CONIC SEMESP

25º Congresso Nacional de Iniciação Científica

TÍTULO: PREVISÃO DA DEMANDA REGIONAL POR VAGAS DE TI NO BRASIL: UMA ANÁLISE COMPARATIVA DE MODELOS DE APRENDIZADO DE MÁQUINA

CATEGORIA: CONCLUÍDO

ÁREA: CIÊNCIAS EXATAS, DA TERRA E AGRÁRIAS

SUBÁREA: Computação e Informática

INSTITUIÇÃO: INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS GERAIS - IFMG

AUTOR(ES): DANILO BATISTA LIMA

ORIENTADOR(ES): CARLOS ALEXANDRE SILVA

COLABORADOR(ES): ISABELLE YASMIN DE ARAUJO MOREIRA

Realização: SEMESP

IES parceiras: PUC CAMPINAS

CATEGORIA CONCLUÍDO

1. RESUMO

Este estudo desenvolve e compara modelos de aprendizado de máquina para a previsão da demanda regional por vagas no setor de Tecnologia da Informação (TI) no Brasil. A partir de um conjunto de dados com mais de 66 mil anúncios de emprego, extraídos do LinkedIn via *web scraping* e estruturados como séries temporais diárias, a análise abrange as cinco macrorregiões brasileiras e uma categoria específica para vagas remotas. Foram avaliados comparativamente três algoritmos de *machine learning* para previsão, sendo eles *Support Vector Regression* (SVR), redes neurais *feedforward* do tipo MLP (*Multi-Layer Perceptron*) e redes neurais *Long Short-Term Memory* (LSTM). O desempenho dos modelos, treinados individualmente por região, foi aferido por meio das métricas MAE (*Mean Absolute Error*) e MDA (*Mean Directional Accuracy*). Os resultados demonstram que o modelo *feedforward* obteve a melhor performance, apesar do modelo SVR ter obtido resultados parecidos. De acordo com os dados, as vagas remotas, somadas às das regiões Sudeste e Sul, representam mais de 90% do total, evidenciando a menor oferta nas demais regiões. Dessa forma, este trabalho demonstra a viabilidade da metodologia para a captura de padrões de demanda regionais, oferecendo um panorama quantitativo robusto do comportamento do setor de TI brasileiro por região e modalidade de trabalho.

2. INTRODUÇÃO

O setor de Tecnologia da Informação (TI) no Brasil é atualmente o maior da América Latina (ABES, 2024), porém exibe uma acentuada desigualdade na distribuição de oportunidades, com forte concentração na região Sudeste[1]. A recente expansão do trabalho remoto tem mudado esse quadro, atuando como um vetor de descentralização que cria um cenário propício para a interiorização de vagas e a

[1] https://escoladanuvem.org/falta-de-vagas-no-norte-e-nordeste-revela-disparidade-no-mercado-de-tecnologia-e-desafio-de-inclusao-para-o-setor/

redução das disparidades regionais[2]. Nesse contexto, a previsão da demanda por profissionais de TI por região e modalidade de trabalho torna-se uma ferramenta estratégica fundamental para uma análise aprofundada deste mercado emergente.

Embora a literatura internacional já traga estudos sobre o uso de aprendizado de máquina para prever tendências no mercado de trabalho (SENTHURVELAUTAM, SENANAYAKE, 2023; GAJEWSKI et al., 2023), ainda são relativamente poucos os trabalhos que exploram esse tema no contexto regional brasileiro. As pesquisas existentes sobre o mercado nacional, em geral, não detalham diferenças regionais ou não utilizam modelos preditivos voltados especificamente para a abertura de vagas no setor de tecnologia (TESSARIN, MORCEIRO, 2022; BRITTO et al., 2023).

