I-912 - MODELOS DE APRENDIZAGEM DE MÁQUINA PARA PREVISÃO DA DEMANDA DE ÁGUA DA REGIÃO METROPOLITANA DE SALVADOR, BAHIA

Edmilson dos Santos de Jesus[(1)]
Graduado em Análise de Sistemas pela UNEB (2003), MBA em Gestão da Informação pela UNIFACS(), Mestre em Computação pela UFBA (2023). Analista de Saneamento da EMBASA, atuando como Analista de BI, líder de projetos e doutorando do Programa de Pós-Graduação em Computação, do Instituto de Computação da UFBA (PGCOMP/IC/UFBA:2024-2027).

Endereço[(1)]**:** 4ª Avenida, 420, Centro Administrativo da Bahia - CAB, 41745-002, Salvador, Bahia, Brasil - Tel: (71) 3372-4213 - e-mail: edmilson.jesus@embasa.ba.gov.br

Gecynalda Oliveira Gomes Silva
Possui graduação em Estatística pela Universidade Federal do Ceará (1999), mestrado em Estatística pela Universidade Federal de Pernambuco - UFPE (2005) e doutorado em Ciência da Computação pela UFPE (2010). Professora Associada III e Coordenador do Colegiado do Curso de Estatística da UFBA.

Endereço[(1)]**:** Av. Milton Santos, s/nº - Ondina, Salvador - BA, 40170-110, Salvador, Bahia, Brasil – Telefone – (71) 3283-5750 - E-mail: gecynaldassg@ufba.br

RESUMO

O A água é essencial para a vida humana e prever sua demanda é um grande desafio, o objetivo deste trabalho foi propor um novo modelo híbrido de inteligência artificial, para previsão de demanda de água, através da decomposição das séries temporais de 10 reservatórios que abastecem a Região Metropolitana de Salvador (RMS). Utilizando dados de vazão dos reservatórios, obtidos junto à Empresa Baiana de Águas e Saneamento (EMBASA), e dados meteorológicos, do site do Instituto Nacional de Meteorologia do Brasil (INMET). Os resultados demonstraram a viabilidade do uso do modelo proposto, comparado a outros modelos tradicionais como o Multilayer Perceptron (MLP), Support Vector Regression (SVR), Short Long Term Memory (LSTM) e Autoregressive and Integrated Moving Average (ARIMA).

PALAVRAS-CHAVE: séries temporais, inteligência artificial, demanda de água, redes neurais, reservatórios de distribuição.

INTRODUÇÃO

A taxa de crescimento populacional global registra um acréscimo aproximado de 80 milhões de indivíduos anualmente, equivalente a mais de 200 mil pessoas diariamente, reforçando a imperatividade de uma oferta crescente de água destinada ao consumo humano (RIPPLE et al., 2019). O fenômeno do aquecimento global e a expansão demográfica global (DINIZ, 2019) resultam no aumento exponencial da demanda por água potável. Nesse contexto, a gestão eficaz dos serviços públicos de abastecimento de água assume papel crucial para assegurar o suprimento ininterrupto de água potável à população. Aprimorar a precisão na previsão da demanda hídrica torna-se cada vez mais essencial, e a aplicação de modelos de Aprendizado de Máquina, como as Redes Neurais Artificiais (ANN, do inglês Artificial Neural Networks) e suas variantes, emerge como contribuição decisiva nesse processo, permitindo predições mais acuradas de cenários futuros de demanda.

A Região Metropolitana de Salvador (RMS) abrange 13 municípios, conforme indicado pelo Instituto de Pesquisa Econômica Aplicada (IPEA, 2015). Dentre esses, seis municípios, a saber: Candeias, Lauro de Freitas, Madre de Deus, Salvador, São Francisco do Conde e Simões Filho, são exclusivamente atendidos pelo Sistema Integrado de Abastecimento da RMS. A gestão operacional desse sistema é atribuída à Empresa Baiana de Águas e Saneamento S.A. (EMBASA), compreendendo 16 reservatórios, cinco captações e quatro estações de tratamento de água. Esse sistema complexo desempenha um papel crucial no fornecimento de

água potável para uma população que ultrapassa os 3 milhões de habitantes, correspondendo a 85% da população total da RMS, conforme estimativas do Instituto Brasileiro de Geografia e Estatística (IBGE, 2021).

