Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Previsão de Preços na Agricultura Familiar: Uma Análise de Modelos Aplicados à Alface Crespa

Patrick Canto de Carvalho

JUIZ DE FORA
AGOSTO, 2025

Previsão de Preços na Agricultura Familiar: Uma Análise de Modelos Aplicados à Alface Crespa

Patrick Canto de Carvalho

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Heder Soares Bernardino

JUIZ DE FORA

AGOSTO, 2025

Previsão de Preços na Agricultura Familiar: Uma Análise de Modelos Aplicados à Alface Crespa

Patrick Canto de Carvalho

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Heder Soares Bernardino
Doutor em Modelagem Computacional

Jairo Francisco de Souza
Doutor em Informática

Luciana Conceição Dias Campos
Doutora em Engenharia Elétrica

JUIZ DE FORA
8 DE AGOSTO, 2025

À minha família e meus amigos.

Resumo

A agricultura é uma atividade essencial para a subsistência humana e de grande relevância econômica no Brasil. Embora soluções tecnológicas venham sendo cada vez mais aplicadas ao setor, pequenos produtores ainda têm acesso limitado a essas inovações. Entre as abordagens computacionais promissoras, destacam-se os modelos preditivos aplicados a séries temporais. No entanto, muitos estudos existentes concentram-se em grandes *commodities* e previsões de curto prazo, o que limita sua utilidade para a agricultura familiar brasileira. Este trabalho busca desenvolver, analisar e comparar modelos preditivos aplicados à previsão de preços agrícolas com foco em previsões de longo prazo (*multistep*) e no uso de variáveis exógenas, como fatores climáticos. Foram utilizados dados semanais de preços de alface crespa (CEPEA) e dados meteorológicos de Teresópolis-RJ (INMET), reamostrados para frequência semanal. Os modelos avaliados foram SARIMAX, LSTM e Prophet. Os resultados indicam que a inclusão de variáveis exógenas melhora o desempenho dos modelos, especialmente do LSTM. O SARIMAX se destaca em horizontes curtos, enquanto o LSTM obtém melhores resultados no longo prazo. Além disso, o erro tende a crescer com o horizonte de previsão, mas de forma não linear nos modelos LSTM e Prophet.

Palavras-chave: Séries temporais multivariadas, redes neurais artificiais, previsão, agricultura familiar.

Abstract

Agriculture is an essential activity for human subsistence and holds great economic importance in Brazil. Although technological solutions have been increasingly applied to the sector, small-scale farmers still have limited access to these innovations. Among the promising computational approaches, predictive models applied to time series stand out. However, many existing studies focus on major commodities and short-term forecasts, which limits their usefulness for Brazilian family farming. This work aims to develop, analyze, and compare predictive models for agricultural price forecasting, focusing on long-term (multistep) forecasts and the use of exogenous variables, such as climatic factors. Weekly price data for lettuce (CEPEA) and meteorological data from Teresópolis-RJ (INMET), resampled to a weekly frequency, were used. The models evaluated were SARIMAX, LSTM, and Prophet. The results indicate that including exogenous variables improves model performance, especially for LSTM. SARIMAX performs better for short-term forecasts, while LSTM shows superior results in the long term. Additionally, the prediction error tends to increase with the forecasting horizon, although not linearly in the case of LSTM and Prophet models.

Keywords: Multivariate time series, artificial neural networks, forecasting, family agriculture.

Agradecimentos

Gostaria de agradecer em primeiro lugar aos meus pais por extraírem da terra, com muito suor, o sustento que me permitiu ocupar um espaço que nunca puderam. Embora tenham tido pouco estudo, sempre entenderam a importância de plantar para colher e assim o fizeram, acreditando em mim.

Agradeço enormemente ao meu orientador, Heder Soares Bernardino, por estar sempre disponível e me guiar com paciência e bom humor durante toda a realização deste e outros trabalhos.

Aos meus amigos e colegas que me apoiaram com palavras de carinho e apoio, acreditando em mim, principalmente quando eu duvidei.

Aos professores que, ao longo da minha vida, compartilharam generosamente seus conhecimentos.

"Não há fins, e nunca haverá fins, para o girar da Roda do Tempo. Mas foi um fim."

Robert Jordan, A Roda do Tempo

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ADF	*Augmented Dickey-Fuller test*
ANN	*Artificial Neural Network*
ARIMA	*Autoregressive Integrated Moving Average*
CEPEA	Centro de Estudos Avançados em Economia Aplicada
CNN	*Convolutional Neural Network*
GRNN	*Generalized Neural Network*
INMET	Instituto Nacional de Meteorologia
KPSS	*Kwiatkowski–Phillips–Schmidt–Shin test*
LSTM	*Long Short-Term Memory*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MSE	*Mean Square Error*
NMAE	*Normalized Mean Absolute Error*
RF	*Random Forest*
RMSE	*Root Mean Squared Error*
SAD	Sistema de Apoio à Decisão
SARIMA	*Seasonal Autoregressive Integrated Moving Average*
SARIMAX	*Seasonal Auto-Regressive Integrated Moving Average with Exogenous Factors*
T-GCN	*Temporal Graph Convolutional Network*
VIF	*Variance Inflation Factor*

1 Introdução

A agricultura familiar é responsável por uma parcela significativa da produção de alimentos no Brasil, especialmente de frutas, legumes e verduras. Segundo dados de 2021, cerca de 67% dos 15 milhões de produtores rurais do país pertencem a esse segmento (CONAFER, 2021), caracterizado pela gestão e força de trabalho majoritariamente familiar, além de limitações de área e renda conforme a legislação vigente (BRASIL, 2006). Apesar de sua importância econômica e social, a agricultura familiar enfrenta desafios estruturais importantes, entre eles o baixo acesso a tecnologias digitais e a limitada escolaridade de seus produtores (BUAINAIN; CAVALCANTE; CONSOLINE, 2021).

Em contraste com os avanços promovidos pela chamada Agricultura 4.0 — que inclui o uso de ferramentas computacionais para análise de dados e tomada de decisão —, a realidade da agricultura familiar ainda é marcada por práticas baseadas na experiência empírica e subjetiva. Uma das áreas com grande potencial de impacto é a previsão de preços de produtos agrícolas, que poderia oferecer suporte estratégico a decisões de plantio e investimento. No entanto, essa funcionalidade segue pouco explorada nesse contexto.

Embora haja uma extensa literatura sobre previsão de preços agrícolas utilizando séries temporais e modelos de aprendizado de máquina, a maioria dos estudos concentra-se em mercados de *commodities* de larga escala, frequentemente em países como China e Índia (BAYONA-ORÉ; CERNA; HINOJOZA, 2021). Esses mercados possuem características distintas do cenário da agricultura familiar brasileira, tanto em escala quanto na variedade de produtos comercializados. Além disso, são raros os trabalhos que abordam previsões de longo prazo: horizonte temporal essencial para agricultores que precisam tomar decisões com meses de antecedência, dado o tempo necessário entre o plantio e a colheita (AMARO et al., 2007). A ausência de modelos computacionais adaptados a essa realidade contribui para a vulnerabilidade econômica dos pequenos agricultores, que muitas vezes tomam decisões de forma subjetiva e sem apoio em dados.

Séries temporais, nas quais se baseiam estes modelos preditivos, são um conjunto de medições tomadas sequencialmente no tempo em intervalos regulares. Podem ser dados

de um fenômeno medido a cada hora, dia, semana ou outra periodicidade (BOX et al., 2015). Para a predição de preços na agricultura, a principal variável a ser observada é o próprio preço dos produtos de interesse. Considerando que a variação no preço de produtos agrícolas é influenciada por diversos fatores, muitas vezes consideram-se outras variáveis que podem, entretanto, ser complexas e difíceis de serem obtidas (BAYONA-ORÉ; CERNA; HINOJOZA, 2021).

Diante desse cenário, este trabalho propõe o desenvolvimento e avaliação de modelos de aprendizado de máquina para séries temporais aplicados à previsão de preços agrícolas com foco na agricultura familiar brasileira. O objetivo central é investigar como tais técnicas podem ser utilizadas para realizar previsões de longo prazo (também chamadas de *multistep*) considerando tanto séries univariadas quanto multivariadas. Além disso, busca-se compreender como os preços se relacionam com outras variáveis, em especial àquelas referentes a condições meteorológicas.

As investigações se concentram nos modelos Prophet, LSTM e SARIMAX, sendo os dois últimos amplamente usados para previsão de séries temporais. Os dados utilizados serão preços da Alface Crespa no município de Teresópolis - RJ [1], obtidos do Centro de Estudos Avançados em Economia Aplicada (CEPEA), utilizando como fatores exógenos dados do mesmo município obtidos junto ao Instituto Nacional de Meteorologia (INMET), como pressão atmosférica, temperatura e velocidade dos ventos.

A agricultura tem papel fundamental na sobrevivência da humanidade e a relevância deste estudo está alinhada com os Objetivos de Desenvolvimento Sustentável propostos pela ONU, que incluem o aumento da produtividade e da renda da agricultura familiar (UNITED NATIONS, 2024). Ao desenvolver modelos computacionais adaptados a esse público, espera-se contribuir fornecendo uma fundação para futuros sistemas acessíveis a estes e reduzir desigualdades no acesso à tecnologia, promovendo maior autonomia, planejamento e estabilidade financeira para pequenos produtores rurais.

[1] ⟨https://www.teresopolis.rj.gov.br/teresopolis-recebe-o-titulo-de-capital-estadual-da-agricultura-familiar⟩

1.1 Objetivos

O objetivo geral deste trabalho é analisar e comparar diferentes modelos para predição de preço futuro de produtos da agricultura familiar utilizando métodos computacionais. Para que o objetivo geral seja atingido, são identificados os seguintes objetivos específicos:

- Identificar dados relevantes, como histórico de preços, condições climáticas etc;
- Realizar a coleta dos dados necessários;
- Tratar os dados para serem utilizados pelos modelos;
- Analisar relações entre a variável de interesse e os fatores exógenos;
- Identificar e aplicar métodos de aprendizagem de máquina para predição sobre os dados coletados;
- Comparar e analisar os modelos utilizados.

1.2 Organização do Trabalho

Este trabalho está organizado em cinco capítulos. O Capítulo 1 apresenta a introdução, contextualizando o problema, os objetivos e a justificativa do estudo. O Capítulo 2 traz a fundamentação teórica, com conceitos e trabalhos relacionados ao tema. O Capítulo 3 descreve a metodologia, abordando os dados, os modelos e os procedimentos adotados. O Capítulo 4 apresenta e discute os resultados obtidos. Por fim, o Capítulo 5 apresenta as conclusões e sugestões para trabalhos futuros.

2 Fundamentação Teórica

Este capítulo apresenta os conceitos fundamentais que sustentam o desenvolvimento deste trabalho, o qual investiga a aplicação de métodos de aprendizagem de máquina para previsão de preços no contexto da agricultura familiar. Dado que a proposta central envolve a modelagem de séries temporais multivariadas com uso de variáveis exógenas, inicia-se com a definição e caracterização desse tipo de dado, abordando sua estrutura, propriedades e principais finalidades (Seção 2.1).

Na sequência, são introduzidos os modelos estatísticos e computacionais utilizados neste trabalho nas tarefas de previsão: o modelo SARIMAX (Seção 2.2), o modelo Prophet (Seção 2.3) e as redes neurais recorrentes do tipo LSTM (Seção 2.4).

Além dos modelos, são discutidos os métodos de avaliação mais adequados para esse tipo de tarefa, incluindo técnicas específicas de validação cruzada para séries temporais (Seção 2.5) e métricas quantitativas amplamente utilizadas na mensuração do desempenho preditivo (Seção 2.6).

Por fim, apresenta-se uma breve discussão sobre o papel dos sistemas de apoio à decisão (Seção 2.7) por sua relevância na aplicação prática dos modelos desenvolvidos.

2.1 Séries Temporais

Uma série temporal é uma sequência de observações tomadas sequencialmente no tempo. Muitos conjuntos de dados podem ser vistos como séries temporais: quantidade mensal de produtos vendidos por uma fábrica, quantidade de algum produto gerado em uma reação química por hora, entre outros. As observações adjacentes no tempo geralmente são dependentes, sendo esta dependência de muito interesse prático (BOX et al., 2015).

Frequentemente é resultado da observação de algum processo do qual são feitas medições em intervalos de tempo uniformes de acordo com uma taxa de amostragem. É uma sequência de instantes de tempo contíguos. Formalmente, uma série temporal pode

ser definida como uma sequência de n variáveis reais (ESLING; AGON, 2012), tal como:

$$T = (t_1, t_2, ..., t_n), t_i \in \mathbb{R} \tag{2.1}$$

Quando uma série possui uma única variável, como a série temporal que representa a quantidade de produtos vendidos por uma empresa a cada dia, ela é chamada univariada. Já quando apresenta múltiplas dimensões, como uma série de dados climatológicos com temperatura, precipitação e velocidade dos ventos, é chamada multivariada.