Nesse sentido, este estudo propõe uma análise preditiva da oferta de vagas de TI por região. Foi utilizado o LinkedIn como fonte de dados, pois a plataforma concentra cerca de 60% da força de trabalho ativa no Brasil[3], o que oferece um volume de dados dinâmico e representativo. São comparados três algoritmos de aprendizado de máquina: *Support Vector Regression* (SVR), redes neurais *feedforward* do tipo MLP (*Multi-Layer Perceptron*) e redes neurais *Long Short-Term Memory* (LSTM). A performance de cada modelo é analisada usando as métricas MAE (*Mean Absolute Error*) e MDA (*Mean Directional Accuracy*). A análise busca identificar o modelo mais eficaz para prever a abertura de vagas, contribuindo com uma análise relevante para a compreensão das dinâmicas atuais do setor tecnológico brasileiro.

3. OBJETIVOS

Objetivo Geral: Desenvolver e comparar modelos de aprendizado de máquina para realizar previsões regionais da abertura de vagas no setor de tecnologia da informação no Brasil, a fim de identificar a abordagem com melhor desempenho preditivo.

Objetivos Específicos

- Realizar a coleta de dados de anúncios de vagas de TI no LinkedIn por meio de técnicas de *web scraping*.

[2]https://hub.laboratoria.la/br/trabalho-remoto-fez-aumentar-em-50-as-contratacoes-de-ti-fora-do-sudeste-no-brasil

[3] https://www.cnnbrasil.com.br/economia/negocios/com-geracao-z-mais-de-60-da-forca-de-trabalho-nacional-tem-um-perfil-no-linkedin/

- Estruturar, limpar e pré-processar os dados coletados, tratando valores ausentes e transformando-os em séries temporais diárias para cada região brasileira (Norte, Nordeste, Centro-Oeste, Sudeste e Sul) e para a modalidade remota.
- Implementar, treinar e ajustar os hiperparâmetros de três modelos de aprendizado de máquina: SVR, rede neural *feedforward* do tipo MLP (*Multi-Layer Perceptron*) e rede neural LSTM.
- Avaliar e comparar o desempenho dos modelos utilizando as métricas MAE (*Mean Absolute Error*) e MDA (*Mean Directional Accuracy*), identificando a abordagem mais precisa e robusta para a previsão da demanda de vagas no contexto analisado.

4. METODOLOGIA

A metodologia do estudo consistiu em três fases principais e sequenciais: coleta de dados, pré-processamento e treinamento. O diagrama da Figura 1 explica, de forma resumida, o processo executado por este estudo.

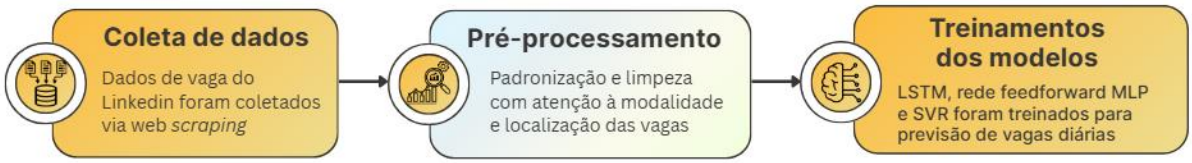

Figura 1. Diagrama da metodologia.

Inicialmente, foi desenvolvida uma ferramenta de *web scraping* para coletar dados públicos de vagas de TI do LinkedIn, definindo palavras-chave e uma estratégia de coleta diária para criar um *dataset* abrangente.

Na fase de pré-processamento, os dados foram padronizados, com atenção especial à localidade e modalidade das vagas. Vagas remotas foram tratadas como uma categoria separada por não terem região definida. Como *features* de entrada, foram usados o dia da semana e *lags* dos dias anteriores.

