Research, Society and Development, v. 12, n. 3, e8112340438, 2023
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v12i3.40438

Uso do ARIMA e SVM para previsão de séries temporais do sistema elétrico brasileiro

Use of ARIMA and SVM for forecasting time series of the Brazilian electrical system

Uso de ARIMA y SVM para pronósticos de series de tiempo del sistema eléctrico brasileño

Recebido: 04/02/2023 | Revisado: 18/02/2023 | Aceitado: 19/02/2023 | Publicado: 25/02/2023

Lucas Renan Maués Nunes
ORCID: https://orcid.org/0000-0001-8657-0183
Universidade do Estado do Pará, Brasil
E-mail: lucasnunes2030@gmail.com
Juam Sousa Veras
ORCID: https://orcid.org/0000-0001-8668-7082
Universidade do Estado do Pará, Brasil
E-mail: juam13pb@gmail.com
João Pedro Ribeiro Silva
ORCID: https://orcid.org/0000-0002-4136-6733
Universidade do Estado do Pará, Brasil
E-mail: kandrocmatunein@gmail.com
Thiago Nicolau Magalhães de Souza Conte
ORCID: https://orcid.org/0000-0002-1288-366X
Universidade do Estado do Pará, Brasil
E-mail: thiagonconte@uepa.br
Wilker José Caminha dos Santos
ORCID: https://orcid.org/0000-0002-5265-583X
Universidade do Estado do Pará, Brasil
E-mail: wilkercaminha@uepa.br
Roberto Célio Limão e Oliveira
ORCID: https://orcid.org/0000-0002-6640-3182
Universidade Federal do Pará, Brasil
E-mail: limao@ufpa.br

Resumo

O presente trabalho se propõe a prever séries temporais do setor elétrico brasileiro. Para tanto, procurou-se realizar previsões para o Preço de Liquidação das Diferenças (PLD) e a velocidade do vento para movimentação dos aerogeradores, que transforma a energia cinética das correntes de ar em energia elétrica, a partir da metodologia ARIMA, baseado na estatística computacional, e o modelo SVM, proveniente da área de inteligência artificial, sendo que o período analisado corresponde de 2001 a 2009 para o PLD e de 2004 a 2017 para o vento. Os resultados fornecem uma ferramenta de análise para o mercado livre de energia, na medida que demonstram tendências de preços e produção elétrica, servindo de auxílio à tomada de decisões, sendo o ARIMA, o modelo preditivo que performou melhor as previsões a curto prazo. Apesar disso, conclui-se que o SVM tem um potencial para produzir resultados mais assertivos para previsões a longo prazo, visto que o modelo tem muitas características que podem ser exploradas e assim potencializar previsões com grandes volumes de dados em situações mais complexas.

Palavras-chave: ARIMA; SVM; Aprendizado de máquina; Séries temporais; Setor Elétrico Brasileiro.

Abstract

The present work proposes to forecast time series of the Brazilian electricity sector. For this purpose, an attempt was made to make predictions for the Settlement Price of Differences (PLD) and the wind speed for moving wind turbines, which transforms the kinetic energy of air currents into electrical energy, based on the ARIMA methodology, based on statistics computational, and the SVM model, from the area of artificial intelligence, and the period analyzed corresponds from 2001 to 2009 for the PLD and from 2004 to 2017 for the wind. The results provide an analysis tool for the free energy market, as they demonstrate price trends and electricity production, serving as an aid to decision-making, with ARIMA being the predictive model that performed best in short-term forecasts. Despite this, it is concluded that the SVM has the potential to produce more assertive results for long-term forecasts, since the model has many characteristics that can be exploited and thus enhance forecasts with large volumes of data in more complex situations.

Keywords: ARIMA; SVM; Machine learning; Time series; Brazilian Electric Sector.

Resumen
El presente trabajo propone pronosticar series de tiempo del sector eléctrico brasileño. Para ello se intentó realizar predicciones para el Precio de Liquidación de las Diferencias (PLD) y la velocidad del viento para los aerogeneradores en movimiento, que transforma la energía cinética de las corrientes de aire en energía eléctrica, con base en la metodología ARIMA, basada en estadísticas computacional, y el modelo SVM, del área de inteligencia artificial, y el periodo analizado corresponde del 2001 al 2009 para el PLD y del 2004 al 2017 para el eólico. Los resultados brindan una herramienta de análisis para el mercado libre de energía, ya que demuestran la evolución de los precios y la producción eléctrica, sirviendo de ayuda para la toma de decisiones, siendo ARIMA el modelo predictivo que mejor se desempeñó en los pronósticos a corto plazo. A pesar de esto, se concluye que el SVM tiene potencial para producir resultados más asertivos para pronósticos a largo plazo, ya que el modelo tiene muchas características que pueden ser explotadas y así mejorar los pronósticos con grandes volúmenes de datos en situaciones más complejas.
Palabras clave: ARIMA; SVM; Aprendizaje automático; Serie de tiempo; Sector Eléctrico Brasileño.

1. Introdução

As mudanças evidenciadas ao longo dos anos, se devem ao fato das grandes transformações proporcionadas pela evolução tecnológica, que tiveram como marco as três revoluções industriais registradas ao final do século XVIII. A partir desse momento, o mundo vivenciou grandes mudanças que impactaram diversos campos da sociedade, estimulada pela exposição a novas tecnologias, e atualmente, se destacando com a escalada da indústria 4.0, considerada a quarta revolução industrial, e sustentada por conceitos que amadureceram com o tempo, bem como, a Internet das Coisas, Segurança Cibernética, Big Data, Computação em Nuvem e a Inteligência Artificial (Sakurai & Zuchi, 2018).

Evidentemente, a tecnologia continua se desenvolvendo em suas particularidades, assim como, as barreiras que cada uma delas encontram para sua adoção, como destaca Iszczuk et al. (2021), no entanto, destacamos o termo que engloba todos esses aspectos, os dados, que desempenham um papel fundamental e vem atuando ativamente em meio a esse processo, haja vista, que eles podem ser descritos, representados e interpretados de formas distintas (Isotani & Bittencourt, 2015). Logo, ao trabalhar as análises e consequentemente a interpretação dos dados, é essencial na resolução de problemas, que apoiado em técnicas como as de aprendizagem de máquina, proporciona ao decisor uma visão mais ampla na definição de movimentos estratégicos.

Apoiado nisso, o cerne de análises será o setor elétrico brasileiro, visto como um gerador de grandes volumes de dados, especialmente o mercado de curto prazo, que diz respeito a compra e venda de energia elétrica, no qual os dados podem ser extraídos, analisados e posteriormente tratados com o intuito de agregar valor e basear tomadas de decisões, como aquelas de ordem econômica que impactam na produtividade, inovação e sustentabilidade (da Silva, 2021). Outrossim, essa abordagem proporciona a adoção de técnicas que corroboram com minimização das emissões de Gases do Efeito Estufa (GEE) para atmosfera, inclusive para substituição dos combustíveis fósseis a partir do emprego de fontes de energia sustentável, o que estimula a produção de energia limpa e contribui com a preservação ambiental.

