CONIC SEMESP

25º Congresso Nacional de Iniciação Científica

TÍTULO: ANÁLISE COMPARATIVA DE ALGORITMOS DE MACHINE LEARNING PARA A PROJEÇÃO DE ÁREAS EM CRESCIMENTO NO SETOR DE TECNOLOGIA DA INFORMAÇÃO NO BRASIL

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): ISABELLE YASMIN DE ARAUJO MOREIRA

ORIENTADOR(ES): CARLOS ALEXANDRE SILVA

COLABORADOR(ES): DANILO BATISTA LIMA

Realização:
SEMESP

IES parceiras:
PUC CAMPINAS

CATEGORIA CONCLUÍDO

1. RESUMO

O presente estudo propõe e avalia uma metodologia híbrida de machine learning para a predição da demanda por profissionais no setor de Tecnologia da Informação (TI) no Brasil. Utilizando um conjunto de dados extraído da plataforma LinkedIn por meio de web scraping, a abordagem consiste em duas etapas. Primeiramente, aplica-se o algoritmo não supervisionado K-means para segmentar as vagas em clusters representativos de áreas de atuação, como Desenvolvimento Web, Mobile, Dados, QA e Infraestrutura. Subsequentemente, três modelos de aprendizado supervisionado são empregados para prever a abertura de novas vagas em cada área de atuação: a rede neural Feedforward Multi-Layer Perceptron (MLP), a rede neural Long Short-Term Memory (LSTM) e o algoritmo eXtreme Gradient Boosting (XGBoost). Os resultados indicam a predominância de três áreas consolidadas no cenário nacional, sendo elas desenvolvimento web Fullstack, Backend e Mobile. Além disso, o XGBoost foi o modelo mais performático, o que se evidencia tanto em métricas de precisão quanto na sua capacidade de capturar os picos e vales da demanda de vagas. Conclui-se que a abordagem híbrida oferece um framework robusto e granular para a análise e previsão da demanda no dinâmico mercado de TI.

2. INTRODUÇÃO

A crescente complexidade e ascensão do setor de Tecnologia da Informação (TI) tornam cada vez mais desafiadora a tarefa de antecipar as demandas por profissionais qualificados (ABES, 2024). Embora a literatura internacional tenha avançado na aplicação de técnicas de machine learning para prever a demanda de trabalho, com modelos como LSTM (Long-Short Term Memory) e XGBoost (eXtreme Gradient Boosting) mostrando resultados promissores (DAWSON, 2020; KIM, 2025), esses estudos frequentemente tratam os mercados de forma agregada. Essa abordagem é uma limitação significativa no contexto brasileiro, onde a escassez de talentos em TI é amplamente reconhecida, mas raramente analisada com rigor preditivo e granular.

Instituições como a Brasscom[1] projetam um déficit de centenas de milhares de profissionais até 2025, mas não especificam quais áreas serão mais impactadas. Estudos nacionais, por sua vez, tendem a focar em outros setores, como varejo e logística (LINHARES, 2022; PINTO, 2021), deixando o mercado de TI subexplorado.

Nesse cenário, o LinkedIn emerge como uma fonte de dados rica, dinâmica e segmentada sobre o mercado de trabalho (FLORENTINA, 2020). Por meio de técnicas de web scraping, é possível extrair e estruturar informações de vagas em larga escala, transformando-as em séries temporais que refletem a evolução da demanda em nichos específicos como desenvolvimento, ciência de dados, segurança e cloud computing. Este estudo propõe utilizar esses dados para criar modelos preditivos segmentados, oferecendo uma visão detalhada das tendências futuras do mercado de TI no Brasil.

3. OBJETIVOS

Objetivo Geral: Desenvolver e comparar o desempenho de diferentes modelos de machine learning para prever, com granularidade por área de atuação, a demanda futura por profissionais no setor de Tecnologia da Informação brasileiro.

