

Predição de Vazão de Rede com Imputação de Dados baseada em Modelos de Regressão

Maria L. Linhares¹, Janaina R. Santos¹, Ismael F. de Castro¹,
Maria C. Mesquita¹, Alex F. R. Trajano², Rafael L. Gomes¹

¹Universidade Estadual do Ceará (UECE)

{malu.linhares, janaina.ribeiro, clara.mesquita,
ismael.fonteles}@aluno.uece.br, rafa.lopes@uece.br

²Instituto Atlântico (IA)

alex.ferreira@atlantico.com.br

Abstract. *Periodic measurements provide essential metrics for network performance monitoring, such as throughput, loss, and latency. However, these measurements are susceptible to failures that result in missing data, hindering critical management tasks like performance prediction. In this context, this work presents an approach for network throughput prediction with data imputation based on regression models. The proposal performs missing value imputation using an Ensemble Stacking Regressor model; subsequently, machine learning models are employed for performance prediction. Experiments using real-world data from RNP demonstrate that high-quality imputation substantially improves prediction accuracy.*

Resumo. *Medições periódicas fornecem métricas essenciais para monitoramento do desempenho da rede, tais como vazão, perda e atraso. Contudo, essas medições são suscetíveis a falhas que resultam em dados ausentes, dificultando ações de gerenciamento primordiais, como a predição do desempenho. Dentro deste contexto, este trabalho apresenta uma abordagem para predição de vazão de rede com imputação de dados baseada em modelos de regressão. A proposta realiza a imputação dos valores faltantes utilizando um modelo Ensemble Stacking Regressor e, em seguida, modelos de aprendizado de máquina são usados para a predição do desempenho. Em experimentos utilizando dados reais da Rede Nacional de Pesquisa (RNP), os resultados mostram que uma imputação de alta qualidade melhora substancialmente a precisão das predições.*

1. Introdução

Os serviços de monitoramento de rede são fundamentais para compreender o comportamento da infraestrutura e apoiar o planejamento estratégico em diversos ambientes operacionais [Souza et al. 2024]. Esses serviços fornecem medições regulares de desempenho, como vazão, atraso e perda de pacotes, essenciais para o planejamento de capacidade, gerenciamento de Qualidade de Serviço (QoS) e conformidade com Acordos de Nível de Serviço (SLA), sendo o desempenho adequado da vazão crítico para garantir QoS e disponibilidade dos serviços de Internet [Ferreira et al. 2024]. Entretanto, a eficácia dessas atividades de monitoramento é frequentemente comprometida por falhas

de medição, resultando em conjuntos de dados incompletos que dificultam análises subsequentes e tarefas de predição [Zhu et al. 2024, Silveira et al. 2023].

As medições de vazão, diferentemente de outras métricas de rede, são realizadas em intervalos mais longos devido ao elevado consumo de recursos que impõem à infraestrutura [Mok et al. 2021, Nobre et al. 2025]. Esses procedimentos envolvem a transferência de grandes volumes de dados entre dois pontos da rede, o que pode gerar sobrecarga nos equipamentos e degradar o desempenho geral da rede [Brito et al. 2026]. Quando executadas com alta frequência, tais medições podem impactar diretamente a Qualidade de Serviço (QoS) percebida pelos usuários, ocasionando aumento de latência, redução de vazão efetiva e interrupções momentâneas no tráfego regular [Gomes et al. 2016, Gomes et al. 2014]. Por esse motivo, essas medições de vazão costumam ser cuidadosamente planejadas para períodos de baixa demanda, de modo a minimizar seus efeitos adversos e garantir maior precisão nos dados coletados [Lopes Gomes and Roberto Mauro Madeira 2012].

Além disso, dados faltantes em medições de rede podem ocorrer por diversos fatores, como falhas de hardware, congestionamento, erros de software ou problemas de transmissão. Essas ausências introduzem desafios relevantes para abordagens orientadas a dados, sobretudo aquelas que dependem de séries temporais contínuas, como os modelos de predição baseados em padrões históricos para antecipar o comportamento futuro da rede [Zhu et al. 2024]. A presença de valores ausentes não apenas compromete a análise precisa do desempenho da rede, mas também afeta negativamente o treinamento e a confiabilidade de modelos de aprendizado de máquina voltados à previsão de condições operacionais [Ferreira et al. 2024].

Técnicas tradicionais de imputação, como substituição pela média e métodos de interpolação, oferecem abordagens diretas para lidar com dados faltantes. No entanto, esses métodos frequentemente falham em capturar as dependências temporais complexas e características dinâmicas inerentes aos padrões de tráfego de rede [Xiao et al. 2025, Zhang et al. 2024]. Assim, a imputação de dados inadequada pode distorcer comportamentos de tráfego subjacentes e degradar significativamente o desempenho dos modelos de predição [Ferreira et al. 2024].

Neste contexto, este artigo apresenta uma solução de dois estágios que avalia o impacto de diferentes estratégias de imputação na predição de desempenho de rede. Diferentemente de trabalhos anteriores que abordam imputação e predição isoladamente, a solução proposta examina explicitamente a conexão entre qualidade de imputação e precisão de predição usando dados reais de rede de pontos de comunicação geograficamente distribuídos.