Há diversos métodos que envolvem a aprendizagem de máquina propostos na literatura com diferentes abrangências, porém se tratando da energia elétrica brasileira, podemos citar Lagasse (2020), que utiliza ferramentas estatísticas para prever o comportamento do PLD a partir da análise de regressão linear múltipla e séries temporais. Além disso, Conte et al. (2021), propõe uma abordagem híbrida para previsão de séries temporais que combina Algoritmo Genético e Long Short Term Memory (AG+LSTM) para predizer o PLD e a velocidade do vento, e igualmente, temos a combinação de técnicas lineares e não lineares para previsão de séries temporais da velocidade do vento com dados de diferentes estados da região Nordeste, porém, envolvendo modelos estatísticos como ARIMA e de inteligência artificial como MLP e LSTM (Assunção et al., 2022). Por fim, (Nascimento, 2022), trabalha métodos de suavização exponencial e modelos de redes neurais artificiais para previsão de séries temporais do preço de energia elétrica no mercado de curto prazo do submercado Sudeste/Centro-Oeste.

Neste sentido, o presente estudo buscou analisar e aplicar o método Auto-Regressivo Integrado de Médias Móveis

(ARIMA) e a Regressão de Máquinas de Vetores de Suporte (SVM) para previsão de séries temporais do setor de energia elétrica no Brasil. Portanto, a finalidade será determinar o modelo com melhor performance preditiva a partir da análise comparativa dos erros de previsão obtidos com os resultados do ARIMA e SVM.

Para alcançar este objetivo será identificado um modelo de previsão a partir dos dados referente ao Preço de Liquidação das Diferenças (PLD) da região Norte e a geração de energia eólica, especificamente da cidade de Macau, no Nordeste brasileiro. Dessa maneira, será necessário cumprir algumas etapas, como a análise e pré-processamento dos dados, estimação e ajuste dos modelos a partir da observação dos seus parâmetros, e por fim, análise dos resultados com a finalidade de obter um diagnóstico na comparação dos modelos apresentados.

2. Metodologia

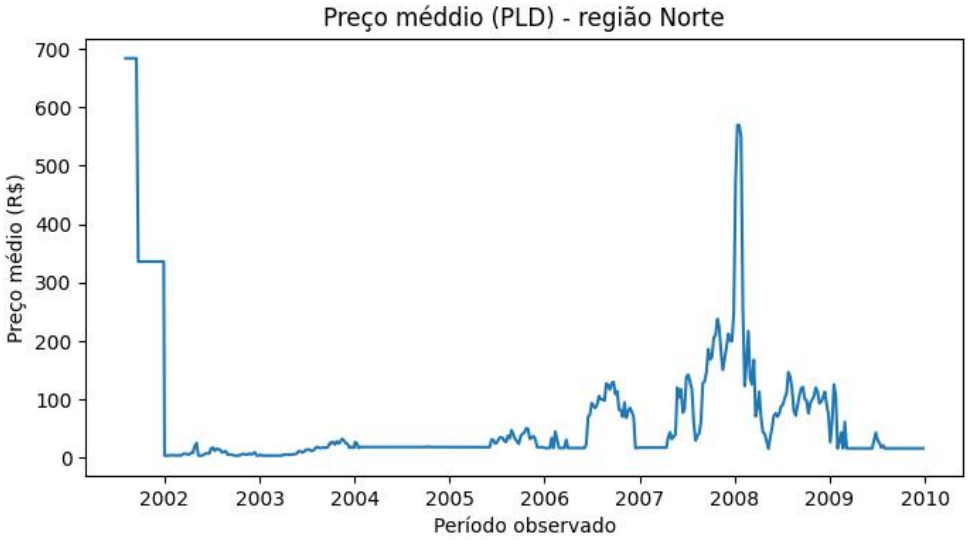
A proposta do trabalho consiste em consumir dados relativos à oferta e demanda de energia elétrica no mercado de curto prazo brasileiro, mais especificamente sobre o PLD, e alternativas de produção elétrica como a velocidade do vento. Para isso, utilizam-se modelagens estatísticas e de machine learning para previsão de séries temporais, sendo respectivamente, o ARIMA e SVM, ademais, todo esse processo de concepção dos modelos foi feito utilizando a linguagem python, assim como as bibliotecas utilizadas.

Consideremos, primeiramente, uma metodologia tradicional da estatística proposta por George Box e Gwilym Jenkins, o método Box-Jenkins, que consiste na utilização de modelos Auto-Regressivos Integrados de Médias Móveis ARIMA(p,d,q), que analisa os dados históricos para determinar os pontos de observações futuros como ressalta Borsato e Corso (2019), e seguidamente, uma abordagem de machine learning para problemas de classificação com suporte para regressão, as Máquinas de Vetores de Suporte (SVM). Nesse caso, a regressão SVM ou somente SVR, acrescenta um hiperparâmetro Epsilon, que controla a largura da via, além daqueles já utilizados para classificação, como C, responsável por equilibrar a largura da via com os limites de violação das margens, além de Gamma, que serve como um hiperparâmetro de regularização que age no estreitamento das curvas, e o Kernel, que ajusta as previsões para resolver problemas não lineares (Geron, 2019).

2.1 Base de dados e pré-processamento

Os dados são referentes ao preço médio do PLD da região Norte, com 440 observações semanais de 2001 a 2009, e da velocidade do vento em aerogeradores de energia, contendo 4900 observações diárias do período de 2004 a 2017 da cidade de Macau no Nordeste brasileiro (Conte et al., 2021). As Figuras 1 e 2 apresentam, respectivamente, as visualizações das bases de dados do PLD semanal médio e da velocidade do vento.

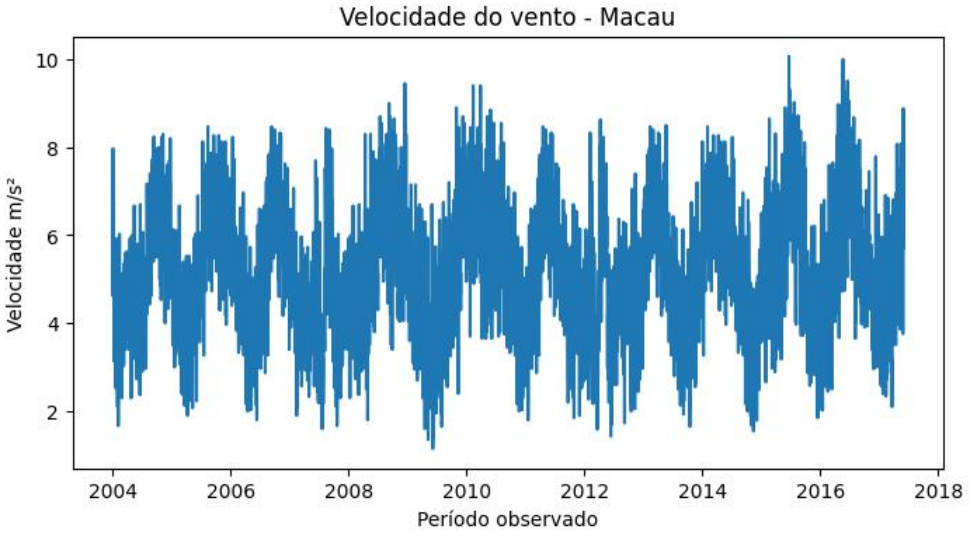
Figura 1 - Dados do preço semanal médio da região Norte.



Fonte: Autores.

A Figura 1 é referente a distribuição dos dados para o Preço da Liquidação das Diferenças nos períodos observados, nesse caso, podemos verificar que ao final de 2002 e no início de 2008, houve altas nos preços de energia registradas para o período, o que pode ter sido ocasionado por diversos fatores, como a queda na produção energética ou a alta na demando por energia elétrica.

