Machine Learning Models for Predicting COVID-19: An Ensemble Approach Applied to the State of Alagoas

José Lucas Bispo dos Santos[1], Elmo Araújo Filho[1], Marília G. F. de M. Oliveira[2] Augusto C. F. de M. Oliveira[3], Gustavo H. F. de M. Oliveira[1]

[1] Sistemas de Informação, Universidade Federal de Alagoas (UFAL)[1].

[2]Universidade Federal Rural de Pernambuco (UFRPE)[2].

[3]Engenharia de Software, Universidade de Pernambuco (UPE)[3].

{jose.santos7, elmo.filho}@arapiraca.ufal.br, marilia.gfmo@gmail.com
augusto.oliveira@upe.br, gustavo.oliveira@penedo.ufal.br

Abstract. *COVID-19 emerged as the most contagious variant of the coronavirus, triggering a pandemic with a global impact. Forecasting strategies based on time series were innovative for predicting cases and supporting government decisions. However, less assisted areas, such as cities in the interior of Alagoas, often do not have access to these options. Therefore, this study proposes a solution to this issue through machine learning models. The results highlight the method's effectiveness as an alternative and its superiority compared to individual models.*

Resumo. *O COVID-19 surgiu como a variante mais contagiosa do coronavírus, desencadeando uma pandemia de impacto global. Estratégias de previsão baseadas em séries temporais foram implementadas para prever os casos e amparar decisões governamentais. Contudo, áreas menos assistidas, como cidades do interior de Alagoas, frequentemente não acessaram essas previsões. Diante disso, este estudo propõe uma solução para este cenário através de um Ensemble de modelos de aprendizagem de máquina. Os resultados destacam a eficácia do método nas previsões e em comparação com modelos individuais.*

1. Introdução

O coronavírus teve sua primeira aparição em meados dos anos 90 e, ao longo do tempo, passou por várias mutações genéticas originárias de animais como roedores, camelos e morcegos [Kwekha-Rashid et al. 2023]. Em 2002, na China, o coronavírus foi associado a grandes calamidades durante a pandemia da Síndrome Respiratória Aguda Grave (SARS). Posteriormente, em 2012 e 2015, ressurgiu através de epidemias da Síndrome Respiratória do Oriente Médio (MERS). Já em dezembro de 2019, surgiu a pior das suas variações, o COVID-19, um vírus altamente contagioso, transmitido pelo ar [Santosh 2020]. Essa nova mutação, teve origem em morcegos e se propagou de maneira exponencial, tornando-se uma ameaça global [Espinosa et al. 2020].

A propagação do vírus ocasionou o surgimento de diversos problemas, como doenças respiratórias potencialmente graves, aumento de transtornos psicológicos e desafios no tratamento adequado aos pacientes por parte dos sistemas de saúde

[Chiattone et al. 2022]. De acordo com [AlJame et al. 2020], após aproximadamente 1 ano de pandemia, em 11 de julho de 2020, o COVID-19 havia gerado 12 milhões de casos confirmados em 216 países ao redor do mundo, resultando em um número expressivo de 500.000 mortes. Esse grande número de casos confirmados resultou na insuficiência dos equipamentos médicos e da capacidade de diagnósticos, sobrecarregando os hospitais [Zhou et al. 2021].

Durante a pandemia de COVID-19, foram implementados diversos métodos para prever a disseminação do vírus para que fosse possível criar estratégias de ação. Entre eles, destacam-se os sistemas de previsão baseados na análise de séries temporais, que visam prever a quantidade de novos casos confirmados para os próximos dias, semanas ou meses. O principal desafio dessas previsões reside na constante mudança nos dados de disseminação do vírus. Segundo [Cramer et al. 2022], essas mudanças podem ser influenciadas por diversos fatores, como decisões governamentais e comportamentais.

Um exemplo dessa complexidade surgiu nos primeiros meses da pandemia, quando os sistemas de previsão apresentaram imprecisões devido às variações nos dados decorrentes das estratégias adotadas pelos governos. Nessa situação, os modelos individuais de previsão são os mais afetados, pois suas habilidades de aprendizado sobre os dados podem não ser capazes de capturar todas as variações de padrões [Acito 2023]. Uma maneira de mitigar esse problema é através da adoção de um comitê de regressores (Ensemble Regressors, em inglês) [Cramer et al. 2022], uma abordagem de aprendizagem de máquina que combina previsões de vários modelos simultaneamente, reduzindo assim as possíveis imprecisões dos modelos individuais.

Embora esses sistemas possam ser utilizados para prever a disseminação do vírus, apenas as grandes cidades e capitais se beneficiariam dessas previsões, pois estas são comumente geradas pelos grandes centros de pesquisa ou universidades. Por outro lado, cidades do interior, por não terem pesquisas nesse nível, não conseguem acesso a previsões para preparar adequadamente seus sistemas de saúde. Diante disso, este trabalho apresenta o produto resultante de um projeto de iniciação científica que visou desenvolver um sistema de Combinação de Modelos de Aprendizado de Máquina (Ensemble) para previsão de dados de COVID-19 para cidades do interior do estado de Alagoas: Penedo, Coruripe, Arapiraca e Maceió (capital do estado). A hipótese que motivou esta pesquisa é que a combinação de modelos individuais distintos pode oferecer uma ampla variedade de resultados em que o Ensemble pode optar para melhorar o seu desempenho de previsão.

