Revista Brasileira de Climatologia
Brazilian Journal of Climatology
OPEN ACCESS
LEGD Editora

PREENCHIMENTO DE FALHAS EM SÉRIES TEMPORAIS DA TEMPERATURA DO AR: UMA COMPARAÇÃO ENTRE MODELOS DE MACHINE LEARNING

Gap filling in air temperature time series: a comparison between machine learning models

Imputación de fallos en series temporales de temperatura del aire: una comparación entre modelos de machine learning

Anisio Alfredo da Silva Junior
Instituto Federal de São Paulo
anisio.silva@ifsp.edu.br

Raphael de Souza Rosa Gomes
Universidade Federal do Mato Grosso
raphael@ic.ufmt.br

Carlo Ralph De Musis
Universidade de Cuiabá
carlo.demusis@gmail.com

Jonathan Willian Zangeski Novais
Instituto Federal de Mato Grosso
jonathan.zangeski@ifmt.edu.br

Daniela Maionchi
Universidade Federal do Mato Grosso
dmaionchi@fisica.ufmt.br

Josiel Maimone de Figueiredo
Universidade Federal do Mato Grosso
josiel@ic.ufmt.br

Resumo: Neste estudo, foi conduzida uma análise comparativa de diferentes algoritmos de Aprendizado de Máquina (ML) para o preenchimento de falhas em dados de temperatura do ar de quatro localizações de estados brasileiros distintos. Seis algoritmos foram avaliados: regressão linear, regressão LASSO, rede elástica, k-vizinhos próximos, árvores de decisão (CART) e regressão de vetor de suporte (SVR). Os resultados, referentes a todas as localizações, mostram que o modelo Support Vector Regression (SVR) foi o mais promissor, com RMSE excepcionalmente baixos, variando entre 0,1712 °C e 0,2062 °C. Isso sugere que o SVR pode ser a melhor escolha para a previsão da temperatura do ar. Enquanto a Árvore de Decisão apresentou resultados sólidos, com RMSE variando entre 0,2198 °C e 0,3746 °C. Os modelos Elastic Net (EN) e LASSO tiveram desempenho inferior, com RMSE entre 1,6935 °C e 2,8555 °C. O modelo K-Nearest Neighbors (KNN) obteve resultados intermediários, com RMSE variando entre 0,5579 °C e 0,7567 °C. A Regressão Linear também apresentou resultados variáveis, com RMSE entre 0,7474 °C e 1,4010 °C.

Palavras-chave: Preenchimento de falhas. Aprendizado de Máquina. árvores de decisão. máquinas de vetores de suporte. SVR. CART. Rede Elaśtica. LASSO. KNN. Regressão Linear.

Abstract: This study conducted a comparative analysis of different machine learning (ML) algorithms for filling gaps in air temperature data from four different locations in Brazilian states. Six algorithms were evaluated: linear regression, LASSO regression, elastic net, k-nearest neighbors, decision trees (CART), and support vector regression (SVR). The results, covering all sites, indicate that the Support Vector Regression (SVR) model was the most promising, with exceptionally low RMSE ranging from 0.1712 °C to 0.2062 °C. This suggests that SVR may be the best choice for predicting air temperature. While Decision Trees showed robust results with RMSE ranging from 0.2198 °C to 0.3746 °C. The Elastic Net (EN) and LASSO models performed poorly, with RMSE ranging from 1.6935 °C to 2.8555 °C. The K-Nearest Neighbors (KNN) model produced intermediate results, with RMSE ranging from 0.5579 °C to 0.7567 °C. Linear Regression also showed variable results, with RMSE ranging from 0.7474 °C to 1.4010 °C.

Keywords: Imputing missing values. Machine Learning. Decision Trees. Support Vector Machines. SVR. CART. Elastic Net. LASSO. KNN. Linear Regression.

Resumen: En este estudio, se llevó a cabo un análisis comparativo de diferentes algoritmos de Aprendizaje Automático (ML) para la imputación de fallos en datos de temperatura del aire de cuatro ubicaciones en distintos estados de Brasil. Se evaluaron seis algoritmos: regresión lineal, regresión LASSO, red elástica, k vecinos más cercanos, árboles de decisión (CART) y regresión de vectores de soporte (SVR). Los resultados, referentes a todas las ubicaciones, muestran que el modelo Support Vector Regression (SVR) fue el más prometedor, con valores de RMSE excepcionalmente bajos, que oscilan entre 0,1712 °C y 0,2062 °C. Esto sugiere que SVR puede ser la mejor opción para la predicción de la temperatura del aire. Mientras tanto, el Árbol de Decisión presentó resultados sólidos, con RMSE que varían entre 0,2198 °C y 0,3746 °C. Los modelos Elastic Net (EN) y LASSO tuvieron un rendimiento inferior, con RMSE entre 1,6935 °C y 2,8555 °C. El modelo K-Nearest Neighbors (KNN) obtuvo resultados intermedios, con RMSE que varían entre 0,5579 °C y 0,7567 °C. La Regresión Lineal también presentó resultados variables, con RMSE entre 0,7474 °C y 1,4010 °C.

Palabras clave: Imputación de fallos. Aprendizaje Automático. árboles de decisión. máquinas de vectores de soporte. SVR. CART. Red Elástica. LASSO. KNN. Regresión Lineal.