Figura 2 - Dados da velocidade do vento em Macau.



Fonte: Autores.

Diferentemente, a Figura 2 nos apresenta a distribuição dos dados da velocidade do vento, que se refere a geração de energia eólica, sendo uma importante fonte na produção energética que se soma a uma parcela de toda eletricidade produzida em território nacional, nesse caso, é importante ressaltar a linearidade nos dados, ou seja, em todo o horizonte observado, não houve grandes picos na variação da velocidade do vento, contribuindo para obtenção de resultados mais assertivos.

Enfim, é importante evidenciar a etapa de pré-processamento dos dados, pois é nesse ponto que eles serão preparados e corrigidos a fim de melhorar a eficácia das previsões, dessa forma, será possível minimizar a influência que os dados inconsistentes exercem sobre os resultados (Conceição et al., 2021). Isso posto, para as referidas bases de dados foram trabalhadas

as conversões dos tipos de dados, a formatação dos valores numéricos e o ajustes das casas decimais, além disso, foi realizada a separação dos dados com 75% para treino e 25% para teste, e a padronização dos valores para aplicar as previsões.

2.2 Análise de séries temporais

Na oportunidade, foi considerado a análise da série temporal, para entender a estrutura e identificar se os dados analisados obedecem às regras de estacionariedade através do teste de Dickey-Fuller, função da biblioteca *statsmodels*, que determina se a média e a variância são constantes, nesse caso, analisa se a série apresenta tendência (Filho, 2022). Sendo assim, entende-se, como uma série de observações presentes ao longo de um determinado intervalo de tempo, tendo como finalidade a análise da correlação a partir da disposição dos dados históricos.

Esse procedimento é importante para a modelagem do ARIMA, haja vista, que o modelo requer, ainda, a definição dos parâmetros "P" responsável pela ordem do modelo autorregressivo, "D" para a quantidade de diferenciações necessárias para tornar a série estacionária e "Q" que concerne a parte da definição da ordem de médias móveis do modelo de séries temporais (Paula et al., 2022).

2.3 Otimização de hiperparâmetros

Para o modelo ARIMA, o processo de definição dos parâmetros pode acontecer de forma manual através da observação dos gráficos *Autocorrelation Function* (ACF) e *Partial Autocorrelation Function* (PACF) para se obter respectivamente os parâmetros "q" e "p". Entretanto, foi utilizado uma abordagem automática, por meio de métodos de otimização de hiperparâmetros, utilizado para avaliar o modelo e facilitar na busca por configurações que apresentam a melhor performance (Kirchoff, 2019). Dito isso, a proposta foi empregar, para o ARIMA, a função Auto ARIMA, da biblioteca *pmdarima*, e o método de pesquisa em grade, já para o SVM, foi utilizado os métodos Random Search e Bayesian Search, importadas respectivamente, das bibliotecas sklearn e skopt.

O Auto ARIMA percorreu um conjunto de parâmetros que melhor se ajusta a um modelo ARIMA para determinar o número de diferenciações "D" e estimar os valores de "P" e "Q", conforme as regras estabelecidas pelo Akaike Information Criterion (AIC), que fornece o melhor modelo para a série temporal (Awan & Aslam, 2020). Em contrapartida, o Grid Search, comumente usado em algoritmos de machine learning, objetiva definir um número de valores para os parâmetros e iterar a maior quantidade possível de combinações a fim de identificar aquela com o menor erro (Kirchoff, 2019).

De acordo com Neves (2020), o Random Search gera de forma aleatória os conjuntos de hiperparâmetros, sem a necessidade de testar todas as combinações para minimizar o tempo de execução do algoritmo. Por outro lado, o Bayes Search agiliza esse processo ao reutilizar as informações nos pontos de iterações passadas, ou seja, o método melhora a busca de hiperparâmetros a partir dos resultados de combinações nas execuções anteriores (Alves et al., 2022). Em ambos os métodos, os hiperparâmetros do modelo SVR otimizados foram, C, Epsilon e Gamma.

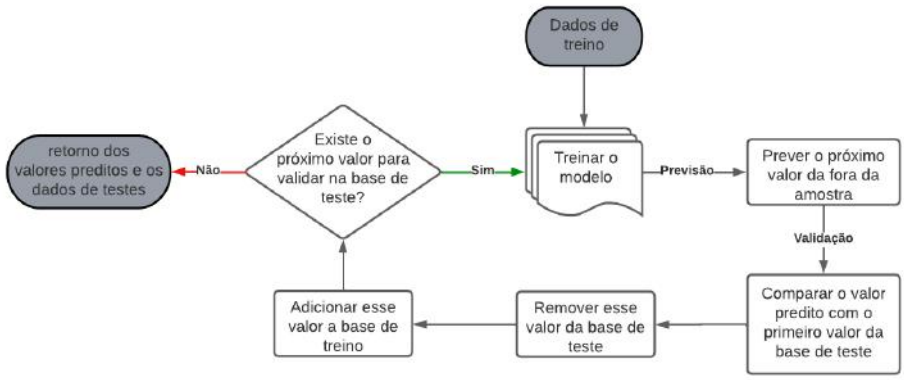
2.4 Autoregressive Integrated Moving Average (ARIMA)

O ajuste dos modelos ARIMA se encarrega de criar as previsões dentro da base amostral, sendo então, responsável por avaliar a eficiência dos modelos gerados pelo Auto ARIMA e o Grid Search, dessa maneira, a combinação de parâmetros que obteve o melhor desempenho, nessa etapa, foi selecionada como o modelo ideal para criar previsões fora da amostra.

O algoritmo responsável por extrapolar as previsões além da base amostral, foi a validação Walk Forward, nesse caso, a técnica fundamenta-se em utilizar as abordagens Expanding Window (EW), que se resume em adicionar novos valores a base de treino conforme o conjunto de teste seja iterado, assim, expandindo o horizonte de treinamento à cada novo valor predito, e o

Sliding Window (SW), que funciona de forma similar ao anterior, porém ele remove da janela de treinamento os valores mais antigos, assim sempre trabalhando com uma faixa de valores mais recentes (de Paula et al., 2020). A Figura 3 apresenta o procedimento adotado pelo algoritmo no processo de previsão fora da amostra.

Figura 3 - Representação da execução do algoritmo Walk Forward.



Fonte: Autores.

Conforme a Figura 3 descreve, o algoritmo consiste em treinar o modelo com a base de treino, validar o valor predito com a base de teste e por fim adicionar o valor real à base de treino, após isso o processo será repetido até que todos os dados da base de teste envolvidos na previsão tenham sido percorridos e adicionados à base de treinamento.

2.5 Support Vector Regression (SVR)

Para a construção e efetivação dos resultados com o SVR foi necessário a utilização da técnica de Feature Engineering, um procedimento fundamental em previsão de séries temporais com machine learning, o qual é a extração de novos recursos a partir da variável observada (Wang et al., 2022). Esse processo também pode ser entendido como uma remodelagem do problema de série temporal para aprendizado supervisionado, em que precisamos inserir os dados nas entradas X e saída y da função fit(X,y) do modelo SVR, encontrada na biblioteca *sklearn.svm.SVR*.

Essa técnica foi responsável por produzir novas variáveis de entrada com atrasos X(t-1), X(t-2), X(t-3)... X(t-n) em relação à variável de saída em seu tempo atual y(t) (Brownlee, 2017). Dessa forma, o modelo recorrerá às variáveis remodeladas como recurso de entrada X para aprender seus padrões e comportamentos, e assim prever as saídas y, e decorrente disso, os modelos serão ajustados para obter previsões com métodos diretos, que usa de todas as entradas disponíveis, e métodos iterativos, que cria previsões em várias etapas de tempo (Lim & Zohren, 2021).