Embora o sistema integrado de abastecimento da RMS inclua 16 reservatórios, apenas 11 recebem água diretamente das estações de tratamento. Os cinco reservatórios restantes são abastecidos de forma indireta pelos primeiros. Este estudo teve como objetivo prever a vazão de água para os 10 reservatórios que recebem água diretamente das estações de tratamento, uma vez que é a partir destes que a água potável é distribuída para zonas, bairros e reservatórios intermediários, garantindo o abastecimento contínuo de água para toda a população. O décimo primeiro reservatório, R23A, não foi objeto de estudo devido à falta de dados suficientes para previsão.

OBJETIVOS

O propósito deste estudo foi examinar a viabilidade de um novo modelo híbrido computacional que emprega o modelo híbrido computacional de Máquinas de Vetores de Suporte para Regressão com Redes Neurais Artificiais (SVR-ANN) para a previsão da demanda hídrica dos 10 reservatórios principais que fornecem água potável à Região Metropolitana de Salvador (RMS). Essa abordagem foi comparada a outros modelos tradicionais de aprendizado de máquina, visando realizar previsões mais precisas, analisando sua efetividade por meio de estatísticas de erro.

METODOLOGIA

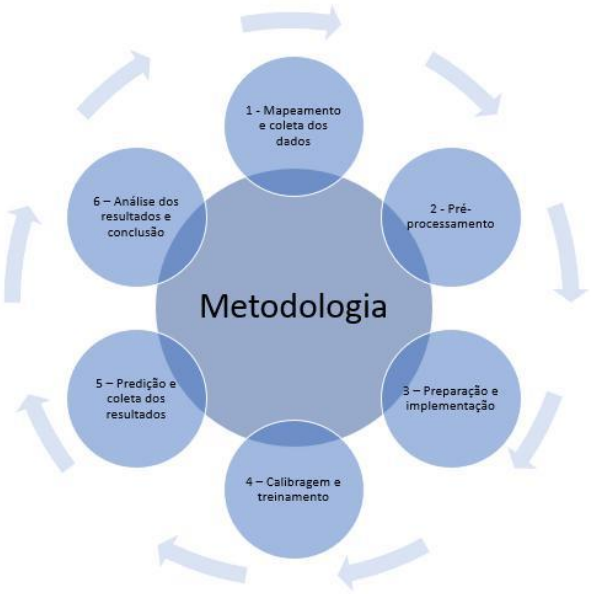

Figura 1: Fluxo proposto na metodologia de trabalho

A metodologia de pesquisa adotada neste trabalho foi organizada em etapas que englobaram: o mapeamento e coleta de dados, pré-processamento, preparação e implementação, calibração e treinamento, predição e recuperação de resultados, seguidos pela análise dos resultados e conclusão. A figura 1 ilustra o fluxo das ações delineadas na metodologia e aplicadas para a execução das atividades de pesquisa, alinhadas com os objetivos estabelecidos no estudo.

Durante a fase de coleta, foram obtidos e carregados dados históricos meteorológicos e de consumo dos reservatórios de água potável que fornecem abastecimento diário aos bairros e localidades da região metropolitana, cobrindo o período de janeiro de 2017 a fevereiro de 2022, fornecidos pela EMBASA. A partir desses dados, foi possível analisar o padrão de consumo de água ao longo do tempo nas regiões abastecidas, identificando sazonalidade e distribuição da demanda. No processo de pré-processamento, ferramentas como o

SAP© Data Services Designer foram utilizadas para a cópia e transformação dos dados extraídos do ambiente de produção da Plataforma de Informações das Plantas de Processos da EMBASA (PIPPE), enquanto a linguagem R foi empregada para o tratamento de dados ausentes e inválidos, resultando nas informações da figura 2 .