Submetido em: 20/10/2023
Aceito para publicação em: 23/09/2024
Publicado em: 16/10/2024

1. INTRODUÇÃO

O estudo da temperatura do ar é uma área de pesquisa de grande relevância, devido à sua estreita relação com as mudanças climáticas e o aquecimento global, fenômenos que têm impactos significativos na vida humana e no meio ambiente (THOBER et al., 2018). O crescente interesse por informações meteorológicas precisas e de curto prazo tem impulsionado o desenvolvimento de novas abordagens e tecnologias para aprimorar a acurácia das previsões meteorológicas (WEN et al., 2020).

Nos últimos anos, o uso de métodos de Aprendizado de Máquina (ML) tem se destacado como uma abordagem promissora nesse contexto. Esses métodos, que englobam uma variedade de algoritmos e técnicas computacionais, têm demonstrado eficácia em diversas aplicações, desde a análise de dados até previsões complexas, frequentemente apresentando baixo custo em comparação com os métodos tradicionais (XU et al., 2021).

No entanto, para alcançar melhores precisões meteorológicas, é necessária uma monitoria climática eficaz, que exige a coleta e o armazenamento de dados micro meteorológicos, essenciais para o estudo das mudanças nas temperaturas do ar, precipitação, umidade e outros fatores climáticos. Contudo, a disponibilidade desses dados é muitas vezes comprometida por falhas e lacunas, prejudicando a precisão das previsões climáticas, levando a análises distorcidas e a tomadas de decisões inadequadas, o que afeta a eficiência das políticas públicas e das medidas de proteção ambiental.

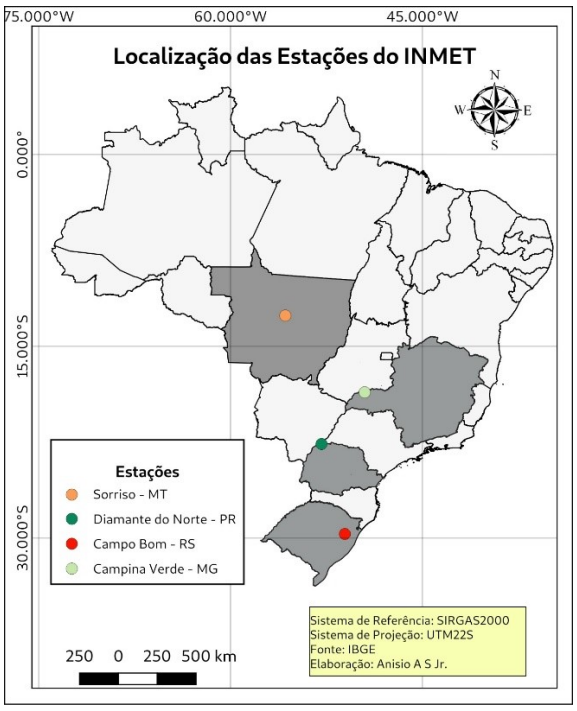
Diversos estudos têm abordado a utilização de ML no preenchimento de dados. Bonfante et al. (2013) e Katipoğlu e Reşat (2021) utilizaram redes neurais artificiais para preencher as lacunas nos dados de temperatura, obtendo resultados bem-sucedidos. Coulibaly e Evora (2007) aplicaram o Perceptron Multicamada (MLP, do inglês Multilayer Perceptron) para o preenchimento de dados de precipitação e temperaturas extremas. Kajewska-Szkudlarek e Stańczyk (2018) investigaram a eficácia do método de Regressão de Vetor de Suporte (SVR, do inglês Support Vector Regression), entre outros, para completar os dados diários ausentes de temperatura e umidade. Tosunoğlu et al. (2020) investigaram a eficiência do algoritmo k-vizinhos mais próximos (KNN, do inglês k-nearest neighbors), entre outros, para prever o fluxo hídrico mensal.

Neste trabalho, são apresentadas algumas abordagens para preencher lacunas em dados micro meteorológicos, visando impulsionar o progresso da área de Ciências Ambientais. Essas estratégias não apenas aprimoram a compreensão dos métodos empregados, mas também ampliam sua aplicabilidade para diversas localidades geográficas, enriquecendo o campo com novos conhecimentos e informações importantes.

2. DADOS E MÉTODOS

Neste estudo, utilizamos dados coletados de quatro estações meteorológicas automáticas (EMAs) administradas pelo Instituto Nacional de Meteorologia (INMET). Estes registros, disponíveis na base de dados oficial do INMET, originam-se das cidades de Campina Verde - MG (Estação A519), Sorriso - MT (Estação A904), Diamante do Norte - PR (Estação A849) e Campo Bom - RS (Estação A884). O período de coleta abrangeu de 03/12/2013 a 01/06/2023. A localização geográfica dessas estações está representada na Figura 1.

Figura 1 - Mapa das localizações



Fonte: Elaborado pelos autores (2023).

Nas estações automáticas, as leituras são realizadas a cada cinco segundos, e ao final de doze leituras consecutivas, a média é armazenada. No contexto deste estudo, adotou-se a

média horária. Os dados são compostos por variáveis microclimáticas que incluem radiação solar (KJ/m²), velocidade do vento (m/s), ponto de orvalho (°C), umidade relativa do ar (%), pressão atmosférica (hPa) e temperatura do ar (°C).