Objetivos Específicos:

- Coletar um grande volume de dados de anúncios de vagas de TI no Brasil a partir da plataforma LinkedIn.
- Aplicar técnicas de aprendizado não supervisionado (K-means) para segmentar as vagas em clusters representativos de diferentes áreas de atuação.
- Implementar e treinar três modelos de aprendizado supervisionado, MLP (Multi-Layer Perceptron), XGBoost e LSTM, para prever a quantidade de vagas em cada cluster.
- Avaliar e comparar a performance dos modelos preditivos utilizando métricas como Erro Médio Absoluto (MAE) e Acurácia Direcional Média (MDA).

4. METODOLOGIA

A metodologia do projeto foi dividida em quatro etapas: coleta de dados de vaga do LinkedIn por web scraping, pré-processamento, segmentação de mercado com

[1] https://www.unialfa.com.br/brasil-precisa-de-quase-800-mil-profissionais-de-ti-ate-2025-afirma-estudo/

K-means e treinamento de modelos preditivos: MLP, XGBoost e LSTM. A Figura 1 ilustra o fluxo metodológico adotado.

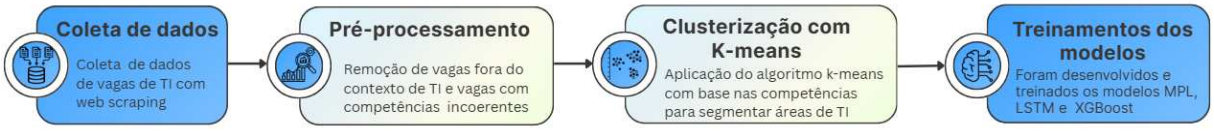

Figura 1. Diagrama da metodologia.

Primeiro, realizou-se a coleta e o tratamento de dados, com extração automatizada de vagas de tecnologia no LinkedIn e posterior limpeza para garantir qualidade. Em seguida, aplicou-se o algoritmo K-means sobre as competências descritas nas vagas para segmentar o mercado em áreas técnicas; o número de clusters foi definido empiricamente, e os grupos foram rotulados com base em palavras-chave predominantes.

Depois dessas etapas, foi conduzida a modelagem preditiva supervisionada, na qual os três modelos, MLP, XGBoost e LSTM, foram treinados e avaliados. A Feedfoward foi incluída como baseline de rede neural simples, capaz de capturar relações não lineares básicas; o XGBoost foi escolhido pela sua eficiência em lidar com dados heterogêneos e interações complexas; e a LSTM pela sua arquitetura sequencial, adequada à captura de padrões temporais e dependências de longo prazo. O desempenho de cada modelo foi mensurado por meio do MAE (Erro Médio Absoluto), que avalia a magnitude média do erro absoluto das previsões, e do MDA (Acurácia Direcional Média), que mede a capacidade do modelo de prever corretamente a direção das variações ao longo do tempo, isto é, se houve aumento ou queda. Desse modo, essas métricas se complementam ao avaliar simultaneamente a precisão das previsões e a captura das tendências do mercado.

5. DESENVOLVIMENTO

O desenvolvimento seguiu as etapas descritas na metodologia. A coleta de dados foi realizada por meio de scraping automatizado no LinkedIn, utilizando Selenium para navegação dinâmica e Beautiful Soup para extração de elementos HTML. Devido à limitação da plataforma, que restringe resultados a 1000 vagas por busca, foram empregadas consultas booleanas[2] combinando cargos, tecnologias e filtros temporais, sendo usado o das últimas 24h. A execução do processo ocorreu

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

de forma automatizada e com frequência diária, compreendendo o trimestre de março a maio de 2025. Após a remoção de vagas irrelevantes e ruídos, obteve-se um total de 66.863 vagas válidas. Todo o processo de coleta respeitou os limites éticos e legais, em conformidade com a Lei Geral de Proteção de Dados (BRASIL, 2018), adotando os princípios de finalidade, necessidade, privacidade e segurança (CARDOSO, 2021). Para cada vaga, foi coletado o título, data da publicação e competências requisitadas.

A etapa de segmentação utilizou o algoritmo K-means com k=10 definido empiricamente, visando a interpretabilidade dos resultados. Testes prévios com os métodos elbow (cotovelo) e silhouette (silhueta) indicaram um k superior a 20, o que dificultaria a análise individual. A clusterização foi aplicada sobre o vetor de competências de cada vaga para identificar aquelas com requisitos similares, e os 10 clusters gerados são exibidos na Figura 2.