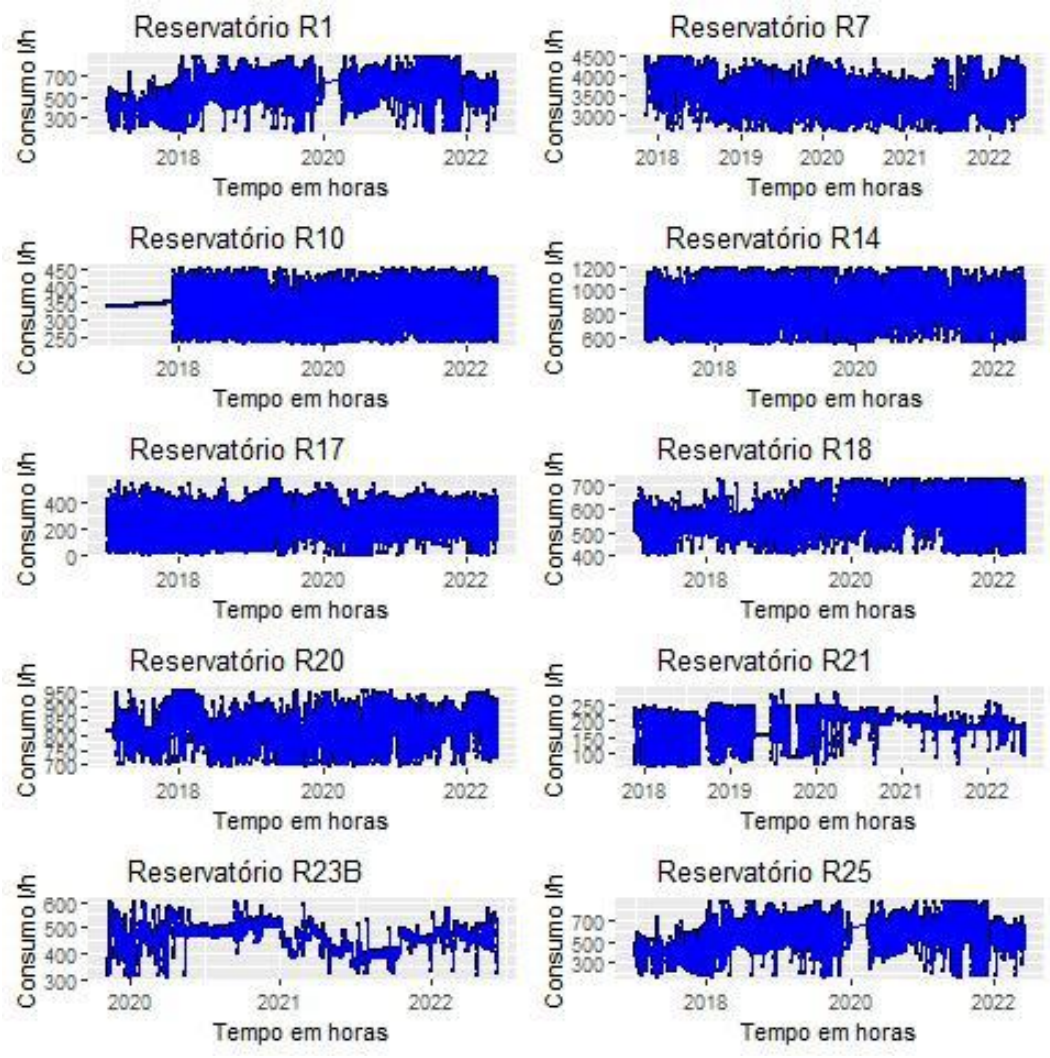

Figura 2 - Histórico de vazão dos reservatórios R1, R7, R10, R14, R17, R18, R20, R21, R23B e R25 após o pré-processamento.