O estágio inicial envolve Imputação de Dados, onde valores faltantes de vazão são preenchidos inteligentemente usando um modelo de *ensemble* do tipo *stacking*. Esta abordagem aproveita métricas de rede leves prontamente disponíveis, isto é, dados de atraso e *traceroute*, que apresentam menor custo e sobrecarga operacional de coleta em comparação a medições diretas de vazão, além de poderem ser obtidas em frequências temporais significativamente maiores, para estimar com precisão as medições esparsas de vazão. O *Ensemble Stacking Regressor*, combinando múltiplos aprendizes base com um meta-aprendiz, demonstra consistentemente robustez e precisão superiores na imputação

através de várias taxas de dados faltantes em comparação com métodos tradicionais. No estágio subsequente de Predição de Dados, os dados de séries temporais imputados são alimentados em modelos avançados de redes neurais recorrentes, especificamente LSTM e GRU, juntamente com um modelo *Random Forest*, para gerar previsões precisas de vazão. A proposta fornece evidências empíricas de que a imputação de alta qualidade se traduz diretamente em melhorias significativas na precisão de predição downstream, estabelecendo uma diretriz crucial para desenvolver pipelines resilientes de monitoramento e análise de rede.

Quanto aos experimentos, utilizamos dados reais da Rede Nacional de Ensino e Pesquisa (RNP) coletados pela ferramenta Monipê¹. A proposta demonstrou um desempenho superior em termos de precisão na imputação (alcançando 57,4% menor erro que os métodos existentes), bem como que essa capacidade de imputação de alta qualidade possibilitou o aumento na eficácia do processo de predição de desempenho subsequente.

De maneira geral, as principais contribuições deste artigo são: (1) a proposição de um método de imputação de dados baseado em *Ensemble Stacking Regressor*, projetado especificamente para lidar com a esparsidade e irregularidade de medições de vazão em redes operacionais; (2) a validação da eficácia do modelo proposto frente a abordagens convencionais, demonstrando sua robustez em diferentes cenários de perda de dados (20% a 50%); e (3) a demonstração empírica da relação causal entre a qualidade da imputação e a precisão da predição de vazão, evidenciando que um pré-processamento inteligente é determinante para o desempenho de modelos de redes neurais em ambientes reais. Ao estabelecer esta conexão, este trabalho fornece uma estratégia prática para construir sistemas de monitoramento resilientes, capazes de manter a precisão analítica mesmo diante de falhas de medição.

O restante deste artigo está organizado da seguinte forma: A Seção 2 apresenta o estado da arte e os trabalhos relacionados ao contexto de imputação, predição e monitoramento de desempenho de rede, a Seção 3 descreve a proposta, enquanto a Seção 4 detalha os experimentos realizados, e a Seção 5 discute os resultados. Finalmente, a Seção 6 conclui o artigo e apresenta diversos trabalhos futuros.

2. Trabalhos Relacionados

Avanços recentes em aprendizado ensemble demonstraram capacidades promissoras no tratamento de tarefas complexas de predição. Métodos ensemble de *stacking*, que combinam múltiplos estimadores base através de uma abordagem de meta-aprendizado, têm mostrado desempenho robusto em vários domínios ao aproveitar as forças complementares de modelos diversos [Ganaie et al. 2022, Amara et al. 2025]. Essas técnicas são particularmente adequadas para aplicações de séries temporais, pois podem se adaptar melhor a padrões não lineares e dependências temporais [Divina et al. 2018]. Não obstante, a aplicação específica de métodos ensemble avançados para imputação em contextos de monitoramento de rede, e seu impacto subsequente na precisão de predição, permanece insuficientemente explorada [Kim et al. 2023, Zhang et al. 2023].

Park et al. [Park et al. 2023] apresentam uma abordagem baseada em modelos de aprendizagem profunda para estimar valores ausentes em dados multivariados de séries

¹<https://memoria.rnp.br/pd/monipe.html>

temporais. O foco do estudo está no preenchimento de lacunas longas e contínuas, como meses de observações diárias ausentes, em vez de lidar com valores ausentes aleatórios. Apesar de sua relevância, o trabalho não aborda diretamente a integração de técnicas de imputação para melhorar a precisão de tarefas de predição. Similarmente, na referência [Ahn et al. 2021], os autores realizam uma análise abrangente das principais técnicas de imputação de dados aplicadas a séries temporais. O estudo avalia métodos como interpolação linear, KNN, e redes neurais para prever valores ausentes, destacando o KNN como o método com melhor desempenho em diversos cenários. Apesar disso, o trabalho enfatiza a importância de métodos que capturem padrões temporais complexos.

Com base nesses trabalhos, nossa proposta busca preencher as lacunas identificadas, combinando a robustez do uso de *Ensemble Stacking Regressor* com a consideração dos aspectos sazonais das séries temporais para criar uma abordagem eficaz tanto para imputação quanto para predição. Esse diferencial permite lidar com dados de rede de forma mais precisa, melhorando a qualidade das análises e previsões no contexto de desempenho de rede.

3. Proposta

O desenvolvimento da proposta seguiu um *pipeline* projetado para demonstrar que a qualidade da imputação se traduz diretamente em melhor desempenho de predição na previsão de vazão de rede. Conforme ilustrado na Figura 1, o pipeline consiste em quatro estágios interconectados:

- **Coleta de Dados:** recuperação de dados de vazão, atraso, e *traceroute* da ferramenta de monitoramento;
- **Pré-processamento:** Inicialmente, foi realizada a remoção de *outliers* com o objetivo de eliminar medições anômalas que poderiam introduzir viés nas análises e nos modelos de predição. Em seguida, a partir do *traceroute*, foram extraídas informações relativas ao número de saltos do caminho e ao link gargalo, definido como o enlace com a menor capacidade de vazão ao longo da rota. Após a extração dessas métricas, os dados de vazão, número de saltos, link gargalo e atraso foram alinhados temporalmente com base no *timestamp*, garantindo consistência entre as variáveis utilizadas.
- **Imputação de Dados:** Nessa etapa, as métricas de atraso, número de saltos e link gargalo são utilizadas como variáveis de entrada de um modelo *Ensemble Stacking Regressor*. Esse ensemble é composto por cinco modelos base e um meta-modelo responsável por combinar suas predições. O objetivo do modelo é estimar os valores de vazão a partir das demais métricas de rede, explorando a complementaridade entre diferentes técnicas de aprendizado. As estimativas de vazão produzidas pelo *ensemble* são então empregadas no processo de imputação dos dados ausentes, caracterizando a abordagem proposta baseada em *Ensemble Stacking Regressor*.
- **Predição:** A série temporal de vazão resultante do processo de imputação é utilizada como entrada para os modelos de predição. Especificamente, modelos de aprendizado de máquina, como LSTM, GRU e *Random Forest*, são empregados para capturar padrões temporais e não lineares, visando a predição futura da vazão a partir da série temporal imputada anteriormente.