Figura 2. Clusters do K-means.

Os clusters foram rotulados por meio de análise quantitativa das palavras-chave e títulos das vagas. Ao final desse processo, as áreas de atuação identificadas na base de dados foram: Desenvolvimento Fullstack (34.6%), Desenvolvimento Backend (23.3%), Desenvolvimento Mobile (16.5%), Dados e IA (6.4%), Desenvolvimento Frontend (4.4%), QA e testes (4%), Infraestrutura e suporte (10.7%).

Em seguida, séries temporais diárias foram geradas para cada área, representando o número de vagas publicadas por dia. O pré-processamento incluiu detecção de outliers via intervalo interquartil (IQR), substituição por média, normalização com Min-Max Scaling e engenharia de atributos, com criação de lags e codificação do dia da semana. Os três modelos preditivos foram implementados: uma rede neural Feedforward do tipo MLP (com duas camadas de 130 neurônios e função de ativação ReLU), XGBoost (com profundidade máxima de três nós e regularização L2) e LSTM (com duas camadas bidirecionais de 64 neurônios). Para ambas as redes foram imputados lags de 7 dias e, para o XGBoost, de 4 dias. A otimização de hiperparâmetros foi realizada por Grid Search, com divisão 80/20 para treino e teste. A avaliação combinou as métricas MAE e MDA, medindo o erro médio e a direção das variações, respectivamente.

6. RESULTADOS

A performance dos modelos foi avaliada separadamente para cada um dos sete grupos. A Tabela 1 resume todas as métricas.

	Feedforward MLP				**LSTM**				**XGBoost**			
	Treino		**Teste**		**Treino**		**Teste**		**Treino**		**Teste**	
	MAE	**MDA (%)**	**MAE**	**MDA (%)**	**MAE**	**MDA (%)**	**MAE**	**MDA (%)**	**MAE**	**MDA (%)**	**MAE**	**MDA (%)**
Dados/IA	7.07	83.3	13.2	60	9.17	70.37	12.9	60	5.67	85.9	12.7	73.3
Infra. e Suporte	15.1	89.2	29.7	75	24.4	82.1	31.4	75	13	93.1	24.4	84.6
Mobile	24	83.9	26.3	80	34.8	62.5	31.4	66.7	15.7	89.8	35.9	73.3
QA/Testes	3.31	85.1	9	57.1	7.97	68.5	9.2	57.1	4.43	85.7	7.75	73.3
Backend	35	85.7	32.1	80	51	80.3	40.8	66.7	22.9	88.1	37.3	73.3
Frontend	4.5	79.2	7.7	64.2	6.93	66	6.69	71.4	3.68	85.7	8.07	78.6
Fullstack	51.6	76.7	55.7	60	67.5	69.6	50.1	86.7	42	86.4	56	86.7
MÉDIA	20	83.3	**24.8**	**62.6**	28.8	71.2	**37.3**	**69**	15.34	85.7	**26.1**	**77.5**

Tabela 1. Desempenho dos modelos por área de atuação.

Em primeira análise, o XGBoost apresentou o melhor desempenho geral, com menor MAE e maior MDA no conjunto de teste, evidenciando maior acurácia e capacidade preditiva. O modelo Feedforward teve desempenho razoável, mas menor generalização, com aumento do MAE e redução do MDA nos testes. O LSTM apresentou desempenho intermediário, com métricas entre os demais modelos.