A etapa de treinamento envolveu o treinamento dos três algoritmos preditivos citados na seção 2. A escolha da LSTM e da SVR fundamenta-se em sua ampla utilização em problemas de séries temporais. A MLP, embora não seja projetada especificamente para esse tipo de dado, possui capacidade de capturar padrões não lineares, o que permite avaliar em que medida uma abordagem dessa natureza pode

superar as alternativas tradicionais. Para cada algoritmo, seis instâncias iguais foram individualmente treinadas (uma para cada região e uma para vagas remotas) para capturar as particularidades de cada série temporal. O desempenho foi avaliado para comparar a eficácia de cada técnica na previsão da demanda de vagas. Foram utilizadas duas métricas: o MAE (*Mean Absolute Error*), que mede o erro absoluto médio entre valores previstos e reais, e o MDA (*Mean Directional Accuracy*), que avalia a capacidade do modelo de prever as variações no conjunto de dados, isto é, a tendência da série temporal. Um MAE baixo indica que, em média, as previsões do modelo estão próximas dos valores reais. Já o MDA, quanto mais próximo de 1 (ou de 100%, em notação percentual), mostra maior capacidade do modelo em identificar corretamente a direção dos dados, isto é, se há aumento ou queda no momento adequado. No entanto, um MAE reduzido não garante um MDA elevado, assim como um MDA elevado não assegura previsões precisas em termos absolutos. Por isso, é essencial considerar ambas as métricas de forma complementar.

5. DESENVOLVIMENTO

O desenvolvimento executou todas as etapas descritas na seção 4. Os dados foram extraídos da página de busca de vagas do LinkedIn por meio de um algoritmo em Python que, para contornar a limitação de 1.000 resultados por busca da plataforma, utilizou 27 palavras-chave de TI em consultas booleanas[4], além de filtros personalizados. O processo foi executado diariamente ao longo de três meses, entre 1 de março de 2025 e 1 de junho de 2025, coletando apenas vagas das últimas 24 horas em uma prática de *web scraping* amparada legalmente por focar em dados públicos e não pessoais (CARDOSO, 2020). A base bruta inicial de 68.743 registros foi submetida a um processo de limpeza que removeu vagas fora do escopo, resultando em 66.863 vagas válidas. De cada vaga, foram coletados o título, a data de publicação, sua modalidade (presencial, híbrido ou remoto), a sua localização e competências associadas. Para a análise, os dados ausentes de modalidade e região foram imputados proporcionalmente, e as vagas remotas foram tratadas como uma categoria distinta.

Os dados foram agregados em seis séries temporais diárias, uma para cada região do Brasil e uma para vagas remotas, gerando seis grupos: Norte (0.9% das

[4] https://www.linkedin.com/help/linkedin/answer/a524335

vagas da base), Nordeste (5.2%), Centro-Oeste (3.7%), Sudeste (32.3%), Sul (20.3%) e Remoto (37.6%). Para tratar os dados, foram removidos *outliers* usando o método do intervalo interquartil (IQR), substituindo-os pela média da série, e as séries foram normalizadas usando a escala Min-Max para o intervalo [0, 1].

Após o tratamento, foi feito o treinamento dos modelos: rede *feedforward* MLP, rede LSTM e SVR. Como *features*, foram usados o dia da semana e a quantidade de vagas abertas em dias anteriores. Ambas as redes *feedforward* e LSTM foram configuradas com duas camadas de 128 neurônios, ativação ReLU, penalização L2 e *lags* de 4 dias. Já a SVR foi definida com *kernel* do tipo RBF (*Real Basis Function*) e parâmetro de regularização igual a 0.5, *epsilon* igual a 0.1 e *lags* de 3 dias. Os hiperparâmetros e a quantidade de lags foram selecionados por *Grid Search*, usando como critério a configuração com menor MAE médio entre as regiões. Cerca de 80% dos dados foram usados para treino e 20% para teste, e o desempenho dos modelos foi avaliado usando o MAE e o MDA.

6. RESULTADOS

Os modelos foram treinados e avaliados. As métricas geradas foram registradas para cada grupo analisado. A Tabela 1 detalha todas elas.

