2. ed. Melbourne: OTexts, 2018. Disponível em: https://otexts.com/fpp2/. Acesso em: 05 jun. 2025.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. ***An Introduction to Statistical Learning: with Applications in R***. New York: Springer, 2013.

KOHAVI, R. ***A study of cross-validation and bootstrap for accuracy estimation and model selection.*** Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, 1995. p. 1137–1143.

MCKINNEY, W. ***Python for Data Analysis.*** 2. ed. Sebastopol: O'Reilly Media, 2017.

MORETTIN, P. A.; BUSSAB, W. O. ***Estatística Básica***. 6. ed. São Paulo: Saraiva, 2010.

PEDREGOSA, F. *et* al. ***Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research***, v. 12, p. 2825-2830, 2011.

RONCORONI, A.; FUSAI, G.; CUMMINS, M. (Eds.). ***Handbook of Multi-Commodity Markets and Products: Structuring, Trading and Risk Management***. John Wiley & Sons, 2015.

SILVA, J. P.; MENDES, A. F. ***Inovação tecnológica na agricultura: desafios e oportunidades.*** *Revista Brasileira de Agrociência*, v. 25, n. 3, p. 50-65, 2019.

TIBSHIRANI, R. ***Regression shrinkage and selection via the*** LASSO***.*** *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 58, n. 1, p. 267-288, 1996.

VANDERPLAS, J. ***Python Data Science Handbook: Essential Tools for Working with Data.*** Sebastopol, CA: O'Reilly Media, 2016.

ZOU, H.; HASTIE, T. ***Regularization and variable selection via the elastic net.*** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, v. 67, n. 2, p. 301-320, 2005.

## APÊNDICE

**Anexo 1 - Resultado dados Ridge *Fold* com *PolynomialFeatures***

```
------------------------------------
Avaliações dos folds:
[[0.00316386 0.04534981 0.         ]
 [0.22606163 0.4213103  0.         ]
 [0.0133119  0.09864084 0.         ]
 [0.05854984 0.19870131 0.         ]
 [0.08119502 0.28317139 0.         ]
 [0.02076471 0.12275737 0.         ]
 [0.09542671 0.29987619 0.         ]
 [0.02567647 0.14118784 0.         ]
 [0.36705179 0.52943797 0.         ]
 [3.04523229 1.14430178 0.         ]]
------------------------------------
```

Fonte: Elaborado pelo autor.

**Anexo 2 - Resultado dados Ridge *Fold* sem *PolynomialFeatures***

```
------------------------------------
Avaliações dos folds:
[[0.49046415 0.69634824]
 [1.01181222 0.98556354]
 [0.05472801 0.20156475]
 [0.06490412 0.20671569]
 [0.74443898 0.85851451]
 [0.91515535 0.95097439]
 [1.23474467 1.10884528]
 [0.47193922 0.65235213]
 [0.66098516 0.70699461]
 [4.35330032 1.91899196]]
------------------------------------
```

Fonte: Elaborado pelo autor.

**Anexo 3 - Resultado dados Ridge *k-Fold* com *Polynomial Features***

```
--------------------------------------------------------------------------------
Modelo treinado com alpha = 0.001 e grau do polinômio = 3
Treinando fold 0
Coeficientes: [ 0.00000000e+00 -5.59275401e-02  3.39380647e-04  2.22698656e-06]
Intercepto: 2.760506781010166
--------------------------------------------------------------------------------
Treinando fold 1
Coeficientes: [ 0.00000000e+00  1.38284793e-02 -6.53603885e-04  6.57319433e-06]
Intercepto: 1.2977420234173094
--------------------------------------------------------------------------------
Treinando fold 2
Coeficientes: [ 0.00000000e+00  1.48105385e-02 -7.33476444e-04  7.12618623e-06]
Intercepto: 1.421856449754204
--------------------------------------------------------------------------------
Treinando fold 3
Coeficientes: [ 0.00000000e+00  1.09654409e-02 -6.62644292e-04  6.78148758e-06]
Intercepto: 1.4499550400020043
--------------------------------------------------------------------------------
Treinando fold 4
Coeficientes: [ 0.00000000e+00  2.23670695e-02 -8.51277504e-04  7.61455553e-06]
Intercepto: 1.3567681914792509
--------------------------------------------------------------------------------
Treinando fold 5
Coeficientes: [ 0.00000000e+00  1.62706965e-02 -7.40744904e-04  7.08156525e-06]
Intercepto: 1.4081296977527709
--------------------------------------------------------------------------------

Treinando fold 6
Coeficientes: [ 0.00000000e+00  1.22602750e-02 -6.41055789e-04  6.49617621e-06]
Intercepto: 1.4372804603892986
--------------------------------------------------------------------------------
Treinando fold 7
Coeficientes: [ 0.00000000e+00  1.51553704e-02 -7.32013728e-04  7.08711763e-06]
Intercepto: 1.422244888857601
--------------------------------------------------------------------------------
Treinando fold 8
Coeficientes: [ 0.00000000e+00  2.36421129e-02 -9.13980018e-04  7.96993087e-06]
Intercepto: 1.3496474883752474
--------------------------------------------------------------------------------
Treinando fold 9
Coeficientes: [ 0.00000000e+00  5.49592274e-02 -1.75065005e-03  1.37755030e-05]
Intercepto: 1.118714991146768
--------------------------------------------------------------------------------
```