Logo, para alcançar o objetivo deste estudo, a seção 2, fundamentação teórica, descreve o que são previsão de séries temporais e a definição de Ensembles. Na seção 3, metodologia, o método proposto é descrito. Na seção 4, os experimentos e resultados são discutidos. Por fim, na seção 5, as principais conclusões deste estudo são enumeradas.

2. Fundamentação Teórica

2.1. Análise e Previsão de Séries Temporais

As séries temporais, também conhecidas como séries históricas, são definidas como uma sequência de observações $X = \{x_1, x_2, \ldots, x_t\}$, obtidas em intervalos regulares de tempo [Latorre and Cardoso 2001]. No contexto da COVID-19, um exemplo de série temporal pode ser a quantidade de casos de pacientes confirmados diariamente, em que o dia 1 é

representado por x_1, o dia 2 por x_2 e assim por diante, até que o dia n seja representado por x_n.

A análise deste tipo de estrutura de dados permite descrever seu comportamento por meio de gráficos, identificar tendências, sazonalidades e realizar previsões de valores futuros [Morettin and Toloi 2018]. As previsões podem ser geradas por meio de métodos estatísticos ou de métodos de inteligência artificial, como os algoritmos de aprendizado de máquina [AlJame et al. 2020].

Mas, para que esses métodos aprendam os padrões dos dados adequadamente, a série temporal precisa ser transformada em um par de dados ($\mathbf{X}$, y). Nesse par, $\mathbf{X}$ representa a matriz dos padrões de entrada e y os alvos que os modelos precisam aprender a prever. Uma abordagem comum para isso é a Janela do Tempo, que extrai pequenas subsequências da série temporal usando a expressão: $Z_t = \{x_{t-w+1}, x_{t-w+2}, \cdots, x_t\}$. Por exemplo, o padrão de entrada $X_1 = \{x_1, x_2, \cdots, x_w\}$ é usado para prever x_{w+1}, $X_2 = \{x_2, x_3, \cdots, x_{w+1}\}$ é usado para prever x_{w+2} e assim por diante. w refere-se à quantidade de dados de entrada utilizado pelo modelo para compreender os padrões existentes. Assim, se torna possível a aplicação de qualquer método de previsão.

Com isso, as previsões geradas, podem embasar as políticas de saúde pública e direcionar a tomada de decisões com intervenções estratégicas, como compra de vacinas, construção de leitos e gerenciamento de casos e óbitos [Carlotto 2021].

2.2. Conjunto de Regressores - Ensemble Regressors

Os avanços na inteligência artificial e no aprendizado de máquina têm proporcionado significativos benefícios aos sistemas de previsão, permitindo o desenvolvimento de algoritmos capazes de identificar padrões em dados históricos de maneira mais precisa e eficiente [Kwekha-Rashid et al. 2023]. Esses algoritmos são projetados para detectar interações complexas entre variáveis, reconhecendo correlações que são fundamentais para realizar previsões com precisão.

No entanto, não há um modelo único que possa garantir um desempenho ótimo em todas as situações [Acito 2023]. A eficácia de um modelo depende da sua capacidade individual de capturar os padrões específicos presentes nas séries históricas. Por exemplo, modelos estatísticos lineares frequentemente falham ao lidar com dados que possuem relações não-lineares complexas [Figueiredo Filho and Silva Júnior 2009].

Para superar essas limitações, [Cramer et al. 2022] recomendam o uso de conjunto de regressores (Ensemble Regressors, em inglês). Essa técnica combina as previsões de múltiplos modelos individuais, aproveitando suas diferentes abordagens e capacidades de generalização [Zhou et al. 2021]. A qualidade da previsão de um Ensemble depende da diversidade entre os modelos individuais, pois modelos muito semelhantes tendem a oferecer previsões parecidas, reduzindo a eficácia da combinação no Ensemble.

Por esse motivo, diversos trabalhos na literatura propuseram o uso de Ensemble para realizar a previsão de dados de COVID-19 em diferentes contextos. Em [Maaliw et al. 2021], foi utilizado um Ensemble com os modelos ARIMA e S-LSTM para previsão de séries temporais de 467 dias para os países Filipinas, Estados Unidos, Índia e Brasil. [Shastri et al. 2022] propôs o CoBiD-Net e também avaliou o LSTM e a LSTM Convolucional para prever o número de casos confirmados e de mortes nos EUA, Índia e

Brasil. [Liapis et al. 2020] propôs o uso de Ensembles para previsão de casos nos países da Europa do Sul e Central. Por último, [Ashofteh et al. 2022] investigou a aplicação de um ensemble com o modelo bayesiano de média (BMA) para séries temporais de 61 países.

No entanto, vale salientar que, na maioria dos trabalhos revisados, o foco está principalmente em países, deixando de lado contextos específicos importantes, como cidades do interior. Cidades menores, como as de Alagoas, que são o foco deste trabalho, são frequentemente negligenciadas. Esse cenário aponta para a necessidade de estudos futuros que explorem a aplicação de técnicas avançadas de previsão em áreas sub-representadas, assim proporcionando uma compreensão mais abrangente e precisa das dinâmicas locais e contribuindo para uma gestão mais eficaz de recursos.

3. Métodos

O método proposto teve como finalidade o desenvolvimento de um Ensemble de aprendizado de máquina para previsão de séries temporais. A sua estrutura, ilustrada na Figura 1, apresenta 5 etapas.