	Feedforward				**LSTM**				**SVR**			
	Treino		**Teste**		**Treino**		**Teste**		**Treino**		**Teste**	
Região	**MAE**	**MDA**	**MAE**	**MDA**	**MAE**	**MDA**	**MAE**	**MDA**	**MAE**	**MDA**	**MAE**	**MDA**
Remoto	68.3	73.3%	45.0	93.3%	91.8	65.0%	55.1	93.3%	74.2	77.0%	47.6	93.3%
Sudeste	76.8	76.7%	63.4	73.3%	92.8	60.0%	67.1	53.3%	69.1	77.0%	57.0	73.3%
Sul	35.4	77.6%	37.8	66.7%	44.9	74.1%	39.1	46.7%	36.0	77.0%	36.2	73.3%
Norte	0.54	76.6%	0.94	60.0%	1.01	66.0%	1.19	46.7%	0.67	78.4%	0.86	66.7%
Nordeste	8.32	86.0%	15.5	60.0%	9.47	84.2%	13.2	66.7%	8.77	86.2%	14.1	73.3%
Centro--Oeste	7.01	87.5%	11.3	71.4%	8.09	76.8%	9.83	66.7%	6.55	79.0%	12.3	64.3%
MÉDIA	**32.7**	**79.6%**	**30.4**	**80.7%**	**41.3**	**67.6%**	**42.5**	**62.2%**	**33.7**	**79.0%**	**29.2**	**74.0%**

Tabela 1. Desempenho dos modelos por região.

O desempenho médio nos dados de teste indica que o modelo *feedforward* alcançou o melhor equilíbrio entre erro absoluto médio (MAE) e acurácia direcional (MDA), superando o SVR na previsão de direção e o LSTM em ambos os aspectos. O SVR teve MAE ligeiramente menor que o *feedforward*, mas MDA inferior, enquanto o LSTM apresentou o pior desempenho geral, especialmente nas regiões mais complexas. Regionalmente, os melhores resultados foram obtidos para Remoto, Sudeste e Sul, enquanto Norte, Nordeste e Centro-Oeste, embora ocasionalmente apresentem métricas elevadas, contam com um volume de registros muito menor, o que limita a confiabilidade dessas estimativas. Sendo assim, as métricas sugerem que os modelos capturam melhor os padrões das regiões com maior volume e variabilidade de dados, enquanto o desempenho em áreas com menos registros deve ser interpretado com cautela para evitar conclusões precipitadas.