Na fase de preparação e implementação, os modelos de Aprendizado de Máquina foram desenvolvidos utilizando as linguagens Python e R, empregando técnicas de otimização para maximizar a eficiência de cada modelo. Em seguida, durante a etapa de predição e recuperação de resultados, as previsões foram realizadas, os resultados foram coletados e as métricas de erro foram calculadas. Por fim, a análise dos resultados foi conduzida, culminando na seleção do modelo mais adequado para cada reservatório.

RESULTADOS E DISCUSSÕES

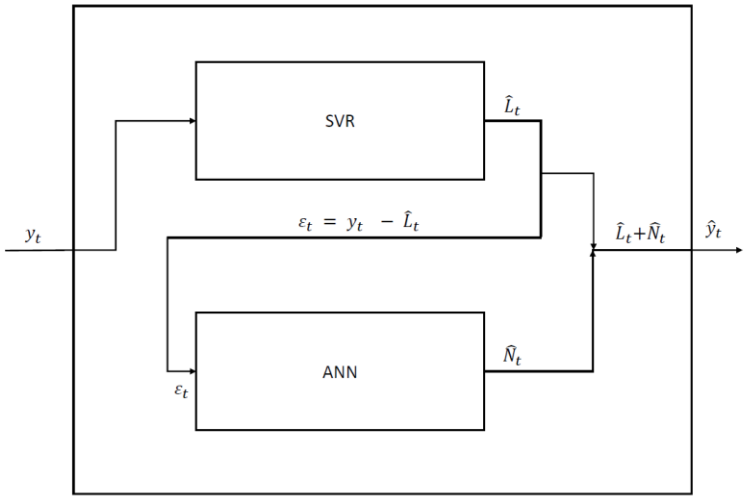

Figura 3 - Arquitetura do modelo SVR-ANN.

O modelo SVR-ANN foi implementado seguindo a metodologia proposta por Zhang (2003) para o modelo aditivo, representado pela Equação (2), onde **Lt** denota a componente linear da série temporal, e **Nt** denota a componente não linear.

A arquitetura da SVR-ANN é apresentada na figura 3. O primeiro passo foi prever a parte linear da série temporal utilizando o modelo SVR, tendo a série **yt** como entrada. A série temporal resultante foi denominada **Lt** e, posteriormente, subtraída da série real **yt** para extrair a componente não linear **εt**, também conhecida como ruído, conforme representado pela Equação (1).

A segunda etapa foi analisar a modelagem da componente não linear, usando a série **εt** como entrada para o modelo da ANN. O resultado da predição realizada pela ANN foi denominado **Nt** e foi utilizado para obter o resultado combinado da predição final, na Equação (2).

$$\varepsilon t = yt - Lt \qquad \textbf{equação (1)}$$

$$yt = Lt + Nt \qquad \textbf{equação (2)}$$

Para análise comparativa sobre o desempenho dos modelos implementados neste trabalho foram utilizadas como referência as métricas de erro MAPE, MAE, MSE e RMSE. A apresentação e análise dos resultados são apresentados a seguir.

Tabela 1 - Menores estatísticas MAPE na previsão horária de cada reservatório (RSVR).

RSVR	SVR-LSTM	SVR-MLP	SVR	LSTM	ARIMA1	ARIMA2	MLP
R1	3,54	3,32	28,94	2,31	6,69	6,64	3,25
R25	10,13	11,96	15,70	4,93	11,88	11,85	9,02
R14	7,17	5,82	9,75	4,85	17,89	20,87	6,03
R17	34,41	34,32	8,00	84,53	38,02	52,47	34,69
R18	6,09	4,70	16,46	3,95	6,51	6,45	5,04
R21	2,27	1,33	3,28	1,20	7,02	3,34	1,35
R23B	2,04	0,95	4,39	0,94	1,05	8,52	0,95
R20	1,82	1,54	6,86	1,53	1,63	3,24	1,51
R10	9,87	8,09	6,33	7,34	12,61	11,71	8,01
R7	5,91	4,33	78,10	3,34	7,44	7,56	3,82

Fonte: Próprio autor.

