

Plataforma Jacarandá: Integração de Dados com Inteligência Artificial para Apoio à Gestão Acadêmica

Ricardo Marçal de Andrade Nascimento¹, Bruno de Araújo Alves,
Jamilly Costa Ribeiro, João Gabriel Prado, Luiz Fernando de Oliveira Corrêa,
Victor Augusto Guimarães, Daniel Xavier de Sousa²

¹Instituto Federal de Educação, Ciência e Tecnologia de Goiás (IFG)
Campus Anápolis – Goiás – Brasil

ricardomarc102@hotmail.com¹, daniel.sousa@ifg.edu.br²

Abstract. *The integration and management of academic data in higher education institutions still represent a significant challenge due to the heterogeneity of data sources and the large volume of information, especially in multi-campus environments. Although Business Intelligence strategies facilitate data access and analysis, there are still limitations regarding the integration of advanced Artificial Intelligence techniques. In this context, this work presents the Jacarandá Platform, a use case that integrates institutional data and combines visualization with Machine Learning and Natural Language Processing.*

Resumo. *A integração e gestão de dados acadêmicos em instituições de ensino superior ainda representam um desafio relevante devido à heterogeneidade das fontes e ao grande volume de dados, especialmente em ambientes multicampi. Embora estratégias de Business Intelligence facilitem o acesso e a análise de informações, ainda há limitações na integração com técnicas avançadas de Inteligência Artificial. Nesse contexto, este trabalho apresenta a Plataforma Jacarandá, um caso de uso que integra dados institucionais e combina visualização com Aprendizado de Máquina e Processamento de Linguagem Natural.*

1. Introdução

Mesmo diante dos avanços recentes em sistemas de bancos de dados, a integração e a gestão eficiente de dados acadêmicos para apoio à tomada de decisão ainda representam um desafio relevante em Instituições de Ensino Superior (IES) [da Conceição et al. 2026, Vieira et al. 2025]. Esse desafio é intensificado pela heterogeneidade das fontes de dados e pelo seu crescente volume. Tal heterogeneidade abrange desde diferentes contextos institucionais — ensino, pesquisa, extensão e gestão — até a diversidade de perfis de milhares de docentes atuando em distintas áreas e linhas de pesquisa. Em instituições distribuídas em múltiplos campi, a construção de sistemas capazes de integrar e suportar análises avançadas torna-se ainda mais necessária.

Para mitigar esses desafios, globalmente as IES têm adotado estratégias baseadas em *Business Intelligence* (BI), cujo objetivo é facilitar o acesso e a análise interativa de dados, fornecendo suporte à tomada de decisão [Mishra 2025]. Apesar dos avanços, em um trabalho recente, os autores [Vieira et al. 2025] observam uma baixa integração entre mecanismos tradicionais de visualização de dados, como *dashboards*, e modelos

avançados de Inteligência Artificial (IA). Algo que poderia, por exemplo, enriquecer a análise ao caracterizar automaticamente entidades institucionais, como o corpo docente, e oferecer serviços analíticos mais especializados.

É nesse contexto que apresentamos a Plataforma Jacarandá¹, recentemente colocada em funcionamento numa instituição pública de ensino e pesquisa com mais de 2 mil servidores em 16 campi, e construída pelos próprios alunos. A ferramenta é uma solução para integração, processamento e análise inteligente de dados institucionais, combinando mecanismos tradicionais de análise, como relatórios e visualizações, coadunando com técnicas de Aprendizado de Máquina (AM) e Processamento de Linguagem Natural (PLN). Nessa linha a ferramenta suporta: (i) extração e visualização de palavras-chave das pesquisas realizadas, com diferente níveis de popularidade, apresentando as competências institucionais específicas por campus e áreas do conhecimento, (ii) classificação automática de professores de acordo com as áreas de conhecimento seguindo o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)², resolvendo o problema de falta de nomenclatura entre diversos campi e instituições, e (iii) recomendação de artigos científicos personalizados, aumentando a efetividade da divulgação acadêmica e incentivando a colaboração entre docentes.