Existe uma grande variedade de propósitos aos quais a análise de séries temporais serve, em campos como meteorologia, marketing, saúde, negócios, mercado de ações e setor bancário (MAHALAKSHMI; SRIDEVI; RAJARAM, 2016). Entre eles estão indexação, classificação, clusterização, segmentação (ou sumarização), detecção de anomalias, descoberta de *motifs* (subsequências que aparecem de forma recorrente na série temporal) e predição (ESLING; AGON, 2012).

Esta última é um dos conceitos centrais deste trabalho e uma importante área de pesquisa, sendo uma das mais extensivamente aplicadas. Como já mencionado, valores subsequentes são dependentes uns dos outros, sendo assim predição tem como objetivo modelar estas dependências para prever os próximos valores da série (ESLING; AGON, 2012).

Um dos principais objetivos deste trabalho é prever preços futuros com base em séries temporais. Neste contexto, é importante, primeiramente, definir e analisar o conceito de predição (ou previsão) antes de seguir para os métodos existentes para a realização de previsões.

Box et al. (2015) define previsão como uma função $\hat{z}_t(l)$, que denota uma predição feita no tempo t sobre algum instante $t + l$ no futuro (chamado *lead time* l) baseada na informação disponível dos valores anteriores z_t, z_{t-1}, ..., z_{t-r}, onde r é o tamanho da janela de dados passados disponíveis. O objetivo é obter uma função tal que o desvio $z_{t+l} - \hat{z}_t(l)$ entre o valor real e o previsto seja o menor possível para cada *lead time* l.

Athiyarath, Paul e Krishnaswamy (2020) categorizam os métodos de análise de séries temporais em 3 classes: regressão, métodos estocásticos e métodos de *soft computing*. Regressão é um dos métodos mais simples e mais amplamente usados para determi-

nar uma relação entre um valor que se deseja prever e variáveis relacionadas. Entre eles estão a regressão linear e a regressão linear múltipla.

Os métodos estocásticos consideram a série como uma realização de um processo estocástico. Eles apresentam alta precisão quando os dados da série não são complexos e satisfazem as condições de estacionariedade (vide Seção 2.1.1). Os métodos de *soft computing* são mais flexíveis, tendo maior capacidade de modelar diferentes tipos de séries temporais.

2.1.1 Estacionaridade e Testes ADF e KPSS

Uma série é considerada estacionária quando suas propriedades, como média e variância, são constantes (KWIATKOWSKI et al., 1992). Uma série com sazonalidade ou tendência, portanto, não é estacionária. Isto é importante pois alguns modelos estatísticos tomam como premissa que a série modelada é estacionária. O modelo SARIMAX, por exemplo, internamente utiliza diferenciação para tornar a série de entrada estacionária (VAGROPOULOS et al., 2016).

Existem alguns testes estatísticos usados para identificar esta característica em séries temporais. Aqui serão utilizados os testes Augmented Dickey-Fuller (ADF) (DICKEY; FULLER, 1979) e Kwiatkowski-Phillips-Schmidt-Shin (KPSS) (KWIATKOWSKI et al., 1992). No ADF, a hipótese nula afirma que os dados não são estacionários e, caso o p-valor seja pequeno, ela é descartada e conclui-se que a série é estacionária. Já o KPSS é um teste onde a hipótese nula é que os dados são estacionários e, caso o p-valor seja pequeno, ela é rejeitada e conclui-se que a série não é estacionária. Observa-se que os testes apresentados baseam-se em hipóteses opostas.

2.2 Modelos para Previsão

2.2.1 SARIMAX

Seasonal Auto-Regressive Moving Average with Exogenous Variables (SARIMAX) (VAGROPOULOS et al., 2016) é um modelo estatístico de predição de séries temporais construído a partir do SARIMA(*Seasonal Auto-Regressive Moving Average*), estendido para

lidar com variáveis exógenas (variáveis externas relacionadas à variável principal de interesse). Este, por sua vez, é uma extensão do ARIMA (*Auto-Regressive Moving Average*), apresentado por Box e Jenkins (1976), com adição de termos referentes à sazonalidade da série temporal.

O modelo SARIMAX pode ser representado pelas equações a seguir, como descrito em Vagropoulos et al. (2016):

$$\varphi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D y_t = \beta_k x_{k,t'} + \theta_q(B)\Theta_Q(B^s)\varepsilon_t \tag{2.2}$$

$$\varphi_p(B) = 1 - \sum_{i=1}^{p}\varphi_i B^i \qquad \Phi_P(B^s) = 1 - \sum_{i=1}^{P}\Phi_i B^{s,i} \tag{2.3}$$

$$\theta_q(B) = 1 - \sum_{i=1}^{q}\theta_i B^i \qquad \Theta_Q(B^s) = 1 - \sum_{i=1}^{Q}\Theta_i B^{s,i} \tag{2.4}$$

$$\nabla^d = (1-B)^d \qquad \nabla_s^D = (1-B^s)^D \tag{2.5}$$

Onde

y_t: variável prevista.

$\varphi_p(B)$: polinômio autorregressivo (AR) de ordem p.

$\theta_q(B)$: polinômio de média móvel (MA) de ordem q.

$\Phi_P(B^s)$: polinômio autorregressivo sazonal de ordem P.

$\Theta_Q(B^s)$: polinômio de média móvel sazonal de ordem Q.

β_k : coeficiente associado à k-ésima variável exógena.

$x_{k,t'}$: vetor das variáveis da k-ésima entrada exógena no tempo t.

ε_t: representa um processo de ruído branco.

s: define o período da sazonalidade.

O operador de diferenciação ∇^d remove a não-estacionaridade não-sazonal da série, enquanto ∇_s^D remove a não-estacionaridade sazonal. O operador B, chamado de operador de defasagem (*backshift*), desloca a série y_t no tempo: $B^k(y_t) = y_{t-k}$.

A notação compacta frequentemente utilizada para o modelo, com todos os seus

parâmetros, tem a forma *SARIMAX(p,d,q), (P,D,Q,s)+X*, onde X representa as variáveis exógenas. A implementação utilizada no presente trabalho está presente na biblioteca *statsmodels*[2] , implementada na linguagem de programação *Python*.

2.2.2 Prophet

O Prophet (TAYLOR; LETHAM, 2017) é um modelo de código aberto desenvolvido pelo Facebook[3] com o objetivo de tornar a previsão de séries temporais mais escalável ao facilitar que o utilizem pessoas não especialistas na área (mas que tenham conhecimento no domínio do problema). Além disso, o modelo busca ser adequado para uma larga gama de problemas de previsão com características comuns e, ainda, fornecer respostas simples e interpretáveis aos humanos envolvidos no processo.

Neste modelo, as séries temporais são modeladas como a soma de três componentes: tendência, sazonalidade e feriados, chamados g(t), s(t) e h(t), respectivamente, como explicitado na equação:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{2.6}$$

Onde ϵ_t representa mudanças que os outros componentes do modelo não acomodam. Assume-se que ϵ_t é normalmente distribuído.

Ainda de acordo com Taylor e Letham (2017), o modelo é descrito como sendo similar a um modelo aditivo generalizado (GAM), podendo acomodar novos componentes, se necessário, mas que não leva em consideração dependências temporais entre os dados, sendo efetivamente um processo de regressão ou ajuste de curvas.

2.2.3 Redes Neurais LSTM

Uma abordagem que tem se tornado popular para diversas aplicações, entre elas a previsão de séries temporais, são as Redes Neurais Artificiais, ou ANN - *artificial neural networks* -, por sua flexibilidade e facilidade de operar com dados não lineares (SANGHANI; BHATT; CHAUHAN, 2016).

[2] ⟨https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html⟩
[3] ⟨https://facebook.github.io/prophet/⟩

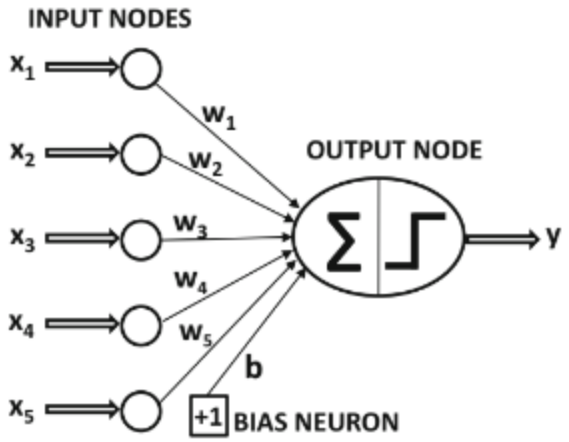
Redes neurais artificiais são algoritmos de aprendizagem de máquina populares que simulam o mecanismo de aprendizagem de organismos biológicos. A base desses sistemas é o neurônio. Estes se conectam por meio de axônios e dendritos, sendo a região entre axônios e dendritos chamada sinapse. A força das conexões sinápticas pode mudar em decorrência de estímulos externos, resultando em um processo de aprendizagem. (AGGARWAL, 2018).

Nas redes neurais artificiais, este mecanismo é simulado. Nelas existe uma unidade computacional também chamada neurônio. Estas unidades são conectadas umas às outras por meio de pesos, que representam o atributo de força das conexões sinápticas nos organismos biológicos. Cada entrada em um determinado neurônio é modificada de acordo com o peso, afetando a função que é computada por ele. Os valores de entrada são propagados pela rede, dos neurônios de entrada até os de saída, usando os pesos como parâmetros intermediários. A aprendizagem ocorre por meio da mudança nos pesos que conectam os neurônios, sendo os estímulos externos os dados de treinamento que contêm exemplos do que deve ser aprendido (AGGARWAL, 2018).

Na Figura 2.1, é apresentado um diagrama de um neurônio do mais simples tipo de rede neural artificial: o *perceptron*. Como exemplo, pode ser considerada a situação em que as instâncias de treinamento são da forma $(\overline{X}, y)$, onde cada $\overline{X} = [x_1...x_d]$ contém d variáveis representando, cada uma, um aspecto (*feature*) da entrada e $y \in \mathbb{R}$ contém o valor observado da variável. O objetivo é que a rede seja capaz de atribuir um valor para instâncias de $\overline{X}$ inéditas.

A camada de entrada (*input layer*) contém d nós que transmitem d variáveis de aspecto (*feature*) $\overline{X} = [x_1...x_d]$ com as arestas de pesos $\overline{W} = [w_1...w_d]$ para um nó de saída (*output node*). A função linear $\overline{W} \cdot \overline{X} = \sum_{i=1}^{d} w_j x_j$ é computada. Após isto, a *função de ativação* desse valor real é computada para predizer a variável dependente de $\overline{X}$. A função de ativação é aplicada à soma dos pesos recebidos pelo neurônio, sendo responsável por introduzir a não linearidade que torna as redes neurais capazes de modelar relações complexas. Podem ser usadas diferentes funções de ativação para simular diferentes modelos. Nos casos em que os dados não estão centrados em 0, o peso *bias* (viés) pode ser acrescentado, tendo o efeito de deslocar o resultado final computado pelo nó de saída.

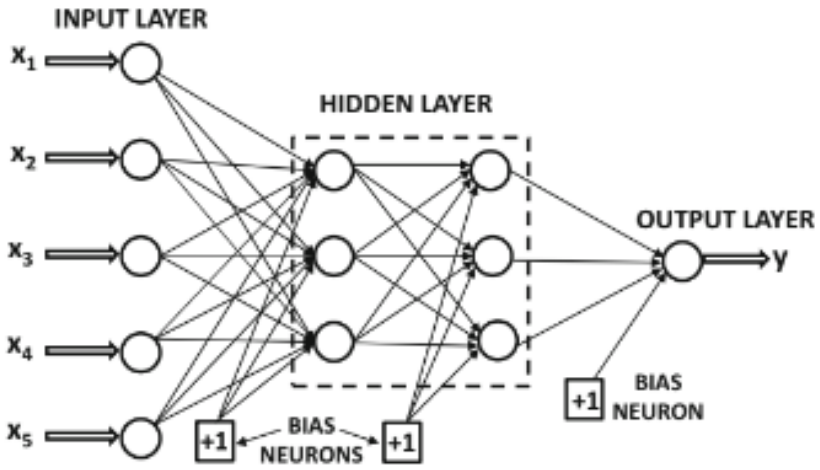
Figura 2.1: Perceptron.



Fonte: Aggarwal (2018)

O perceptron possui apenas uma camada com um neurônio computacional. No entanto, é possível combinar muitos neurônios, resultando em uma rede de múltiplas camadas (Figura 2.2). Assim, aumenta-se a capacidade da rede de modelar funções mais complexas dos dados de entrada. A forma com a qual essas unidades são combinadas, a arquitetura, também é importante. Além disso, é necessário um conjunto de dados de treinamento suficientemente grande (AGGARWAL, 2018).