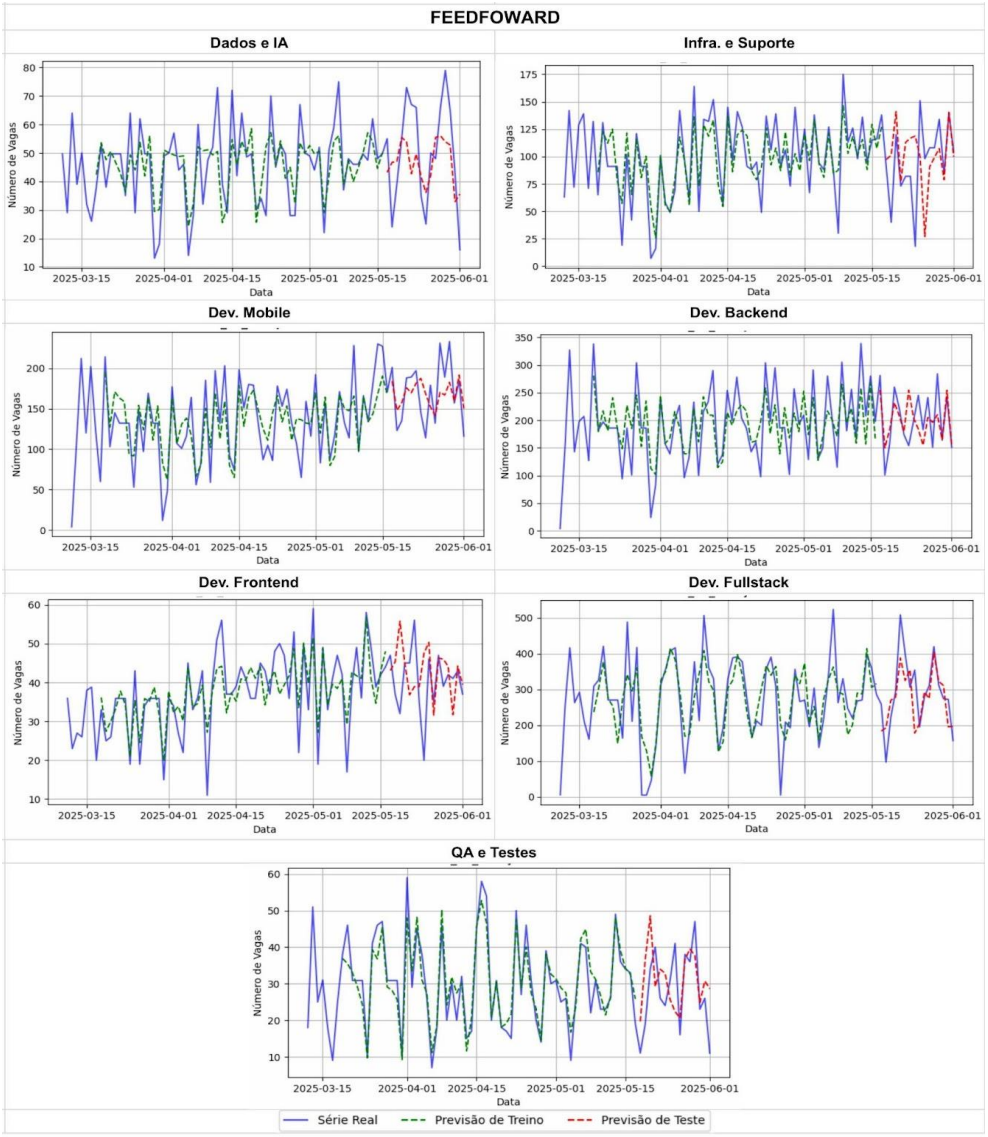

Figura 3. Previsões realizadas pela rede Feedforward.

A Figura 3 é referente às previsões geradas pelo modelo Feedforward. O modelo Feedforward apresentou MAE de treino de 20 e MDA de 83.3%, métricas que, no teste, caíram para 24.8 e 62.6%, respectivamente. Essa queda indica dificuldade de generalização, parcialmente atribuível à arquitetura do modelo. Apesar de receber lags de 7 dias, a ausência de camadas recorrentes impede a construção de memória de longo prazo, fazendo com que o MLP processe cada

conjunto de lags de forma independente. Como resultado, suas previsões seguem a tendência geral da série, mas não capturam picos e vales complexos, explicando sua instabilidade e menor capacidade de generalização.