Dentre os modelos tradicionais implementados, o LSTM foi um dos que apresentou melhor desempenho na previsão horária, enquanto os modelos ARIMA se destacaram nas previsões diárias e semanais, porém, não houve unanimidade em relação a um modelo específico, o que sugere que são as características individuais de cada série temporal que determinam o melhor modelo para sua previsão. Dentre os modelos escolhidos para previsão (diária, horária e semanal) de cada reservatório, o modelo LSTM foi o que apresentou a melhor estatística MAPE, com 0,9% para o reservatório R23B na previsão horária, e o modelo SVR foi o que registrou a pior estatística com 22,1% na previsão semanal também para o reservatório R23B.

A figura 4 mostra o gráfico *boxplot* dos resultados MAPE alcançados por cada modelo, na previsão horária. Onde é possível observar que os maiores outliers foram registrados nos modelos LSTM, SVR e ARIMA2, respectivamente. Enquanto os modelos MLP, SVR-MLP e SVR-LSTM apresentaram a menor variabilidade. Embora a implementação do MLP seja mais simples em comparação com os modelos híbridos, o MLP sozinho pode não ser adequado para aplicação em séries temporais com comportamento dependente, ao contrário do modelo híbrido onde o MLP é aplicado aos resíduos que possuem um comportamento mais aleatório.

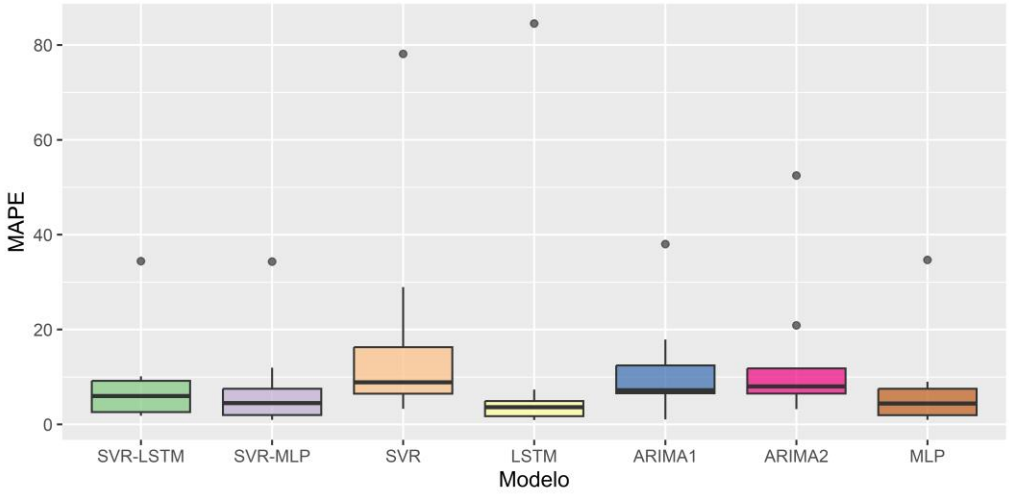

Figura 4 - Gráfico *boxplot* das estatísticas MAPE por modelo implementado.

CONCLUSÃO

Foram realizadas as implementações dos modelos tradicionais MLP, SVR e ARIMA, que alcançaram resultados satisfatórios na previsão de demanda de água, medidos através da métrica estatística conhecida como média de erro percentual absoluto (MAPE), com valores menores que 3%. Depois disso, foram introduzidos modelos mais complexos de aprendizado de máquina, como os modelos SVR-LSTM e SVR-MLP. Esses modelos demonstraram estatísticas MAPE também competitivas na previsão horária das séries temporais dos reservatórios R21, R23B e R20, alcançando percentuais de 2,27%, 2,04% e 1,82% para o SVR-LSTM, e 1,33%, 0,95% e 1,54% para o SVR-MLP, respectivamente. Esses resultados evidenciam que o modelo híbrido SVR-ANN, seja SVR-LSTM ou SVR-MLP, seguindo a metodologia proposta por Zhang (2003), pode ser uma solução viável para a predição de séries temporais relacionadas à demanda de água.