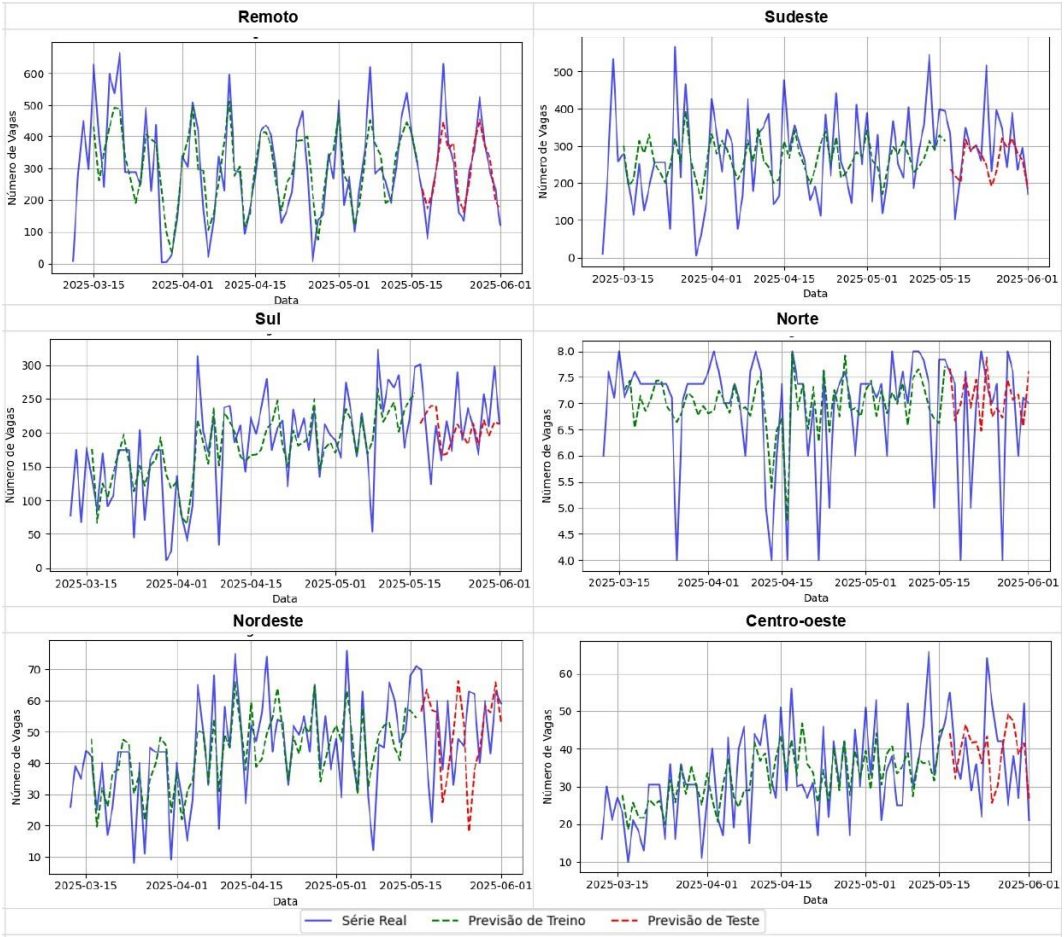

Figura 2. Previsão da rede neural *feedforward*.

A Figura 2 apresenta as previsões geradas pela rede neural *feedforward*. Os gráficos demonstram uma distinção clara no desempenho do modelo entre as

regiões. Para as séries temporais com maior volume de dados (Remoto, Sudeste e Sul) a previsão no conjunto de teste exibe uma forte correspondência com os dados observados, capturando de forma consistente a dinâmica dos picos e vales. Em contraste, nas demais regiões as previsões do modelo apresentam uma menor aderência à série real, com desvios mais acentuados. A capacidade de generalização do modelo, portanto, é diretamente influenciada pela qualidade e amplitude da série temporal utilizada no treinamento, o que é um fator relevante para a aplicação do modelo em diferentes contextos geográficos.