É nesse ponto que trabalhamos a definição dos modelos SVR com base nos métodos Random Search e Bayes Search, pois para avaliá-los, será necessário apresentar as entradas e saídas para os métodos de busca, sendo então, crucial a aplicação da técnica de Feature Engineering para concepção desses recursos, representados pelos atrasos. Entretanto, antes de prosseguir com a otimização de hiperparâmetros, devemos encontrar o número de atrasos necessários em nossos dados, visto que esses recursos serão úteis para avaliação dos parâmetros. Para essa finalidade, será construído um algoritmo responsável por criar essas variáveis gradativamente e treiná-las em um modelo base, por meio do método de previsão direta, e ao final, será retornado a quantidade de atrasos que melhor se ajustou, ou seja, o número de entradas necessárias para prever as saídas (Brownlee, 2017).

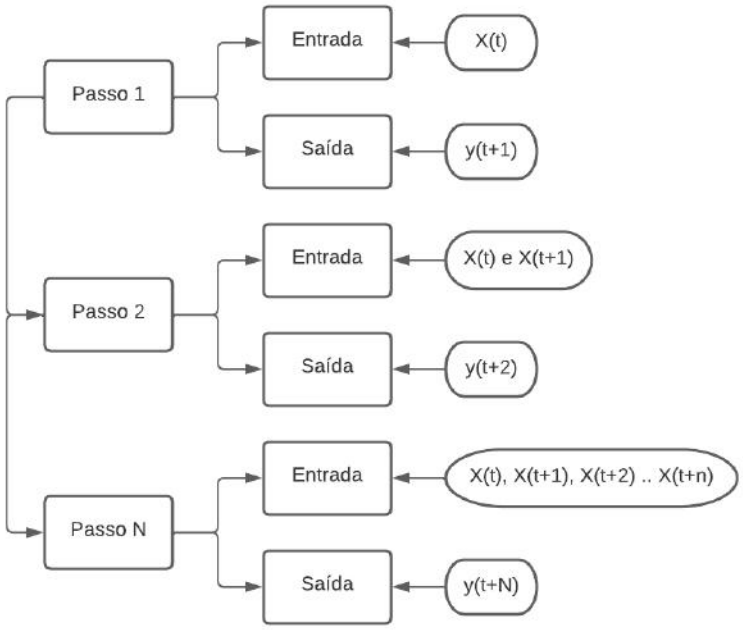
Esse procedimento acontece da seguinte forma, será primeiramente realizada a construção de uma nova variável com atraso X(t-1) que servirá como entrada do modelo para prever a saída y(t) em seu tempo atual, após isso, um modelo SVR base

será treinado, ou seja, um modelo que dispensa a necessidade de definição de parâmetros, visto que o objetivo é apenas verificar o comportamento das entradas e saídas em relação aos atrasos, e ao final, será criada a previsão por meio do método direto. Por conseguinte, o algoritmo retorna uma pontuação de acuracidade que indica o número de atrasos em cada iteração, e assim, o processo será repetido até que todo o intervalo definido tenha sido percorrido e a melhor pontuação seja obtida, assim como seu número de atrasos, e nessa ocasião, foram obtidos 88 atrasos para os dados sobre a velocidade do vento e 13 atrasos para o preço médio do PLD.

Após essa etapa, foi possível iniciar a otimização de hiperparâmetros, por meio dos métodos Random Search e Bayes Search, que utilizará as novas variáveis de entrada, encontradas anteriormente, para otimizar os hiperparâmetros SVR e encontrar o melhor modelo em cada método. Em seguida, será utilizado novamente o método de previsão direta, agora com os modelos encontrados na busca, que permitirá a análise do comportamento desses modelos diante aos dados utilizados, haja vista, que esse tipo de previsão utiliza todas as entradas disponíveis, em outras palavras, esse processo contém apenas o treinamento do modelo, sem a necessidade de validação dos resultados (Lim & Zohren, 2021).

Finalmente, o método iterativo será utilizado para produzir previsões em etapas de tempo, isto é, possibilitando o processo de previsão fora da amostra, em que foram apresentadas ao modelo as entradas e saídas em seus respectivos passos, e para isso, o modelo é treinado diversas vezes com entradas de tamanhos diferentes em cada passo no tempo. Nesse caso, ao invés de utilizar um número fixo de atrasos, o processo de previsão foi feito de forma escalonada, em que o número de atrasos cresce à medida que os passos de tempo se alargam (Macêdo, 2022). A Figura 4 exemplifica esse processo de previsão iterativa.

Figura 4 - Procedimento de previsão iterativa.



Fonte: Autores.

Consoante a Figura 4, foram utilizados os dados de entrada X(t) e as saídas y(t+1) para prever um passo à frente, depois disso, são apresentadas as entradas X(t) e X(t+1) para prevermos a saída y(t+2) dois passos à frente, e assim, esse processo será repetido até prevermos todos os 12 passos de tempo, e dessa forma, o objetivo será analisar os erros nas etapas 1, 3, 8 e 12.

Para finalizar, é válido destacar a etapa responsável pela acurácia dos modelos, que compara os valores previstos com os reais, e assim, indica o quanto o modelo errou na previsão (Macêdo, 2022). As métricas de erro utilizadas para avaliar o

desempenho dos modelos são a Raiz do Erro Médio Quadrático (RMSE) e o Erro Médio Quadrático (MSE).

3. Resultados e Discussão

Consoante os ajustes realizados nos dados, foi possível analisar o comportamento da série temporal e posteriormente criar as previsões dentro e fora da base amostral. O Quadro 1 apresenta os valores do teste de Dickey-Fuller.

Quadro 1 - Resultados do teste de Dicker-Fuller.

	Teste Estatístico	P-Value	Valores Críticos		
			1%	5%	10%
PLD	-4.458580	0.000233	-3.445794	-2.868349	-2.570397
Vento	-5.147908	0.000011	-3.431694	-2.862134	-2.567086

Fonte: Autores.

Com base no teste de estacionariedade observado no Quadro 1, o p-value está abaixo dos 5% nas duas variáveis analisadas, além disso, o teste estatístico é inferior aos valores críticos 1%, 5% e 10%, nesse caso, rejeitando a hipótese nula para a série temporal não estacionária, visto que o teste assumiu a hipótese alternativa que indica a série como estacionária (Silva, 2020). A vista disso, dispensa-se a transformação dos dados para ajustar o modelo, o que foi sustentado com os resultados da otimização de hiperparâmetros para o ARIMA, haja vista, que as melhores configurações encontradas não trouxeram diferenciações no parâmetro "D", responsável pela transformação da série temporal. O Quadro 2 apresenta os modelos encontrados pelos métodos de otimização para o ARIMA com base nos dados do PLD e vento.

Quadro 2 - Resultado da otimização dos hiperparâmetros para o modelo ARIMA.

	Auto ARIMA		Grid Search	
Dados	Vento	PLD	Vento	PLD
Modelo	(10,0,5)	(1,0,1)	(23,0,25)	(16,0,24)
RMSE	1.025	42.786	1.018	42.173
MSE	1.050	1830.600	1.036	1778.597

Fonte: Autores.