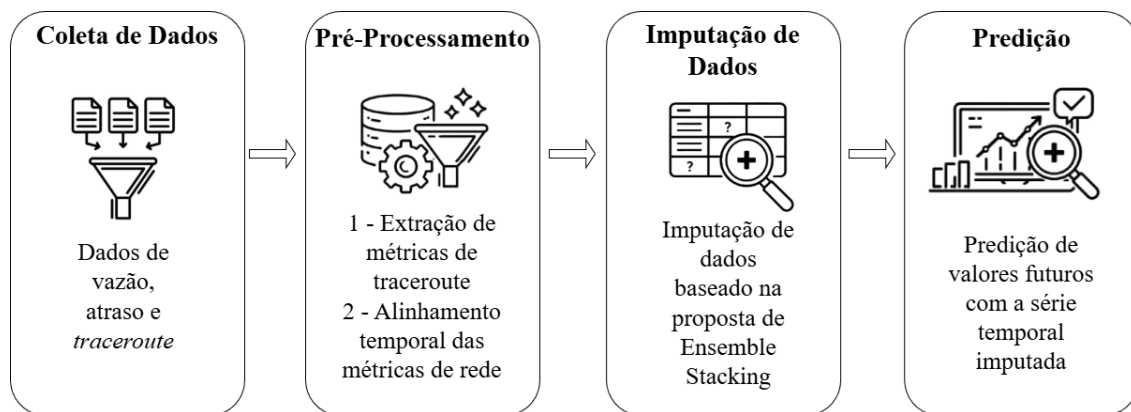


Figure 1. Pipeline de predição de vazão com imputação de dados.

Inicialmente, a solução considera dados de medição provenientes de serviços de monitoramento de rede existentes, independentemente da ferramenta utilizada. A partir desses dados, são aplicadas as etapas do pipeline apresentado, abrangendo o pré-processamento, a imputação dos valores ausentes e a posterior construção dos modelos de predição. As ações realizadas em cada um desses estágios são detalhadas nas seções seguintes.

Em síntese, A proposta adota uma abordagem focada nas duas últimas etapas do *pipeline*: a imputação de dados, realizada por meio de um modelo baseado em *Ensemble Stacking Regressor*, e a predição de vazão, empregando modelos de aprendizado de máquina. Essa avaliação em dois estágios permite estabelecer uma relação causal clara, na qual uma maior precisão no processo de imputação se reflete diretamente em melhor desempenho dos modelos de predição.

3.1. Garantia de Qualidade dos Dados

As medições brutas de rede passam por um pré-processamento abrangente para garantir a qualidade dos dados. Implementamos detecção e remoção sistemática de *outliers* utilizando o critério estatístico do *Interquartile Range* (IQR), no qual observações fora do intervalo foram descartadas. Após essa etapa, aplicamos normalização dos dados por meio do *Standard Scaling*, transformando as variáveis para média zero e desvio padrão unitário, utilizando o *StandardScaler* da biblioteca *scikit-learn*.

A regularização temporal aborda padrões de amostragem irregular por meio de um processo de alinhamento baseado em *timestamp*, no qual a variável de vazão é utilizada como referência temporal. Para cada observação de vazão, são associadas as métricas auxiliares de atraso e as métricas derivadas do *traceroute*, especificamente número de saltos e identificação do link gargalo, selecionando-se as observações temporalmente mais próximas dentro de uma janela máxima de 10 minutos. Caso não existam medições dentro dessa janela, a instância é descartada, evitando a associação de dados inconsistentes.

A escolha da vazão como referência se deve ao fato de ser a variável alvo e possuir a menor frequência de amostragem (a cada 6 horas), enquanto as demais métricas são coletadas em frequências mais altas (1 minuto para atraso e 10 minutos para métricas derivadas do *traceroute*), permitindo sua associação de forma consistente. A janela de 10 minutos foi definida como um compromisso entre garantir a disponibilidade de dados e

preservar a coerência temporal entre as variáveis, além de estar alinhada com a frequência de coleta das métricas derivadas do *traceroute*.

Esse intervalo também atua como um filtro para flutuações momentâneas, preservando padrões mais estáveis e relevantes da rede. No entanto, é importante destacar que esse processo pode suavizar fenômenos transitórios de curta duração, uma vez que variações rápidas podem não ser completamente capturadas no alinhamento. Ainda assim, essa limitação é mitigada pelo fato de a variável alvo apresentar baixa frequência de amostragem, de modo que o modelo prioriza tendências agregadas do comportamento da rede.

Do *traceroute*, extraímos duas características críticas do caminho: contagem de saltos, determinada pela quantidade de roteadores intermediários entre origem e destino, e identificação do link gargalo, obtida a partir da análise de medições históricas de vazão e informações de capacidade dos enlaces, permitindo identificar o segmento com menor largura de banda ao longo da rota. Essas características capturam aspectos estruturais relevantes do caminho de rede e são utilizadas como entrada no processo de imputação com o *Ensemble Stacking Regressor* proposto.