As estações foram selecionadas com o objetivo de assegurar diversidade nas séries temporais. Isso leva em consideração fatores climáticos que desempenham um papel significativo na influência sobre o comportamento das variáveis analisadas, característica fundamental para a avaliação do desempenho de cada modelo. Portanto, cada localidade possui seu próprio conjunto de dados, permitindo a análise individual de cada uma delas.

2.1. Métodos de Regressão

Os métodos de regressão desempenham um papel fundamental na modelagem de relações entre variáveis em ecossistemas, permitindo a previsão de variáveis dependentes com base em variáveis independentes. Esses métodos abrangem uma ampla gama de técnicas, cada uma com suas próprias características e aplicações específicas.

A Regressão Linear (LR) destaca-se como um dos métodos mais amplamente adotados, buscando estabelecer e quantificar a relação entre uma variável dependente contínua e um ou mais preditores independentes.

Quando se trata de seleção e regularização de variáveis, a Regressão LASSO surge como uma modalidade de regressão linear que visa prevenir o sobreajuste ao incorporar um termo de penalidade à função de custo do modelo. Essa abordagem permite identificar as variáveis mais relevantes e controlar a complexidade do modelo resultante.

A Rede Elástica (EN) apresenta-se como uma combinação das características da regressão Ridge e LASSO, oferecendo uma solução robusta para lidar com multicolinearidade e moderar a complexidade do modelo. Essa técnica é especialmente útil quando se trabalha com conjuntos de dados com alta dimensionalidade e correlações entre as variáveis.

O método dos k-vizinhos mais próximos (KNN) destaca-se por sua intuitividade, estimando o valor de uma variável alvo com base nos valores de observações vizinhas no espaço de atributos. Essa abordagem não paramétrica é capaz de capturar relações não lineares e é frequentemente utilizada em tarefas de classificação e regressão.

As Árvores de Classificação e Regressão (CART) oferecem uma abordagem baseada em árvores de decisão para prever uma resposta com base em entradas. Esses métodos adotam o erro quadrático médio como métrica para estabelecer divisões eficazes, permitindo a construção de modelos interpretáveis e de fácil visualização.

Por fim, a Regressão de Vetor de Suporte (SVR) apresenta-se como uma abordagem de aprendizado de máquina específica para tarefas regressivas, fundamentada na teoria das Máquinas de Vetores de Suporte (SVM). Essa técnica busca determinar uma função que se ajuste de forma otimizada aos dados de treinamento, levando em consideração a margem de erro e a complexidade do modelo.

2.2. Configuração dos modelos

Nos modelos EN e LASSO, foi utilizado um valor de $\lambda = 1$, realizando 1000 iterações (p) e com uma tolerância de 10^{-3} para o algoritmo de otimização. Essa configuração permite controlar a intensidade da penalização L1 aplicada aos coeficientes do modelo, garantindo que alguns coeficientes sejam reduzidos a zero e, assim, realizando a seleção de variáveis de forma automática. O número de iterações e a tolerância são ajustados para garantir que o algoritmo de otimização convirja eficientemente e forneça uma solução satisfatória.

No modelo KNN, foi aplicado o número de vizinhos igual a 5 associado à distância euclidiana.

Para o modelo CART utilizamos o MSE como critério para medição da impureza de um nó, com um número mínimo de amostras igual a 2 e número mínimo para estar em uma folha igual a 1.

2.3. Pré-processamento, reamostragem e avaliação

O pré-processamento dos dados envolveu inspeção e exclusão de linhas com dados ausentes ou inválidos e ajuste de dimensionalidade dos mesmos (Equação 2.0), uma vez que as variáveis independentes apresentam grandezas incongruentes (GARRETA; MONCECCHI, 2013).

$$z = \frac{X - u}{s}$$

Equação 1.0

onde: X representa a amostra de treinamento, u é a média das amostras e s o desvio padrão. Essa padronização dos valores não altera a distribuição inicial de X.

No experimento foram empregadas duas técnicas de validação, sendo elas a validação cruzada e a reamostragem dos dados. Na validação cruzada foi empregada a técnica *k-fold* com *k* = 10, na qual os dados são divididos em *k* grupos de tamanhos aproximadamente iguais. O modelo é treinado *k* vezes, sendo que em cada iteração, *k -1* grupos são utilizados como conjunto de treinamento, enquanto um grupo é usado como conjunto de teste. Isso permite que cada grupo seja utilizado como conjunto de teste uma vez, proporcionando uma avaliação mais abrangente do desempenho do modelo.

Na aleatoriedade na divisão dos *folds*, utilizou-se uma semente de aleatoriedade, a fim de garantir que a divisão seja reproduzível, ou seja, resulte sempre na mesma distribuição dos dados quando a mesma semente é utilizada. Ressalta-se que, nesse contexto, a aleatoriedade é baseada em números pseudoaleatórios (MATSUMOTO; NISHIMURA, 1998).

A reamostragem de dados constitui uma técnica importante no âmbito da análise de dados e modelagem estatística, proporcionando uma partição eficaz do espaço de recursos de maneira mais flexível e possibilitando a aprendizagem de interações de ordem superior entre os recursos (MORRIS; YANG, 2021). A Tabela 1 apresenta três configurações distintas, exemplificando a divisão dos dados entre as bases de treinamento e validação, fundamentais para o processo de modelagem.