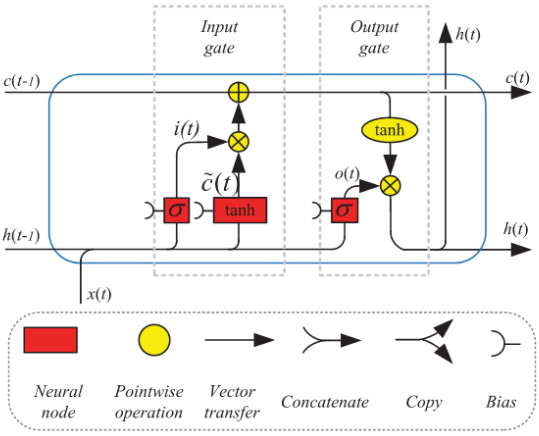
Figura 2.2: Rede neural com múltiplas camadas.



Fonte: Aggarwal (2018)

Redes neurais convencionais enfrentam problemas quando se trata de entradas de tamanho variável, além de não fornecerem informações úteis a respeito de dependências sequenciais. Este tipo de dependência é muito comum em aplicações de processamento

Figura 2.3: Arquitetura da célula LSTM original, onde $h(t)$ é a saída, $x(t)$ é a entrada e $c(t)$ é o estado interno da célula para um dado momento t da série temporal.



Fonte: Yu et al. (2019)

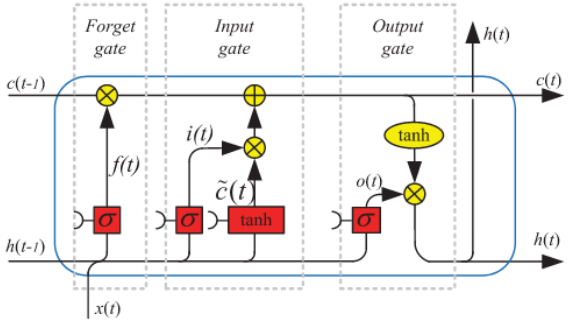
de linguagem natural e análise de séries temporais (AGGARWAL, 2018). Para isto, surge como alternativa a rede neural recorrente (RNN - *recurrent neural network*). Em uma rede neural recorrente, a camada oculta tem seu estado alterado com base na última informação da sequência processada, gerando um efeito de aprendizagem sequencial (ELMAN, 1990).

Quando o intervalo entre dados em uma sequência é grande, as RNNs tradicionais sofrem do chamado problema do gradiente desaparecido (BENGIO; SIMARD; FRASCONI, 1994), apresentando dificuldade em aprender as dependências temporais de longo prazo. Por esse motivo, foram criadas as redes LSTM (*Long Short-Term Memory*), um tipo de RNN que possui propriedades de memória que as tornam muito úteis em tarefas como predição de séries temporais, reconhecimento de fala, entre outros (YU et al., 2019).

Este efeito de memória foi um avanço introduzido por Hochreiter e Schmidhuber (1997) pela criação de uma nova porta dentro da célula, resultando na chamada célula LSTM, representada na Figura 2.3. No entanto, o termo LSTM normalmente se refere a células com uma porta *forget*, introduzida mais tarde por Gers, Schmidhuber e Cummins (2000), que determina se a informação armazenada na célula deve ser mantida ou *esquecida* (ver Figura 2.4).

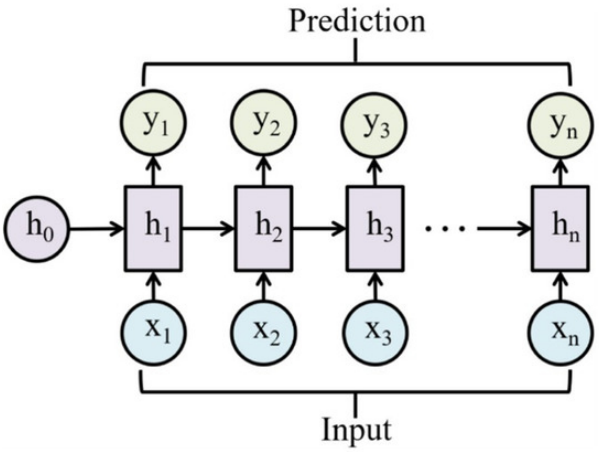
As células LSTM, ao serem combinadas, formam redes que podem ser estruturadas em diferentes arquiteturas, como LSTM empilhada, bidirecional, convolucional, entre outras (YU et al., 2019). A Figura 2.5 ilustra a arquitetura de uma rede LSTM básica.

Figura 2.4: Célula LSTM com porta *forget*, onde $h(t)$ é a saída, $x(t)$ é a entrada e $c(t)$ é o estado interno da célula para um dado momento t da série temporal



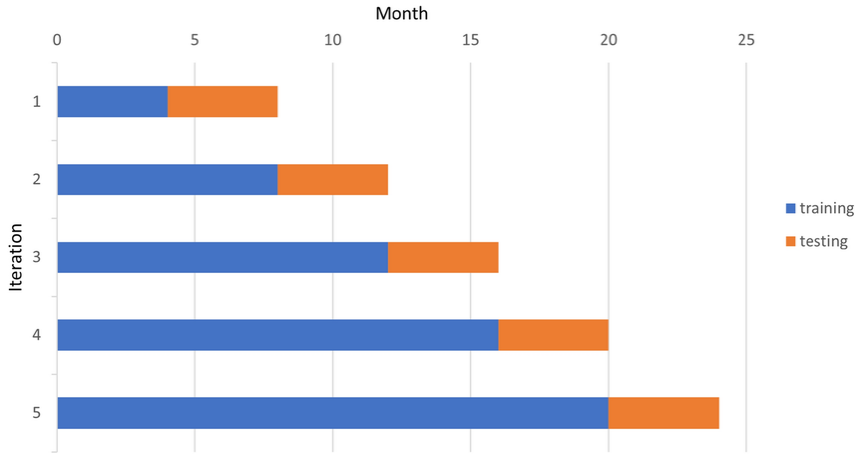
Fonte: Yu et al. (2019)

Figura 2.5: Arquitetura de uma rede LSTM básica, onde os elementos h_t representam o estado oculto, x_t representa a entrada e y_t representa a saída, ou predição, para cada passo t.



Fonte: Mienye, Swart e Obaido (2024)

Figura 2.6: Validação cruzada *forward-chaining*.



Fonte: Phumchusri, Chewcharat e Kanokpongsakorn (2024)

2.3 Métodos de Avaliação

2.3.1 Validação Cruzada

A abordagem mais tradicional para avaliação de modelos de séries temporais é simplesmente particionar a série em duas partes e tomar a última para calcular o erro do modelo, não fazendo uso do conjunto de dados inteiramente (BERGMEIR; BENíTEZ, 2012)

A fim de aproveitar todos os dados disponíveis, a validação cruzada (*cross-validation* ou CV) é um método amplamente utilizado para avaliar algoritmos de classificação e regressão (BERGMEIR; HYNDMAN; KOO, 2018). No entanto, quando se trata de séries temporais, o método de validação cruzada tradicional não é adequado, visto que, neste tipo de dado, existe uma dependência temporal entre os valores da sequência. Sendo assim, torna-se mais apropriado utilizar a validação cruzada por *forward-chaining*, onde o conjunto de treino contém dados até um certo ponto t e o conjunto de teste contém os pontos *t+1, t+2, ..., t+n*. A cada iteração, o conjunto de treino é ampliado, passando a incluir dados até *t+n* e o conjunto de teste passa a ser de $t + n$ até $t + n + s$, onde s é o tamanho definido para o conjunto de teste (BERGMEIR; BENíTEZ, 2012). É importante ressaltar que a cada iteração, ou *fold* (configuração de divisão treino-teste), o modelo é treinado e testado novamente. Este processo, ilustrado na Figura 2.6, simula o uso real do modelo, sendo capaz de avaliar a sua estabilidade e evitar vazamento de dados futuros.

2.3.2 Métricas

Para avaliar o desempenho de técnicas de predição de forma objetiva, é necessário que se use métricas de avaliação. As principais delas são detalhadas a seguir, de acordo com Mahalakshmi, Sridevi e Rajaram (2016). Considere que em uma série com N valores, p_x e $\hat{p}_x$ são o valor real e o valor previsto, respectivamente, e $p_{\max}$ e $p_{\min}$ são os valores máximo e mínimo da previsão obtida.

- Erro Médio Absoluto - *Mean Absolute Error*(MAE):

$$MAE = \frac{1}{N}\sum_{x=1}^{N}|p_x - \hat{p}_x| \tag{2.7}$$

- Erro Médio Absoluto Normalizado - *Normalized Mean Absolute Error* (NMAE):

$$NMAE = \frac{MAE}{p_{\max} - p_{\min}} \tag{2.8}$$

- Erro Médio Absoluto Percentual - *Mean Absolute Percentage Error* (MAPE):

$$\text{MAPE } = \frac{1}{N}\sum_{x=1}^{N}\left|\frac{p_x - \hat{p}_x}{p_x}\right| * 100 \tag{2.9}$$

- Erro Médio Quadrático - *Mean Square Error* (MSE):

$$MSE = \frac{1}{N}\sum_{x=1}^{N}(p_x - \hat{p}_x)^2 \tag{2.10}$$

- Raiz do Erro Médio Quadrático - *Root Mean Square Error* (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{x=1}^{N}(p_x - \hat{p}_x)^2} \tag{2.11}$$

- Coeficiente de Determinação - R^2:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{p}_i - \bar{p})^2}{\sum_{i=1}^{n}(p_i - \bar{p})^2} \tag{2.12}$$

A métrica MAPE possui uma limitação no caso em que o denominador é zero, no entanto isto não apresenta problemas para este trabalho, visto que os valores considerados são maiores que zero.

2.4 Sistemas de Apoio à Decisão

Segundo Bonczek, Holsapple e Whinston (2014), a decisão pode ser vista como um produto de um sistema de processamento de informações humano-máquina. Um sistema que processa informações e produz como resultado uma decisão pode ser chamado de sistema de tomada de decisões. Um sistema de processamento de informações inserido em um sistema de tomada de decisões é referido como um sistema de apoio à decisão (SAD). Estes podem ser humanos, mecânicos ou sistemas humano-máquina inseridos na estrutura de tomada de decisões de uma organização e possuir diferentes tipos de habilidades de processamento de informações e fazê-lo em diferentes graus.

Algumas definições já foram propostas ao conceito de apoio à decisão. Segundo os autores previamente citados, ao examinar sistemas existentes, é possível observar algumas características em comum, detalhadas a seguir: 1) sistemas de apoio à decisão são aqueles que dão suporte ao tomador de decisão ao resolver problemas não programados, não estruturados ou semiestruturados. Além disso, 2) oferecem algum tipo de linguagem ou interface para que o usuário faça solicitações ao sistema para busca de informações, análises e relatórios. O sistema pode permitir análises *ad hoc* (não padronizadas, personalizas).

3 Trabalhos Relacionados

Neste capítulo são apresentados trabalhos que estão relacionados ao tema desta monografia, observando seus objetivos, principais componentes e contribuições. São considerados relacionados todos aqueles trabalhos que apliquem técnicas de predição a preços de produtos agrícolas considerando fatores exógenos e façam previsões a longo prazo (múltiplos passos à frente), ou seja, usem séries multivariadas e realizem previsões *multi-step*. Ao final do capítulo é apresentada a Tabela 3.1, que sumariza as características observadas nos trabalhos citados.

Madaan et al. (2019) buscou prever preços da cebola e batata na Índia e também identificar eventos anômalos que podem influenciar a variação dos preços, como a acumulação de produtos pelos vendedores. Foram coletados dados sobre eventos climáticos ou acumulação a partir de notícias presentes na internet. Os autores realizaram comparações entre modelos de previsão univariados (ARIMA e SARIMA) e modelos multivariados (SARIMAX e LSTM), fazendo previsões de forma recursiva para um horizonte de 30 dias. A análise dos resultados indicou que os modelos multivariados obtiveram um resultado superior aos univariados, sendo o SARIMAX o melhor. Além disso, as melhorias foram percebidas ao incorporar, progressivamente, duas variáveis exógenas: preço de centro comercial vizinho mais correlacionado (identificado por meio de uma análise de *shifted correlation*) e, posteriormente, o CPI (*Consumer Price Index*). Os dados climáticos foram usados apenas para fins de análise e classificação de anomalias, não contribuindo para a previsão dos preços.

Harrykissoon e Hosein (2023) compararam a qualidade de previsões *mustistep* recursivas com previsões diretas utilizando o modelo SARIMAX para produtos agrícolas em Trinidad e Tobago. Os produtos cujo preço foi utilizado foram: tomate, banana, laranja, gengibre, pimenta, alface, abóbora, repolho, pepino e mandioca. As variáveis exógenas incluem volume de vendas, precipitação, temperaturas máxima e mínima. Foi observado que a previsão direta obteve, em geral, melhores resultados do que a recursiva, visto que nesta existe um acúmulo progressivo de erros de predição a cada passo. No

entanto, foi apontado que a previsão recursiva pode dar resultados razoáveis para séries com comportamento mais linear e que sejam estacionárias. Por fim, notou-se que a previsão recursiva de um passo à frente foi melhor que o *Naive Seasonal Mean Model*.