Diante dessas circunstâncias, as informações apresentadas no Quadro 2 nos permitem analisar os resultados da previsão com o ARIMA a partir dos métodos de otimização Auto ARIMA e Grid Search, inicialmente dentro da amostra (Barbosa et al., 2020). Nesse caso, os resultados são apresentados através dos erros de previsão, com as métricas RMSE e MSE, utilizadas para testar a acurácia dos modelos ajustados.

Nesse ponto, cabe salientar a importância da otimização dos parâmetros *p*, *d* e *q*, pois os erros obtidos com os métodos de otimização se mostraram relativamente menores, em comparação aos modelos selecionados manualmente por meio dos gráficos ACF e PACF, que nessa ocasião, resultaram em modelos autoregressivos, como o ARIMA(1,0,0) para o PLD médio, e ARIMA(6,0,0) para a velocidade do vento, como aborda (Nascimento, 2022).

Adiante, observamos que os modelos ARIMA(23,0,25) para a velocidade do vento e o ARIMA(16,0,24) para o preço médio do PLD, obtiveram os melhores ajustes na previsão dentro da amostra, sendo o Grid Search, o método de otimização com o menor erro de previsão. Nesse ponto, os dados foram previstos fora da amostra para 12 passos à frente, sendo um salto de 12 dias para o vento e 12 semanas para o PLD (Conte et al., 2021). As Figuras 5 e 6 apresentam respectivamente os resultados da

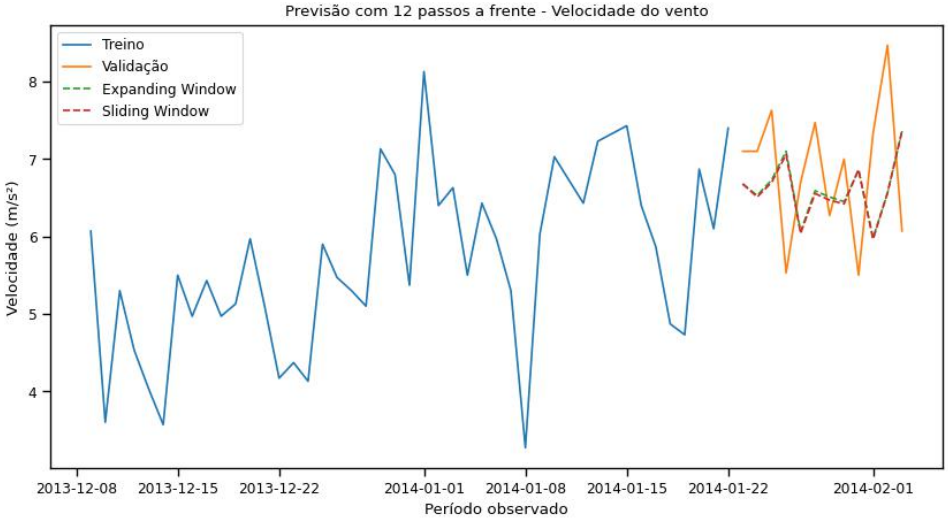
previsão do preço médio do PLD e velocidade do vento com a validação Walk Forward para o modelo ARIMA.

Figura 5 - Previsão de 12 passos à frente para o preço médio do PLD.

Fonte: Autores.

Na previsão do PLD médio, é importante destacar que o conjunto reservado para o treinamento do modelo ARIMA vai até o final de 2007, nesse caso, todas as ocorrências após esse período pertence à base de teste, ou seja, será o conjunto de dados usado para validar os resultados, conforme expressado na Figura 5.

Figura 6 - Previsão de 12 passos à frente para a velocidade do vento.



Fonte: Autores.

Da mesma forma, a previsão da velocidade do vento reserva parte dos dados para o treinamento do modelo, que nesse caso, vai até o início de 2014, sendo a maior parcela utilizada no treino, como representado na Figura 6. Logo, foram feitas previsões para 12 passos à frente nas duas bases de dados, e assim, para avaliar o desempenho dos modelos, foi realizado um comparativo entre as abordagens Expanding Window e Sliding Window da técnica de validação apresentada anteriormente, a qual permitiu treinar e validar os modelos ARIMA em diferentes subconjuntos de dados (de Paula et al., 2020). Essa metodologia, permitiu entender o comportamento dos dados no método de validação para os modelos selecionados, em geral, foi observado pouca diferença entre ambos, sendo que nos dois casos as tendências se equiparam, exceto nas curvas de crescimento e

decaimento, em que as abordagens destoam, como na base de dados do PLD, no qual o Expanding Window se aproximou um pouco mais dos pontos de validação. Os Quadros 3 e 4 apresentam os erros de previsão em cada abordagem.

Quadro 3 - Erros de previsão para o preço médio da região Norte com o modelo ARIMA(16,0,24).

	Expanding Window		Sliding Window	
Passos	RMSE	MSE	RMSE	MSE
1	8.966	80.389	8.966	80.389
3	5.824	33.914	8.854	78.397
8	27.170	738.202	54.637	2985.186
12	54.805	3003.568	46.425	2155.289
Erro Geral	69.516	4832.523	75.138	5645.726

Fonte: Autores.

Quadro 4 - Erros de previsão para a velocidade do vento em Macau com o modelo ARIMA(23,0,25).

	Expanding Window		Sliding Window	
Passos	RMSE	MSE	RMSE	MSE
1	0.421	0.177	0.421	0.177
3	0.902	0.813	0.932	0.869
8	0.550	0,302	0.579	0.335
12	1.290	1.664	1.293	1.671
Erro Geral	1.092	1.192	1.096	1.201

Fonte: Autores.

Em suma, os erros individuais de previsão dos Quadros 3 e 4 indicam uma pequena vantagem do Expanding Window no processo de validação dos dados nas etapas 3, 8 e 12, nos dados sobre o vento, e somente as etapas 2 e 8, nos dados sobre o PLD, sendo que no processo de previsão de 1 passo a frente o resultado se manteve em ambas as abordagens de validação. Nesse caso, a validação com Expanding Window se ajustou melhor aos modelos utilizados, quando analisamos o erro geral obtido, indicando um domínio da abordagem no ajuste dos modelos.

Em contrapartida, no modelo SVR, os métodos responsáveis por encontrar as melhores configurações de ajuste aos dados, foram o Bayes Search (BS) e o Random Search (RS). Os Quadros 5 e 6 apresentam os resultados da otimização nas duas bases de dados.

Quadro 5 - Resultado da otimização de hiperparâmetros para o PLD.

	Hiperparâmetros			Avaliação	
Métodos	C	Epsilon	Gamma	RMSE	MSE
Bayes Search (BS)	1000	1	9.70e-07	24.437	597.152
Random Search (RS)	70.645	0.230	4.64e-06	23.425	548.750

Fonte: Autores.

Quadro 6 - Resultado da otimização de hiperparâmetros para a velocidade do vento.

	Hiperparâmetros		Avaliação	
Métodos	C	Epsilon	RMSE	MSE
Bayes Search (BS)	5.202	0.061	0.740	0.548
Random Search (RS)	31.623	1e-05	0.748	0.559

Fonte: Autores.