Tabela 1- Reamostragem dos dados em bases de treinamento e validação com três diferentes configurações.

Reamostragem	Treinamento (%)	Validação (%)
A	80	20
B	60	40
C	40	60

Fonte: Elaborado pelos autores (2023).

A Tabela 1 apresenta três configurações distintas para a reamostragem dos dados em bases de treinamento e validação, visando uma análise mais abrangente do desempenho do modelo. Na configuração A, 80% dos dados são alocados para o treinamento, enquanto os 20% restantes são destinados à validação. A configuração B, por sua vez, utiliza 60% dos dados para treinamento e 40% para validação. Já a configuração C inverte essa proporção, com 40% dos dados sendo utilizados para treinamento e 60% para validação. Essa abordagem permitiu avaliar a capacidade de generalização dos modelos, ajustando-os de forma a evitar o sobreajuste e garantindo sua aplicabilidade em dados não vistos anteriormente.

Para a avaliação dos erros, adotou-se a raiz do erro quadrático médio (RMSE, do inglês *Root Mean Squared Error*) (Equação 2.1), o valor do viés (Equação 2.2) e o coeficiente de linearidade de Pearson (r).

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

Equação 2. 1

$$\mathrm{bias} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

Equação 2. 2

Onde y_i representa o valor real do alvo e $\hat{y}_i$ o valor previsto pelo modelo para a amostra i, e n o número total de amostras.

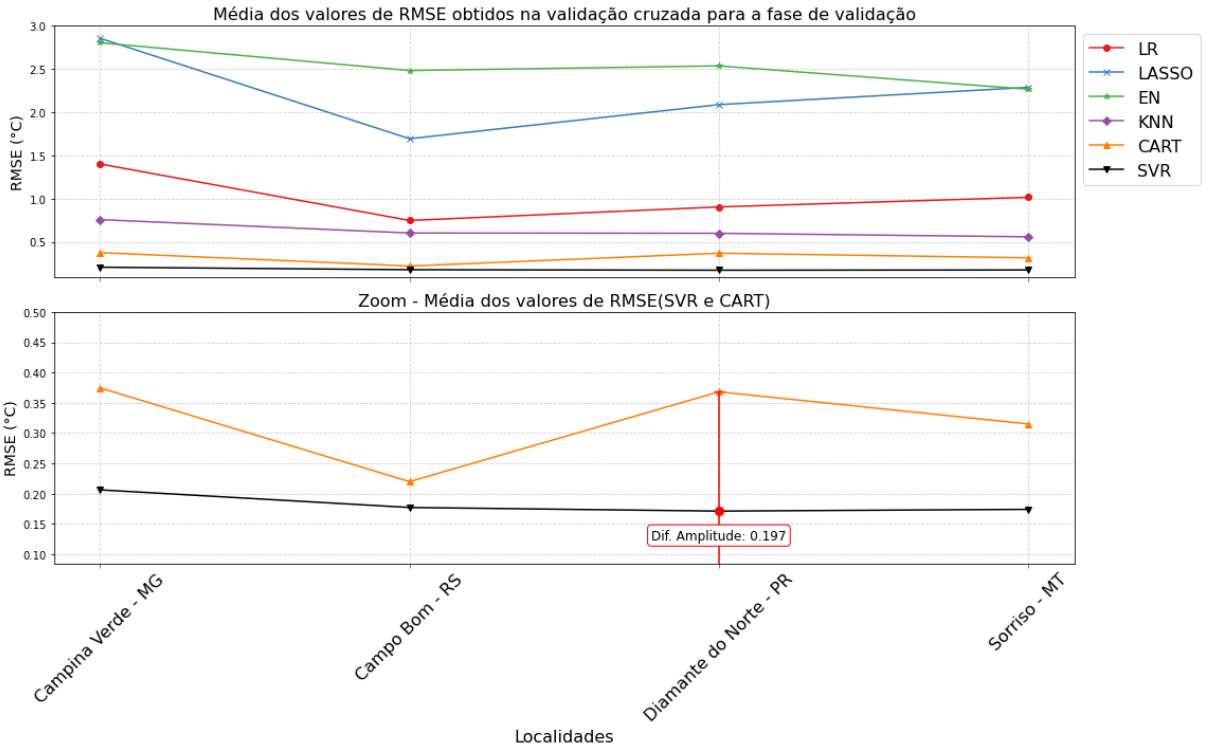
3. RESULTADOS E DISCUSSÃO

A Figura 2 ilustra as médias dos valores de RMSE (°C) obtidos durante a validação. Analisando os resultados, observa-se que:

- O modelo CART demonstrou ser bastante eficaz, com valores de RMSE oscilando entre 0,2198 °C e 0,3746 °C em todas as localizações. Isso sugere que a Árvore de Decisão pode ser uma alternativa apropriada para prever a temperatura do ar nas áreas estudadas.
- O EN teve um desempenho inferior, com valores que vão de 2,2662 °C a 2,8038 °C.
- O modelo LASSO mostrou uma performance variável, com valores de RMSE oscilando entre 1,6935 °C e 2,8555 °C.