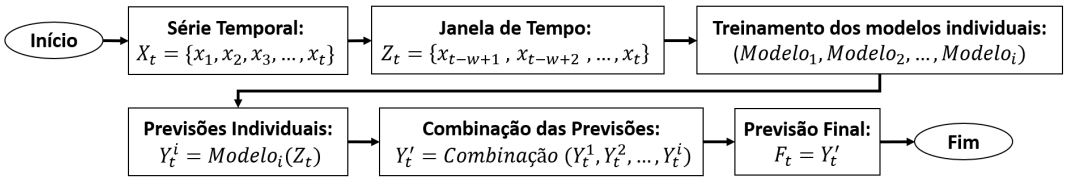


Figura 1. Estrutura geral do sistema Ensemble.

Inicialmente, o sistema recebe uma série temporal representada por X_t , ou seja, uma sequência de observações ordenadas ao longo do tempo. Após obter a série temporal, a janela de tempo foi utilizada para ajustar os dados para os modelos conforme discutido na subseção 2.1. A janela de tempo, extrai várias subsequências de dados da série temporal, na qual chamamos de Z_t, e os transforma em vetores de entrada e saída (**X**, y) para que seja possível aprender a correlação entre os dados.

Em seguida, os modelos individuais são treinados e usados para gerar previsões para os dados, conforme a expressão: $Y_t^i = Modelo_i(Z_t)$, onde Y_t^i é a previsão feita pelo modelo i para a observação no tempo t e Z_t são os dados da janela de tempo usados como entrada do modelo. Por exemplo, o $Modelo_1$ pode ser uma regressão linear, o $Modelo_2$ uma árvore de decisão e o $Modelo_3$ uma rede neural. Juntos, esses modelos geram um conjunto de previsões individuais para a observação no tempo t, representadas por: Y_t^1, Y_t^2, Y_t^3.

O Método de Combinação é aplicado sobre as previsões individuais para formar uma única previsão, representada pela expressão: $Y_t' = Combinacao(Y_t^1, Y_t^2, \cdots, Y_t^3)$. Para exemplificar, considere (3, 4, 5) como as previsões de três modelos para o tempo t. Nesse caso, pode-se utilizar qualquer uma das medidas de tendência central, como média, moda e mediana. Para o exemplo discutido, se a média for utilizada, a previsão final seria 4.

A média é útil por fornecer o comportamento mais comum entre as previsões, ajudando a entender de forma simples e clara a distribuição dos dados. A mediana, pode

ser usada para identificar o valor do meio das previsões ordenadas, sendo especialmente útil na presença de outliers ou à assimetria dos dados. Por fim, a moda pode ser usada para identificar e destacar os valores mais frequentes das previsões.

Dessa maneira, a previsão final será o resultado do método de combinação para cada observação de entrada, gerando um vetor com a previsão final, determinada pela expressão, $F_t = Y'_t$. Por fim, Essa previsão será utilizada para análises futuras e tomadas de decisões.

4. Experimentos

Nesta seção serão discutidas as séries temporais que foram utilizadas para os experimentos, seus tratamentos, as métricas usadas para avaliação e a configuração experimental.

4.1. Bases de Dados Utilizadas

O conjunto de dados utilizado neste estudo foi obtido da plataforma do Ministério da Saúde que possui acesso público[1]. O conjunto de dados foi coletado de 26 de março de 2020 a 31 de fevereiro de 2023, gerando um quantitativo de 335.570 casos de pacientes infectados pela COVID-19. Os experimentos foram realizados para quatro cidades do estado de Alagoas: Maceió, Coruripe, Arapiraca e Penedo. As variáveis utilizadas foram: i) os casos de pacientes recuperados e ii) aqueles que vieram a óbito, totalizando ao todo dez bases. A Figura 2 representa a esquematização das bases.

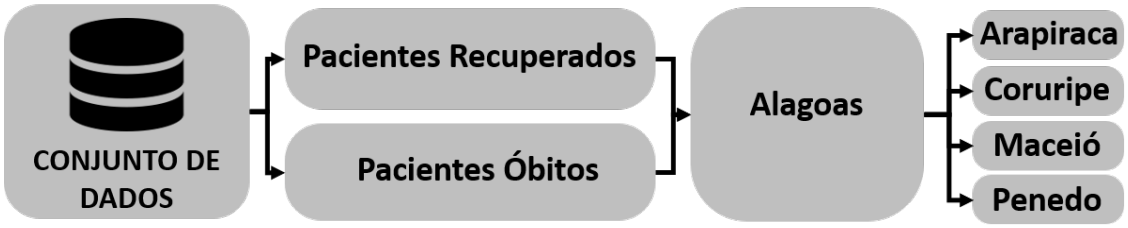


Figura 2. Esquematização das bases de dados.

4.2. Tratamento dos Dados

As séries temporais foram fracionadas utilizando a técnica holdout em que os 80% primeiros dados foram utilizados para treinamento dos modelos, enquanto que os 20% últimos foram utilizados para testar a capacidade de generalização dos modelos em dados diferentes dos treinados. Os 80% foram estipulados para que os modelos pudessem capturar um conjunto suficientemente grande de exemplos para conseguir identificar as variabilidades e padrões existentes.

Para garantir a convergência do treinamento dos modelos e obter um melhor desempenho, os dados passaram pelo processo de normalização, que consiste no ajuste dos dados da série temporal para a escala de 0 e 1. O método Min-Max Scaling foi utilizado para essa tarefa, representado pela expressão: $Xnorm = (X - Xmin)/(Xmin - Xmax)$. Onde, $Xnorm$ é o conjunto de dados normalizado, $Xmin$ é o valor mínimo em X e $Xmax$ é o valor máximo em X.