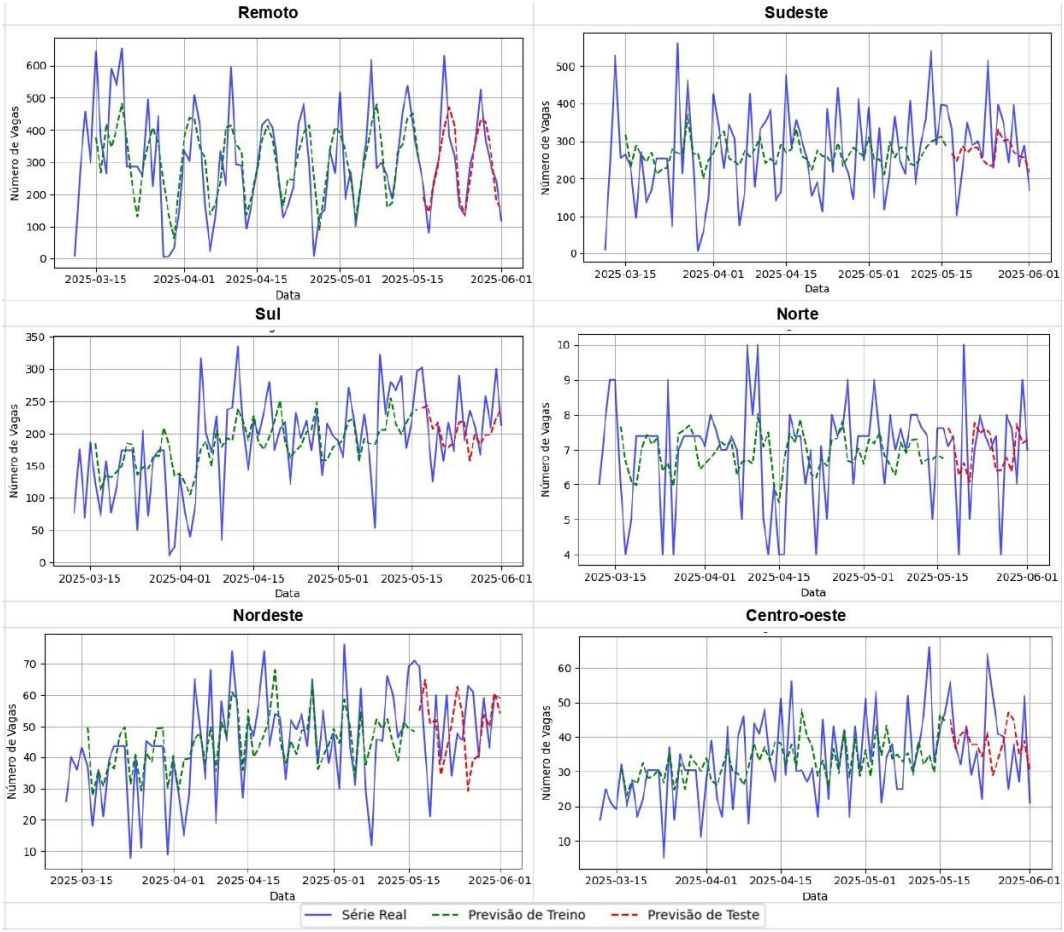

Figura 3. Previsão da rede neural LSTM.

Por outro lado, os gráficos das previsões da LSTM, exibidos na Figura 3, mostram que o modelo tem dificuldade em replicar a volatilidade das séries temporais em diferentes regiões. No conjunto de dados, a linha de previsão tende a se suavizar, falhando em capturar os picos e oscilações observados nos dados reais, especialmente nas regiões Sudeste e Sul. É plausível que isto esteja relacionado à própria arquitetura da LSTM, que privilegia a retenção de informações de longo prazo. Consequentemente, o modelo pode estar dando peso excessivo a

padrões históricos menos relevantes para a dinâmica recente, resultando em previsões menos sensíveis a variações de curto prazo. Essa limitação indica que redes de memória de longo prazo podem não ser ideais para séries temporais com alta frequência e grande dinamismo.