- Os resultados obtidos com o KNN foram intermediários, evidenciados por valores de RMSE na faixa de 0,5579 °C a 0,7567 °C.
- Para a regressão linear resultou em valores entre 0,7474 °C e 1,4010 °C.
- Notavelmente, o SVR se destacou como o modelo com o melhor desempenho, exibindo valores extremamente baixos, que variam de 0,1712 °C a 0,2062 °C em todas as localidades.

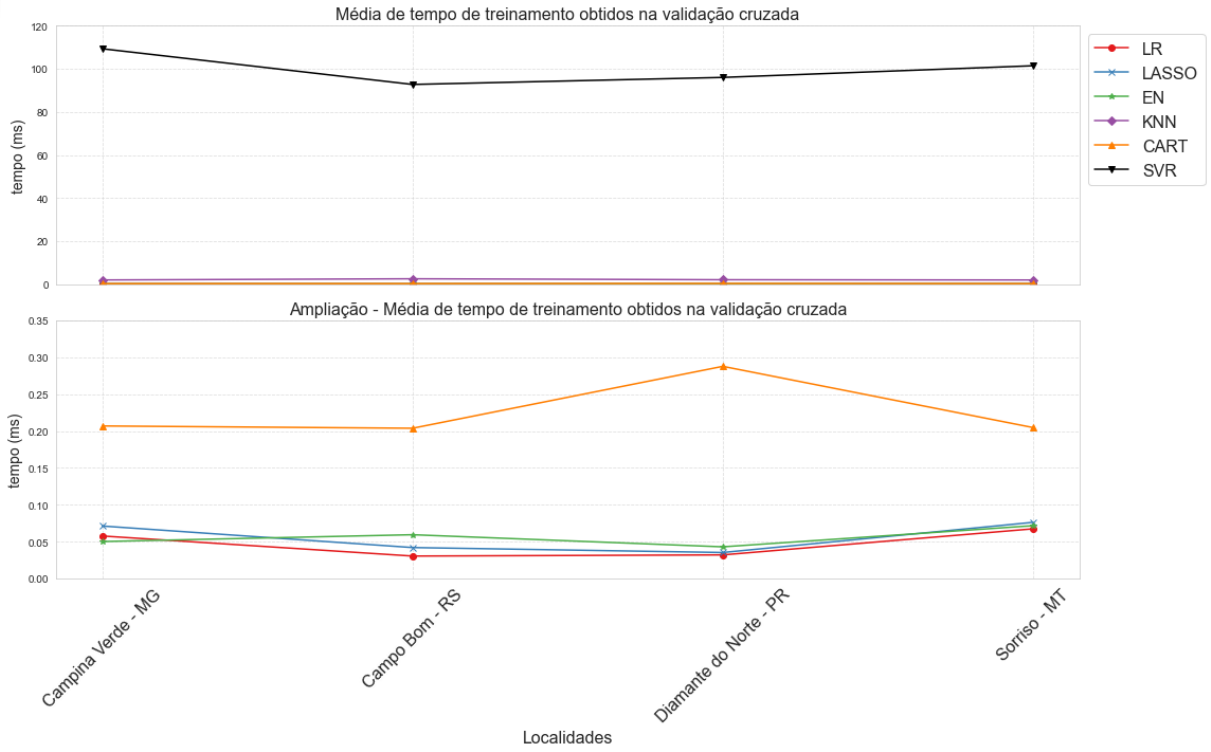
Figura 2 - Média dos valores de RMSE (°C) resultantes da validação cruzada



Fonte: Elaborado pelos autores (2023).

A Figura 3 expõe a média de tempo de treinamento obtida na validação cruzada, os modelos LR, LASSO, EN e KNN demonstraram um excelente desempenho, exigindo menos de 0,1 segundos (s) para serem treinados. O modelo CART apresentou valores inferiores a 0,3 segundos, o modelo SVR obteve valores inferiores a 120 segundos.

Figura 3 - Média do tempo (s) de execução na etapa de treinamento



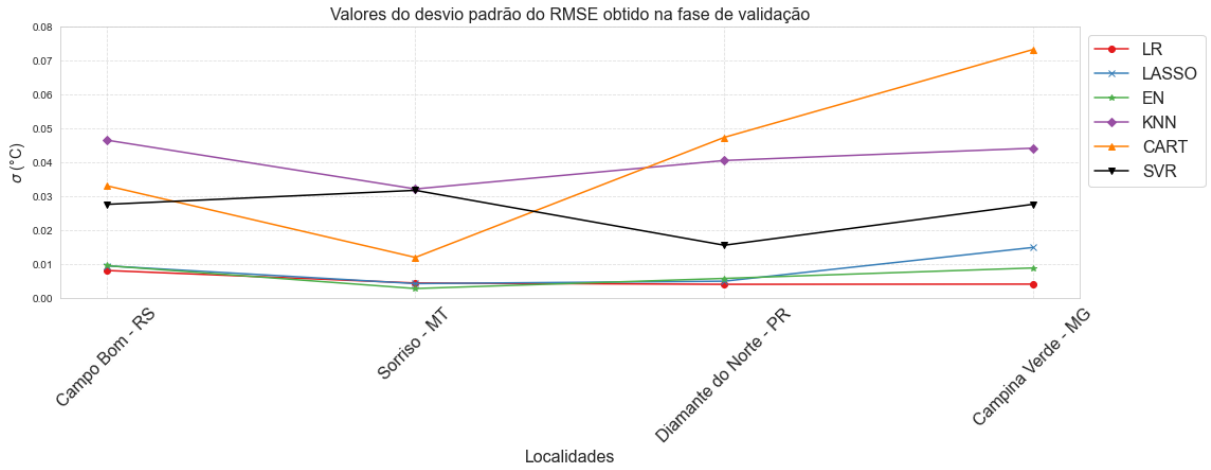
Fonte: Elaborado pelos autores (2023).