[1] `https://www.gov.br/saude/pt-br/composicao/seidigi/demas/covid19`

4.3. Métricas de Avaliação

Para avaliar a qualidade das previsões, a métrica do Erro Quadrático Médio (MSE) foi utilizada, conforme definida abaixo:

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

O MSE calcula a média das diferenças absolutas entre os valores previstos $\hat{y}_i$ e os valores reais y_i. Ele reflete a precisão média do modelo de previsão ao obter a soma de todas as diferenças absolutas e dividir pelo número total n de observações avaliadas. Quando o MSE está próximo de 0 indica que o modelo está, em média, muito próximo dos valores reais, significando boa precisão. Por outro lado, MSE próximo de 1 indica grandes discrepâncias entre os valores previstos e reais, sugerindo que o modelo pode não ser muito preciso.

Por fim, para atestar os resultados obtidos, foi utilizado a biblioteca AutoRank proposta por [Herbold 2020]. O AutoRank se baseia no protocolo proposto por [Demšar 2006] que avalia inicialmente a normalidade das amostras antes da aplicação de um teste estatístico. Se os dados não forem normais, correções de Cohen's são realizadas para garantir a veracidade dos resultados. Em seguida, o Teste de Friedman é utilizado para detectar diferenças significativas de múltiplos algoritmos em vários conjuntos de dados. A hipótese nula do teste considera que não há diferença significativa entre os algoritmos. Uma vez que a hipótese nula é rejeitada, ou seja, após detectar diferenças significativas entre os algoritmos, o teste post-hoc de Nemenyi é utilizado para identificar quais algoritmos diferem entre si. Ambos os testes foram utilizados com nível de significância de $\alpha = 0,05$.

4.4. Configuração dos Experimentos

Para o desenvolvimento dos experimentos, foram utilizados seis algoritmos, disponíveis na plataforma Scikit-Learn[2]. Para garantir o bom desempenho dos algoritmos, os mesmos foram submetidos a técnica de otimização de hiperparâmetros GridSearch que avalia todas as combinações possíveis de parâmetros de modo a encontrar aquele que resulta no menor erro de treinamento. A Tabela 1 mostra os parâmetros avaliados e aqueles que geraram os melhores desempenhos para os experimentos.

Para o método de Ensemble proposto, apenas os algoritmos LarsLasso, Lasso e LinearRegression exibidos na Tabela 1 foram utilizados para popular o Ensemble. A escolha destes modelos se dá pelo fato de serem simples e fornecerem perspectivas distintas para a definição da previsão final.

Para geração dos resultados, foi utilizada a plataforma Google Colab e a linguagem de programação Python. Todos os algoritmos acima descritos foram executados 50 vezes, por ser uma amostra significativa que permite compreender a variação de diferentes inicializações dos algoritmos e com isso garantir a aplicação adequada do teste de Friedman com o pós-teste de Nemenyi.

[2]`https://scikit-learn.org/stable/index.html`

Tabela 1. Gridsearch e melhores parâmetros por modelo.

Algoritmos	Grid Search	Melhores Parâmetros
LinearRegression	fit_intercept: [True, False], copy_X: [True, False]	copy_X: True, fit_intercept': True
Lasso	alpha: [0.001, 0.1, 1, 10, 100], max_iter:[1000,5000, 10000]	alpha: 0.001, max_iter': 1000
LassoLars	——	——
DecisionTreeRegressor	max_depth:[10,20,30], min_samples_split: [1,2,4], min_samples_leaf: [7, 8, 9, 10, 11,15,20]	max_depth: 10, min_samples_leaf: 4, min_samples_split: 2
MLPRegressor	hidden_layer_sizes: [(5,), (10,), (15,), (20,), (30,)], activation: ['relu', 'tanh'], alpha: [0.0001, 0.001], max_iter': [50, 80, 100, 150]	activation: tanh, alpha: 0.0001, hidden_layer_sizes: (30,), max_iter: 80
SVR	kernel: ['rbf', 'sigmoid'], C: [0.1, 1, 10, 100]	C: 0.1, kernel: linear

5. Resultados

Os resultados dos experimentos estão apresentados na Tabela 2, por meio da média e desvio padrão para o Erro Quadrático Médio (MSE). Os resultados em negrito representam o algoritmo que obteve o melhor desempenho para cada respectiva série temporal. Para afirmar esses resultados, o teste estatístico de Friedman com o pós-teste de Nemenyi é ilustrado na Figura 3. As discussões dos resultados serão organizadas em duas subseções, iniciando pela subseção 5.1, que trata da análise do desempenho do ensemble, e pela subseção 5.2, que avalia todos os algoritmos experimentados.

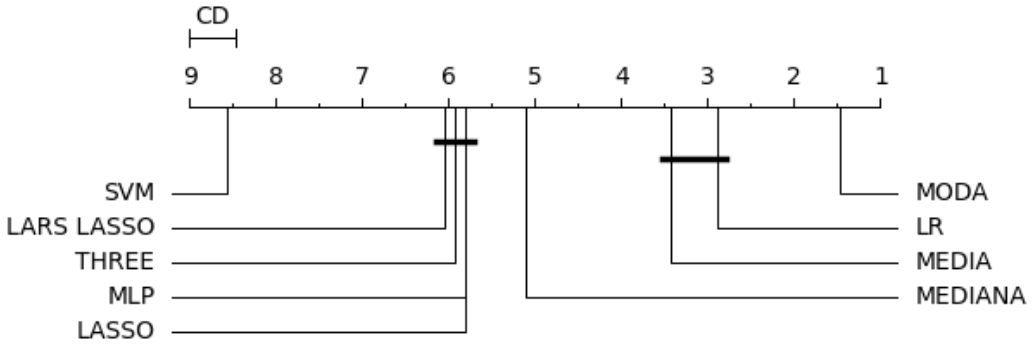


Figura 3. Ranking de Friedman com pós-teste de Nemenyi.