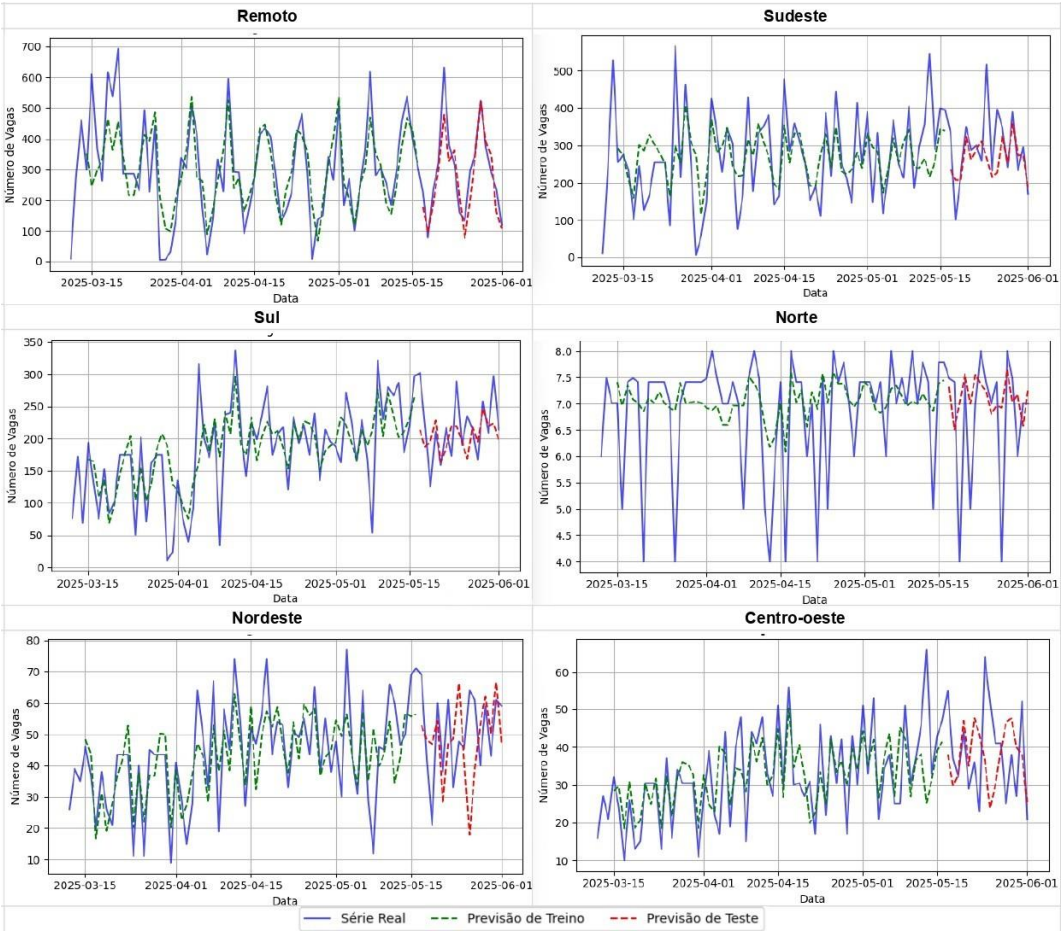

Figura 4. Previsão do modelo SVR.

As previsões do modelo SVR são exibidas na Figura 4. Percebe-se que o modelo SVR e a rede neural *feedforward* obtiveram desempenhos muito próximos. Isso se justifica pelo uso do kernel RBF no SVR, que confere ao modelo a capacidade de capturar padrões não lineares presentes nos dados. Essa característica o torna comparável, em termos de expressividade, a uma rede neural *feedforward*, o que explica a proximidade dos desempenhos observados. Embora tenha apresentado o menor MAE, indicando uma maior precisão na minimização do erro, a rede neural *feedforward* superou-o no MDA, mostrando uma melhor capacidade de prever a tendência geral dos dados. Essa distinção crucial revela um *trade-off*: enquanto o SVR é a melhor opção para minimizar o erro de previsão, a

rede neural *feedforward* se destaca quando a preocupação principal é capturar a direcionalidade e a tendência dos dados.

Sendo assim, de acordo com os resultados obtidos, a rede neural *feedforward* demonstrou ser a mais robusta. Ela apresentou o melhor equilíbrio entre a minimização do erro (MAE) e a acurácia direcional (MDA), superando os outros modelos e mostrando uma capacidade superior de generalização. O modelo SVR, embora tenha obtido um MAE ligeiramente menor, teve um MDA inferior, indicando uma limitação na previsão de tendências. Já a rede LSTM foi a que apresentou o pior desempenho, falhando em capturar a volatilidade das séries. Assim, conclui-se que a rede neural *feedforward* se apresenta como a alternativa mais adequada e equilibrada para a tarefa de previsão considerada neste estudo, ainda que o SVR também tenha demonstrado desempenho satisfatório.