Considerando as limitações de modelos tradicionais como o SARIMAX, Özden (2023) compara os modelos convolutional neural network (CNN),*Random Forest* (RF) e LSTM para previsão dos preços da batata, cebola e alho na Turquia, utilizando como variável exógena o volume comercializado destes produtos. O estudo considerou um horizonte de previsão de 10 dias. De acordo com as métricas MAE e RMSE obtidas, concluiu-se que o resultado do modelo LSTM foi um pouco superior (mas muito próximo) do RF, enquanto o CNN demonstrou taxas de erro expressivamente menores que as dos outros dois citados. No entanto, foram utilizadas, para todos os modelos, arquiteturas básicas.

Utilizando a mesma base de dados do estudo mencionado acima, Özden e Bulut (2024) introduzem, no contexto de predição de preços de produtos agrícolas, o uso da *Spectral Temporal Graph Neural Network* (StemGNN): arquitetura que modela relações temporais e espaciais. O desempenho preditivo é realizado em dois horizontes diferentes (5 e 10 dias) e a avaliação deste é feita por meio das métricas MAE e RMSE. O modelo proposto é comparado com os modelos RF, CNN e LSTM, com a mesma arquitetura empregada no estudo anterior, obtendo os melhores resultados para o StemGNN.

Min et al. (2025), dando sequência à discussão sobre redes StemGNN, faz uma análise comparativa abrangente entre modelos baseados em GNN, quais sejam StemGNN e *temporal graph convolutional network* (T-GCN), e aqueles baseados em RNN, neste caso Stacked LSTM multivariado e univariado. Os dados utilizados dizem respeito aos preços da batata, cebola, alface e pepino na Coreia do Sul e incluem, também, variáveis exógenas meteorológicas. O estudo compara os resultados dos modelos citados para previsão utilizando diferentes janelas de suavização de dados (7, 14, 21 e 28 dias), tamanhos de janela de dados de entrada (15, 30, 45 e 60 dias), horizontes de previsão (7 e 14 dias) e, além de comparar modelos multivariados e univariados, avalia as principais variáveis que afetam o preço dos produtos estudados.

A partir das análises apresentadas, os autores demonstram que a suavização contribui de forma eficaz para a melhoria do desempenho nos modelos multivariados (espe-

cialmente para predição de séries mais voláteis) e que, ademais, as variáveis exógenas contribuem significativamente para a predição dos preços. A correlação entre a variável-alvo e as exógenas (verificada por uma análise de correlação cruzada) pode variar para diferentes culturas e a correta identificação e uso de variáveis altamente correlacionadas ao que se deseja prever pode ajudar a prevenir o fenômeno chamado *time-shift*. Por fim, o trabalho conclui que os modelos baseados em GNN apresentados são mais eficazes no contexto agrícola com variáveis exógenas.

Waeodi, Boongasame e Thammarak (2025) avaliaram o desempenho de modelos LSTM e CNN na previsão de preços da glória-da-manhã (vegetal folhoso amplamente usado na culinária asiática) utilizando variáveis meteorológicas e de eventos como feriados, sábados e domingos. Com uma ênfase na preparação e tratamento de dados, o estudo realizou um processo de seleção de *features* baseado no VIF (Fator de Inflação da Variância), uma medida estatística que avalia a multicolinearidade entre variáveis explicativas, permitindo a remoção de atributos redundantes e potencialmente danosos ao modelo.

Ajustes finos de hiperparâmetros foram realizados com base diferentes configurações de modelos e avaliados meio de testes estatísticos. Além disso, foi realizada uma análise para determinar qual o impacto de cada variável envolvida na predição utilizando o processo chamado Análise de Sensibilidade de Sebol (*Sebol Sensitivity Analysis*). Mostrou-se, por meio das métricas MSE, RMSE, MAPE e MAE, que os modelos que passaram pelo processo de seleção de variáveis obtiveram resultados superiores aos que não passaram e que, os modelos LSTM obtiveram, de forma consistente, resultados superiores aos dos modelos CNN.

O estudo reforça a importância da engenharia de variáveis e da análise estatística no aprimoramento da acurácia preditiva em contextos agrícolas e contribui ao apresentar configurações otimizadas de modelos LSTM e CNN específicas para a previsão de preços da glória-da-manhã, além de empregar ferramentas estatísticas como o VIF para seleção de variáveis e a Análise de Sensibilidade de Sobol para avaliar a influência relativa dos fatores preditivos.

A Tabela 3.1 resume aspectos relevantes dos trabalhos analisados neste capítulo.

Em comum, todos os estudos contribuem significativamente para a previsão *multistep* de preços de produtos agrícolas com o uso de variáveis exógenas, tanto no que diz respeito às técnicas preditivas adotadas quanto ao emprego de ferramentas estatísticas para análise e avaliação. Observa-se, no entanto, que a análise do comportamento dos modelos em diferentes horizontes de previsão, embora presente em alguns trabalhos, ainda pode ser aprofundada, especialmente no que diz respeito à evolução dos erros e à adequação de modelos a horizontes mais longos.

Além disso, os estudos analisados são, em geral, voltados a contextos geográficos específicos, com características mercadológicas e climáticas próprias, o que limita sua generalização para outras regiões. Essa limitação reforça a necessidade de análises localizadas, como a que se propõe neste trabalho, com foco no contexto regional específico dos dados utilizados.

Outro ponto observado é a subutilização de técnicas modernas de avaliação, como a validação cruzada para séries temporais, que pode melhorar o aproveitamento de dados e fornecer avaliações mais robustas, especialmente em conjuntos reduzidos.

Por fim, destaca-se que modelos mais recentes e promissores, como o Prophet, não foram considerados nos estudos analisados. Trata-se de um modelo com baixa complexidade e mínima necessidade de ajuste de parâmetros, o que o torna uma alternativa interessante para comparação com modelos mais sofisticados, sobretudo em contextos com restrições de dados ou infraestrutura. A exploração de seu desempenho neste cenário representa, portanto, uma contribuição adicional deste trabalho.

Tabela 3.1: Trabalhos relacionados sobre previsão de preços agrícolas.

#	País	Produtos	Métodos	Horizontes	Variáveis Exógenas
1	Índia	cebola batata	ARIMA SARIMA SARIMAX LSTM	30 dias	preços vizinho CPI
2	Trinidad e Tobago	tomate banana laranja gengibre pimenta alface abóbora repolho pepino mandioca	SARIMAX	12 meses	vol. vendas clima
3	Turquia	batata alho cebola	CNN LSTM RF	10 dias	vol. vendas
4	Turquia	pimentão tomate abóbora pepino	StemGNN CNN LSTM	5 dias 10 dias	vol. vendas
5	Coreia do Sul	batata cebola alface repolho	Stacked LSTM T-GCN StemGNN	7 dias 14 dias	clima
6	Tailândia	glória-da-manhã	LSTM CNN	5 dias 7 dias 14 dias 21 dias	clima
7	**Brasil**	**alface crespa**	**LSTM** **SARIMAX** **Prophet**	**1 a 8 semanas**	**clima**

Fonte: Elaborado pelo autor.

Referências: [1] Madaan et al. (2019). [2] Harrykissoon e Hosein (2023), [3] Özden (2023), [4] Özden e Bulut (2024), [5]Min et al. (2025), [6] Waeodi, Boongasame e Thammarak (2025), [7] este trabalho.

4 Metodologia

O presente trabalho é uma pesquisa quantitativa sobre modelos computacionais para previsão de séries temporais de preços na agricultura familiar. O processo metodológico detalhado neste capítulo envolve coleta de dados de séries temporais junto às instituições competentes, análise e tratamento dos dados, implementação de modelos de aprendizado de máquina e avaliação do desempenho dos mesmos.

4.1 Conjunto de Dados

Para este estudo, foram utilizados tanto dados de preços de hortaliças quanto de variáveis meteorológicas. Foram utilizados dados de preços semanais de hortaliças na cidade de Teresópolis - RJ de janeiro de 2016 até dezembro de 2023, obtidos a partir do Centro de Estudos Avançados em Economia Aplicada (CEPEA)[4]. A base possui dados de apenas alguns produtos, quais sejam: alface crespa, alface lisa e alface americana. Para alface americana e alface crespa, há também séries de preços de atacado (preços ao consumidor final), mas essas séries se encontravam vazias, sendo aproveitadas para este trabalho apenas as variáveis apresentadas na Tabela 4.1, referentes aos preços dos produtos vendidos pelos produtores (na base possuem sufixo "Roça").

Tabela 4.1: Tabela de Variáveis

Nome da Variável
Alface Americana
Alface Crespa
Alface Lisa

Fonte: Elaborado pelo autor.

O conjunto de dados meteorológicos foi obtido a partir do Instituto Brasileiro de Meteorologia (INMET)[5]. Ele cobre o mesmo período que as séries de preços (janeiro de 2016 a dezembro de 2023) e contém medições realizadas a cada hora por uma estação

[4] ⟨https://www.hfbrasil.org.br/br/banco-de-dados-precos-medios-dos-hortifruticolas.aspx⟩
[5] ⟨https://portal.inmet.gov.br/dadoshistoricos⟩

meteorológica automática presente na cidade de Teresópolis - RJ e as variáveis registradas pela estação são apresentadas na Tabela 4.2. Para maior facilidade de leitura, os nomes das variáveis foram abreviados.

Tabela 4.2: Tabela de Variáveis

Nome da Variável	**Nome simplificado**
PRECIPITAÇÃO TOTAL, HORÁRIO (mm)	PREC_HOR
PRESSÃO ATMOSFÉRICA AO NÍVEL DA ESTAÇÃO, HORÁRIA (mB)	PRES_EST_HOR
PRESSÃO ATMOSFÉRICA MÁX. NA HORA ANT. (AUT) (mB)	PRES_MAX_ANT
PRESSÃO ATMOSFÉRICA MÍN. NA HORA ANT. (AUT) (mB)	PRES_MIN_ANT
RADIAÇÃO GLOBAL (Kj/m²)	RAD_GLOB
TEMPERATURA DO AR - BULBO SECO, HORÁRIA (°C)	TEMP_AR_HOR
TEMPERATURA DO PONTO DE ORVALHO (°C)	TEMP_ORV
TEMPERATURA MÁXIMA NA HORA ANT. (AUT) (°C)	TEMP_MAX_ANT
TEMPERATURA MÍNIMA NA HORA ANT. (AUT) (°C)	TEMP_MIN_ANT
TEMPERATURA ORVALHO MÁX. NA HORA ANT. (AUT) (°C)	ORV_MAX_ANT
TEMPERATURA ORVALHO MÍN. NA HORA ANT. (AUT) (°C)	ORV_MIN_ANT
UMIDADE REL. MÁX. NA HORA ANT. (AUT) (%)	UMID_MAX_ANT
UMIDADE REL. MÍN. NA HORA ANT. (AUT) (%)	UMID_MIN_ANT
UMIDADE RELATIVA DO AR, HORÁRIA (%)	UMID_AR_HOR
VENTO, DIREÇÃO HORÁRIA (gr) (°(gr))	VENT_DIR_HOR
VENTO, RAJADA MÁXIMA (m/s)	VENT_RAJA_MAX
VENTO, VELOCIDADE HORÁRIA (m/s)	VENT_VEL_HOR

Fonte: Elaborado pelo autor.

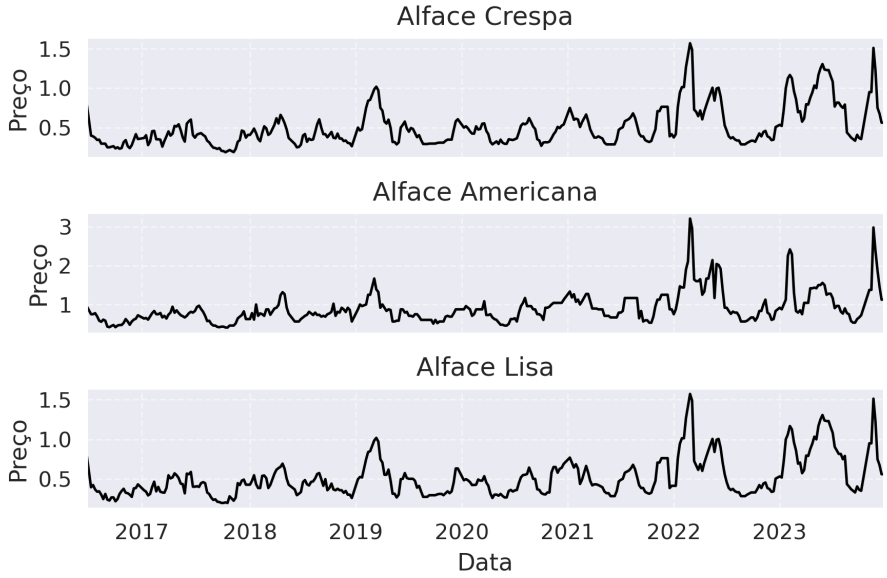
4.2 Análise e Tratamento dos Dados

4.2.1 Conjunto de Dados de Preço

Ao analisar os dados ausentes das variáveis de preço, percebeu-se que as séries “Alface Americana - Atacado” e “Alface Crespa - Atacado” não possuíam dados, logo, foram descartadas. As demais variáveis possuíam cerca de 5% de dados ausentes e a estratégia utilizada para o seu preenchimento foi a interpolação linear, tradicionalmente usada na literatura e facilmente aplicada. A Figura 4.1 apresenta um gráfico destas variáveis.