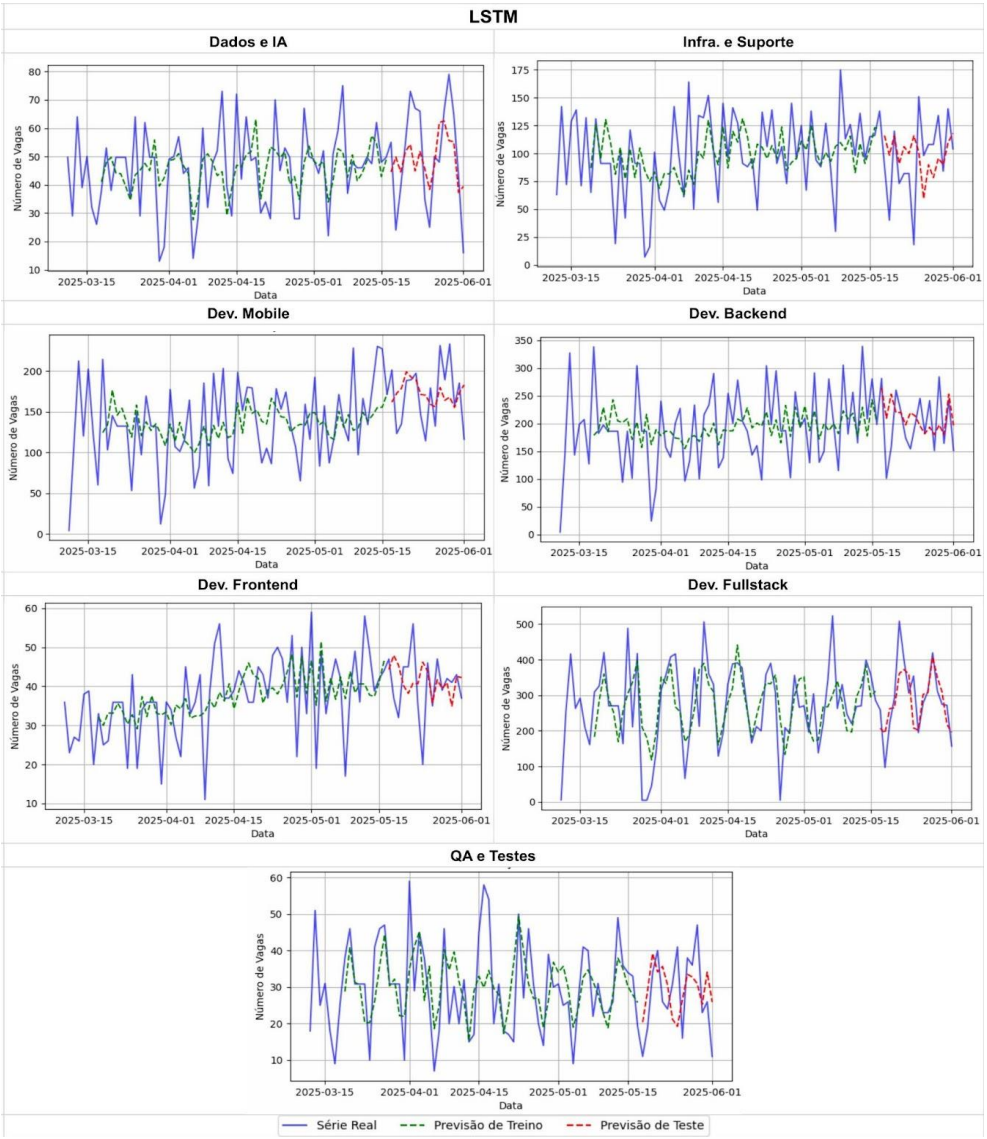

Figura 4. Previsões realizadas pela rede LSTM.

As previsões do modelo LSTM são exibidas na Figura 4. A rede apresentou desempenho intermediário, com MAE de treino de 28.8 e MDA de 71.2%, mudando para MAE 37.3 e MDA 69% no teste. Apesar de ter um MAE maior que o modelo Feedforward, a estabilidade relativa do MDA indica uma capacidade de generalização ligeiramente maior que o Feedforward, graças à sua arquitetura recorrente. Este fator permite capturar tendências e variações da série temporal de forma mais consistente em comparação com Feedforward, apesar de não superar o XGBoost. Entretanto, ao se apoiar fortemente em padrões históricos antigos, a

LSTM pode ponderar excessivamente informações de menor relevância, fator que explica sua dificuldade em prever picos e vales acentuados como se percebe pelos gráficos.

