# Suporte à Elaboração de Matriz Curricular e Disciplinas com Inteligência Artificial Generativa e Grafos para Cursos de Graduação

**Néfi de Medeiros Fernandes**[1], **Maria da Conceição Moraes Batista**[1], **Roberta Macedo Marques Gouveia**[1], **Rafael Ferreira Mello**[1] **e Gabriel Alves** [1]

[1]Universidade Federal Rural de Pernambuco (UFRPE)
Recife – PE – Brasil

{nefi.fernandes, maria.cmbatista, roberta.gouveia,
rafael.mello, gabriel.alves}@ufrpe.br

***Abstract.*** *This work proposes a method to support the development of curricular matrices by integrating Generative Artificial Intelligence (GenAI), Natural Language Processing (NLP), and Graph Analysis techniques. The approach employs embeddings and cosine similarity to identify similar courses within heterogeneous university curricula, generate new courses by merging content, and structure the curriculum based on centrality metrics (Degree, Betweenness, Closeness). The results demonstrate a high similarity between the generated and original courses, with an average similarity of 0.91, and effectiveness in promoting interdisciplinary integration. The matrix generated in the case study for a Data Science program adheres to the guidelines of the Brazilian Computer Society (SBC) and the existing curricula of universities (UFC/UFPB), in addition to presenting a coherent pedagogical structure validated by a curricular graph structure with 40 courses and 311 connections, as well as by the validation of the relationships between courses in the matrix.*

***Resumo.*** *Este trabalho propõe um método para suporte à elaboração de matrizes curriculares, integrando técnicas de Inteligência Artificial Generativa (IAGen), Processamento de Linguagem Natural (PLN) e Análise de Grafos. A abordagem utiliza embeddings e similaridade de cosseno para identificar disciplinas semelhantes em matrizes universitárias heterogêneas, gerar novas disciplinas por fusão de conteúdos e estruturar o currículo com base em métricas de centralidade (Grau, Betweenness, Closeness). Os resultados demonstram alta similaridade entre disciplinas criadas e as disciplinas originais, com similaridade média de 0,91 e eficácia na integração interdisciplinar. A matriz gerada no estudo de caso para um curso de Ciência de Dados respeitando as diretrizes da Sociedade Brasileira de Computação (SBC) e matrizes existentes de universidades (UFC/UFPB), além de uma estrutura pedagógica coerente, validada por uma estrutura de grafo curricular com 40 disciplinas e 311 conexões, e pela validação das relações entre as disciplinas na matriz.*

## 1. Introdução

A construção de matrizes curriculares para cursos de graduação é um processo complexo e demorado, exigindo a análise de diretrizes normativas e a adequada ordenação de disciplinas por níveis e eixos formativos [Bianchessi 2021]. Esse desafio é amplificado pela crescente diversidade de áreas do conhecimento e pela necessidade de atualização constante dos conteúdos acadêmicos [da Silveira et al. 2024]. Tradicionalmente, essa análise é realizada manualmente

DOI: 10.5753/sbie.2025.12485 

por comitês, um processo subjetivo e de baixa escalabilidade. Abordagens computacionais existentes buscaram automatizar partes desse processo, utilizando técnicas de PLN para identificar similaridades entre disciplinas [Al-Omari et al. 2020] ou análise de redes para mapear fluxos de pré-requisitos já definidos [Stavrinides and Zuev 2023]. Contudo, essas abordagens geralmente se limitam a analisar estruturas existentes, sem a capacidade de sintetizar e gerar novos componentes curriculares coerentes para preencher lacunas identificadas. Recentemente, técnicas de Inteligência Artificial Generativa (IAGen) demonstraram potencial para apoiar esse processo, especialmente na análise e reorganização de grandes volumes de ementas e matrizes existentes. Relatórios internacionais, como o da UNESCO [Holmes et al. 2023], e trabalhos acadêmicos reconhecem esse potencial, mas alertam para riscos relacionados a vieses e desigualdades