Código Fonte: <https://github.com/bccanapolis/ia-sbbsd26>

2. Trabalhos Relacionados

O trabalho recente apresentado em [Vieira et al. 2025] realiza um mapeamento sistêmico sobre o uso de BI em instituições de ensino superior a partir da análise de 422 estudos. Os autores evidenciam que abordagens de BI e mineração de dados desempenham papel fundamental no aprimoramento da análise e da gestão institucional. Porém, o estudo destaca desafios recorrentes relacionados à heterogeneidade e integração de dados, apontando a necessidade de soluções automatizadas apoiadas por técnicas de IA.

No Brasil, o governo federal disponibiliza plataformas de BI voltadas à análise de dados educacionais, como a Plataforma Nilo Peçanha (PNP)³ e a plataforma Universidade 360°⁴, apresentando institutos e universidades federais, respectivamente. Ambas consolidam informações acadêmicas, administrativas e de gestão, contribuindo para a transparência e integração de dados em nível nacional. No entanto, tais plataformas possuem um escopo de análise macro, combinando diversas instituições, e não contemplam demandas específicas de gestão local, como análises em nível de curso/disciplina, caracterização do perfil do corpo docente ou suporte à colaboração entre pesquisadores.

De forma mais isolada, diversos trabalhos exploram o uso de técnicas de IA para caracterização e análise de dados acadêmicos [Escarassatti 2022, Viegas et al. 2022, da Conceição et al. 2026]. Apesar dos avanços, esses trabalhos focam em tarefas específicas, sem abordar de forma integrada a combinação entre análise de dados, caracterização de perfis e suporte à decisão em um ambiente institucional unificado.

¹<http://jacaranda.ifg.edu.br/video>

²<https://lattes.cnpq.br/documents/11871/24930/TabeladeAreasdoConhecimento.pdf/>

³<https://www.gov.br/mec/pt-br/npn>

⁴<https://www.gov.br/mec>

3. Aplicação das Técnicas de AM e PLN

A Plataforma Jacarandá se apresenta como uma ferramenta que utiliza tecnologias recentes de integração de dados, a partir de diversas fontes institucionais, entre elas: Sistema Unificado de Gestão Pública (SUAP)⁵, Plataforma Lattes, Plataforma Sucupira e Diretório de Grupos de Pesquisa do CNPq. Ainda, apresenta mais de 30 diferentes gráficos e relatórios, descrevendo informações sobre servidores, discentes, cursos/disciplinas, atuação docente, gestão, pesquisa e extensão. Porém, aqui destacamos a integração dessas funcionalidades com tarefas que exploram algoritmos de AM e PLN.

3.1. Extração de Palavras-Chave nos Currículos

Em um contexto de diversidade de atuação dos docentes em instituições multicampi, extrair os milhares tópicos de pesquisa no imenso espaço de publicações e trabalhos existentes não é uma tarefa trivial. Dado um problema de cauda-longa, os gestores precisam conhecer as palavras-chave mais populares, mas também mapear aquelas abordadas por alguns poucos professores, e que ajuda a entender as especificidades da instituição.

Assim, após avaliar e testar diversas estratégias, aplicamos uma adaptação ao algoritmo *KeyBERT* [Grootendorst 2020]. De forma nativa, primeiro o algoritmo gera uma representação *embeddings* para um documento, e na sequência gera a mesma representação para diversas palavras-chave candidatas, selecionando as que tiverem maior similaridade com a representação do documento. Em nosso pipeline, consideramos como documento o conjunto de títulos extraído das publicações bibliográficas registradas no currículo Lattes de um docente. Ainda, definimos um **conjunto T** de termos menos representativos (como “caso-de-uso”, “estudo”, “problema”, ...), obtidos a partir de alta frequência nos dados de todos os docentes. Depois excluimos do nosso corpus os termos pertencentes ao conjunto **T**. Os termos resultantes são selecionados considerando até 4-gramas de termos. Para representação vetorial utilizamos modelo *Sentence-BERT* [Reimers e Gurevych 2019]⁶.