5.1. Análise do Desempenho dos Modelos Ensemble

Nessa primeira análise, ao observar a Tabela 2, percebe-se que as abordagens baseadas em Ensemble combinadas obtiveram os melhores resultados para 5 das 10 bases. Segundo o ranking de Friedman (Figura 3), o Ensemble com a combinação por moda obteve o melhor desempenho, sendo estatisticamente diferente das demais abordagens, visto que a sua distância crítica (CD) de Nemenyi é superior.

Em relação aos outros métodos de combinação, a combinação por média foi a que mais se aproximou dos melhores resultados, enquanto a mediana figurou na pior colocação entre as abordagens de combinação. Para compreender esses resultados, pode-se analisar a previsão dos sistemas para as séries temporais de Arapiraca e Penedo (Figura

5c), pois foram as mais desafiadoras para este estudo, apresentando grandes períodos de tempo sem nenhuma observação. Essa característica cria algumas retas na série temporal, o que dificulta bastante a previsão dos modelos.

Tabela 2. Resultados dos modelos para o MAE

Séries Temporais	MLP	DecisionTree	SVM	Lasso	LassoLars	LinearRegression	Ensemble (Média)	Ensemble (Mediana)	Ensemble (Moda)
Alagoas (Óbitos)	0.0105 (0.00)	0.0132 (4.753)	0.0062 (8.673)	0.0062 (8.673)	0.0062 (8.673)	**0.0062 (0.000)**	0.0067 (0.000)	0.0062 (0.001)	0.0104 (0.001)
Alagoas (Confirmados)	0.0074 (0.002)	0.0092 (0.001)	0.0078 (0.000)	0.0070 (8.673)	0.0070 (8.673)	0.0078 (1.734)	**0.0067 (0.000)**	0.0068 (0.000)	0.0077 (0.000)
Maceió (Óbitos)	0.0108 (0.003)	0.015 (3.102)	0.0082 (0.000)	0.0074 (0.000)	**0.0074 (8.673)**	0.0083 (0.000)	0.0079 (0.001)	0.0075 (0.001)	0.0091 (0.001)
Maceió (Confirmados)	0.0039 (0.000)	0.0042 (5.346)	0.0054 (0.000)	0.0028 (8.673)	0.0028 (4.336)	0.0038 (0.000)	0.0028 (0.000)	**0.0027 (0.000)**	0.0039 (0.000)
Arapiraca (Confirmados)	0.0044 (0.001)	0.0066 (1.117)	0.0081(0.000)	0.0036 (0.000)	0.0036 (4.336)	**0.0032 (0.000)**	0.0035 (0.000)	0.0034 (0.000)	0.0042 (0.000)
Arapiraca (Óbitos)	0.0212 (0.003)	0.0318 (0.000)	0.0176 (0.000)	0.0203 (0.000)	0.02030 (0.000)	**0.0174 (0.000)**	0.0187 (0.000)	0.0186 (0.000)	0.0203 (0.000)
Coruripe (Confirmados)	0.0267 (0.006)	0.0330 (0.000)	0.0249 (3.469)	0.0254 (0.000)	0.0254 (0.000)	0.0251 (0.000)	0.0251 (0.000)	**0.0251 (0.000)**	0.0254 (0.000)
Coruripe (Óbitos)	0.0365 (0.054)	0.0022 (8.892)	0.0110 (1.734)	0.00164 (2.168)	0.00164 (2.168)	0.0191 (0.000)	0.0049 (2.168)	0.0026 (2.168)	**0.0016 (2.168)**
Penedo (Óbitos)	0.0381 (0.017)	0.0368 (0.001)	**0.0221 (3.469)**	0.0236 (6.938)	0.0236 (6.938)	0.0251 (3.469)	0.0225 (6.938)	0.0221 (6.938)	0.0236 (6.938)
Penedo (Confirmados)	0.0159 (0.007)	0.0258 (3.469)	0.0091 (0.000)	0.0063 (8.673)	0.0063 (8.673)	0.0158 (0.000	0.0090 (8.673)	0.0083 (8.673)	**0.0063 (8.673)**

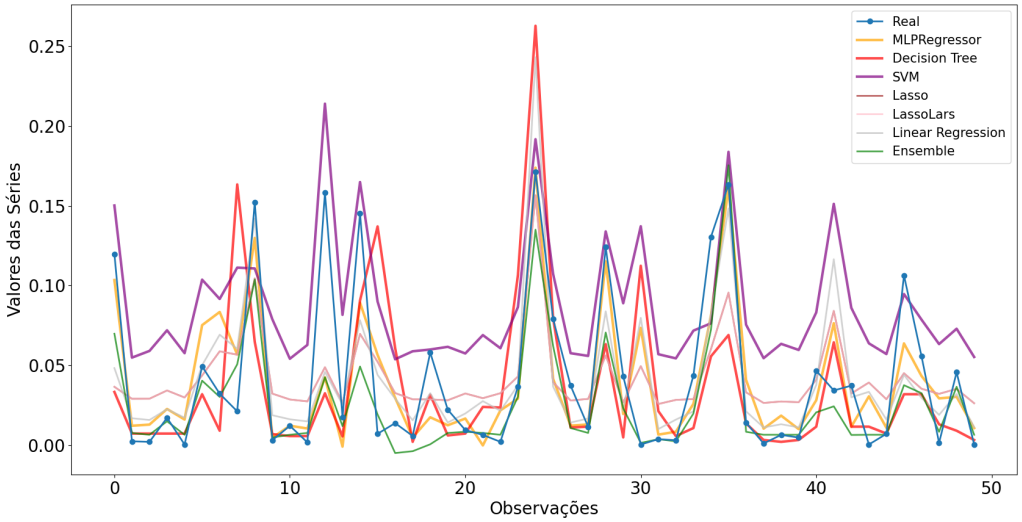
Ainda assim, percebe-se que a abordagem de Ensemble (Moda) foi o modelo que manteve sua previsão mais próxima dos valores reais. Isso se deve ao método de combinação por moda, que seleciona como previsão final apenas os valores mais frequentes das previsões, assim reduzindo a variância. A média, por combinar todas as previsões, acaba sendo sensível a previsões muito ruins, o que prejudica a previsão final. A mediana seleciona a previsão do meio, mas, se muitos modelos tiverem previsões prejudicadas, a previsão final do Ensemble também acaba sendo ruim.