A análise dos resultados da reamostragem das configurações A, B e C proporcionou uma compreensão mais aprofundada acerca do desempenho dos modelos de aprendizado de máquina empregados neste estudo. De modo geral, variações nas proporções de dados de treinamento e teste possuem o potencial de influenciar a performance do modelo, uma vez que diferentes divisões podem conter padrões ou diversidades distintas nos dados. No entanto, os resultados obtidos nesta pesquisa indicam que, independentemente das configurações de reamostragem utilizadas, todos os modelos mantiveram performances de RMSE similares àquelas obtidas na etapa inicial de validação. Este comportamento sugere que os modelos foram treinados de forma a se adaptarem eficientemente às variações nos dados de reamostragem. Em contrapartida, no trabalho de Bonfante et al. (2013), observou-se um comportamento distinto, no qual o aumento da quantidade de falhas na série de dados impactou negativamente a eficiência do preenchimento.

A estabilidade desses modelos é evidenciada pelo desvio padrão observado nos valores de RMSE, que variaram entre 0,002 °C a 0,073 °C, em todas as configurações de

reamostragem, indicando que os modelos oferecem um nível de consistência na sua capacidade de predição.

Figura 4 - Desvio padrão da média de RMSE (°C)



Fonte: Elaborado pelos autores (2023).

A análise da Figura 5 revela *insights* importantes sobre o comportamento dos diferentes modelos de aprendizado de máquina quando comparados aos dados originais em um cenário de reamostragem C. Nesse contexto, com 40% dos dados destinados ao treinamento, houve uma distinção clara nos níveis de desempenho dos modelos.

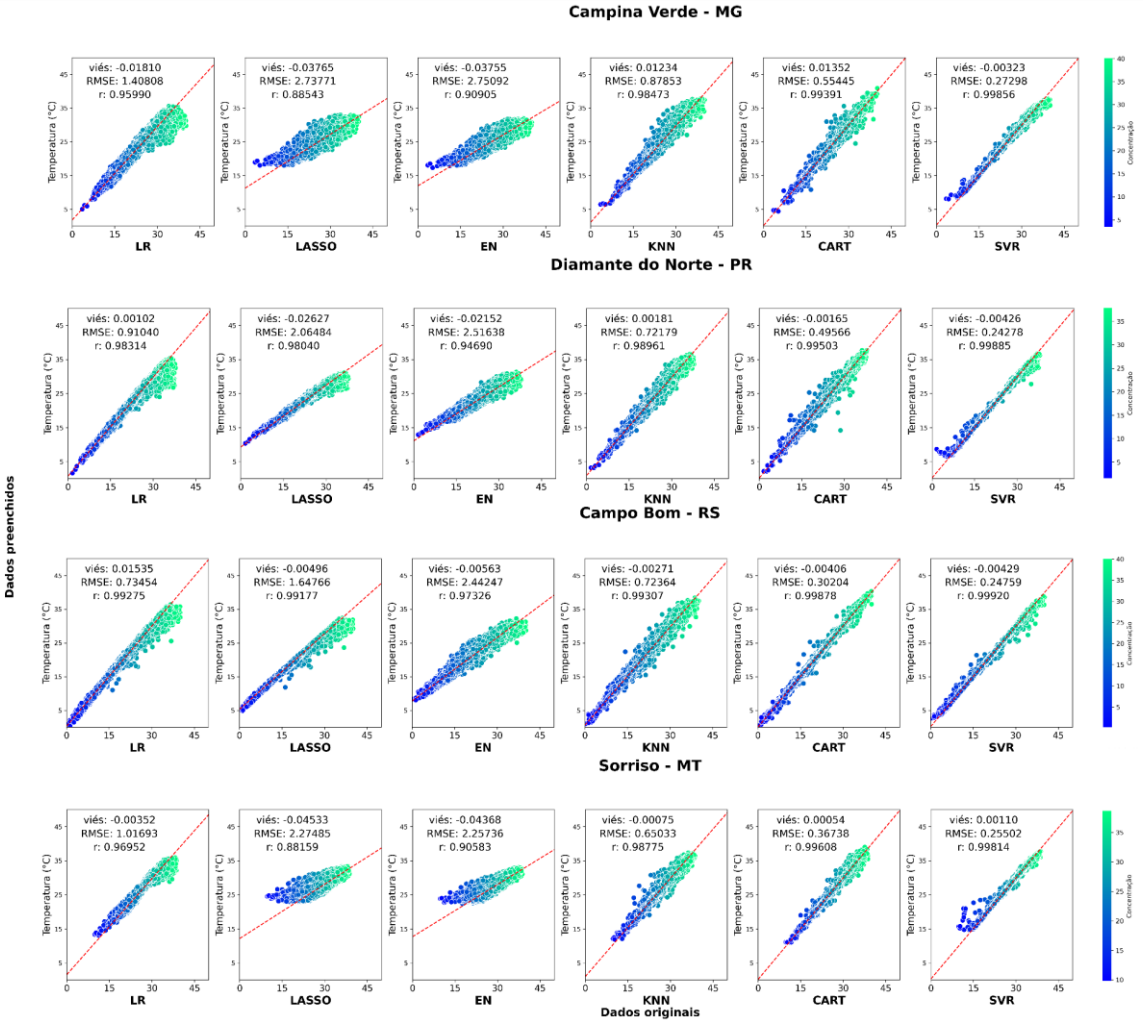
Os coeficientes de Pearson, que avaliam a linearidade entre os dados originais e os valores previstos, indicam que a maioria dos modelos foi capaz de capturar com precisão a relação subjacente dos dados, com exceção do modelo LASSO. Os modelos KNN, CART e SVR apresentaram uma correlação com coeficientes lineares superiores a 0,94. Para esses modelos, os valores de viés são baixos, inferiores a 0,013 °C, sugerindo que os erros sistemáticos nas previsões são pequenos.