Nota-se que os preços dos produtos, embora numericamente diferentes, seguem padrões extremamente semelhantes, logo, neste estudo, será utilizada apenas a variável “Alface Crespa”, referida a partir de agora simplesmente como “preço”.

Figura 4.1: Preços dos produtos da base de dados



Fonte: Elaborado pelo autor.

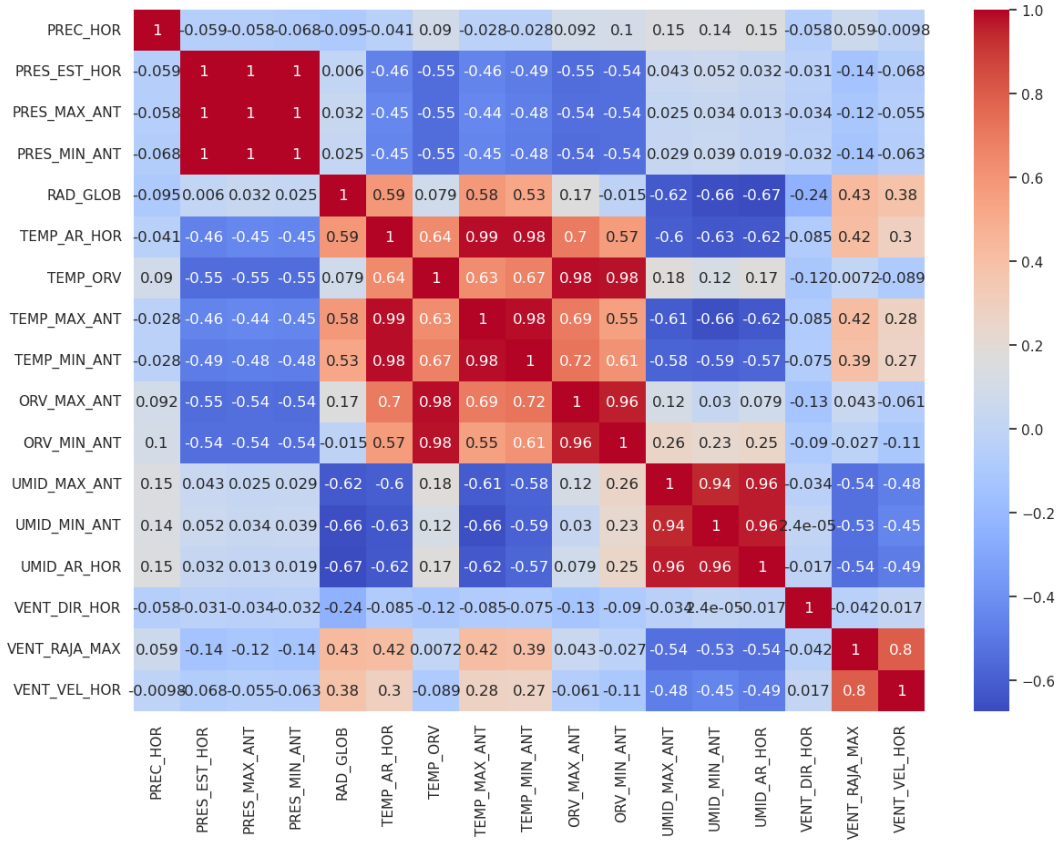
4.2.2 Conjunto de Dados Meteorológicos

Algumas séries do conjunto de dados meteorológicos possui alguns valores negativos (abaixo de -900), o que não é coerente, visto que elas representam grandezas físicas. Assim, estes valores foram entendidos como uma representação de valores ausentes e removidos da base. Verificou-se, então, que a base de dados meteorológicos possui dados ausentes, sendo necessária uma estratégia para seu preenchimento. Para tanto, a estratégia adotada foi a interpolação linear.

4.2.3 Redução do Conjunto de Dados

Devido ao grande número de variáveis presentes na base, faz-se necessário selecionar quais delas são mais relevantes para este trabalho. Foi realizada uma análise de correlação (Figura 4.2) entre as variáveis climáticas e notou-se grande correlação entre algumas delas, especialmente aquelas que se referem a medidas relacionadas ao mesmo aspecto climático, por exemplo, variáveis associadas à temperatura. Cada conjunto de variáveis altamente correlacionadas — consideradas aqui com coeficiente de correlação de Spearman (WISSLER, 1905) maior ou igual a 0.9 — teve todos os elementos menos um descartado. A Tabela 4.3 apresenta os grupos de variáveis altamente correlacionadas e, dentro de cada grupo, qual variável foi selecionada para prosseguir no estudo.

Figura 4.2: Mapa de correlação entre as variáveis meteorológicas.



Fonte: elaborado pelo autor.

4.2.4 Conjunto de Dados Unificado

Tendo em vista que os dados desta base têm periodicidade horária e os da base de preços têm periodicidade semanal, optou-se por fazer uma reamostragem. Este é um processo de modificação da frequência de uma série temporal por meio de funções de agregação. Ou seja, no caso deste conjunto de dados, tomaram-se os dados horários de cada semana e foi aplicada uma função de agregação, tornando a sua periodicidade semanal.

Existem algumas funções de agregação que podem ser usadas e, para entender qual seria a mais adequada, os dados foram reamostrados usando diferentes funções: média, soma, mínimo e máximo. Min et al. (2025) apontou efeitos positivos da suavização das séries temporais sobre o desempenho dos modelos; então, a partir de todas as séries obtidas na etapa anterior, foram incluídas também suas versões suavizadas, utilizando médias móveis com uma janela de tamanho 4 (semanas). Este valor foi escolhido a partir das análises feitas por Min et al. (2025), que obteve os melhores resultados ao suavizar as suas séries por meio de médias móveis com janela de 28 dias.

Tabela 4.3: Grupos de variáveis altamente correlacionadas e variáveis selecionadas.

Grupo	**Variáveis**	**Selecionada**
1	PREC_HOR	PREC_HOR
2	PRES_EST_HOR, PRES_MAX_ANT, PRES_MIN_ANT	PRES_MIN_ANT
3	RAD_GLOB	RAD_GLOB
4	TEMP_AR_HOR, TEMP_MAX_ANT, TEMP_MIN_ANT	TEMP_AR_HOR
5	TEMP_ORV, ORV_MAX_ANT, ORV_MIN_ANT	TEMP_ORV
6	UMID_MAX_ANT, UMID_MIN_ANT, UMID_AR_HOR	UMID_AR_HOR
7	VENT_DIR_HOR	VENT_DIR_HOR
8	VENT_RAJA_MAX, VENT_VEL_HOR	VENT_VEL_HOR

Fonte: elaborado pelo autor

Assim, para cada variável original com frequência horária, foram obtidas 8 diferentes variáveis com frequência semanal. Estas séries foram alinhadas às séries de preços com base em identificadores de semana. Dessa forma, cada ponto da série de preços foi associado aos dados climáticos correspondentes à mesma semana, viabilizando a análise conjunta das duas séries.

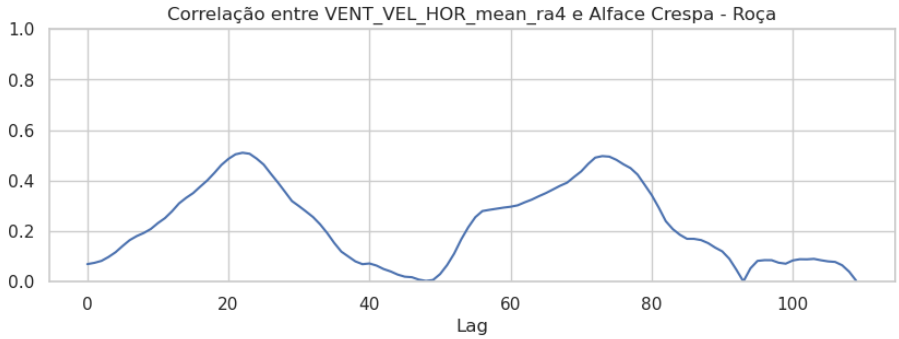
4.2.5 Seleção de Variáveis

Além da análise de correlação entre as variáveis exógenas, foi realizada uma análise de correlação deslocada entre as variáveis exógenas e a variável-alvo de modo a determinar quais delas seriam mais apropriadas a compor os modelos preditivos, visto que um grande número de variáveis, especialmente quando não trazem informações úteis sobre a variável predita, pode introduzir ruído e degradar o desempenho do modelo. *Correlação deslocada* indica a correlação entre uma série temporal e outra com diferentes defasagens (*lags*). Isto, em outras palavras, mede quanto uma série y_k se correlaciona com uma outra série x_{t+k} ao longo de vários *lags* k.

A Figura 4.3 mostra um exemplo de correlação deslocada entre a variável VENT_VEL_HOR_mean_ra4 e o preço. Pode-se ver que, neste caso, a série de preços tem uma correlação moderada com o valor dessa variável deslocada por aproximadamente 20 passos. Em outras palavras, o preço em um certo momento tem uma correlação moderada com o valor desta variável cerca de 20 semanas antes deste momento.

A Tabela 4.4 traz os valores máximos (em módulo) das correlações deslocadas

Figura 4.3: Exemplo de correlação deslocada (em módulo) entre uma variável e o preço.



Fonte: elabordo pelo autor.

observadas para as variáveis com os maiores valores nesse critério. Foi acrescentado um sufixo aos nomes de todas as variáveis indicando a função de agregação e se houve suavização (sufixo "_ra4", quando positivo).

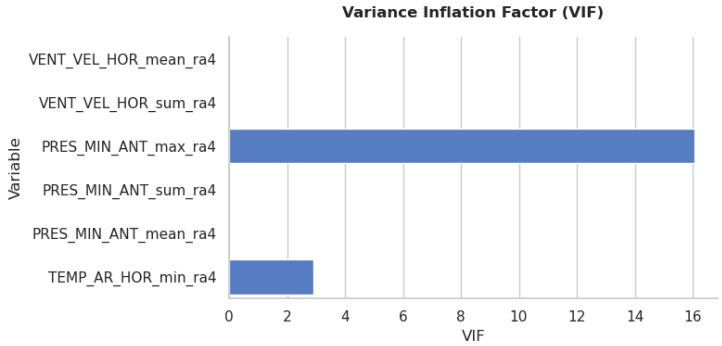
Tabela 4.4: Correlação (em módulo) deslocada máxima entre variáveis exógenas e o preço

Variável	**Correlação Máxima**
VENT_VEL_HOR_mean_ra4	0.531
VENT_VEL_HOR_sum_ra4	0.531
PRES_MIN_ANT_max_ra4	0.476
PRES_MIN_ANT_sum_ra4	0.459
PRES_MIN_ANT_mean_ra4	0.459
TEMP_AR_HOR_min_ra4	0.449

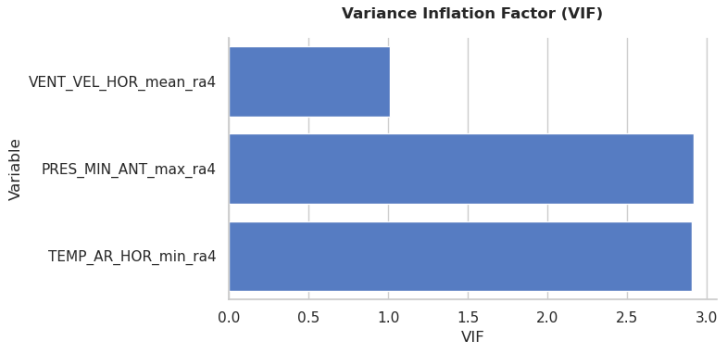
Fonte: elaborado pelo autor.

Dentre as variáveis selecionadas, foi realizada, ainda, uma análise de multicolinearidade – dependência quase linear entre os regressores (MONTGOMERY; JENNINGS; KULAHCI, 2015) –, visto que a presença de variáveis com multicolinearidade pode introduzir ruído no modelo ao incorporar informações redundantes. A Figura 4.4a exibe os valores calculados do VIF *Variance Inflation Factor* (VIF) para cada variável (os elementos onde não aparece barra indicam que o valor calculado do VIF foi igual a infinito). São removidas progressivamente as séries de maior VIF até que todas apresentem um valor menor que 5, como feito por Waeodi, Boongasame e Thammarak (2025), sendo obtido o resultado apresentado na Figura 4.4b. As variáveis escolhidas para os próximos passos são: VENT_VEL_HOR_mean_ra4, PRES_MIN_ANT_max_ra4 e TEMP_AR_HOR_min_ra4.