Este trabalho propõe um método original que analisa e gera conteúdo curricular. A abordagem integra fusão de disciplinas similares e conteúdos de normas e diretrizes curriculares via IAGen e organização semântica por grafos, contemplando: (i) extração de palavras-chave; (ii) agrupamento de disciplinas por similaridade de *embeddings*; (iii) síntese de novas disciplinas por fusão de conteúdos correlatos; (iv) filtragem e alinhamento a documentos de referência; e (v) estruturação da matriz curricular com base em métricas de centralidade de grafos. Para assegurar conformidade normativa e reduzir vieses, a saída de cada etapa do método proposto deve ser validada por um ser humano, seguindo uma abordagem *Human-In-The-Loop* (HITL), a fim de evitar vieses e garantir que decisões importantes sejam tomadas de forma responsável e crítica, proposta como protocolo para produção com papéis definidos e critérios de intervenção, enquanto na prova de conceito serviu apenas para calibrar *prompts* e limiares.

O método é validado em um estudo de caso para Ciência de Dados, com uma avaliação quantitativa que utiliza similaridade de cosseno para medir a coesão, testes de hipótese para a aderência normativa, e métricas de modularidade para a validação estrutural. Como principal contribuição, apresentamos um fluxo de trabalho original e reprodutível que oferece suporte à criação e estruturação de matrizes curriculares.

As perguntas de pesquisa que orientam este estudo são:

- **PP 01:** De que forma a combinação de IAGen e similaridade semântica permite gerar e validar novos componentes curriculares coesos a partir da fusão de conteúdos de fontes heterogêneas?
- **PP 02:** Qual o desempenho da utilização de IAGen e Análise de Grafos para elaborar matrizes curriculares interdisciplinares alinhadas a referenciais formativos?

## 2. Trabalhos Relacionados

A construção de matrizes curriculares que conciliem diretrizes educacionais, flexibilidade e aderência ao mercado é um desafio recorrente para instituições de ensino. Nesse cenário, técnicas de IAGen e Análise de Grafos vêm sendo exploradas como alternativas promissoras, permitindo identificar similaridades entre disciplinas, propor novos conteúdos e estruturar percursos formativos mais dinâmicos e personalizados.

[Bahroun et al. 2023] revisam o uso crescente da IAGen na educação, com foco em avaliação, tutoria automatizada e geração de materiais didáticos. Já [Qu et al. 2024] analisam o uso de grafos na educação, destacando sua aplicação na construção de ontologias, representação vetorial (*embedding*) e sistemas de recomendação.

Estudos recentes exploram o uso de IAGen para gerar sequências de aprendizado, como no sistema *CurricuLLM* [Mehta et al. 2023], e para o design de cursos [Chris and Sheriffdeen 2024,

Kumar et al. 2024, Lu and Zoghi 2024]. Tais trabalhos demonstram ganhos de eficiência, mas também alertam para riscos como a geração de informações incorretas e a necessidade de fomentar a colaboração humano-IA [Padovano and Cardamone 2024]. O método proposto neste trabalho avança ao integrar a fusão de conteúdos via IAGen com a modelagem estrutural via Análise de Grafos, superando a decomposição puramente sequencial e atribuindo papéis pedagógicos às disciplinas com base em métricas de centralidade.

Paralelamente, a Análise de Grafos tem sido consistentemente aplicada para modelar e compreender a estrutura de currículos. [Tuzón et al. 2024] utilizam redes complexas para analisar a estrutura de currículos de Física e Matemática, identificando conceitos centrais e a coerência de blocos de conteúdo, tanto isoladamente quanto de forma integrada. De forma similar, [Stavrinides and Zuev 2023] estudam redes de pré-requisitos de cursos para visualizar, analisar e otimizar currículos, aplicando medidas de centralidade e estratificação topológica para identificar cursos importantes e o fluxo de conhecimento. Esses estudos reforçam a utilidade da análise de grafos para desvendar interdependências e hierarquias em estruturas curriculares existentes.

Apesar dos avanços, a literatura carece de um método integrativo que abranja desde a geração de novos componentes curriculares até sua validação e estruturação. Nossa proposta busca preencher essa lacuna, através de um método original que combina IAGen para a síntese de conteúdo e a Análise de Grafos para a validação estrutural.