Porém, mesmo tentando retirar diversos termos menos significativos, percebemos uma análise muito poluída. Para melhorar a extração, calculamos uma penalidade para candidatos menos representativos. Ou seja, dada a similaridade S entre um candidato e o documento, subtraindo de S a similaridade entre o candidato e o conjunto **T** (conjunto de termos menos representativos). A depender da similaridade entre o candidato e o conjunto **T**, a penalização pode ser maior. Para cada título de publicação bibliográfica, selecionamos os candidatos (ou palavras-chave) mais próximos, com um máximo de quatro por documento. Para evitar repetições e contabilizar a frequência ou popularidade, dado dois candidatos A e B , se $A \subset B$, nós retiramos B . Por fim, consideramos candidatos iguais aqueles que compartilham o mesmo radical entre seus termos.

O conjunto de palavras-chave resultante é inserido no módulo de análise visual da Jacarandá. A solução projeta as palavras-chave em um mapeamento gráfico interativo, onde cada palavra-chave selecionada é representada por uma circunferência, em que o tamanho reflete a quantidade de docentes envolvidos com a palavra-chave e a intensidade cromática indica o volume de publicações tendo a palavra chave. Como descrito na Figura 1, a plataforma permite aumentar ou diminuir o filtro de popularidade.

⁵<https://suap.ifg.edu.br/>

⁶Usamos o modelo *paraphrase-multilingual-MiniLM-L12-v2*.

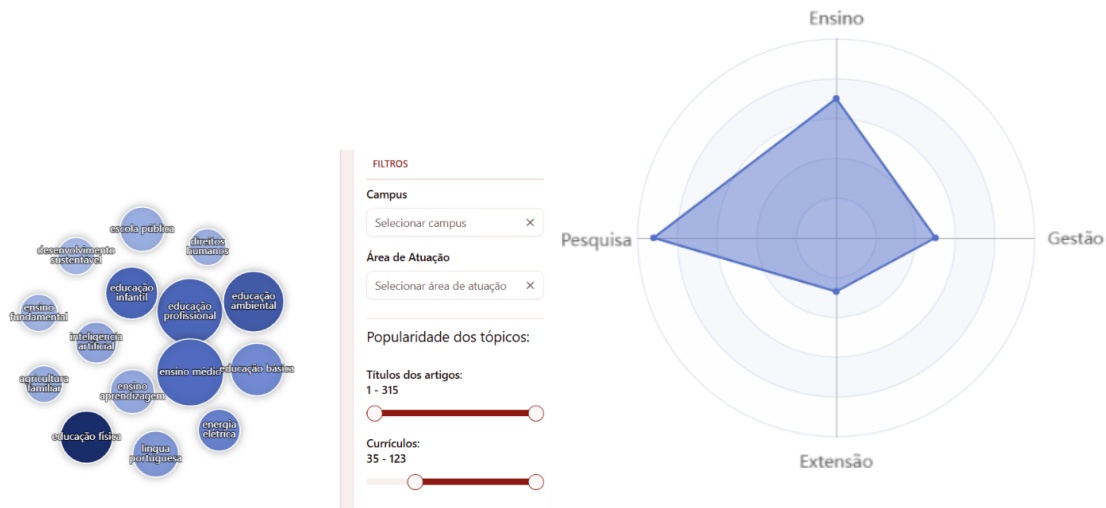


Figura 1. Extração de palavras-chave a partir dos Currículos Lattes.

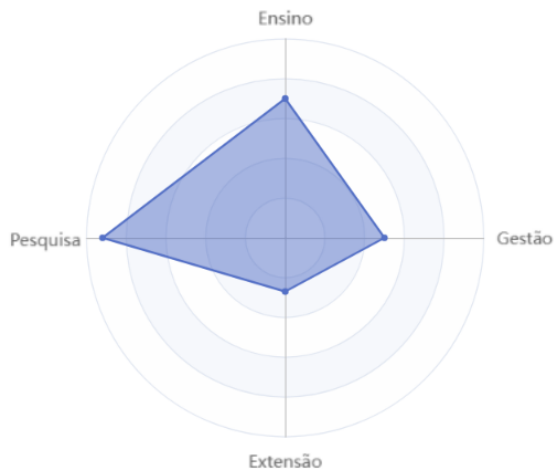


Figura 2. Mapa de Radar para Análise do perfil de Atuação dos docentes e grupos de docentes.