3.2. Imputação de Dados: *Ensemble Stacking Regressor*

O *Ensemble Stacking Regressor* (S) estima valores faltantes de vazão ($\tilde{Y}_{\text{imputed}}$) aproveitando métricas de rede leves prontamente disponíveis X (atraso e características de *traceroute*). O *Ensemble Stacking Regressor* opera em dois níveis, combinando o poder preditivo de K modelos base distintos (f_1, \dots, f_K) através de um meta-aprendiz (g):

- **Nível 0 - Aprendizes Base:** Múltiplos modelos de regressão geram previsões out-of-fold via validação cruzada de 5 *folds*, cada um capturando diferentes padrões e relações das características de entrada. Para um dado conjunto de características de entrada X , cada aprendiz base f_k gera uma previsão $p_k(X)$:

$$p_k(X) = f_k(X), \quad \text{para } k = 1, \dots, K$$

Os aprendizes base utilizados incluem *XGBoost*, *Random Forest*, *Gradient Boosting* e *K-Nearest Neighbors* (KNN). Esses modelos são treinados nos dados observados, não faltantes ($X_{\text{obs}}, Y_{\text{obs}}$), e utilizam validação cruzada de 5 *folds* para gerar previsões out-of-fold robustas, que servem como entrada para o próximo nível.

- **Nível 1:** O meta-aprendiz *Ridge Regression* combina as previsões base de forma otimizada. As K previsões dos aprendizes base são concatenadas em uma nova matriz de características, $P = [p_1(X), p_2(X), \dots, p_K(X)]$. O meta-aprendiz g , definido como um modelo *Ridge Regression* nesta implementação, é treinado em P para produzir a previsão ensemble final $\tilde{Y}_{\text{imputed}}$:

$$\tilde{Y}_{\text{imputed}} = S(X) = g(P) = g(f_1(X), f_2(X), \dots, f_K(X))$$

Essa escolha é empiricamente e teoricamente justificada, pois as previsões dos modelos base em métodos de *stacking* tendem a ser altamente correlacionadas, caracterizando um cenário de multicolinearidade. Trabalhos clássicos mostram que meta-aprendizes lineares regularizados são particularmente adequados nesse contexto, uma vez que a regularização L2 estabiliza a estimação dos coeficientes

e reduz a variância do modelo final [Wolpert 1992, Breiman 1996]. Além disso, a regressão ridge é amplamente reconhecida por sua eficácia em lidar com multicolinearidade, produzindo estimativas mais estáveis e com melhor capacidade de generalização [Hoerl and Kennard 1970, Hastie et al. 2009].

Especificamente, para *Ridge Regression*, a predição final assume a forma:

$$\tilde{Y}_{\text{imputed}} = w_0 + \sum_{k=1}^K w_k \cdot p_k(X)$$

onde w_k são os pesos ótimos aprendidos pelo meta-aprendiz ao minimizar uma função de perda regularizada (norma L2).

Assim, o objetivo do *Ensemble Stacking Regressor* é estimar com precisão os valores faltantes de vazão ($\tilde{Y}_{\text{imputed}}$) de modo que, quando combinados com os valores observados (Y_{obs}), uma série temporal robusta e completa \hat{Y} seja gerada:

$$\hat{Y} = \{Y_{\text{obs}}, \tilde{Y}_{\text{imputed}}\}$$

A eficácia do ensemble é medida principalmente pela minimização do RMSE da imputação em um conjunto de validação reservado:

$$\min_S \text{RMSE}_{\text{Imputation}}(S) = \sqrt{\frac{1}{N_{\text{val}}} \sum_{t \in \text{Validation}} (\tilde{Y}_t - Y_{\text{true},t})^2}$$

Quanto aos hiperparâmetros dos modelos, eles foram otimizados via busca em grade com validação cruzada para maximizar a precisão da imputação enquanto previnem *overfitting*. Essa abordagem transforma a desvantagem temporal das medições esparsas de vazão em uma vantagem ao aproveitar a rica informação temporal de métricas leves, permitindo avaliação contínua de largura de banda mantendo precisão.

3.3. Predição de Dados

Após a imputação, avaliamos como a qualidade da imputação afeta o desempenho de predição usando três modelos: *Long Short-Term Memory* (LSTM), *Gated Recurrent Unit* (GRU) e *Random Forest*. Esses modelos foram selecionados para representar paradigmas de aprendizado tanto sequenciais (LSTM, GRU) quanto *ensemble* (*Random Forest*) para predição de séries temporais.

LSTM e GRU são arquiteturas de redes neurais recorrentes especificamente projetadas para aprender dependências de longo prazo em dados sequenciais. LSTM emprega uma célula de memória com três portas especializadas (entrada, esquecimento e saída) que regulam o fluxo de informação, permitindo manter informações relevantes através de intervalos de tempo estendidos enquanto mitiga problemas de desaparecimento de gradiente [Hochreiter and Schmidhuber 1997]. GRU simplifica esta arquitetura usando apenas duas portas (reset e atualização) e um estado oculto unificado, frequentemente resultando em treinamento mais rápido e melhor generalização em conjuntos de dados com amostras limitadas [Rafi et al. 2021].

Random Forest fornece uma perspectiva de aprendizado não sequencial ao agregar predições de múltiplas árvores de decisão treinadas em subconjuntos aleatórios de dados. Sua robustez contra *overfitting* e capacidade de capturar relações não lineares o tornam uma linha de base sólida para avaliar os benefícios relativos de modelos temporais.

4. Experimentos

Esta seção detalha a configuração experimental projetada para avaliar a proposta. Para apoiar a reprodutibilidade e pesquisas futuras, o conjunto de dados usado em nossos experimentos, juntamente com o código-fonte, está disponível². Para realizar experimentos realistas, utilizamos os dados do serviço de monitoramento (MonIPÊ) da RNP³, que emprega o padrão internacional de monitoramento perfSONAR.