No entanto, ao avaliar o RMSE, há nuances a considerar. O KNN apresentou valores de RMSE entre 0,65 °C e 0,87 °C, enquanto o CART, embora tenha apresentado correlações semelhantes, mostrou-se mais eficaz, com RMSE entre 0,30 °C e 0,55 °C.

O SVR apresentou resultados mais eficientes neste cenário, apresentando um desempenho notavelmente superior em relação aos demais, com um RMSE entre 0,24 °C e 0,27 °C e um coeficiente de correlação de Pearson superior a 0,993, o que reafirmou sua capacidade de capturar a relação linear nos dados. A consistência demonstrada pelo SVR em

diferentes localizações ressalta sua adaptabilidade e precisão nas previsões, corroborando os resultados obtidos por Kajewska-Szkudlarek e Stańczyk (2018), que também identificaram o SVR como um modelo eficaz e confiável para o preenchimento de dados de parâmetros térmicos e de umidade do ar. A concordância entre os resultados desta pesquisa e do estudo mencionado evidencia a eficiência do SVR como uma alternativa viável para a previsão de temperatura e umidade.

Figura 5 - Gráfico de dispersão entre os dados originais de temperatura e os previstos na reamostragem C



Fonte: Elaborado pelos autores (2023).

O estudo realizado por Katipoğlu e Reşat (2021), que empregou Redes Neurais Artificiais (RNA) e dados de estações adjacentes para a interpolação de dados faltantes, obteve um coeficiente de determinação (R^2) de 0,992 e um erro médio absoluto (MAE) de 0,03. Apesar

da elevada acurácia alcançada, é fundamental considerar que a utilização de RNAs acarreta um considerável dispêndio computacional quando comparada aos modelos utilizados no presente estudo. Essa observação reforça a ideia de que, embora os métodos aqui apresentados não atinjam a precisão de uma RNA, aproximam-se consideravelmente desta, com a vantagem de requererem recursos computacionais menores.

4. CONCLUSÃO

A avaliação da eficácia dos modelos de aprendizado de máquina na previsão de temperaturas revelou insights significativos. O modelo SVR emergiu como a escolha primordial, exibindo robustez nas previsões em todas as localizações, com um RMSE inferior a 0,27 e um coeficiente de correlação (r) superior a 0,993. Esse desempenho notável destaca a capacidade do SVR em capturar com precisão os padrões subjacentes nos dados de temperatura. Além disso, o modelo CART também se destacou favoravelmente, especialmente quando se considera o RMSE, que ficou abaixo de 0,55. Essa métrica indica a habilidade do CART em minimizar a discrepância entre os valores previstos e os valores reais, o que o torna uma alternativa válida em cenários onde a redução de erros é crítica.

No entanto, o modelo KNN, apesar de exibir um alinhamento linear satisfatório com os dados observados, não foi tão eficaz quanto o CART em termos de RMSE, sugerindo que, enquanto o KNN pode capturar a tendência geral dos dados, pode não ser tão apto em previsões refinadas ou em contextos em que a precisão é de extrema importância.

Um aspecto deste estudo foi a análise das reamostragens. Independentemente da proporção de dados usados para treinamento e teste, o desempenho dos modelos manteve-se estável. Isso é uma indicação positiva da robustez dos modelos e do conjunto de dados em si. Sugerindo que, dentro das configurações testadas, a quantidade de dados destinados ao treinamento foi suficiente para garantir previsões precisas, independentemente da variação na distribuição dos dados, ressaltando a capacidade de obter resultados consistentes sem necessariamente depender de grandes conjuntos de dados, permitindo flexibilidade na aplicação prática destes modelos em diferentes cenários.