Figura 4.4: Comparação das análises de multicolinearidade entre as variáveis por meio do VIF.



(a) Análise inicial.



(b) Após remoção de variáveis com VIF alto.

Fonte: elaborado pelo autor.

4.2.6 Normalização dos Dados

Todos os dados foram normalizados pelo método MinMax (HAN; KAMBER, 2012), aplicados após a repartição de conjuntos de treino e teste gerada pela validação cruzada, visto que, a fim de reduzir o vazamento de informação entre o conjunto de treino e teste, deve-se aplicar a normalização apenas ao conjunto de treino e utilizar a mesma escala no conjunto de teste.

4.3 Seleção dos Modelos

Os modelos selecionados para esta comparação foram: SARIMAX, LSTM e Prophet. Os dois primeiros são amplamente utilizados na literatura para predição de séries temporais e o último é um modelo mais recente, escolhido por sua simplicidade de configuração e uso.

4.3.1 SARIMAX

Primeiramente, foi necessário selecionar os parâmetros do modelo SARIMAX. Conforme discutido na Subseção 2.2.1, o modelo pode ser escrito como SARIMAX (p,d,q)(P,D,Q,S)+X, tornando-se necessário definir estes parâmetros. Para definir os parâmetros d e D, foram utilizados os testes estatísticos ADF e KPSS. A Tabela 4.5 indica os resultados dos testes.

Como um dos testes indicou que a série não é estacionária, foi realizado um processo de diferenciação na série e o teste refeito. Os resultados do segundo teste são apresentados na Tabela 4.6.

Tabela 4.5: Resultados do primeiro teste de estacionaridade

Teste	**p-valor**	**Nível de Significância**	**Conclusão**
ADF	0.000013	5%	Rejeita H_0 (série estacionária)
KPSS	0.010	5%	Rejeita H_0 (série não estacionária)

Fonte: Elaborado pelo autor.

Tabela 4.6: Resultados do segundo teste de estacionaridade

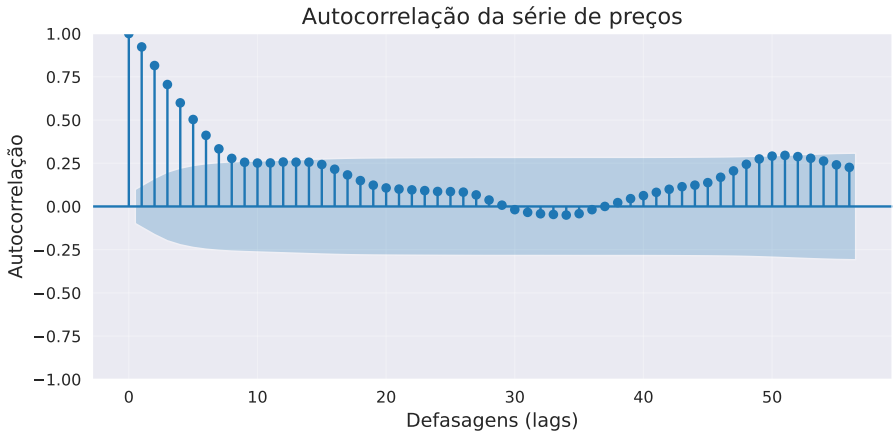
Teste	**p-valor**	**Nível de Significância**	**Conclusão**
ADF	5.98×10^{-18}	5%	Rejeita H_0 (série estacionária)
KPSS	0.100	5%	Não rejeita H_0 (série estacionária)

Fonte: Elaborado pelo autor.

Como, na primeira execução, o teste KPSS indicou não estacionaridade e discordou do teste ADF, a série pode ser entendida como estacionária por diferenciação[6]. Visto que, após uma diferenciação, os testes concordam em indicar estacionaridade, define-se o parâmetro $d = 1$. Pelo gráfico de autocorrelação (Figura 4.5), série de preços não demonstra uma forte sazonalidade, no entanto, de modo a não desconsiderar esta possibilidade, deixa-se o valor do parâmetro D em aberto. Com e $S = 52$ (número de semanas em um ano e supondo que possa haver sazonalidade anual), restam apenas os parâmetos D, p, P, Q e q a serem determinados. Para tal foi feito um *Grid-Search* utilizando o *Akaike Information Criterion* (AIC) como parâmetro de comparação. A partir disto, os parâmetros foram definidos da seguinte forma: $D = 0$, $p = 0$, $P = 0$ e $q = 1$, ficando a forma do modelo definida como SARIMAX(0, 1, 1),(0, 0, 0, 52)+X.

[6] ⟨https://www.statsmodels.org/dev/examples/notebooks/generated/stationarity_detrending_adf_kpss.html⟩

Figura 4.5: Autocorrelação da série de preços.



Fonte: elaborado pelo autor.

4.3.2 LSTM

Foi utilizada uma arquitetura LSTM básica, cujos parâmetros foram tomados a partir de Murugesan, Mishra e Krishnan (2021), que realizam a comparação de diferentes arquiteturas LSTM aplicadas à predição de preços de produtos agrícolas. Estes parâmetros são definidos como mostrado na Tabela 4.7.

Tabela 4.7: Hiperparâmetros

Camada	**Parâmetro**	**Valor**
LSTM	Unidades (units)	`100`
	Função de ativação	`ReLU`
Dropout	Taxa de dropout	`0.2`
Dense	Unidades	`80`
	Função de ativação	`ReLU`
Output	Unidades	`8`
	Função de ativação	`Linear`
Compilação	Função de perda	`MSE`
	Otimizador	`adam`

Fonte: Elaborado pelo autor.

4.3.3 Prophet

Conforme Subseção 2.2.2, o modelo Prophet foi desenvolvido para que tenha facilidade de uso e exija pouca configuração manual de hiperparâmetros. Sua abordagem estruturada para a decomposição da série temporal em componentes de tendência, sazonalidade e

feriados permite obter bons resultados com uma configuração mínima.

Por essa razão, optou-se por utilizar os parâmetros padrão fornecidos pela biblioteca, uma vez que o Prophet foi projetado para ser robusto mesmo com essa configuração.

4.4 Descrição dos Experimentos

Assim, com os modelos e parametrizações explicitados na Seção 4.3 , foi estruturado um processo geral que consiste na normalização dos dados, execução do modelo e denormalização do resultado. Foi utilizada validação cruzada (Seção 2.3.1) com 10 blocos (valor frequentemente utilizado na literatura) e o horizonte de predição é de 8 semanas, ou seja, dado um momento *t*, deseja-se que o modelo preveja *t+1, ..., t+8* e o erro é calculado sobre todos os pontos previstos. O horizonte de previsão de 8 semanas foi escolhido por ser próximo ao tempo indicado por Amaro et al. (2007) para início da colheita após o plantio.

Os modelos foram executados, primeiramente, como modelos univariados. Em outras palavras, utilizou-se apenas a variável-alvo (preço). Após isto, foram acrescentadas progressivamente três variáveis exógenas, partindo das mais correlacionadas ao preço, a fim de analisar o impacto destas no desempenho preditivo do modelo. Para o treinamento da rede neural LSTM foram utilizadas 200 épocas, com tamanho de lote igual a 20.

5 Resultados e Discussão

Este capítulo apresenta os resultados obtidos a partir do processo descrito no capítulo anterior, buscando realizar uma comparação entre os diferentes métodos aplicados à predição de preços. Primeiro, são analisadas e discutidas métricas gerais de erro de cada técnica utilizada, considerando também as variáveis exógenas e, após isto, é feita uma análise sobre os erros obtidos para cada horizonte de previsão por cada modelo. O ambiente de desenvolvimento utilizado foi baseado na linguagem Python 3.10.16, com a biblioteca *SciKit Learn*[7] para o modelo LSTM, *statsmodels*[8] para o SARIMAX e, para o Prophet, a sua biblioteca própria: *fbprophet*[9]. O código do projeto está disponível em repositório no GitHub[10].

A Tabela 5.1 apresenta os resultados dos diferentes modelos utilizados. Para cada modelo foram acrescentadas progressivamente novas variáveis exógenas a fim de analisar o impacto destas na qualidade das soluções. As variáveis exógenas acrescentadas foram, em ordem, após o preço: VENT_VEL_HOR_mean_ra4, PRES_MIN_ANT_max_ra4 e TEMP_AR_HOR_min_ra4.

Nesta análise geral, o Prophet obteve o pior desempenho, o LSTM com três variáveis exógenas obteve os melhores resultados segundo as métricas R2 e MAPE, enquanto o SARIMAX com duas variáveis exógenas aparece como superior segundo o MSE, MAE e RMSE. No entanto, percebe-se que o desempenho destes foi bastante próximo na maioria das métricas utilizadas.

Os modelos LSTM parecem se beneficiar da adição de novas variáveis exógenas enquanto o Prophet e o SARIMAX não apresentam melhora nem piora significativa. O comportamento dos erros do modelo LSTM em comparação com os demais destaca a capacidade de modelos baseados em redes neurais de extrair relações não lineares complexas a partir dos dados de entrada, indicando que este pode ser um bom caminho para inves-

[7] ⟨https://scikit-learn.org/⟩
[8] ⟨https://www.statsmodels.org/⟩
[9] ⟨https://facebook.github.io/prophet/⟩
[10] ⟨https://github.com/Patrick448/time-series-analysis⟩

Tabela 5.1: Desempenho dos modelos para previsão de preços de Alface Crespa para cada conjunto de variáveis.

Nome	**# Variáveis**	**MSE**	$\boldsymbol{R^2}$	**MAE**	**MAPE**	**RMSE**
lstm1	1	0.102	-0.032	0.249	0.420	0.319
lstm2	2	0.088	0.143	0.235	0.347	0.297
lstm3	3	0.091	0.124	0.230	0.354	0.301
lstm4	4	0.085	**0.159**	0.215	**0.295**	0.291
prophet1	1	0.095	0.038	0.255	0.462	0.308
prophet2	2	0.100	-0.018	0.259	0.467	0.316
prophet3	3	0.101	-0.033	0.259	0.469	0.318
prophet4	4	0.104	-0.061	0.265	0.485	0.322
sarimax1	1	0.076	0.078	**0.203**	0.339	0.277
sarimax2	2	0.078	0.063	0.204	0.342	0.279
sarimax3	3	**0.075**	0.103	**0.203**	0.341	**0.273**
sarimax4	4	**0.075**	0.093	0.204	0.344	0.274

Fonte: elaborado pelo autor.

Nota: Na coluna Variáveis, 1 indica dados apenas de preço, sendo acrescentados nos demais, de forma progressiva, as seguintes: 2) VENT_VEL_HOR_mean_ra4, 3) PRES_MIN_ANT_max_ra4 e 4) TEMP_AR_HOR_min_ra4

tigações futuras, podendo incluir outras parametrizações ou mesmo outras arquiteturas baseadas em redes LSTM.

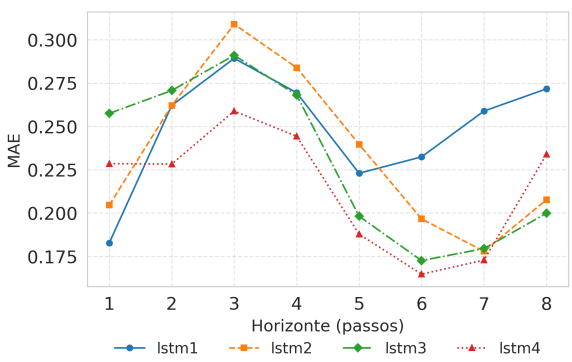
As análises apresentadas acima dão uma visão geral dos modelos, porém deve-se notar que as previsões realizadas são de longo prazo e os erros não necessariamente são os mesmos para todos os passos preditos. A Figura 5.1 apresenta o comportamento do erro (a partir da métrica MAE) dos diferentes modelos para cada horizonte de previsão, de 1 até 8 semanas (ou passos) no futuro. Todos os modelos SARIMAX utilizados (Figura 5.1b) apresentam desempenhos piores conforme o horizonte de previsão aumenta, não sendo observada mudança nesta tendência relacionada ao acréscimo de variáveis exógenas.

Quanto ao LSTM (Figura 5.1a), observa-se uma piora no desempenho progressiva até o passo 3 e, posteriormente, uma queda no erro ao redor dos passos 6 e 7 para todos os modelos exceto o lstm4 (LSTM com quatro variáveis). Neste o erro é menor no passo 5, voltando a crescer posteriormente mais rapidamente do que os outros modelos. Este comportamento indica que as variáveis exógenas, para este modelo, podem contribuir para mitigar erros em previsões de longo prazo.