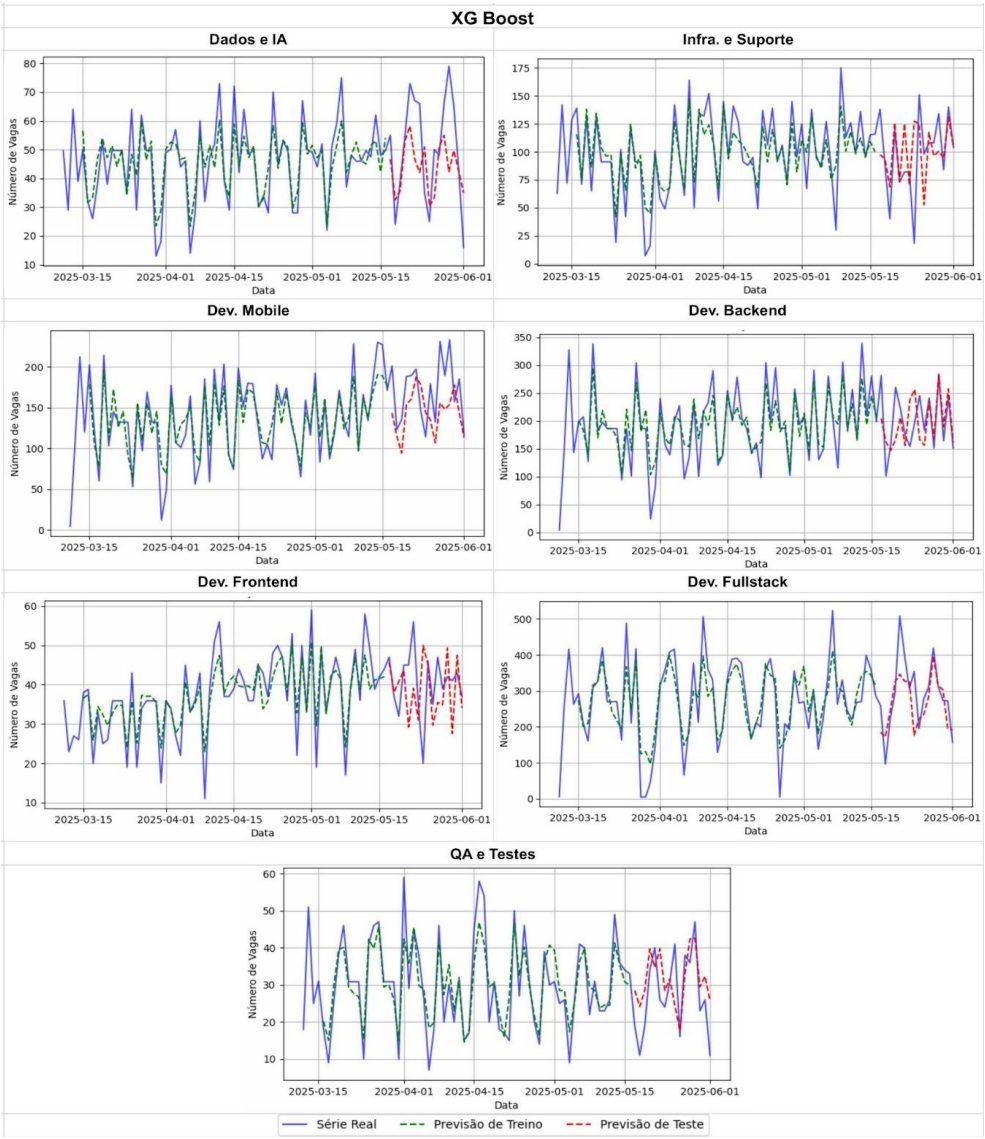

Figura 5. Previsões realizadas pelo XGBoost.

Por fim, as previsões do modelo XGBoost são apresentadas na Figura 5. O XGBoost apresentou o melhor desempenho geral, com MAE de 26.1 e MDA de 77.5% no teste. Sua performance se deve à sua arquitetura de boosting, que constrói uma sequência de árvores de decisão onde cada modelo corrige os erros do anterior. Esse processo o torna particularmente eficaz em capturar padrões complexos e flutuações, explicando sua capacidade de prever picos e vales com maior precisão do que MLP e LSTM. Comparado à LSTM, o XGBoost não sofre do

mesmo viés por padrões históricos antigos, equilibrando melhor a relevância dos dados passados.