3.2. Classificação de Áreas de Atuação

Um outro problema recorrente em instituições multicampi é a ausência de classificação dos docentes por área de conhecimento ou, quando essa classificação existe, a falta de nomenclatura padrão para definir as áreas. Nesse cenário, a realização de um processo manual de categorização para todos os docentes pode se tornar significativamente onerosa. Para mitigar esse problema, aplicamos algoritmos supervisionados de Aprendizado de Máquina capazes de classificar automaticamente os docentes, especialmente nos casos em que o modelo apresenta elevada pontuação de confiança (*score*). Dessa forma, a necessidade de classificação manual é reduzida, ficando restrita apenas às situações mais desafiadoras, como a distinção entre áreas semanticamente próximas, a exemplo de “Engenharia Elétrica” e “Robótica, Mecatrônica e Automação”.

Para essa tarefa, foram consideradas 80 diferentes áreas do conhecimento, seguindo a classificação estabelecida pelo CNPq. Foi realizado um esforço institucional para a classificação manual de aproximadamente 1.200 docentes em diferentes áreas, caracterizando um cenário naturalmente desbalanceado entre as classes. Além dos rótulos, a base de dados contempla atributos como títulos de artigos científicos, áreas de atuação descritas no Lattes (que nem sempre seguem a nomenclatura institucional), formação acadêmica e nomes de disciplinas ministradas. Seguindo a literatura apresentada em [Cunha et al. 2023], foram investigadas diferentes estratégias para avaliação dos modelos de Aprendizado de Máquina, incluindo distintas representações vetoriais, tais como TF-IDF, BERT e abordagens com ajuste fino do BERT. Adicionalmente, foram incorporados atributos derivados de medidas de distância vetorial entre cada instância e as classes⁷. Também foram exploradas diferentes estratégias de construção dos modelos, algoritmos como Random Forest, SVM linear, variações de *Multilayer Perceptron* (MLP) e diferentes combinações baseadas em técnicas de *stacking*. Em todos os experimentos, foi empregada validação cruzada estratificada em 5 *folds*.

⁷Esses atributos calculam a distância em relação aos centroides das classes (*Nearest Centroid*) e à vizinhança local (*K-Nearest Neighbors*).

A partir do maior Macro-F1 e menor variância entre os folds, o modelo MLP (duas camadas de tamanho 200 e 100, e função de ativação tangente hiperbólica, combinado com representação BERT) apresentou os melhores resultados, descrito na Tabela 1⁸. A tabela descreve o limiar do score de confiança, porcentagem de instâncias (cobertura) com score superior ao limiar, e a taxa de erro. A escolha do limiar irá depender da taxa tolerável de erro na classificação, e do trabalho manual aceitável. Em um caso mais restrito, por exemplo, considerando um limiar de 99%, reduzimos o custo de classificação manual para quase 77%, com um erro de 1.18%.

Uma aplicação prática dessa classificação, além de reduzir o esforço manual da tarefa, pode ser visto na Figura 2. Nela analisamos a atuação dos docentes da área de Ciência da Computação sobre os eixos: ensino, pesquisa, extensão e gestão. Sem esse método de classificação, para alguns campus, essa análise seria feita somente considerando todos os docentes de um departamento ou de forma individual, sem permitir análises mais específicas por área.

Tabela 1. Taxa de erro por score de confiança.

Limiar	Cobertura	Erro
≥ 90%	89.77%	2.58%
≥ 95%	86.40%	2.10%
≥ 98%	82.22%	1.47%
≥ 99%	76.88%	1.18%

3.3. Sistema de Recomendação de Artigos Científicos

A Plataforma Jacarandá adota uma estratégia ativa e personalizada de divulgação acadêmica. A ferramenta envia notificações por e-mail aos pesquisadores, recomendando artigos com base na similaridade com suas produções anteriores. Assim, além de favorecer a identificação de interesses comuns, a plataforma incentiva a colaboração e fortalece a integração acadêmica no contexto institucional.