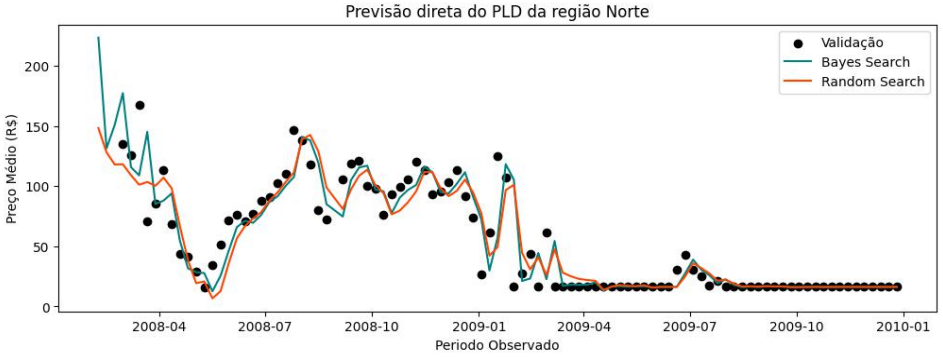
Primeiramente, vamos analisar os resultados das otimizações para o PLD médio apresentado no Quadro 5, em que observamos três parâmetros do modelo SVR que sofreram ajustes, sendo eles C, Epsilon e Gamma, além disso, analisamos os erros de previsão com as métricas RMSE e MSE para selecionar a configuração que melhor se ajustou aos dados. Em virtude disso, vale salientar que durante a otimização, os melhores ajustes foram encontrados com testes nas seguintes combinações, C variando entre os valores de 100 e 1000 para BS e sempre abaixo de 100 para RS, Epsilon entre 1 e 10 para BS e valores entre 0.1 a 1 para RS, já Gamma sempre testou valores entre 5 e 10, o que resultou nas informações do Quadro 5, o que nesse caso, o método Random Search apresentou uma melhor performance na análise dos erros da previsão direta para os dados sobre o PLD.

Por outro lado, o Quadro 6, nos apresenta a otimização para a velocidade do vento, contemplando os parâmetros C e Epsilon, entretanto, analisando os resultados, percebemos que o parâmetro Gamma não sofreu ajustes, nesse caso, ele assume automaticamente o valor padrão do SVR, definido como *scale*. Em relação aos testes de otimização, C retornou valores com melhores ajustes de 1 a 10 para BS e de 10 a 100 para RS, ou seja, quanto mais o valor de C se afastava dessas margens o erro tendia a aumentar, já o Epsilon se manteve com valores sempre abaixo de 0.1 para ambos os métodos, sendo o Bayes Search o método que errou menos na previsão direta para os dados do vento.

Acrescenta-se que, assim como no parâmetro Gamma, igualmente acontece no Kernel, que possibilita a separação dos dados não linearmente separáveis, segundo Acosta e Amoroso (2021), nesse sentido, o parâmetro recebe como valor padrão *Radial Basis Function* (RBF), pois a definição do parâmetro em ambos os métodos acrescia o tempo que os algoritmos ficavam em execução, visto que o Kernel requer um poder computacional maior que os demais parâmetros otimizados, como ressalta Géron (2019), e por esse motivo, optou-se por não incluí-lo na busca dos hiperparâmetros.

Em relação aos resultados da previsão, ficou segmentado em dois tipos, sendo o primeiro, a previsão direta, que possibilitou analisar o comportamento das configurações encontradas na etapa de otimização por meio dos métodos Bayes Search e Random Search, e para cumprir esse procedimento, foi utilizado todo o horizonte de teste, em que foram apresentadas todas as variáveis de entrada geradas anteriormente, e seus resultados preditivos foram comparados com a nossa variável de saída, ou seja, aquela que queremos prever (Géron, 2019). As Figuras 7 e 8 apresentam a plotagem da previsão direta com a comparação dos dois métodos com a variável de saída.

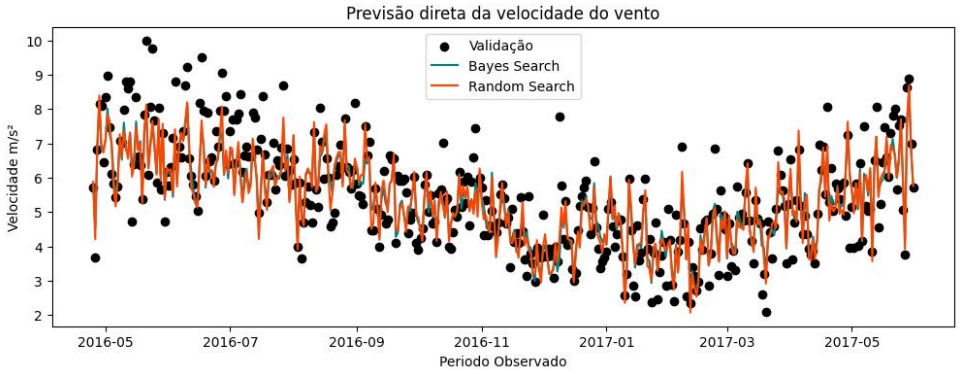
Figura 7 - Comparação da previsão direta para o PLD com os métodos de otimização.



Fonte: Autores.

Na Figura 7, exibimos as saídas dos modelos na base de dados do PLD utilizando os dois métodos de busca e comparamos com os dados de validação, nesse caso, foi aplicado a previsão com método direto, que utiliza todo horizonte de teste como entrada para os modelos otimizados.

Figura 8 - Comparação da previsão direta para a velocidade do vento com os métodos de otimização.

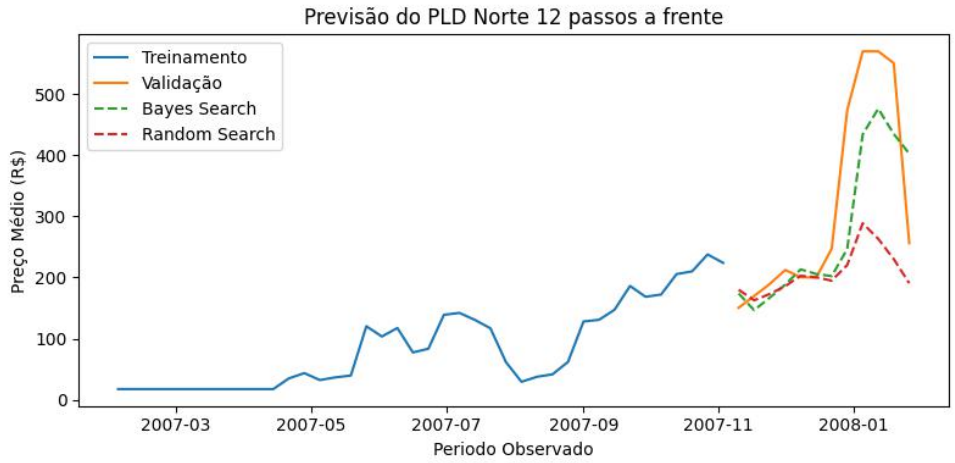


Fonte: Autores.

Da mesma forma, a Figura 8 apresenta as saídas obtidas com o método direto que testou os modelos otimizados para a velocidade do vento e compara com o conjunto de dados reservados para o teste, que nesse caso, serão nossos dados de validação.

Em seguida, modelamos a previsão iterativa, responsável por produzir as previsões em 12 passos no tempo, ou seja, nessa etapa foi realizado o treinamento e validação das previsões com base nos dados de teste. Nesse caso, como resultado, obtivemos em cada processo de iteração um valor predito que utilizou as entradas criadas, assim, gradualmente o problema foi remodelado para gerar uma nova variável, possibilitando o crescimento exponencial dos atrasos, e dessa forma, o valor atual das entradas foi apresentado ao modelo treinado para gerar um valor de previsão, ou seja, a cada passo de tempo, os dados de saída previstos são utilizados como variável de entrada para prever o passo seguinte (Nascimento, 2022). As Figuras 9 e 10 apresentam uma plotagem comparativa dos resultados da previsão com método iterativo para 12 passos à frente.