Vale mencionar que a RNP possui pontos de comunicação em todos os 26 estados da República Federativa do Brasil, incluindo o Distrito Federal, e que esses pontos de comunicação têm comportamento diferente devido às suas particularidades. Seleccionamos cinco pares estratégicos de ponto de presença (PoP) da RNP com foco intencional em enlaces para Natal-RN: SC-RN, AM-RN, AC-RN, MT-SC e PR-GO. Esta seleção prioriza Natal-RN como um *hub* central devido ao seu papel como ponto de aterrissagem para múltiplos cabos submarinos internacionais, tornando-o um ponto de interconexão crítico para cenários de descarregamento de dados internacionais. O conjunto de dados foi coletado no período de janeiro de 2024 a julho de 2025. Três tipos de medição foram utilizados: atraso, calculado em média em intervalos de 1 minuto; métricas derivadas do *traceroute*, como o número de saltos e a identificação do link gargalo, coletadas a cada 10 minutos; e vazão, medida a cada 6 horas como variável alvo. Essa seleção estratégica garante uma avaliação abrangente por meio de diferentes caminhos de rede, com características geográficas e infraestruturais variadas dentro do ambiente operacional da RNP.

Para estabelecer uma base comparativa rigorosa, foram selecionados cinco métodos de *baseline* amplamente utilizados na literatura para lidar com dados ausentes em séries temporais: interpolação linear, mediana, *k-Nearest Neighbors* (kNN), *forward fill* (ffill) e *backward fill* (bfill). É importante destacar que esses *baselines* foram aplicadas de forma consistente em dois estágios do experimento: primeiro, na etapa de imputação de dados para completar as séries históricas de vazão; e, posteriormente, como base de comparação para avaliar o ganho de desempenho obtido pelo método de imputação proposto (baseado em *stacking*) frente a essas abordagens convencionais durante a predição.

Quanto ao treinamento do modelo de predição, a série temporal completa imputada obtida de cada técnica de imputação foi usada para treinar todos os três modelos. Alocamos 80% de cada série para treinamento e 20% para teste, permitindo avaliação sistemática de como diferentes estratégias de imputação influenciam a qualidade da predição. A otimização de hiperparâmetros foi conduzida usando *Optuna* com uma estratégia de busca bayesiana para garantir comparação justa entre todos os métodos de imputação, onde as configurações de parâmetros com melhor desempenho identificadas pelo *Optuna* estão resumidas na Tabela 1.

Para avaliar de forma abrangente a qualidade de nossa metodologia em ambos os estágios de imputação e predição, empregamos métricas de avaliação específicas por estágio. Primeiro, a **Avaliação de Imputação** analisa a qualidade dos valores preenchidos. Para isso, seleccionamos a maior sequência contínua (sem falhas) de cada série temporal de vazão e removemos aleatoriamente amostras em proporções de 20%, 30%, 40% e

²<https://github.com/LarcesUece/predicao-com-imputacao-avancada>

³www.rnp.br/en/ipe-network

Table 1. Configurações Otimizadas de Hiperparâmetros (Busca via *Optuna*)

Model	Parâmetros	Baseline	Stacking
Random Forest	n_estimators	121	119
	max_depth	15	5
	min_samples_split	4	3
LSTM	window	26	24
	hidden_size	44	57
	num_layers	2	2
	dropout	0.09	0.08
GRU	window	64	64
	hidden_size	41	64
	num_layers	3	3
	dropout	0.10	0.09

50%. Em seguida, aplicamos o método proposto e os *baselines* para imputar esses valores ausentes. A qualidade da imputação é avaliada comparando os valores imputados com os valores reais originais através de métricas estabelecidas: NRMSE (*Normalized Root Mean Square Error*), RMSE (*Root Mean Square Error*), R^2 (Coeficiente de Determinação) e MAPE (*Mean Absolute Percentage Error*). Em segundo lugar, a **Avaliação de Predição** quantifica como a qualidade da imputação se traduz em desempenho de predição usando RMSE e Percentual de Melhoria. A métrica de Percentual de Melhoria quantifica a redução relativa de erro alcançada por nossa imputação baseada em *stacking* comparada ao melhor método *baseline* de acordo com:

$$\text{Melhoria (\%)} = \frac{\text{RMSE}_{\text{Baseline}} - \text{RMSE}_{\text{Stacking}}}{\text{RMSE}_{\text{Baseline}}} \times 100$$

onde $\text{RMSE}_{\text{Baseline}}$ representa o menor RMSE obtido entre os métodos de imputação considerados como *baseline*: interpolação linear, mediana, *k-Nearest Neighbors* (kNN), *forward fill* (ffill) e *backward fill* (bfill). Esta comparação é realizada para cada par de comunicação e modelo de predição. Valores positivos indicam que a imputação baseada em *stacking* melhorou a precisão da predição, demonstrando o impacto direto da qualidade superior de imputação no desempenho final do modelo.

5. Resultados

Esta seção apresenta os resultados obtidos da análise de experimentos conduzidos com um conjunto de dados real, discutindo os principais pontos referentes às ações de imputação de dados e desempenho de predição.

5.1. Desempenho de Imputação

O *Ensemble Stacking Regressor* demonstrou desempenho superior comparado a todos os métodos *baseline* na fase de imputação, conforme detalhado de forma abrangente nas Figuras 2 e 3. Os valores mostrados representam o desempenho médio através de todos os cinco enlaces de rede (SC-RN, AM-RN, AC-RN, MT-SC e PR-GO) para cada taxa de dados faltantes.