Como o Ensemble obteve a melhor colocação no Ranking de Friedman, esses resultados apontam para a confirmação da hipótese de pesquisa deste trabalho, que sugere que combinar as previsões dos modelos individuais pode ser uma alternativa eficaz para superar as dificuldades de prever dados de COVID-19 e, assim, melhorar o desempenho de previsão.

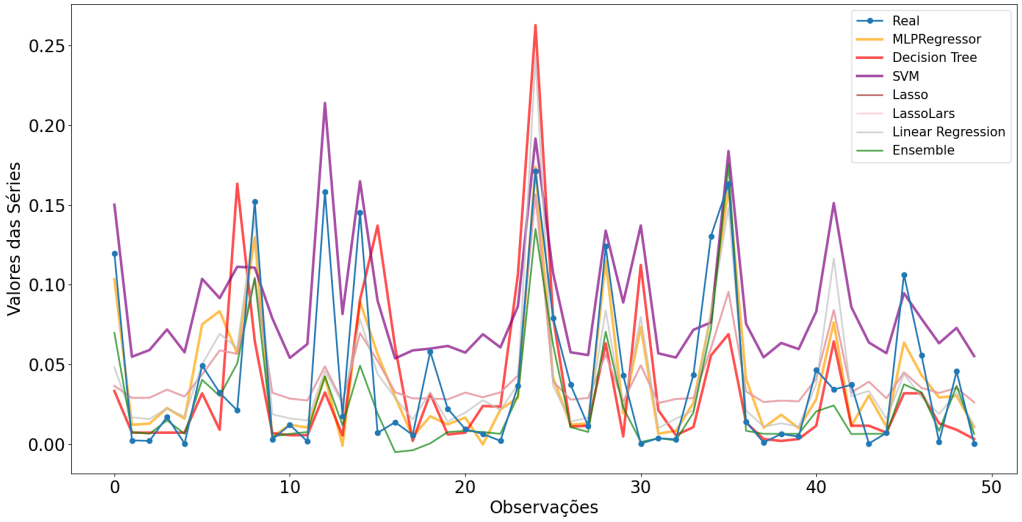
5.2. Comparação entre Algoritmos Lineares e Não-lineares

Nessa segunda análise, as demais abordagens de aprendizado de máquina são avaliadas. Ao observar os resultados da Tabela 2, percebe-se que a Linear Regression obteve os melhores desempenhos em três das séries temporais. Segundo o ranking de Friedman (Figura 3), ela ficou em segundo lugar, superando estatisticamente a mediana do Ensemble. Abordagens com maior viés linear, como LASSO, LARS LASSO e Linear Regression, se destacaram, enquanto as não-lineares, como SVM, Árvore de Decisão (THREE) e MLP, alcançaram os piores rankings de Friedman. Isso sugere que as séries temporais de COVID-19 do estado de Alagoas possuem padrões lineares mais predominantes.

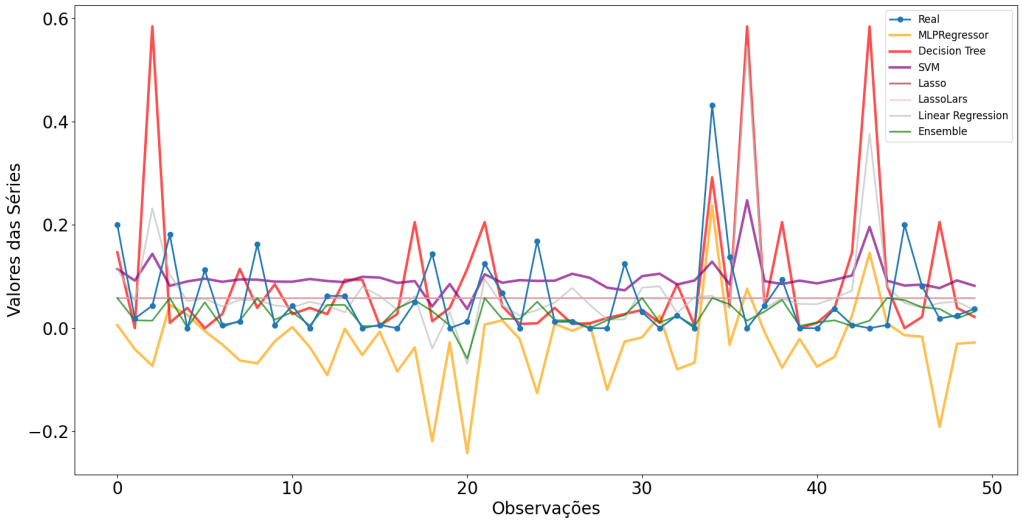
Para compreender visualmente o desempenho desses algoritmos, pode-se analisar as séries temporais (i) do estado de Alagoas (Figuras 5a e 4a); (ii) da cidade de Arapiraca (Figuras 5b e 4b); e (iii) da cidade de Penedo (Figuras 5c e 4c). Percebe-se que, em situações com mais dados, as abordagens MLP e Árvore de Decisão tiveram um bom desempenho (ex: Figuras 4a e 5a), mas em casos com poucos dados (ex: Figuras 4c e 5c), o desempenho de previsão foi bem ruim. Esses resultados também fortalecem a hipótese deste trabalho, indicando que modelos individuais não são tão robustos quando avaliados em muitas bases de dados.