Todos os códigos-fonte implementados para a construção, teste e validação dos modelos, bem como as tabelas de resultados, estão disponíveis no repositório GitHub (https://github.com/edmilsondejesus/waterdemand) para possibilitar a verificação e replicação da pesquisa. Como direção para futuras pesquisas, recomenda-se a incorporação das variáveis tipo de dia e estação do ano para acentuar o impacto da sazonalidade nas previsões. Sugere-se também a exploração da dependência georreferenciada entre os reservatórios, a experimentação do modelo SVR com outras configurações de kernel, como *sigmoid* e *precomputed*, e a implementação de outros

modelos híbridos. Além disso, a inclusão de atrasos para variáveis meteorológicas também pode ser explorada como uma alternativa de desenvolvimento futuro.

AGRADECIMENTOS

Agradeço ao Instituto de Computação da UFBA, EMBASA, Universidade Corporativa – UCE e a Unidade de Suporte Operacional - MSSO, pelo apoio no desenvolvimento da pesquisa, pelas informações fornecidas, paciência e contribuições imprescindíveis para o sucesso da pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

1. BOX, G.; JENKINS, G. Time Series Analysis, Forecasting and Control. 1970.
2. DINIZ, José Eustáquio. Dois mil anos de crescimento demoeconômico global. EcoDebate. 2019. ISSN 2446-9394. Disponível em < https://www.ecodebate.com.br/2019/06/03/dois-mil-anos-de-crescimento-demoeconomico-global-artigo-de-jose-eustaquio-diniz-alves/>. Acessado em 18 mai. 2024.
3. Instituto Brasileiro de Geografia e Estatística (IBGE). Estimativas da população residente no Brasil e unidades da federação com data de referência em 1º de julho de 2021. 2021. Disponível em: <https://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2021/estimativa_dou_2021.pdf>. Acessado em 04 ago. 2022.
4. Instituto de Pesquisa Econômica Aplicada (IPEA). Governança metropolitana no brasil - relatório de pesquisa. 2015. Disponível em: <https://www.ipea.gov.br/redeipea/images/pdfs/governanca_metropolitana/relatorio_1.1_revisao_final_salvador.pdf>. Acessado em 03 ago. 2022.
5. JESUS, E. d. Santos de; GOMES, G. S. d. S. Machine learning models for forecasting water demand for the metropolitan region of salvador, Bahia. Neural Computing and Applications, v. 35, p.19669–19683, 2023. ISSN 1433-3058. Disponível em < http://dx.doi.org/10.1007/s00521-023-08842-0>. . Acessado em 22 mar. 2023.
6. REES, P.; CLARK, S.; NAWAZ, R. Household forecasts for the planning of long-term do-mestic water demand: Application to london and the thames valley. Wiley Online Library, 2020. Disponível em: <https://doi-org.ez10.periodicos.capes.gov.br/10.1002/psp.2288 >. . Acessado em 23 mar. 2022.
7. RIPPLE, W. J. et al. World Scientists' Warning of a Climate Emergency. BioScience, v. 70, n. 1, p. 8–12, 11 2019. ISSN 0006-3568. Disponível em: <https://doi.org/10.1093/biosci/biz088>. . Acessado em 16 abr. 2023.
8. ZHANG, G. Time series forecasting using a hybrid arima and neural network model. Neurocomputing, v. 50, p. 159–175, 2003. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231201007020>. Acessado em 23 mar. 2023.