Fonte: Elaborado pelo autor.

**Anexo 4 - Resultado dados Ridge *k-Fold* sem *Polynomial Features***

```
--------------------------------
Treinando fold 0
Modelo treinado com alpha = 0.001
Coeficientes: [0.02349397]
Intercepto: 0.41878733830756953
--------------------------------
Treinando fold 1
Coeficientes: [0.02340929]
Intercepto: 0.4037428318348921
--------------------------------
Treinando fold 2
Coeficientes: [0.02052347]
Intercepto: 0.659672110874461
--------------------------------
Treinando fold 3
Coeficientes: [0.0199658]
Intercepto: 0.7256276890174811
--------------------------------
Treinando fold 4
Coeficientes: [0.0197319]
Intercepto: 0.8103842244204362
--------------------------------
Treinando fold 5
Coeficientes: [0.02068581]
Intercepto: 0.7623305618031639
--------------------------------
Treinando fold 6
Coeficientes: [0.02174776]
Intercepto: 0.708197289578459
--------------------------------

Treinando fold 7
Coeficientes: [0.02148062]
Intercepto: 0.6744794984442035
--------------------------------
Treinando fold 8
Coeficientes: [0.01737034]
Intercepto: 0.7878043087309967
--------------------------------
Treinando fold 9
Coeficientes: [0.01195281]
Intercepto: 1.0048400194253024
--------------------------------
```

Fonte: Elaborado pelo autor.

**Anexo 5 - Resultado dados** LASSO ***Fold* com *Polynomial Features***

```
----------------------------------------
Avaliações dos crossfolds:
[[0.05619017 0.1929937  0.        ]
 [5.90905171 1.96702736 0.        ]
 [0.48108276 0.54438797 0.        ]
 [6.77582851 2.2421807  0.        ]
 [0.91920253 0.85975693 0.        ]
 [0.02446046 0.10050099 0.        ]
 [0.05529756 0.17545427 0.        ]
 [0.09260474 0.26613497 0.        ]
 [0.43498633 0.57661901 0.        ]
 [0.62081416 0.67329476 0.        ]]
Média: [1.53695189 0.75983507 0.        ]
Variância: 2.5259304961582782
Desvio Padrão: 1.5893176196589145
----------------------------------------
```

Fonte: Elaborado pelo autor.

**Anexo 6 - Resultado dados** LASSO ***Fold* sem *Polynomial Features***

```
----------------------------------------
Avaliações dos folds:
[[0.00532186 0.05709472 0.        ]
 [0.13393663 0.3099704  0.        ]
 [0.57854787 0.74539148 0.        ]
 [0.02842285 0.14625993 0.        ]
 [0.02176974 0.11651686 0.        ]
 [0.0454741  0.20808908 0.        ]
 [0.04635034 0.18243226 0.        ]
 [0.15820252 0.34381044 0.        ]
 [0.28541179 0.44331249 0.        ]
 [1.09674111 0.67950999 0.        ]]
Média: [0.24001788 0.32323876 0.        ]
Variância: 0.0717650408317073
Desvio Padrão: 0.2678899789684327
----------------------------------------
```

Fonte: Elaborado pelo autor.