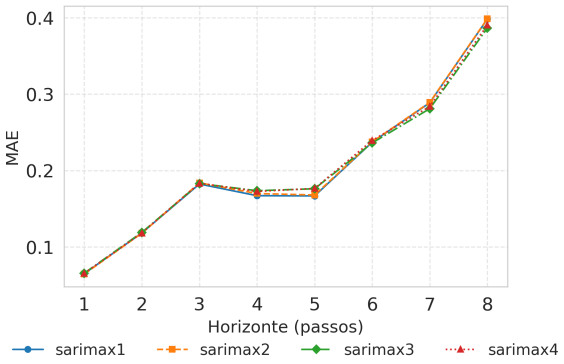
O modelo Prophet (Figura 5.1c) apresenta comportamento semelhante ao do LSTM, mas observa-se menos influência das variáveis exógenas. Além disso, os dois

modelos têm uma piora no desempenho progressiva até o passo 3 e, posteriormente, uma melhora, atingindo o menor erro no passo 5.

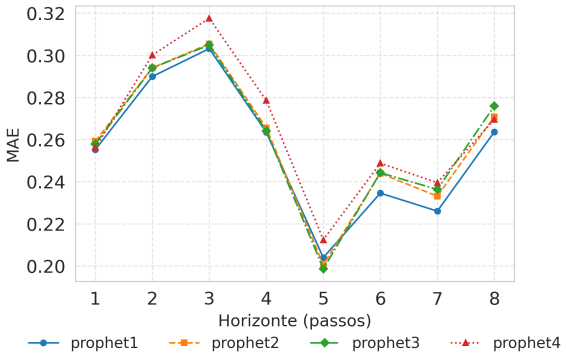
Figura 5.1: Comparação do MAE por horizonte entre modelos



(a) LSTM



(b) SARIMAX



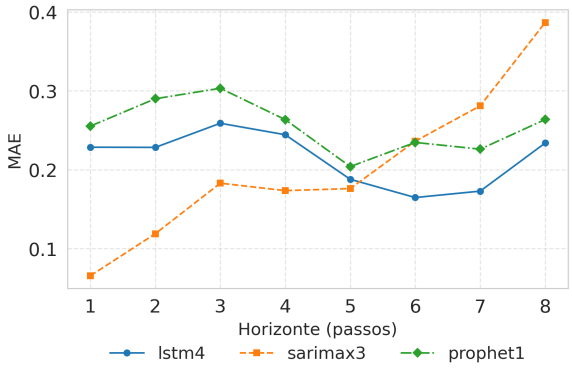
(c) Prophet

Fonte: elaborado pelo autor.

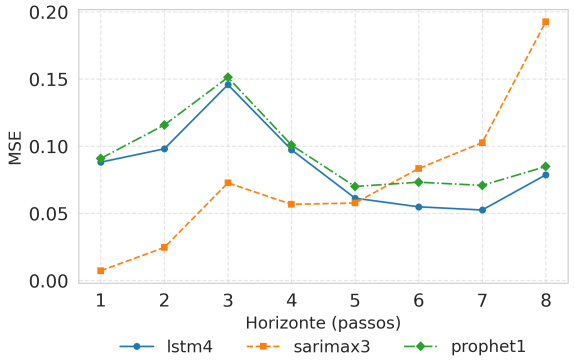
Foram selecionados os melhores modelos de cada técnica de previsão de modo a melhor visualizar de forma comparativa seus respectivos desempenhos em relação a diferentes métricas (MAE, MSE e R2), resultando no que é apresentado na Figura 5.2. A escolha dos melhores modelos para cada técnica se deu da seguinte forma: o modelo lstm4 foi considerado o melhor entre os LSTM pois obteve o melhor desempenho em todas as métricas, o mesmo vale para o prophet1, já o sarimax3 obteve o melhor resultado em três das cinco métricas consideradas, ficando as demais muito próximas. As métricas MAE, MSE e R2 foram escolhidas para simplificar os resultados, visto que as métricas MAE e MAPE são semelhantes, bem como MSE e RMSE.

Verifica-se que aproximadamente até o quinto passo de previsão o modelo SARIMAX obtém o melhor desempenho em todas as métricas aqui analisadas e que, após

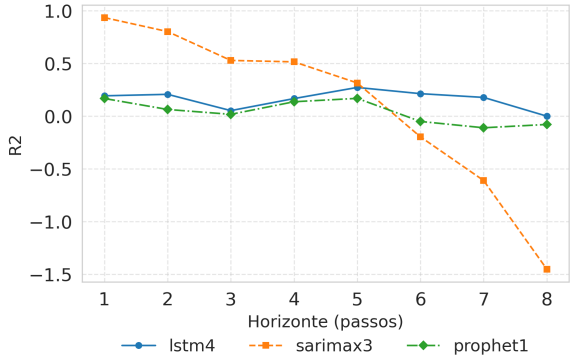
Figura 5.2: Comparação entre os melhores modelos de cada técnica pelas métricas MAE, MSE e R2.



(a) MAE



(b) MSE



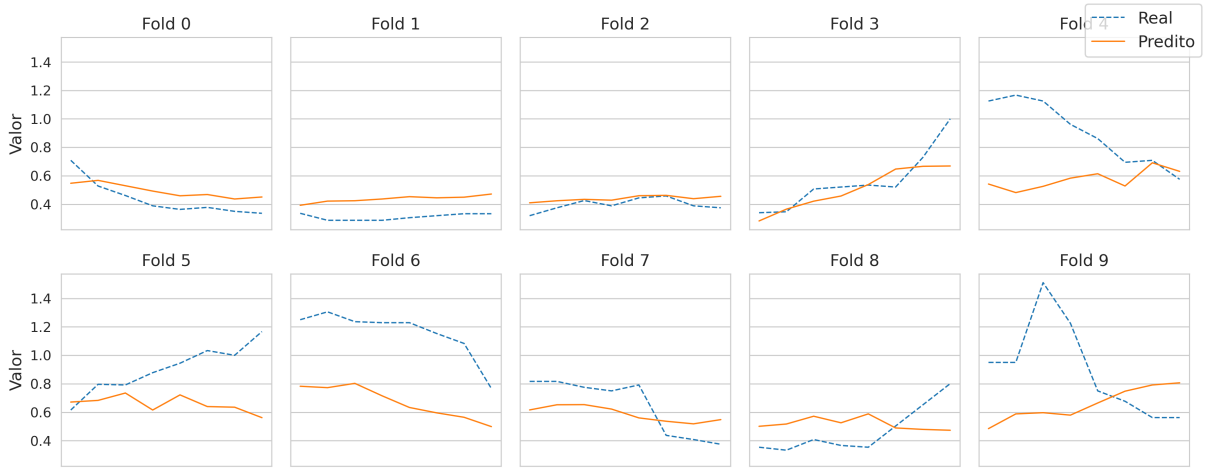
(c) R2

Fonte: elaborado pelo autor.

isto, é ultrapassado pelos outros modelos. Embora o modelo Prophet produza resultados consistentemente piores do que o LSTM, o comportamento geral e os valores dos erros destes dois modelos são semelhantes. Estes resultados apontam que o SARIMAX pode ser mais adequado para a produção de previsões a curto prazo e os demais, especialmente o LSTM, a longo prazo.

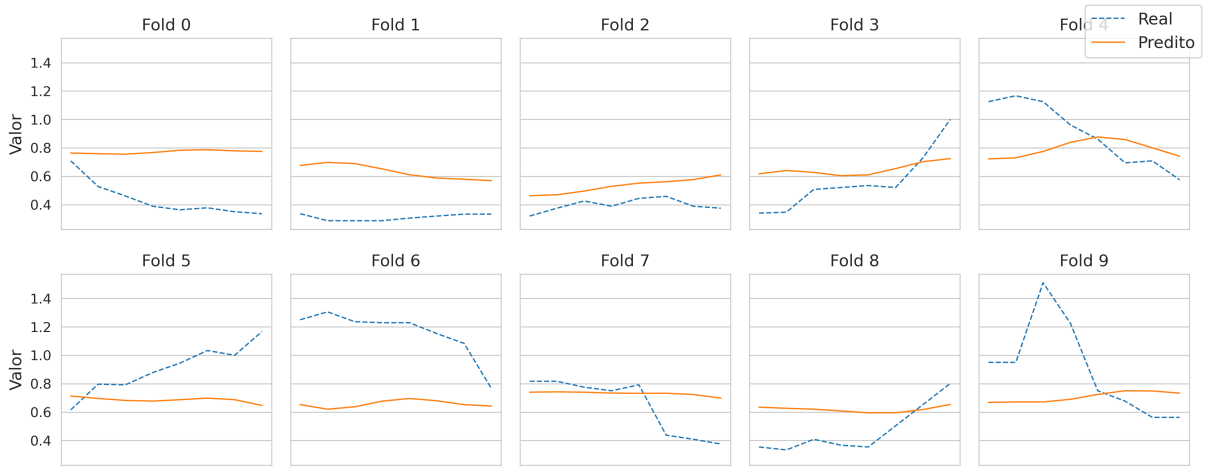
Ao analisar de forma detalhada as previsões de cada modelo ao longo de todos os conjuntos de teste (cada *fold* possui um conjunto de treino e teste), representada nas Figuras 5.3, 5.4 e 5.5, é possível notar que o modelo LSTM aparenta ter capturado melhor o comportamento da série, quando comparado aos demais modelos. O modelo SARIMAX, em especial, demonstra desempenho insatisfatório. Nota-se também variação abrupta na série no *fold* 9 que não foi prevista por nenhum dos modelos, indicando que tal comportamento pode ser um desafio para os modelos preditivos.

Figura 5.3: Valor real e valor predito por *fold* para o modelo LSTM4.



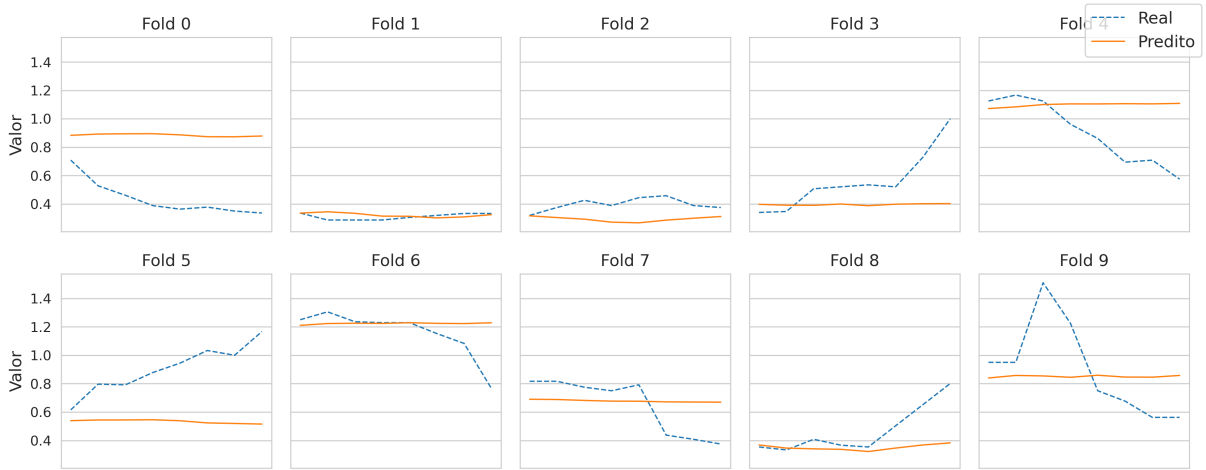
Fonte: elaborado pelo autor.

Figura 5.4: Valor real e valor predito por *fold* para o modelo Prophet1.



Fonte: elaborado pelo autor.

Figura 5.5: Valor real e valor predito por *fold* para o modelo SARIMAX3.



Fonte: elaborado pelo autor.

Nota-se, de forma geral, que os erros dos modelos foram altos, visto que os preços no intervalo estudado variam entre aproximadamente R$0,20 e R$ 1,60 e o melhor MAE obtido é superior a R$0,20. Deve-se considerar, no entanto, que as bases de dados disponíveis não são extensas e que, além disso, não foi feito um ajuste fino sobre os hiperparâmetros dos modelos ou uma investigação mais abrangente sobre outros modelos e arquiteturas. Ainda, a presença de outras variáveis poderia contribuir para o desempenho do modelo, como preços de produtos mais diversos ou o CPI, como feito por Madaan et al. (2019).

É importante reforçar que o horizonte de previsão adotado está relacionado ao tempo entre plantio e colheita geralmente observado para o produto considerado (Alface Crespa). A depender do produto, este tempo pode ser diferente e também o horizonte de previsão, de modo a fornecer previsões úteis ao tomador de decisões.

6 Conclusão

A agricultura é crucial para a manutenção da vida humana e tecnologias que venham a torná-la mais produtiva, rentável e sustentável são importantes para garantir um futuro a todos. No Brasil, em especial, a agricultura familiar tem papel fundamental e é responsável por grande parte do que é produzido, gerando renda e suprindo uma grande demanda por alimentos. No entanto, a adoção de novas tecnologias pelos pequenos agricultores ainda é baixa devido, principalmente, a questões sociais e educacionais. Este trabalho buscou preencher uma lacuna no que diz respeito à avaliação de métodos que forneçam aos agricultores previsões úteis sobre preços de produtos agrícolas aplicados a este contexto específico. Para tal, este trabalho comparou os modelos LSTM, SARIMAX e Prophet aplicados a previsão de séries temporais de um produto típico da agricultura familiar, analisando a qualidade dos seus resultados para diferentes horizontes de previsão.