Figura 9 - Previsão com 12 passos à frente para o PLD médio.

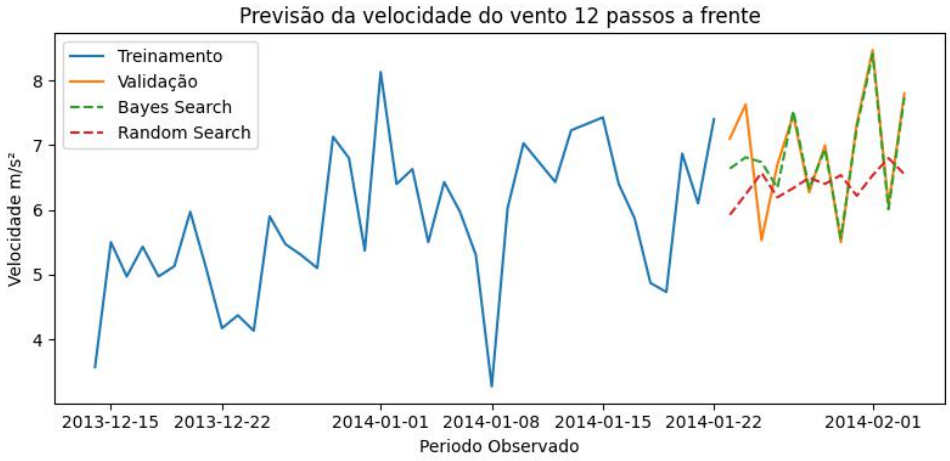


Fonte: Autores.

Consoante ao apresentado na Figura 9, constatamos como se deu o processo de treino e validação das previsões com 12 passos à frente para o SVR na base de dados do PLD, nesse caso, os dados de treino do modelo estão contidos na mesma faixa de tempo que aqueles utilizados no ARIMA, exceto na validação dos resultados, que sofreu um pequeno ajuste por conta dos

atrasos empregados para gerar as novas features, e diante disso, podemos distinguir as performances dos métodos BS e RS em relação ao conjunto de validação.

Figura 10 - Previsão com 12 passos à frente para a velocidade do vento.



Fonte: Autores.

Semelhantemente, a Figura 10 exibe as mesmas características adotadas para o PLD, entretanto utilizando os dados sobre o vento, e nesse ponto, é importante ressaltar que o horizonte de treino reservado para essa base de dados é mais abrangente, uma vez que o intervalo de tempo observado é maior em relação ao PLD, dessa maneira, o modelo possui mais recursos para produzir resultados mais assertivos.

Segundo os resultados que se configuraram até aqui, precisamos dividir a análise e entender como cada modelo se comportou em suas respectivas bases de dados. Analisando, primeiramente, os resultados referente ao PLD, observamos que no Quadro 5, o Random Search apresentou uma pequena vantagem no desempenho para a previsão com método direto, quando analisamos os erros retornados, porém, ao verificarmos a Figura 9, percebemos que o método Bayes Search performou melhor na previsão iterativa. É importante destacar, que esse comportamento, pode estar relacionado com a questão da validação, ou seja, na previsão direta é apresentado todo o conjunto de teste para previsão dos dados com apenas um treinamento do modelo, o que não acontece na previsão iterativa, visto que o modelo é treinado e avaliado com diferentes tamanhos e valores de entradas em cada etapa de tempo (Lim & Zohren, 2021). Em seguida, os resultados do Quadro 6, referente aos dados da velocidade do vento, condiz com o apresentado na Figura 10, que contemplou o método Bayes Search como aquele que apresentou o melhor ajuste nos dois tipos de previsão.

Baseado nessas condições, o modelo SVR eleito como o ideal será aquele proveniente da otimização bayesiana, visto que precisamos considerar o processo de validação exposto anteriormente. O Quadro 7 apresenta um comparativo entre os melhores resultados dos modelos SVR e ARIMA a partir da análise dos erros nas etapas 1, 3, 8 e 12.

Quadro 7 - Análise comparativa do erro RMSE na previsão iterativa nos passos 1, 3, 8 e 12.

	ARIMA		SVM	
Passos	Vento	PLD	Vento	PLD
1	0.421	8.966	0.463	23.306
3	0.902	5.824	1.208	22.460
8	0.550	27.170	0.061	226.877
12	1.290	54.805	0.061	146.293
Erro Geral	1.092	69.516	0.457	99.101

Fonte: Autores.

Consoante o Quadro 7, identificamos que o modelo ARIMA errou menos em algumas etapas de tempo em comparação ao SVM, entretanto, na análise final dos resultados da previsão para todos os 12 passos de tempo, o modelo SVM apresentou um desempenho melhor na análise do RMSE nos dois conjuntos de dados, o que nos leva a considerar que para um processo de previsão a longo prazo, o SVM se torna uma alternativa mais viável.

É importante salientar a presença de outliers na base de dados sobre o PLD, existentes entre os períodos de 2001 a 2002 e no início de 2008, o que contribuiu com o aumento dos erros de previsão, diferente dos dados sobre a velocidade do vento que apresenta valores mais lineares. Um aspecto importante a ser destacado, é a influência que o MSE sofreu desses valores extremos nos resultados sobre o PLD observado nos Quadros 3 e 5, visto que ele eleva a média dos erros ao quadrado (Fontana, 2021). Em contrapartida, o RMSE consegue minimizar bem mais os erros, mesmo essa métrica não sendo a mais adequada para trabalhar com outliers, no entanto, o RMSE ainda é uma métrica de erro padrão para muitos modelos de séries temporais (Kirchoff et al., 2019).

4. Considerações Finais

Foi apresentado a aplicação dos modelos ARIMA e SVM para previsão de séries temporais do sistema elétrico brasileiro, com base nos dados referentes ao Preço de Liquidação das Diferenças da região Norte (PLD) e a velocidade do vento em aerogeradores da cidade de Macau.

Os dados foram tratados, analisados e processados para a estimação dos parâmetros com os métodos Auto ARIMA e Grid Search, para o modelo ARIMA, e seguidamente os métodos Bayes Search e Random Search para o modelo SVM, em que se constatou que os métodos Grid Search e Bayes Search desempenharam os melhores ajustes nos respectivos modelos. Por conseguinte, foi construído o processo de previsão em etapas de tempo, em que foi feita a extrapolação da base amostral em ambos os modelos, e utilizada a técnica de validação Walk Forward, em suas abordagens Expanding Window e Sliding Window para o modelo ARIMA, e a previsão iterativa, que necessitou da utilização de técnicas de Feature Engineering para o modelo SVM, que nesse caso foi o recurso de atraso.

Contudo, vale ressaltar, que foi constatado uma certa influência dos outliers na análise dos erros para a base de dados sobre o PLD, o que dificultou o diagnóstico de desempenho do modelo para os dados em questão com métricas de erro como MSE, que se mostrou bastante sensível a valores discrepantes, mesmo assim, não impossibilitou análise comparativa entre o modelo estatístico e o de machine learning.