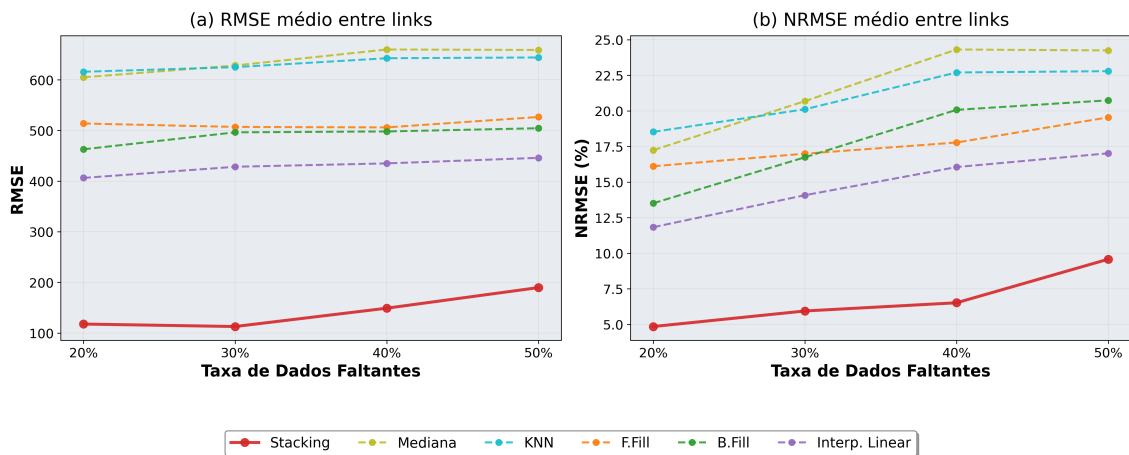


Figure 2. Comparação do RMSE e do NRMSE por taxa de dados faltantes.

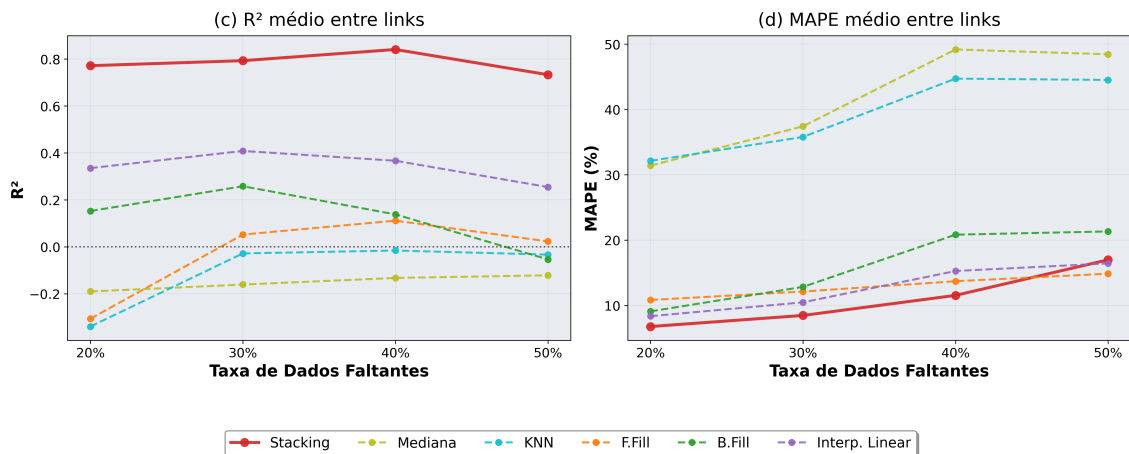


Figure 3. Comparação de desempenho de R^2 e MAPE por taxa de dados faltantes.

Conforme demonstrado nas Figuras 2 e 3, o *Ensemble Stacking Regressor* apresentou desempenho competitivo e consistente em comparação com os métodos *baseline*, especialmente em cenários mais desafiadores. Diferentemente das abordagens tradicionais baseadas em estatísticas simples (Median, KNN), os melhores *baselines* observados foram métodos temporais, como *Forward Fill*, *Backward Fill* e *Linear Interpolation*.

Em termos de RMSE e NRMSE, esses métodos temporais frequentemente apresentaram os menores erros em cenários com menor taxa de dados faltantes, refletindo sua capacidade de explorar dependências locais da série temporal. No entanto, o *Ensemble Stacking Regressor* demonstrou maior robustez à medida que a taxa de dados faltantes aumenta ou em enlaces com maior variabilidade, mantendo erros controlados e desempenho mais consistente entre diferentes cenários. Essa tendência também é evidenciada pelos valores de R^2 . Enquanto alguns *baselines* temporais atingem bons resultados em cenários específicos, o *Ensemble Stacking Regressor* mantém valores consistentemente positivos e elevados, chegando a ultrapassar 0,9 em diversos enlaces, indicando forte capacidade de explicação da variância dos dados mesmo em condições adversas. Em contraste, métodos mais simples (Median e KNN) frequentemente apresentam valores negativos de R^2 , indicando desempenho inferior a uma previsão baseada na média.

Em relação ao MAPE, os métodos temporais novamente apresentam bom desempenho em cenários com baixa taxa de ausência de dados. Entretanto, o *Ensemble Stacking Regressor* mantém erros relativos mais estáveis à medida que a complexidade do problema aumenta, reforçando sua capacidade de generalização. Além disso, o *Ensemble Stacking Regressor* apresentou baixo tempo médio de predição por amostra (1,22 ms), evidenciando sua viabilidade para aplicações em tempo real. Este desempenho foi avaliado em um hardware equipado com processador Intel Core i5-13420H de 13ª Geração (4,6 GHz) e 16 GB de memória RAM DDR5.