## 3. Método e Ferramentas

Para a elaboração dos *prompts* destinados aos modelos de IAGen, adotou-se o método de Engenharia de *Prompt* com Papel, Tarefa e Formato de Saída (RTF, do inglês *Role, Task, Format*) [White et al. 2023]. Este *framework* foi aplicado de forma específica em cada etapa que demandou o uso da IAGen, como a extração de palavras-chave, geração de novas disciplinas e a seleção final de disciplinas, conforme detalhado nas seções subsequentes. Todos os *prompts* utilizados estão disponibilizados em repositório público[1]. A estrutura do RTF é composta por três elementos fundamentais:

- **Papel**:Define o papel ou função que o modelo deve assumir durante a interação;
- **Tarefa**:Especifica a atividade que o modelo deverá executar;
- **Formato de saída**:Determina o formato esperado para a resposta.

O método proposto segue um fuxo sistemático e reprodutível (Figura 1), iniciado pelo pré-processamento das matrizes curriculares heterogêneas, unificando nome e ementa das disciplinas. Em seguida, realiza-se a extração de palavras-chave, utilizadas com os dados das disciplinas na construção de uma matriz de similaridade por cosseno. A partir dela, são geradas novas disciplinas por fusão de conteúdos correlatos. Essas disciplinas passam por duas fases de validação com IAGen — classificação individual e análise consolidada — em relação ao documento de referência. Por fim, a matriz curricular é estruturada com base na análise de grafos, utilizando métricas de centralidade para definir o papel de cada disciplina.

### 3.1. Dados Utilizados

O conjunto de dados utilizado neste estudo é composto por matrizes curriculares de cursos de graduação, extraídas manualmente dos portais oficiais de instituições de ensino superior. Foram selecionadas as matrizes dos cursos de Bacharelado em Administração [2], Engenharia da

[1] `https://github.com/nefif/SBIE_2025.git`
[2] `https://www.ufpe.br/administracao-bacharelado-ccsa`

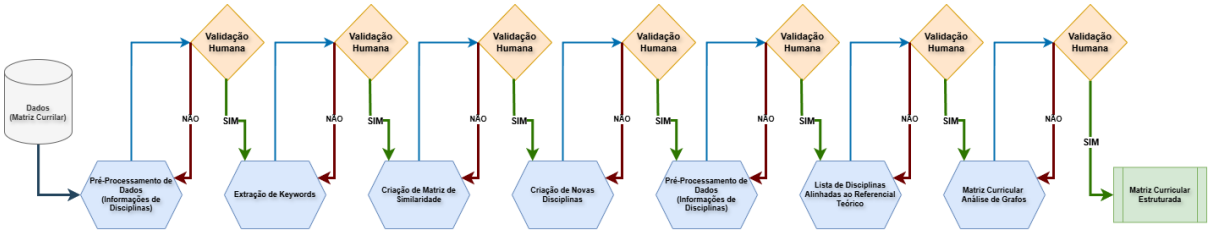


**Figura 1. Fluxo das etapas do modelo**

Computação [3], Ciências Biológicas [4], Sistemas da Informação [5] e Ciências da Computação [6] da Universidade Federal de Pernambuco (UFPE), além dos cursos de Bacharelado em Ciência de Dados da Universidade Federal da Paraíba (UFPB) [7] e da Universidade Federal do Ceará (UFC) [8].

Inicialmente, foram coletadas 416 disciplinas. Em seguida, foi aplicada uma filtragem, mantendo-se apenas as disciplinas obrigatórias que possuíam ementas registradas. Após essa etapa, o conjunto final foi reduzido para 236 disciplinas.

As matrizes curriculares coletadas apresentavam formatos heterogêneos, o que exigiu um processo de padronização. Diante dessa variação estrutural, optou-se por utilizar apenas os campos comuns à maioria das disciplinas. Assim, os atributos considerados no conjunto de dados foram: Curso de Origem, Nome da Disciplina e Ementa. Esses campos serviram como base para as etapas posteriores de pré-processamento, vetorização e análise de similaridade.

