APLICAÇÃO E COMPARAÇÃO DE MODELOS DE MACHINE LEARNING PARA A IDENTIFICAÇÃO DE ÁREAS DE CRESCIMENTO EM TI NO BRASIL

Isabelle Yasmin de Araujo Moreira[1], Danilo Batista Lima[2], Carlos Alexandre Silva[3]

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.

Palavras-chave: Machine Learning. Previsão de Demanda. Mercado de TI.

1 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 et al., 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.

[1] Bacharelado em Sistemas de Informação, Campus Sabará, IFMG
[2] Bacharelado em Sistemas de Informação, Campus Sabará, IFMG
[3] Doutor em Ciência da Computação e Matemática Computacional, Campus Sabará, IFMG

Instituições como a Associação das Empresas de Tecnologia da Informação e Comunicação e de Tecnologias Digitais (Brasscom) projetam um déficit de centenas de milhares de profissionais até 2025 (DINO, 2022), mas não especificam quais áreas serão mais impactadas. Estudos nacionais, por sua vez, tendem a focar em outros setores, como logística (PINTO, 2021). Nesse cenário, o LinkedIn emerge como uma fonte de dados rica e dinâmica. Por meio de técnicas de web scraping, é possível extrair informações de vagas em larga escala, transformando-as em séries temporais que refletem a evolução da demanda em nichos específicos.

Para atingir esse objetivo, a metodologia adotada nesse estudo parte da coleta de dados de vagas no LinkedIn por meio de web scraping. Em seguida, aplica-se o algoritmo de aprendizado não supervisionado K-means para segmentar o mercado em áreas de atuação distintas. Por fim, são desenvolvidos e comparados três modelos de aprendizado supervisionado: uma rede neural Feedforward Multi-Layer Perceptron (MLP), uma rede LSTM e um algoritmo XGBoost, com o objetivo de prever a demanda futura em cada um dos segmentos identificados.

2 DESENVOLVIMENTO

2.1 Metodologia

A metodologia foi organizada em quatro etapas: coleta, pré-processamento, segmentação e treinamento dos modelos. Na coleta, utilizou-se scraping automatizado do LinkedIn, com Selenium para navegação e Beautiful Soup para extração de HTML. Para contornar a limitação de 1000 resultados por busca, foram aplicadas consultas booleanas com filtros temporais. A execução do processo ocorreu de forma automatizada e com frequência diária, compreendendo o trimestre de março a maio de 2025. É importante ressaltar que a coleta esteve em conformidade com a Lei Geral de Proteção aos Dados (BRASIL, 2018), uma vez que não houve coleta de dados pessoais. Após a remoção de vagas irrelevantes, obteve-se um total de 66.863 vagas válidas. Em seguida, aplicou-se o algoritmo K-means (com k=10) sobre os vetores de competências, agrupando vagas segundo habilidades em comum (apêndice A). A análise quantitativa de palavras-chave e títulos das vagas permitiu rotular os clusters em sete áreas de atuação distintas: Desenvolvimento Fullstack (34,6%), Backend (23,3%), Mobile (16,5%), Infraestrutura e Suporte (10,7%), Dados e IA (6,4%), Frontend (4,4%) e QA/Testes (4,0%).

Em seguida, foram construídas séries temporais diárias para cada área, representando o número de vagas publicadas. Nesta etapa, realizou-se a detecção e tratamento de outliers, a normalização Min-Max e a engenharia de atributos, na qual foram usados lags e codificação do dia da semana. Por fim, implementaram-se três modelos preditivos supervisionados: rede neural MLP, XGBoost e rede recorrente LSTM. Os hiperparâmetros foram otimizados por Grid Search, e o desempenho de cada modelo foi mensurado por MAE (Erro Médio Absoluto) e MDA (Acurácia Direcional Média). Cerca de 20% dos dados foram usados para teste.

2.2 Resultados e Discussão

A performance dos modelos preditivos foi avaliada conforme resume o Quadro 1. Os dados são referentes ao desempenho médio dos modelos no conjunto de teste.

Feedforward MLP		LSTM		XGBoost	
MAE	MDA (%)	MAE	MDA (%)	MAE	MDA (%)
24,8	62,6	37,3	69,0	26,1	77,5

Quadro 1- Métricas obtidas pelo treinamento dos algoritmos de machine learning.
Fonte: Elaborado pelos autores (2025).

Essas métricas indicam que o XGBoost apresentou o melhor desempenho no conjunto de teste, com MAE de 26,1 e MDA de 77,5%. O LSTM apresentou desempenho intermediário, com dificuldade em prever variações abruptas, refletindo sua ênfase em padrões históricos mais antigos. Já o MLP Feedforward obteve a performance mais modesta, com MAE de 24,8 e MDA de 62,6%, devido à ausência de memória de longo prazo em sua arquitetura.