O presente estudo apresenta limitações, visto que a análise se restringiu unicamente às temperaturas, desconsiderando a influência de outras variáveis meteorológicas e suas

interações, as quais poderiam contribuir para uma previsão mais acurada. Destarte, sugere-se que pesquisas futuras explorem essas variáveis adicionais, bem como otimizem os parâmetros dos modelos de aprendizado de máquina e investiguem novas arquiteturas de redes neurais, visando aprimorar substancialmente a precisão das previsões. Tais avanços metodológicos têm o potencial de fornecer conhecimentos mais robustos e confiáveis, ampliando a compreensão dos fenômenos meteorológicos e subsidiando tomadas de decisão mais embasadas em diversos setores que dependem de previsões climáticas assertivas.

REFERÊNCIAS

AWAD, Mariette; KHANNA, Rahul**. Efficient learning machines: theories, concepts, and applications for engineers and system designers**. Springer nature, 2015.

BONFANTE, Andreia Gentil et al. **Uma abordagem computacional para preenchimento de falhas em dados micro meteorológicos**. Revista Brasileira de Ciências Ambientais (RBCIAMB), n. 27, p. 61-70, 2013.

BREIMAN, Leo et al. **Classification and regression trees**. CRC press, 1984.

CONNELLY, Lynne**. Logistic regression. Medsurg Nursing**, v. 29, n. 5, p. 353-354, 2020.

COULIBALY, P.; EVORA, N. D**. Comparison of neural network methods for infilling missing daily weather records**. Journal of hydrology, v. 341, n. 1-2, p. 27-41, 2007.

CHATTERJEE, Soumyadeep et al**. Sparse group lasso for regression on land climate variables**. In: 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE, 2011. p. 1-8.

FIX, Evelyn; HODGES, Joseph Lawson. **Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review**/Revue Internationale de Statistique, v. 57, n. 3, p. 238-247, 1989.

FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Rob. **Regularization paths for generalized linear models via coordinate descent**. Journal of statistical software, v. 33, n. 1, p. 1, 2010.

GARRETA, Raul; MONCECCHI, Guillermo**. Learning scikit-learn: machine learning in python**. Packt Publishing Ltd, 2013.

KAJEWSKA-SZKUDLAREK, Joanna; STAŃCZYK, Justyna. **Filling missing meteorological data with Computational Intelligence methods**. In: ITM web of conferences. EDP Sciences, 2018. p. 00015.

KATİPOĞLU, Okan Mert; REŞAT, A. C. A. R. **Estimation of missing temperature data by Artificial Neural Network (ANN)**. Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, v. 12, n. 2, p. 431-438, 2021.

LATIF, Sarmad Dashti et al**. Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches**. Alexandria Engineering Journal, v. 82, p. 16-25, 2023.

MATSUMOTO, Makoto; NISHIMURA, Takuji. **Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator**. ACM Transactions on Modeling and Computer Simulation (TOMACS), v. 8, n. 1, p. 3-30, 1998.

MEGETO, Guilherme AS et al. **Decision tree for classification of soybean rust occurence in commercial crops based on weather variables**. Engenharia Agrícola, v. 34, p. 590-599, 2014.

MOHAMMADI, Kasra et al. **Extreme learning machine based prediction of daily dew point temperature. Computers and Electronics in Agriculture**, v. 117, p. 214-225, 2015.

MORI, Hiroyuki; TAKAHASHI, Akira. **A data mining method for selecting input variables for forecasting model of global solar radiation**. In: PES T&D 2012. IEEE, 2012. p. 1-6.

MORRIS, Clint; YANG, Jidong J. **Effectiveness of resampling methods in coping with imbalanced crash data: Crash type analysis and predictive modeling.** Accident Analysis & Prevention, v. 159, p. 106240, 2021.

PATRICK, Edward A.; FISCHER III, Frederic P. **A generalized k-nearest neighbor rule. Information and control**, v. 16, n. 2, p. 128-152, 1970.

PEDRO, Hugo TC; COIMBRA, Carlos FM. **Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances**. Renewable Energy, v. 80, p. 770-782, 2015.

RAOUHI, El Mehdi; LACHGAR, Mohamed; KARTIT, Ali**. Comparative Study of Regression and Regularization Methods: Application to Weather and Climate Data**. In: WITS 2020: Proceedings of the 6th International Conference on Wireless Technologies, Embedded, and Intelligent Systems. Springer Singapore, 2022. p. 233-240.

SMOLA, Alex J.; SCHÖLKOPF, Bernhard**. A tutorial on support vector regression**. Statistics and computing, v. 14, p. 199-222, 2004.

TCHAKONTE, Siméon et al**. Using machine learning models to assess the population dynamic of the freshwater invasive snail Physa acuta Draparnaud**, 1805 (Gastropoda: Physidae) in a tropical urban polluted streams-system. Limnologica, v. 99, p. 126049, 2023.

THOBER, Stephan et al**. Multi-model ensemble projections of European river floods and high flow**s at 1.5, 2, and 3 degrees global warming. Environmental Research Letters, v. 13, n. 1, p. 014003, 2018.

TOSUNOĞLU, Fatih et al. **Monthly streamflow forecasting using machine learning**. Erzincan University Journal of Science and Technology, v. 13, n. 3, p. 1242-1251, 2020.

WEN, Jiabao et al**. Big data driven marine environment information forecasting: a time series prediction network**. IEEE Transactions on Fuzzy Systems, v. 29, n. 1, p. 4-18, 2020.

XU, Yongjun et al. **Artificial intelligence: A powerful paradigm for scientific research**. The Innovation, v. 2, n. 4, 2021.