(a) Alagoas.

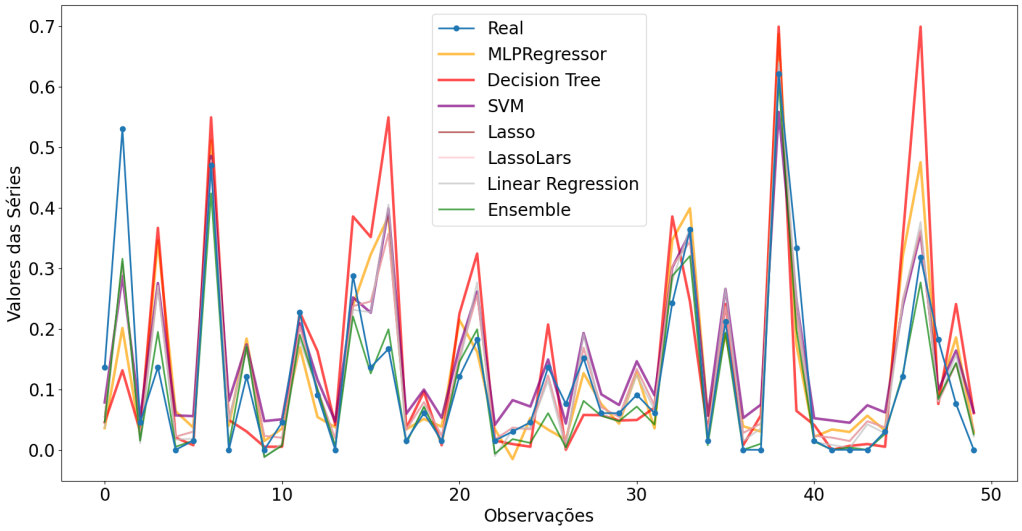


(b) Arapiraca.

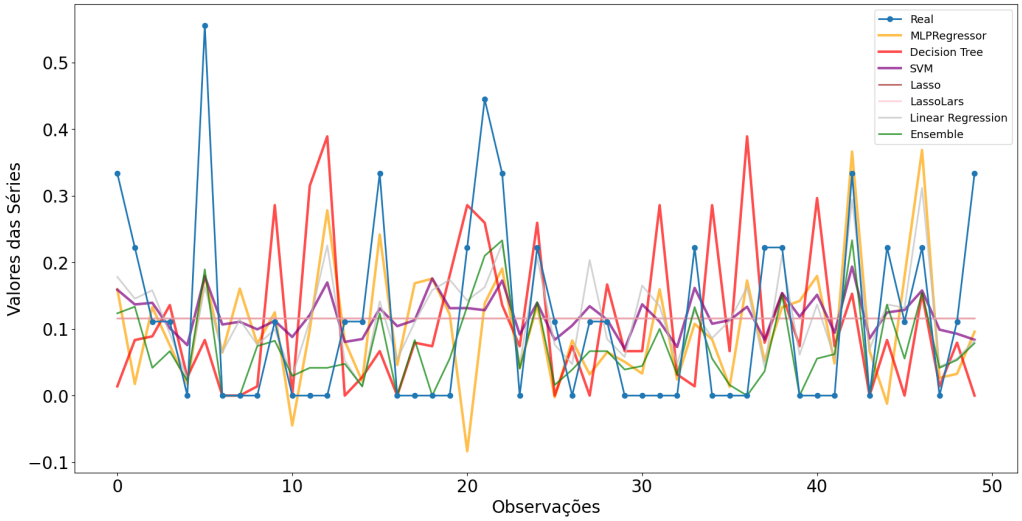


(c) Penedo.

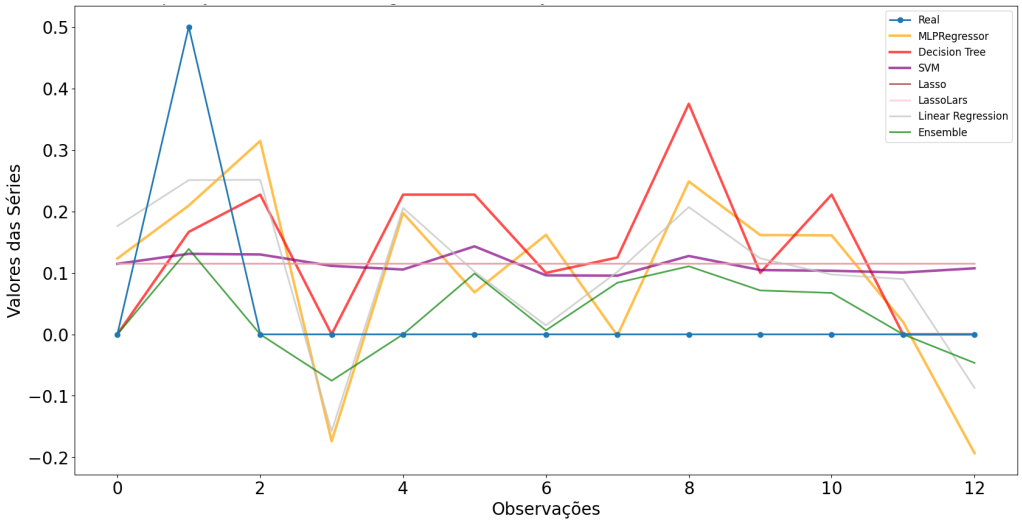
Figura 4. Previsões para os casos confirmados de COVID-19.