**Anexo 7 - Resultado dados** LASSO ***k-Fold*** **com *Polynomial Features***

```
-------------------------------------------------------------------------------
K-crossfold = 0
Modelo treinado com alpha = 0.001 e grau do polinômio = 3
Coeficientes: [ 0.00000000e+00 -5.60917258e-02  3.42215292e-04  2.21288147e-06]
Intercepto: 2.763024797664748
-------------------------------------------------------------------------------
K-crossfold = 1
Coeficientes: [ 0.00000000e+00  9.94325462e-03 -5.76510515e-04  6.16235185e-06]
Intercepto: 1.3377335873240308
-------------------------------------------------------------------------------
K-crossfold = 2
Coeficientes: [ 0.00000000e+00  8.58215118e-03 -6.10089350e-04  6.47239076e-06]
Intercepto: 1.482799681046974
-------------------------------------------------------------------------------
K-crossfold = 3
Coeficientes: [ 0.00000000e+00  4.90007441e-03 -5.45103414e-04  6.16637788e-06]
Intercepto: 1.5126411328259324
-------------------------------------------------------------------------------
K-crossfold = 4
Coeficientes: [ 0.00000000e+00  1.63411621e-02 -7.35004001e-04  7.00540888e-06]
Intercepto: 1.4219277970071245
-------------------------------------------------------------------------------
K-crossfold = 5
Coeficientes: [ 0.00000000e+00  1.10218863e-02 -6.37157875e-04  6.52952330e-06]
Intercepto: 1.4650025303330418
-------------------------------------------------------------------------------
K-crossfold = 6
Coeficientes: [ 0.00000000e+00  6.72271353e-03 -5.29029209e-04  5.89102878e-06]
Intercepto: 1.4957369409381602
-------------------------------------------------------------------------------

K-crossfold = 7
Coeficientes: [ 0.00000000e+00  9.12735480e-03 -6.10547478e-04  6.43541097e-06]
Intercepto: 1.4855943661161866
-------------------------------------------------------------------------------
K-crossfold = 8
Coeficientes: [ 0.00000000e+00  1.96822372e-02 -8.36152754e-04  7.55873216e-06]
Intercepto: 1.3924598450361598
-------------------------------------------------------------------------------
K-crossfold = 9
Coeficientes: [ 0.00000000e+00  4.57700647e-02 -1.55035988e-03  1.25948269e-05]
Intercepto: 1.2085613533279165
-------------------------------------------------------------------------------
```

Fonte: Elaborado pelo autor.

**Anexo 8 - Resultado dados** LASSO ***k-Fold*** **sem** ***Polynomial Features***

```
-----------------------------------
K-crossfold = 0
Modelo treinado com alpha = 0.001
Coeficientes: [0.02349297]
Intercepto: 0.418854400639161
-----------------------------------
K-crossfold = 1
Modelo treinado com alpha = 0.001
Coeficientes: [0.02340844]
Intercepto: 0.4037983762173407
-----------------------------------
K-crossfold = 2
Modelo treinado com alpha = 0.001
Coeficientes: [0.02052269]
Intercepto: 0.6597220340105765
-----------------------------------
K-crossfold = 3
Modelo treinado com alpha = 0.001
Coeficientes: [0.01996506]
Intercepto: 0.725674027499219
-----------------------------------
K-crossfold = 4
Modelo treinado com alpha = 0.001
Coeficientes: [0.01973118]
Intercepto: 0.810428431240025
-----------------------------------
K-crossfold = 5
Modelo treinado com alpha = 0.001
Coeficientes: [0.02068509]
Intercepto: 0.7623738054592557
-----------------------------------
```

```
K-crossfold = 6
Modelo treinado com alpha = 0.001
Coeficientes: [0.02174702]
Intercepto: 0.7082407611487285
------------------------------------
K-crossfold = 7
Modelo treinado com alpha = 0.001
Coeficientes: [0.02147983]
Intercepto: 0.6745244450149173
------------------------------------
K-crossfold = 8
Modelo treinado com alpha = 0.001
Coeficientes: [0.01736947]
Intercepto: 0.7878523268562793
------------------------------------
K-crossfold = 9
Modelo treinado com alpha = 0.001
Coeficientes: [0.01195183]
Intercepto: 1.0048935352287423
------------------------------------
```

Fonte: Elaborado pelo autor.