Em síntese, o modelo ARIMA apresentou mais consistência na previsão em etapas de tempo, entretanto, sua modelagem requer um domínio maior sobre as particularidades encontradas nos dados, dessa forma, determinados aspectos podem comprometer na análise final do modelo. Por outro lado, o SVM expôs erros menores na análise geral dos resultados, se tornando uma alternativa bem mais flexível na análise de séries temporais, visto que sua modelagem está muito centrada na otimização

dos seus hiperparâmetros e na construção de recursos de treinamento do modelo, o que nesse caso, se ajustou bem aos dados apresentados.

No mais, os resultados encontrados mostraram-se satisfatórios no estudo de séries temporais do setor elétrico brasileiro, e inclusive, apresenta uma ferramenta útil na análise de tendências do mercado livre de energia, o que pode se tornar um caminho viável no auxílio à tomada de decisões e ajudar a prevenir perdas, principalmente no mercado de prazo, onde os preços de energia elétrica sofrem constantes oscilações por conta da oferta e demanda, além das questões climáticas que afetam diretamente a produção energética. Dessa forma, independente do modelo utilizado, sendo os mais clássicos, como os provenientes da estatística, ou os mais modernos, como os de machine learning, ramificado da inteligência artificial, que dependem muito do poder computacional empregado, desempenham bem suas funções dentro de suas limitações, ou seja, cada abordagem pode oferecer melhores resultados dependendo da regra de negócio, e não necessariamente um modelo será mais adequado que outro, mas cabe avaliar qual abordagem pode ser utilizada naquele momento, visto que essa decisão está intrinsecamente relacionada ao problema que será trabalhado.

Ademais, como sugestão para trabalhos futuros, recomenda-se a atualização das bases de dados, visto que as séries históricas utilizadas avaliam apenas períodos mais antigos, o que pode influenciar nos resultados preditivos, além disso, ainda existem possibilidades de melhorias a serem exploradas na modelagem do SVM, tais como, abordar outras técnicas de Features Engineering, como construir novas variáveis de entrada por recursos de tempo e estatísticos, e por fim, é válido experimentar outros métodos de otimização de hiperparâmetros que não foram abordados no trabalho, o que pode possibilitar a checagem de mais parâmetro em intervalos maiores.

Referências

Acosta, S. M. & Amoroso, A. L. (2021). Aplicação da regressão por vetores de relevância na modelagem de um processo produtivo. *engenharia de produção: planejamento e controle da produção em foco-volume 1*, *1*(1), 37-52.

Alves, P. F., De Negri, J. A., & Cavalcante, E. J. (2022). Utilizando aprendizado de máquina para estimação do spread das instituições financeiras nos empréstimos do BNDES.

Assunção, A., de Mattos Neto, P. S., & Vasconcelos, E. (2022). Um Sistema Baseado Em Combinação de Modelos para Previsão de Velocidade do Vento. *Revista de Engenharia e Pesquisa Aplicada*, *7*(2), 1-11.

Awan, T. M., & Aslam, F. (2020). Prediction of daily COVID-19 cases in European countries using automatic ARIMA model. *Journal of public health research*, *9*(3), jphr-2020.

Barbosa, R. B., Ferreira, R. T., & Silva, T. M. D. (2020). Previsão de variáveis macroeconômicas brasileiras usando modelos de séries temporais de alta dimensão. *Estudos Econômicos (São Paulo)*, *50*, 67-98.

Borsato, R., & Corso, L. L. (2019). Aplicação de Inteligência Artificial e ARIMA na Previsão de Demanda no setor metal mecânico. *Scientia cum Industria*, *7*(2), 165-176.

Brownlee, J. (2017). *Introduction to time series forecasting with python: how to prepare data and develop models to predict the future*. Machine Learning Mastery.

Conceição, R. M., Santos, S. R., do Nascimento, F. B., dos Santos, W. J. C., & Conte, T. N. M. (2021). Método de aprendizagem supervisionada para a identificação de rastros de cyberbullying. *International Association for Development of the Information Society*. IADES.

Conte, T. N. M. de S., Conte, B. N. M. de S., & Oliveira, R. C. L. (2021). Aplicação Híbrida com Redes Neurais Profundas e Algoritmo Genético para Previsão de Séries Temporais do Sistema de Energia Elétrica Brasileira. *Anais Do 15. Congresso Brasileiro de Inteligência Computacional*. https://doi.org/10.21528/cbic2021-104.

da Silva, F. C. C. (2021). *Gestão de dados científicos*. Interciência.

de Paula, D. M., Júnior, J. C. X., & Miranda, K. F. (2020). Aplicação de Séries Temporais para Previsão de Despesas de Energia Elétrica do Tribunal Regional Eleitoral do Rio Grande do Norte. *Brazilian Journal of Development*, *6*(11), 87089-87112.

Filho, F. L. da S. (2022). Aplicação do modelo de séries temporais para previsão do número de passageiros de uma companhia aérea. https://doi.org/10.31235/osf.io/gmyaj.

Fontana, M. (2021). Modelo de predição de dados baseado em redes neurais recorrentes integrado com historiador industrial. *Revista de Engenharia e Tecnologia*, *13*(4).

Géron, A. (2019). *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. Alta Books.

Isotani, S., & Bittencourt, I. I. (2015). *Dados abertos conectados: em busca da web do conhecimento*. Novatec Editora.

Iszczuk, A. C. D., Ventris, K. F. D., Pinto, G. B., Shirabayashi, J. V., dos Santos, M. A. R., de Souza, R. C. T., & Dal Molin Filho, R. G. (2021). Evoluções das tecnologias da indústria 4.0: dificuldades e oportunidades para as micro e pequenas empresas. *Brazilian Journal of Development*, *7*(5), 50614-50637.

Kirchoff, D. F. (2019). Avaliação de técnicas de aprendizado de máquina para previsão de cargas de trabalho aplicadas para otimizar o provisionamento de recursos em nuvens computacionais.

Lagasse, W. (2020). Previsão do comportamento do preço de liquidação das diferenças (PLD) com ferramentas estatísticas.

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200209.

Macêdo, A. C. C. D. (2022). *Comparando modelos clássicos de séries temporais e aprendizagem de máquina para previsão de demanda na indústria de bebidas* (Bachelor's thesis).

Nascimento, R. de A. (2022). Estudo de métodos de previsão de séries temporais aplicados ao preço da energia elétrica no mercado de curto prazo brasileiro. *Repositorio.ufmg.br*. http://hdl.handle.net/1843/44150.

Neves, J. M. M. (2020). *Otimização de hiperparâmetros em machine learning utilizando uma surrogate e algoritmos evolutivos* (Bachelor's thesis, Universidade Tecnológica Federal do Paraná).

Paula, J. de S., Teixeira, L. L., Rodrigues, S. B., Hickmann, T., Correa, J. M., & Ribeiro, L. da S. (2022). Aplicação de técnicas de aprendizado de máquina e estatística na previsão da demanda de biocombustíveis. *Revista de Gestão E Secretariado*, *13*(4), 2559–2572. https://doi.org/10.7769/gesec.v13i4.1488.

Sakurai, R., & Zuchi, J. D. (2018). As revoluções industriais até a indústria 4.0. *Revista Interface Tecnológica*, *15*(2), 480-491.

Silva, J. S. S. (2020). Modelo de previsão de bolsas de sangue baseado em aprendizado de máquina.

Wang, C., Baratchi, M., Bäck, T., Hoos, H. H., Limmer, S., & Olhofer, M. (2022). Towards Time-Series Feature Engineering in Automated Machine Learning for Multi-Step-Ahead Forecasting. *Engineering Proceedings*, *18*(1), 17.