A partir disto, embora possamos considerar que os diferentes modelos aqui desenvolvidos obtiveram, em média, taxas elevadas de erro, os resultados obtidos demonstram que pode ser interessante adotá-los em cenários distintos, especialmente quando se leva em conta diferentes horizontes de previsão. O modelo SARIMAX demonstrou desempenho superior em previsões de curto prazo (1 ou 2 semanas à frente), enquanto o LSTM obteve erros menores que os demais no longo prazo (8 semanas à frente). Já o modelo Prophet obteve resultados, em geral e considerando todas as métricas, inferiores aos dos outros dois modelos. Além disso, foi possível perceber que o acréscimo de variáveis exógenas pode contribuir para a qualidade das previsões geradas, em particular para as redes neurais LSTM.

Deve-se considerar, no entanto, algumas limitações do presente estudo. As bases utilizadas possuem tamanho reduzido (dados semanais de janeiro de 2016 a dezembro de 2023), o que constitui um grande desafio para alguns métodos de previsão, em especial os baseados em redes neurais. Além disso, embora haja uma grande variedade de produtos agrícolas, esta não pôde ser explorada devido a limitações da fonte de dados, que continha dados apenas de alguns produtos (todos variedades do alface). Outra questão

a ser observada é que outros fatores exógenos podem influenciar os preços de produtos agrícolas, mas neste estudo foram contemplados apenas aqueles relacionados a condições meteorológicas.

Sendo assim, podem ser vislumbrados alguns caminhos para trabalhos futuros com vistas a tornar mais precisas as previsões geradas, bem como compreender melhor os fatores que levam a variações nos preços de produtos agrícolas. Podem ser estudados outros produtos, com seus respectivos ciclos de plantio, considerando fatores macroeconômicos, como, por exemplo, o CPI (*consummer price index*); explorar de forma mais abrangente diferentes modelos preditivos e parâmetros, incluindo modelos mais complexos ou híbridos; e também trazer análises como a realizada pelo presente estudo a diferentes regiões e contextos, levando a uma compreensão mais profunda deste mercado. Por fim, é importante traduzir o conhecimento produzido nestas investigações em sistemas acessíveis àqueles que mais se beneficiariam.

Em suma, entende-se que este trabalho alcançou os objetivos propostos, ao realizar um estudo sobre previsão de preços com um produto tipicamente produzido por pequenos agricultores no Brasil, iniciando a discussão sobre técnicas computacionais avançadas aplicadas a um contexto geralmente negligenciado. Ao estudar previsões a longo prazo, contribui-se para a compreensão sobre as capacidades e limitações de modelos computacionais aplicados à previsão de preços e acrescenta-se peso ao corpo de trabalhos que buscam produzir previsões que possuem utilidade prática na vida dos agricultores, dando bases a futuros sistemas computacionais que possam atingir este público e melhorar suas condições de vida e trabalho.

Bibliografia

AGGARWAL, C. C. *Neural Networks and Deep Learning*: A textbook. Cham: Springer, 2018. 497 p. ISBN 978-3-319-94462-3.

AMARO, G. B.; SILVA, D. M. d.; MARINHO, A. G.; NASCIMENTO, W. M. Recomendações técnicas para o cultivo de hortaliças em agricultura familiar. 2007. ISSN 1415-3033. Disponível em: ⟨http://www.infoteca.cnptia.embrapa.br/handle/doc/781607⟩.

ATHIYARATH, S.; PAUL, M.; KRISHNASWAMY, S. A comparative study and analysis of time series forecasting techniques. *SN Computer Science*, Springer, v. 1, n. 3, p. 175, 2020.

BAYONA-ORÉ, S.; CERNA, R.; HINOJOZA, E. Machine learning for price prediction for agricultural products. *WSEAS TRANSACTIONS ON BUSINESS AND ECONOMICS*, v. 18, p. 969–977, 06 2021.

BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, v. 5, n. 2, p. 157–166, mar. 1994. ISSN 1045-9227, 1941-0093. Disponível em: ⟨https://ieeexplore.ieee.org/document/279181/⟩.

BERGMEIR, C.; BENíTEZ, J. M. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, v. 191, p. 192–213, maio 2012. ISSN 00200255. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S0020025511006773⟩.

BERGMEIR, C.; HYNDMAN, R. J.; KOO, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, v. 120, p. 70–83, abr. 2018. ISSN 01679473. Disponível em: ⟨https://sci-hub.se/https://doi.org/10.1162/neco.1997.9.8.1735⟩.

BONCZEK, R. H.; HOLSAPPLE, C. W.; WHINSTON, A. B. *Foundations of decision support systems*. [S.l.]: Academic Press, 2014.

BOX, G. E.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. *Time series analysis: forecasting and control*. [S.l.]: John Wiley & Sons, 2015.

BOX, G. E. P.; JENKINS, G. M. *Time series analysis: forecasting and control*. Rev. ed. San Francisco: Holden-Day, 1976. (Holden-Day series in time series analysis and digital processing). ISBN 978-0-8162-1104-3.

BRASIL. Lei nº 11.326, de 24 de julho de 2006. estabelece as diretrizes para a formulação da política nacional da agricultura familiar e empreendimentos familiares rurais. *Diário Oficial da União*, Brasília, DF, 2006. Disponível em: ⟨https://www.planalto.gov.br/ccivil_03/_Ato2004-2006/2006/Lei/L11326.htm⟩.

BUAINAIN, A. M.; CAVALCANTE, P.; CONSOLINE, L. Estado atual da agricultura digital no brasil: Inclusão dos agricultores familiares e pequenos produtores rurais. *Documentos de Projetos (LC/TS.2021/61)*, Santiago, Comissão Econômica para a América Latina e o Caribe (CEPAL), 2021.

CONAFER. *Agricultores Familiares São Os Maiores Produtores De Hortaliças E Frutas Do Brasil*. CONAFER, 2021. Disponível em: ⟨https://conafer.org.br/agricultores-familiares-sao-os-maiores-produtores-de-hortalicas-e-frutas-do-brasil/⟩.

DICKEY, D. A.; FULLER, W. A. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, v. 74, n. 366, p. 427, jun. 1979. ISSN 01621459. Disponível em: ⟨https://www.jstor.org/stable/2286348?origin=crossref⟩.

ELMAN, J. L. Finding Structure in Time. *Cognitive Science*, v. 14, n. 2, p. 179–211, mar. 1990. ISSN 0364-0213, 1551-6709. Publisher: Wiley. Disponível em: ⟨https://onlinelibrary.wiley.com/doi/10.1207/s15516709cog1402_1⟩.

ESLING, P.; AGON, C. Time-series data mining. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 45, n. 1, dec 2012. ISSN 0360-0300. Disponível em: ⟨https://doi.org/10.1145/2379776.2379788⟩.

GERS, F. A.; SCHMIDHUBER, J.; CUMMINS, F. Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, v. 12, n. 10, p. 2451–2471, out. 2000. ISSN 0899-7667, 1530-888X. Publisher: MIT Press. Disponível em: ⟨https://direct.mit.edu/neco/article/12/10/2451-2471/6415⟩.

HAN, J.; KAMBER, M. *Data mining: concepts and techniques*. 3rd ed. ed. Burlington, MA: Elsevier, 2012. 114 p. ISBN 978-0-12-381479-1.

HARRYKISSOON, K.; HOSEIN, P. Recursive vs. Direct Forecasting of Crop Prices. In: *2023 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*. Rabat, Morocco: IEEE, 2023. p. 1–6. ISBN 979-8-3503-1335-2. Disponível em: ⟨https://ieeexplore.ieee.org/document/10438140/⟩.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, nov. 1997. ISSN 0899-7667, 1530-888X. Publisher: MIT Press. Disponível em: ⟨https://sci-hub.se/https://doi.org/10.1162/neco.1997.9.8.1735⟩.

KWIATKOWSKI, D.; PHILLIPS, P. C.; SCHMIDT, P.; SHIN, Y. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, v. 54, n. 1-3, p. 159–178, out. 1992. ISSN 03044076. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/030440769290104Y⟩.

MADAAN, L.; SHARMA, A.; KHANDELWAL, P.; GOEL, S.; SINGLA, P.; SETH, A. Price forecasting & anomaly detection for agricultural commodities in india. In: *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*. [S.l.: s.n.], 2019. p. 52–64.

MAHALAKSHMI, G.; SRIDEVI, S.; RAJARAM, S. A survey on forecasting of time series data. In: IEEE. *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*. [S.l.], 2016. p. 1–8.

MIENYE, I. D.; SWART, T. G.; OBAIDO, G. Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. *Information*, v. 15, n. 9, p. 517, ago. 2024. ISSN 2078-2489. Disponível em: ⟨https://www.mdpi.com/2078-2489/15/9/517⟩.

MIN, Y.; KIM, Y. R.; HYON, Y.; HA, T.; LEE, S.; HYUN, J.; LEE, M. R. RNN and GNN based prediction of agricultural prices with multivariate time series and its short-term fluctuations smoothing effect. *Scientific Reports*, v. 15, n. 1, p. 13681, abr. 2025. ISSN 2045-2322. Disponível em: ⟨https://www.nature.com/articles/s41598-025-97724-7⟩.

MONTGOMERY, D. C.; JENNINGS, C. L.; KULAHCI, M. *Introduction to time series analysis and forecasting*. [S.l.]: John Wiley & Sons, 2015.

MURUGESAN, R.; MISHRA, E.; KRISHNAN, A. H. Deep learning based models: Basic lstm, bi lstm, stacked lstm, cnn lstm and conv lstm to forecast agricultural commodities prices. 2021.

PHUMCHUSRI, N.; CHEWCHARAT, T.; KANOKPONGSAKORN, S. Price promotion optimization model for multiperiod planning: a case study of beauty category products sold in a convenience store chain. *Journal of Revenue and Pricing Management*, v. 23, n. 2, p. 164–178, abr. 2024. ISSN 1477-657X. Disponível em: ⟨https://doi.org/10.1057/s41272-023-00438-6⟩.

SANGHANI, A.; BHATT, N.; CHAUHAN, N. A review of soft computing techniques for time series forecasting. *Indian Journal of Science and Technology*, v. 9, n. 1, p. 1–5, 2016.

TAYLOR, S. J.; LETHAM, B. *Forecasting at scale*. 2017. Disponível em: ⟨https://peerj.com/preprints/3190v2⟩.

UNITED NATIONS. *SUSTAINABLE DEVELOPMENT GOALS REPORT 2024*. [S.l.]: UNITED NATIONS, 2024. OCLC: 1492039046. ISBN 9789210031356.

VAGROPOULOS, S. I.; CHOULIARAS, G. I.; KARDAKOS, E. G.; SIMOGLOU, C. K.; BAKIRTZIS, A. G. Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting. In: *2016 IEEE International Energy Conference (ENERGYCON)*. Leuven, Belgium: IEEE, 2016. p. 1–6. Disponível em: ⟨http://ieeexplore.ieee.org/document/7514029/⟩.

WAEODI, K.; BOONGASAME, L.; THAMMARAK, K. Thai Morning Glory Price Forecasting Using Deep Learning. *Appl. Comput. Intell. Soft Comput.*, v. 2025, p. null, 2025. Disponível em: ⟨https://www.semanticscholar.org/paper/fdd88043df54df0f5ba72592794c285df73cdfb2⟩.

WISSLER, C. The Spearman Correlation Formula. *Science*, v. 22, n. 558, p. 309–311, set. 1905. ISSN 0036-8075, 1095-9203. Disponível em: ⟨https://www.science.org/doi/10.1126/science.22.558.309⟩.

YU, Y.; SI, X.; HU, C.; ZHANG, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, v. 31, n. 7, p. 1235–1270, jul. 2019. ISSN 0899-7667, 1530-888X. Disponível em: ⟨https://direct.mit.edu/neco/article/31/7/1235-1270/8500⟩.

ÖZDEN, C. Comparative Analysis of CNN, LSTM And Random Forest for Multivariate Agricultural Price Forecasting. *Black Sea Journal of Agriculture*, v. 6, n. 4, p. 422–426, jul. 2023. ISSN 2618-6578. Disponível em: ⟨http://dergipark.org.tr/en/doi/10.47115/bsagriculture.1304625⟩.

ÖZDEN, C.; BULUT, M. Spectral temporal graph neural network for multivariate agricultural price forecasting; [rede neural de gráfico temporal espectral para previsão de preços agrícolas multivariados]. *Ciencia Rural*, v. 54, n. 1, 2024. Cited by: 1; All Open Access, Gold Open Access. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85164470010&doi=10.1590%2f0103-8478cr20220677&partnerID=40&md5=cb0f61a6a09afb708704d58f099bd3a3⟩.