Os gráficos de previsão das três áreas com maior porcentagem de dados (apêndice B) reforçam a análise das métricas. Observa-se que o XGBoost consegue acompanhar de forma mais precisa tanto os picos quanto os vales das séries temporais, refletindo sua capacidade de capturar padrões complexos. Isso se justifica pela sua arquitetura de boosting, que constrói árvores de decisão sequenciais para corrigir erros anteriores. O LSTM apresenta desempenho intermediário, com boa aderência às tendências gerais, mas dificuldade em antecipar flutuações abruptas, devido à sua memória que privilegia padrões históricos em detrimento de variações de curto prazo. Já o MLP Feedforward mostra maior dispersão em relação à série real,

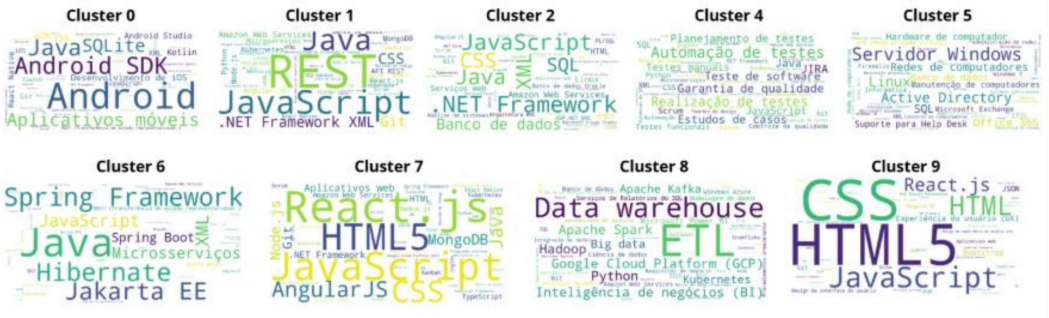
especialmente em períodos de variações intensas, evidenciando sua limitação em lidar com dependências temporais.

3 CONCLUSÃO

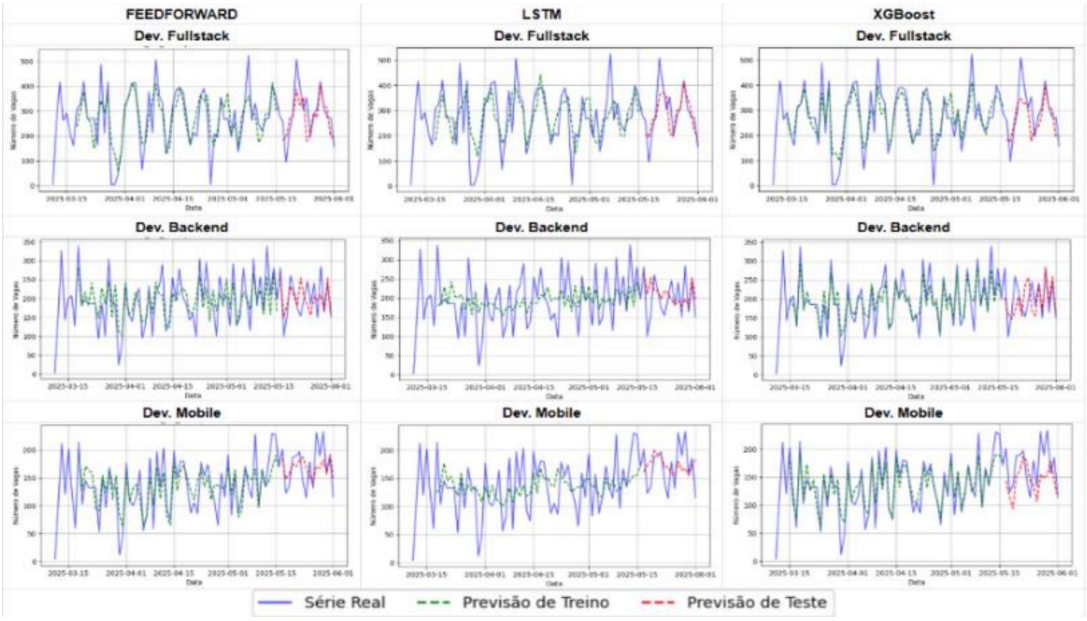
Este estudo demonstrou a viabilidade de uma abordagem híbrida de machine learning para a análise segmentada e predição da demanda no mercado brasileiro de TI. A metodologia de coleta e clusterização foi fundamental para desagregar o mercado, permitindo previsões específicas por nicho e superando as análises generalistas tradicionais. A segmentação com K-means revelou uma forte concentração de vagas nas áreas de desenvolvimento Fullstack, Backend e Mobile, além de segmentos emergentes de grande relevância, como Dados e IA, e QA. Em termos de performance, o modelo XGBoost revelou-se como o mais eficiente, apresentando um equilíbrio superior entre precisão (MAE) e capacidade de prever a direção das variações do mercado (MDA). O LSTM obteve um desempenho intermediário, enquanto o MLP registrou a performance mais modesta, evidenciando as limitações de arquiteturas sem memória temporal para este tipo de problema.

Sendo assim, a principal contribuição deste trabalho é fornecer uma análise robusta do mercado de TI segmentada por áreas de atuação. Para trabalhos futuros, sugere-se a expansão da coleta de dados para um período superior a seis meses, visando construir uma base de dados mais robusta. Adicionalmente, a incorporação de variáveis exógenas, como indicadores econômicos, e a previsão de competências específicas dentro de cada segmento poderiam enriquecer os modelos. Os autores agradecem o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo fomento que tornou viável esta pesquisa.

APÊNDICE A – Nuvens de palavras de competência dos Clusters gerados pelo algoritmo K-means



APÊNDICE B – Gráficos das previsões para as áreas Fullstack, Backend e Mobile



REFERÊNCIAS

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.

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.

DINO. Tecnologia demandará cerca de 800 mil profissionais até 2025. **Valor Econômico**, 26 set. 2022. Disponível em: https://valor.globo.com/patrocinado/dino/noticia/2022/09/26/tecnologia-demandara-cerca-de-800-mil-profissionais-ate-2025.ghtml. Acesso em: set. 2025.

KIM, K. **Forecasting Labor Demand: Predicting JOLT Job Openings using Deep Learning Model**. 2025. Preprint, submetido 24 de Março de 2025. https://arxiv.org/abs/2503.19048

PINTO, Isadora Boldrini. 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