De forma geral, os resultados indicam que, embora métodos temporais simples sejam fortes *baselines* em cenários com baixa perda de dados, o *Ensemble Stacking Regressor* oferece maior robustez, estabilidade e capacidade de generalização, sendo particularmente vantajoso em cenários de alta taxa de dados faltantes ou enlaces com maior variabilidade.

5.2. Desempenho de Predição com Dados Imputados

Os experimentos de predição foram realizados usando três modelos: LSTM, GRU e *Random Forest* treinados nos conjuntos de dados imputados tanto pelos métodos *baseline* quanto pelo *Ensemble Stacking Regressor*. O ajuste de hiperparâmetros foi conduzido usando o *framework Optuna*, que explora eficientemente o espaço de parâmetros através de amostragem bayesiana para identificar a configuração ótima para cada modelo e estratégia de imputação. Em termos de precisão preditiva, modelos treinados em **conjuntos de dados imputados por Stacking** alcançaram métricas de erro substancialmente menores comparados àqueles treinados em imputações *baseline*.

O RMSE médio diminuiu de 0,0273 para 0,0126, representando uma melhoria média de aproximadamente **61,84%**, através de todos os enlaces de rede. Entre os modelos avaliados, o **Random Forest** exibiu a maior melhoria relativa, com um ganho médio de **+73,49%**, seguido pelo **LSTM** com **+61,98%**, e o **GRU** com **+50,06%** de melhoria sobre suas respectivas linhas de base (Tabela 2).

Table 2. Resultados de Predição por Modelo.

Modelo	<i>Baseline</i> RMSE	<i>Stacking</i> RMSE	Melhoria (%)
Random Forest	0.000685	0.000181	73.49
LSTM	0.024395	0.009273	61.98
GRU	0.056710	0.028316	50.06
Média	0.027263	0.012590	61.84

Conforme mostrado na Tabela 2 e Figura 4, o percentual médio de melhoria através de todos os pontos de comunicação avaliados foi **61,84%**, com melhorias de enlaces individuais variando de **+9,39%** (PR–GO) a **+73,49%** (SC–RN). A variabilidade observada através dos enlaces destaca como diferentes topologias de rede e condições de completude de dados influenciam os benefícios preditivos alcançados por estratégias aprimoradas de imputação.

Os resultados da predição de redes neurais, usando os cinco pontos de comunicação selecionados, demonstram que a qualidade da imputação impacta significa-

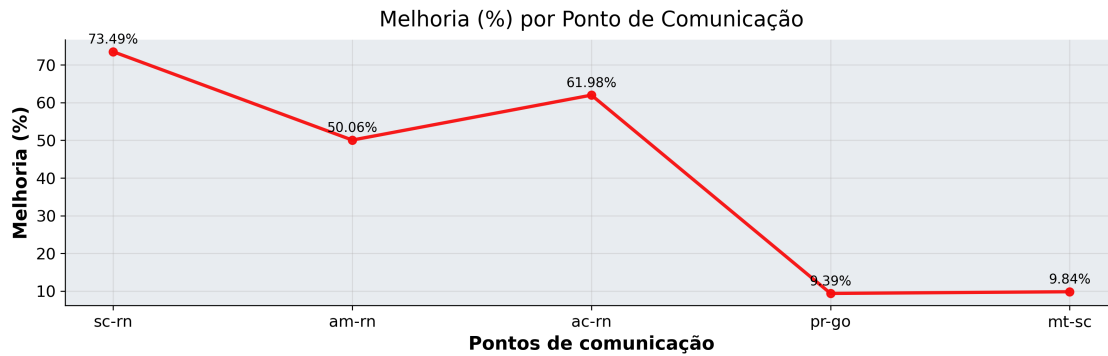


Figure 4. Melhoria (%) por pontos de comunicação e modelo.

tivamente a precisão de predição. Nossa análise revelou que séries temporais imputadas usando o *Ensemble Stacking Regressor* consistentemente produziram predições mais precisas comparadas àquelas geradas pelos métodos de referência. A correlação entre qualidade de imputação e precisão de predição foi forte para as métricas NRMSE e R^2 , com coeficientes de correlação de Pearson de **0,78** e **0,82**, respectivamente, através de todos os cinco enlaces selecionados. Notavelmente, os enlaces que exibiram melhor desempenho de imputação também mostraram resultados de predição mais estáveis e confiáveis, com menor variância nos erros de predição. Esta consistência enfatiza a importância do pré-processamento de dados de alta qualidade para tarefas de predição subsequentes em operações de rede. No geral, estes achados confirmam que a imputação robusta de dados melhora diretamente os modelos de predição.

O *Ensemble Stacking Regressor* não apenas reduz incerteza nas sequências de entrada, mas também amplifica o desempenho preditivo de métodos tanto neurais quanto ensemble. As arquiteturas LSTM e GRU se beneficiaram mais de imputações temporalmente consistentes, enquanto o *Random Forest* serviu como uma forte linha de base não temporal confirmando a generalização de melhorias através de famílias de modelos.

6. Conclusão

Este estudo demonstra a eficácia do *Ensemble Stacking Regressor* para imputação de dados de tráfego de rede e seu impacto nas tarefas de predição subsequentes. Embora métodos *baseline* temporais, como *Forward Fill*, *Backward Fill* e *Linear Interpolation*, apresentem bom desempenho em cenários com baixa taxa de dados faltantes, o modelo proposto se destaca pela maior robustez e consistência ao longo de diferentes enlaces e níveis de ausência de dados (20% a 50%).

O *Ensemble Stacking Regressor* mantém valores de R^2 consistentemente positivos e métricas de erro estáveis (RMSE, NRMSE e MAPE), mesmo em cenários mais desafiadores, nos quais métodos simples tendem a degradar. Esses resultados evidenciam a forte relação entre qualidade de imputação e desempenho de predição, permitindo estimativas mais confiáveis de vazão e, conseqüentemente, apoiando decisões mais eficientes de gerenciamento de recursos de rede, como planejamento de capacidade, alocação de banda e garantia de QoS.