A partir do nosso trabalho descrito em [dos Santos et al. 2025], construímos uma aplicação que pudesse dar maior relevância para os trabalhos mais recentes de um pesquisador. Inicialmente, cada título de artigo no currículo do pesquisador é representado por dois vetores, esparso e denso. Em seguida, fazemos uma comparação de similaridade híbrida (combinando métodos léxicos e semânticos, respectivamente usando BM25 e *SentenceBERT*) com vetores de artigos dos demais pesquisadores. As similaridades avaliadas para os artigos do pesquisador são ponderadas cronologicamente com pesos que variam de 1 (para o mais recente) até $1/k$ (para o mais antigo), em que k é o número total de artigos. A fusão dos rankings produzidos por ambas as abordagens é realizada utilizando o método *Reciprocal Rank Fusion* (RRF), com constante $k = 60$. Destacamos que a combinação híbrida foi motivada pelos resultados em [dos Santos et al. 2025], que em função da quantidade de textos existentes nos dados, modelos baseado em conteúdo apresentam recomendações mais relevantes.

Como resultado, a partir da avaliação de relevância realizada pelos próprios docentes, de aproximadamente 5 mil avaliações observou-se que 91% das recomendações foram classificadas como altamente relevantes, pontuações 4 ou 5. Evidenciando a efetividade da aplicação de técnicas de PLN na recomendação de artigos científicos.

⁸Para uma comparação mais detalhada dos modelos, favor acessar material suplementar em <https://github.com/bccanapolis/ia-sbbd26>

4. Conclusão

A Plataforma Jacarandá é uma aplicação que integra técnicas tradicionais de *BI* e abordagens de Inteligência Artificial para a gestão acadêmica. Os resultados evidenciam a viabilidade de aplicações como extração de palavras-chave, classificação automática de docentes e recomendação personalizada de artigos científicos. Podendo ainda ser adaptadas para diferentes cenários, incluindo mecanismos de busca acadêmica e sistemas de apoio à distribuição de trabalhos para revisão. Além de reduzir o esforço operacional associado à análise e organização de dados institucionais, a solução contribui para a análise estratégica e estimula a colaboração acadêmica.

Agradecimentos Esse trabalho foi apoiado pelo Instituto Federal de Goiás e Pontifícia Universidade Católica de Goiás, conforme projeto publicado no DOU (01/11/2024—Edição: 212— Seção: 3).

Referências

- Cunha, W., Viegas, F., França, C., Rosa, T., Rocha, L., e Gonçalves, M. A. (2023). A comparative survey of instance selection methods applied to non-neural and transformer-based text classification. *ACM Comput. Surv.*, 55(13).
- da Conceição, F. J. C., das Dores, C. C. C., de Campos, G. A. L., e Gomes, R. L. (2026). Modelo híbrido de inteligência artificial para análise de eficiência educacional: Integração de aprendizado não supervisionado e análise envoltória de dados. *Revista Brasileira de Informática na Educação*, 34:429–457.
- dos Santos, J. V. F., de Andrade Nascimento, R. M., de Melo Camargo, A. C., Canuto, S. D. C., de Assis Costa, G., e de Sousa, D. X. (2025). Nova base de dados brasileira para sistemas de recomendação de artigos científicos. Em *Brazilian Symposium on Databases*, Brazil.
- Escarassatti, P. S. (2022). Representação visual dos dados de produção bibliográfica da plataforma Lattes. Dissertação de mestrado, Universidade de São Paulo, São Paulo, Brasil. Programa de Pós-Graduação em Sistemas de Informação.
- Grootendorst, M. (2020). KeyBERT: Minimal keyword extraction with BERT.
- Mishra, M. V. (2025). Data integration and feature engineering for supply chain management: Enhancing decision making through unified data processing. *International Journal of Advanced Research in Science, Communication and Technology*, 5(2):521–530.
- Reimers, N. e Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-Networks. Em *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Viegas, F., Pereira, A., Cecílio, P., Tuler, E., Meira, W. J., Gonçalves, M. A., e Rocha, L. (2022). Semantic academic profiler (SAP): a framework for researcher assessment based on semantic topic modeling. *Scientometrics*, 127(8):5005–5026.
- Vieira, A., Garcés, L., e Seabra, R. (2025). Business intelligence in higher education management systems: A systematic mapping study. Em *Anais do XXXVI Simpósio Brasileiro de Informática na Educação*, p. 69–84, Porto Alegre, RS, Brasil. SBC.