(a) Alagoas.



(b) Arapiraca.



(c) Penedo.

Figura 5. Previsões para os casos de óbitos de COVID-19.

6. Conclusão

Este estudo relatou os resultados de um projeto de iniciação científica que propôs o desenvolvimento de um sistema de previsão baseado em um Ensemble de aprendizagem de máquina para prever casos de COVID-19 no estado de Alagoas. Os resultados destacaram a eficácia do sistema na realização de previsões epidemiológicas em áreas menos assistidas, como nos municípios do interior de Alagoas.

Os resultados obtidos apontam para a confirmação da hipótese de pesquisa deste trabalho que diz que a combinação de modelos individuais distintos podem melhorar o desempenho de previsão de modelos combinados. Os resultados revelaram que a combinação mais eficiente de modelos do Ensemble foi determinada pela moda, indicando que agrupar as previsões mais frequentes apresentaram melhor desempenho. Além disso, observou-se a predominância de padrões lineares nas séries temporais, uma vez que os modelos lineares obtiveram as melhores classificações no ranking de Friedman.

Como próximos passos, recomenda-se explorar técnicas adicionais de Ensemble, como a técnica de poda que visa excluir os modelos do conjunto que apresentam o menor desempenho, assim conservando apenas os modelos que ofereçam previsões mais precisas. Adicionalmente, sugere-se ampliar este estudo para outras regiões geográficas ou para previsão de outras doenças.

Referências

Acito, F. (2023). Ensemble models. In *Predictive Analytics with KNIME: Analytics for Citizen Data Scientists*, pages 255–265. Springer.

AlJame, M., Ahmad, I., Imtiaz, A., and Mohammed, A. (2020). Ensemble learning model for diagnosing covid-19 from routine blood tests. *Informatics in Medicine Unlocked*, 21:100449.

Ashofteh, A., Bravo, J. M., and Ayuso, M. (2022). An ensemble learning strategy for panel time series forecasting of excess mortality during the covid-19 pandemic. *Applied Soft Computing*, 128:109422.

Carlotto, G. B. (2021). Previsão da evolução da covid-19 utilizando métodos de machine learning.

Chiattone, H., Teixeira, H., Vasques, M., Caldeira, L., Izzo, L., Lorandi, A., Rodrigues, A., Gatti, M., and Ripardo, J. (2022). A atuação do psicólogo no pronto socorro adulto e o apoio aos pacientes e familiares na pandemia covid-19. *Hematology, Transfusion and Cell Therapy*, 44:S598.

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y., et al. (2022). Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the united states. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

Espinosa, M. M., de Oliveira, E. C., Melo, J. S., Damaceno, R. D., and Terças-Trettel, A. C. P. (2020). Predição de casos e óbitos de covid-19 em mato grosso e no brasil. *Journal of Health & Biological Sciences*, 8(1):1–7.

Figueiredo Filho, D. B. and Silva Júnior, J. A. (2009). Desvendando os mistérios do coeficiente de correlação de pearson (r). *Revista Política Hoje*, 18(1):115–146.

Herbold, S. (2020). Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, 5(48):2173.

Kwekha-Rashid, A. S., Abduljabbar, H. N., and Alhayani, B. (2023). Coronavirus disease (covid-19) cases analysis using machine-learning applications. *Applied Nanoscience*, 13(3):2013–2025.

Latorre, M. d. R. D. d. O. and Cardoso, M. R. A. (2001). Análise de séries temporais em epidemiologia: uma introdução sobre os aspectos metodológicos. *Revista brasileira de epidemiologia*, 4:145–152.

Liapis, C. M., Karanikola, A., and Kotsiantis, S. (2020). An ensemble forecasting method using univariate time series covid-19 data. In *Proceedings of the 24th Pan-Hellenic Conference on Informatics*, pages 50–52.

Maaliw, R. R., Ballera, M. A., Mabunga, Z. P., Mahusay, A. T., Dejelo, D. A., and Seño, M. P. (2021). An ensemble machine learning approach for time series forecasting of covid-19 cases. In *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0633–0640. IEEE.

Morettin, P. A. and Toloi, C. M. (2018). *Análise de séries temporais: modelos lineares univariados*. Editora Blucher.

Santosh, K. (2020). Covid-19 prediction models and unexploited data. *Journal of medical systems*, 44(9):170.

Shastri, S., Singh, K., Deswal, M., Kumar, S., and Mansotra, V. (2022). Cobid-net: a tailored deep learning ensemble model for time series forecasting of covid-19. *Spatial Information Research*, 30(1):9–22.

Zhou, T., Lu, H., Yang, Z., Qiu, S., Huo, B., and Dong, Y. (2021). The ensemble deep learning model for novel covid-19 on ct images. *Applied soft computing*, 98:106885.