Trabalhos futuros podem explorar abordagens de *stacking* adaptativo que se ajustam dinamicamente às características da rede e padrões de dados faltantes. Adi-

cionalmente, pretende-se investigar as correlações específicas entre a topologia da rede, a volatilidade do tráfego em enlaces geográficos distintos e o desempenho da imputação, visando explicar os fatores causais da variabilidade observada entre os pares de comunicação. Investigar a transferibilidade desses achados para outros tipos de métricas de rede e diferentes escalas geográficas também é uma direção promissora. Por fim, explorar as compensações entre complexidade computacional e precisão de imputação pode orientar cenários operacionais em tempo real, validando a generalizabilidade desses achados em um conjunto ainda mais amplo de enlaces.

Agradecimentos

Os autores gostariam de agradecer ao CNPq (N° 305946/2025-0 e N° 405940/2022-0) e CAPES (N° 88887.954253/2024-00 e N° 88887.972043/2024-00) pelo apoio financeiro.

References

- Ahn, H., Sun, K., and Kim, K. (2021). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials and Continua*, 70:767–779.
- Amara, M., Smairi, N., and Jaballah, M. (2025). Stacked ensemble deep learning for robust intrusion detection in iot networks. pages 1146–1153.
- Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24(1):49–64.
- Brito, M. L. L., Ferreira, M. C. M., Portela, A. L. C., and Gomes, R. L. (2026). Ai-based estimation of bandwidth availability for data offloading in edge-cloud computing. *IEEE Networking Letters*, 8:69–73.
- Divina, F., Gilson, A., Gómez-Vela, F., García Torres, M., and Torres, J. F. (2018). Stacking ensemble learning for short-term electricity consumption forecasting. *Energies*, 11(4):949.
- Ferreira, M. C., Ribeiro, S. E., Nobre, F. V., Linhares, M. L., Araújo, T. P., and Gomes, R. L. (2024). Mitigating measurement failures in throughput performance forecasting. In *2024 20th International Conference on Network and Service Management (CNSM)*, pages 1–7.
- Ganaie, M., Hu, M., Malik, A., Tanveer, M., and Suganthan, P. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151.
- Gomes, R. L., Bittencourt, L. F., and Madeira, E. R. (2014). A bandwidth-feasibility algorithm for reliable virtual network allocation. In *2014 IEEE 28th International Conference on Advanced Information Networking and Applications*, pages 504–511.
- Gomes, R. L., Bittencourt, L. F., Madeira, E. R., Cerqueira, E., and Gerla, M. (2016). Bandwidth-aware allocation of resilient virtual software defined networks. *Computer Networks*, 100:179–194.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.
- Kim, J., Park, S., and Lee, H. (2023). Impact of data imputation on machine learning model performance: A case study in network analysis. *IEEE Transactions on Network Science and Engineering*, 10(4):2341–2354.
- Lopes Gomes, R. and Roberto Mauro Madeira, E. (2012). A traffic classification agent for virtual networks based on qos classes. *IEEE Latin America Transactions*, 10(3):1734–1741.
- Mok, R. K. P., Zou, H., Yang, R., Koch, T., Katz-Bassett, E., and Claffy, K. C. (2021). Measuring the network performance of google cloud platform. In *Proceedings of the 21st ACM Internet Measurement Conference, IMC '21*, page 54–61, New York, NY, USA. Association for Computing Machinery.
- Nobre, F. V. J., Silva, D. d. S., Ferreira, M. C. M. M., Brito, M. L. M. L., de Araújo, T. P., and Gomes, R. L. (2025). Time-weighted correlation approach to identify high delay links in internet service providers. *Journal of Internet Services and Applications*, 16(1):419–430.
- Park, J., Müller, J., Arora, B., Faybishenko, B., Pastorello, G., Varadharajan, C., Sahu, R., and Agarwal, D. (2023). Long-term missing value imputation for time series data using deep neural networks. *Neural Computing and Applications*, 35(12):9071–9091.
- Rafi, S. H., Nahid-Al-Masood, Deeba, S. R., and Hossain, E. (2021). A short-term load forecasting method using integrated cnn and lstm network. *IEEE Access*, 9:32436–32448.
- Silveira, M. M., Portela, A. L., Menezes, R. A., Souza, M. S., Silva, D. S., Mesquita, M. C., and Gomes, R. L. (2023). Data protection based on searchable encryption and anonymization techniques. In *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, pages 1–5. IEEE.
- Souza, M. S., Ribeiro, S. E. S. B., Lima, V. C., Cardoso, F. J., and Gomes, R. L. (2024). Combining regular expressions and machine learning for sql injection detection in urban computing. *Journal of Internet Services and Applications*, 15(1):103–111.
- Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.
- Xiao, H., Shen, X., Li, J., and Yang, X. (2025). A method for filling traffic data based on feature-based combination prediction model. *Scientific Reports*, 15(8441).
- Zhang, W., Liu, Q., and Chen, Y. (2023). On the quality of imputed data for downstream prediction tasks. In *NeurIPS Workshop on Time Series*.
- Zhang, Y., Kong, X., Zhou, W., Liu, J., Fu, Y., and Shen, G. (2024). A comprehensive survey on traffic missing data imputation. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):19252–19275.
- Zhu, Y., Jiang, B., Jin, H., Zhang, M., Gao, F., Huang, J., Lin, T., and Wang, X. (2024). Networked time-series prediction with incomplete data via generative adversarial network. *ACM Transactions on Knowledge Discovery from Data*, 18(5):115.