7. CONSIDERAÇÕES FINAIS

Com base na análise realizada, este estudo evidenciou a viabilidade da aplicação de modelos de aprendizado de máquina para a previsão da demanda por vagas em tecnologia da informação no Brasil em nível regional. A metodologia adotada, que envolveu a coleta sistemática de dados provenientes do LinkedIn e o treinamento de modelos específicos para cada região, mostrou-se adequada para a identificação de padrões complexos. Os resultados obtidos indicaram que tanto a rede neural *feedforward* quanto o SVR apresentam capacidade de tratar séries temporais, oferecendo subsídios relevantes para a compreensão da dinâmica do mercado nacional.

Com base nos resultados, o mercado de TI no Brasil exibe uma forte concentração de vagas, com a modalidade remota, a região Sudeste e a Sul respondendo, em conjunto, por mais de 90% das oportunidades analisadas. Essa expressiva centralização consolida essas três categorias como os principais polos de demanda por profissionais no setor. Em contraste, as demais regiões apresentam uma participação muito menor no mercado nacional de tecnologia, revelando uma acentuada desigualdade regional.

No que concerne ao desempenho, a rede neural *feedforward* apresentou-se como o modelo mais equilibrado, destacando-se pelos valores de MAE e MDA, superando o SVR na previsão de direcionalidade e a rede LSTM em ambos os critérios. O SVR, embora tenha alcançado menor MAE, mostrou menor precisão na detecção de tendências. A rede LSTM, por sua vez, apesar de sua arquitetura

voltada ao tratamento de dados temporais, apresentou limitações em relação à volatilidade, com dificuldades em capturar picos e oscilações relevantes.

Dessa forma, este trabalho contribui para a literatura ao explorar a previsão regional da demanda no mercado de TI brasileiro. Para investigações futuras, recomenda-se a ampliação da base de dados para um horizonte temporal mais extenso, bem como a incorporação de variáveis externas, tais como indicadores econômicos, a fim de aprimorar a capacidade preditiva dos modelos.

8. FONTES CONSULTADAS

ABES: Associação Brasileira das Empresas de Software. **Mercado brasileiro de software: panorama e tendências**, 2024. Disponível em: https://abes.org.br/dados-do-setor/. Acesso em 18 set. 2024.

BRITTO, J.; URRACA-RUIZ, A.; FERRAZ, J.; TORRACCA, J.; LACERDA, H. El impacto de la digitalización sobre empleo y las habilidades por estadios de adopción en Brasil y Argentina. **Revista Brasileira de Inovação**, 2023.

CARDOSO, O. V. **O web scraping viola a proteção de dados pessoais?** JusBrasil, 2021**.** Disponível em: https://www.jusbrasil.com.br/artigos/o-web-scraping-viola-a-protecao-de-dados-pessoais/1152362639. Acesso em 22 dez. 2024.

SENTHURVELAUTHAM, Sharanaja; SENANAYAKE, Nipuna. A machine learning-based job forecasting and trend analysis system to predict future job markets using historical data. **2023 IEEE 8TH INTERNATIONAL CONFERENCE FOR CONVERGENCE IN TECHNOLOGY (I2CT), 8.,** 2023**.** p. 1-7. DOI: 10.1109/I2CT57861.2023.10126233.

TESSARIN, Milene; MORCEIRO, Paulo César. Labour market transformations in the era of new technologies: an analysis by regions, gender and industries in Brazil. **Economic Research Southern Africa,** 2022. DOI: 10.54223/uniwitwatersrand-10539-33455

GAJEWSKI, Przemysław; ČULE, Bojan; RANKOVIC, Nikola. Unveiling the Power of ARIMA, Support Vector and Random Forest Regressors for the Future of the Dutch Employment Market. **Journal of Theoretical and Applied Electronic Commerce Research**, v. 18, n. 3, p. 1365-1403, 2023. DOI: 10.3390/jtaer18030069.