Em síntese, o XGBoost apresentou o melhor desempenho geral, captando tendências, picos e vales de forma mais precisa. A LSTM mostrou melhor generalização que o MLP, beneficiando-se da memória de longo prazo. O MLP, por sua vez, seguiu apenas a tendência geral da série e teve dificuldade em capturar variações mais complexas, resultando em menor estabilidade e capacidade de generalização.

7. CONSIDERAÇÕES FINAIS

Este estudo demonstrou a viabilidade de uma abordagem híbrida de machine learning para a análise segmentada e predição da demanda no mercado brasileiro de Tecnologia da Informação (TI), agregando uma nova perspectiva aos estudos da área. A metodologia de coleta possibilitou a construção de um dataset robusto, enquanto a clusterização com K-means foi fundamental para desagregar o mercado, permitindo previsões específicas por nicho e superando as análises tradicionais generalistas. A análise do mercado de TI no Brasil revela uma forte concentração de vagas na área de desenvolvimento. Especificamente, as especialidades de desenvolvimento web, com ênfase em Fullstack e Backend, e o desenvolvimento Mobile se destacam como os principais impulsionadores da demanda por profissionais. Adicionalmente, foram identificados segmentos emergentes de grande relevância, como Dados e Inteligência Artificial (IA), Frontend, Infraestrutura/Suporte Técnico e Controle de Qualidade (QA).

Quanto à performance, o modelo XGBoost emergiu como o mais eficiente, apresentando um equilíbrio entre precisão e a capacidade de capturar as flutuações da série temporal. O LSTM obteve um desempenho intermediário, utilizando sua memória de longo prazo para seguir tendências de maneira consistente, embora com limitações na previsão de eventos extremos. O MLP, por sua vez, registrou o desempenho mais modesto, acompanhando apenas a tendência geral da série e demonstrando menor estabilidade, o que evidencia as restrições de sua arquitetura sem memória temporal.

Para trabalhos futuros, visa-se a expansão da coleta de vagas diárias por um período superior a seis meses, visando a construção de uma base de dados mais robusta e com maior volume de informações em cada cluster. Adicionalmente, sugere-se a incorporação de variáveis exógenas, como indicadores econômicos,

para enriquecer os modelos preditivos e a expansão da análise para incluir a previsão de competências específicas dentro de cada segmento de atuação. Tais avanços, somados à abordagem e aos resultados iniciais, consolidam um framework robusto para o monitoramento e a previsão estratégica do dinâmico mercado de TI no Brasil.

8. FONTES CONSULTADAS

ABES: ASSOCIAÇÃO BRASILEIRA DAS EMPRESAS DE SOFTWARE. **Mercado brasileiro de software: panorama e tendências.** São Paulo, 2024. Disponível em: https://abes.org.br/dados-do-setor/. Acesso em 18 set. 2024.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF, 14 ago. 2018.

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.

DAWSON, N. et al. Predicting skill shortages in labor markets: a machine learning approach. **IEEE INTERNATIONAL CONFERENCE ON BIG DATA**, Atlanta, 2020. DOI: 10.1109/BigData50022.2020.9377773.

FLORENTINA, M. Web data extraction with robot process automation: Study on linkedin web scraping using uipath studio. **Annals of 'Constantin Brancusi' University of Targu-Jiu. Engineering Series**, 2020.

KIM, K. Forecasting Labor Demand: Predicting JOLT Job Openings using Deep Learning Model. 2025. DOI: 10.48550/arXiv.2503.19048

LINHARES, E. C. L. **Previsão de demanda através de redes neurais em um ambiente de omnicanalidade no varejo.** Tese (Mestrado em Engenharia de Produção e Sistemas). Pontifícia Universidade Católica de Goiás, Goiânia, 2023.

PINTO, I. B.; ZORZO, A.; SCHLÜTER, M. R. Aplicação do aprendizado de máquinas na previsão de demanda. **Congresso de Logística das Faculdades de Tecnologia do Centro Paula Souza – FatecLog.** Mogi das Cruzes, 2021