### 3.2. Extração de Palavras-chave

Para enriquecer a representação semântica das ementas, extraímos dez termos específicos que capturam o conteúdo de cada disciplina, mesmo quando não aparecem literalmente no texto. A extração de dez palavras-chave por disciplina foi adotada para enriquecer e padronizar a representação semântica das ementas sucintas. A entrada é composta pelo nome da disciplina e pela respectiva ementa em texto bruto; a saída consiste em uma lista de palavras-chave em letras minúsculas, evitando termos genéricos (por exemplo, “introdução” ou “avançado”). Esse vetor de palavras-chave corrige descrições genéricas e fornece insumos mais precisos para os cálculos de similaridade semântica.

O *prompt* RTF atribuiu ao modelo o Papel de “especialista em currículo acadêmico” com a Tarefa de extrair as dez palavras-chave em Formato de lista, separadas por vírgula. Segue como foi estruturado o *prompt* segundo o modelo RTF.

**Validação Humana (HITL) -** Após cada extração, um especialista acadêmico deve revisar a lista de palavras-chave:Se os termos forem considerados adequados, a lista segue para a próxima etapa do método; caso identifique-se algum termo irrelevante ou a ausência de termos essenciais, o revisor deve ajustar brevemente o *prompt* e a extração é executada novamente até atingir uma qualidade satisfatória.

[3] https://www.ufpe.br/engenharia-da-computacao-bacharelado-cin
[4] https://www.ufpe.br/ciencias-biologicas
[5] https://www.ufpe.br/sistemas-de-informacao-bacharelado-cin
[6] https://www.ufpe.br/ciencia-da-computacao-bacharelado-cin
[7] https://sigaa.ufpb.br/sigaa/public/curso/portal.jsf?id=14289031&lc=pt_BR
[8] https://si3.ufc.br/sigaa/public/curso/curriculo.jsf?lc=pt_BR&id=72460878

### 3.3. Montagem da Matriz de Similaridade

O objetivo desta etapa é quantificar a similaridade semântica entre todas as disciplinas para gerar uma matriz simétrica $M_{n\times n}$. Para isso, o texto consolidado de cada disciplina (Nome, Ementa e palavras-chave) passou por um *pipeline* de pré-processamento, que incluiu a remoção de *stopwords*, normalização de caixa e a lematização. Optou-se pela lematização em vez da stematização por sua capacidade de preservar a forma canônica das palavras, o que melhora a interpretabilidade dos dados em domínios especializados como o educacional. Estudos prévios indicam que essa escolha pode gerar ganhos em tarefas de Processamento de Linguagem Natural [Balakrishnan and Lloyd-Yemoh 2014]. Ao final do processo, os textos foram convertidos em vetores numéricos através de *embeddings*.

Com os vetores representativos de cada disciplina, foi calculada a similaridade de cosseno, métrica ideal para espaços vetoriais de alta dimensionalidade por avaliar a orientação dos vetores independentemente de sua magnitude [Yamagiwa et al. 2024]. A similaridade entre dois vetores $v_i$ e $v_j$, representando as disciplinas $D_i$ e $D_j$, é dada por:

$$sim(D_i, D_j) = \frac{v_i \cdot v_j}{|v_i|, |v_j|} \tag{1}$$

O resultado de cada comparação é um valor no intervalo ([0,1]), onde 1 representa a semelhança máxima. Todos os resultados foram organizados na matriz de similaridade $M$, que serve como base para a criação de novas disciplinas e para a construção do grafo curricular nas etapas seguintes.

### 3.4. Criação de Disciplinas

Nesta seção descreve-se a agregação e reformulação de disciplinas com base em conteúdos equivalentes para gerar novas disciplinas em quatro fases: (i) identificação de pares com alta similaridade mediante um limiar definido pelo usuário; (ii) formação de grupos de equivalência extraídos como componentes conectados de um grafo não direcionado via Busca em Largura (BFS); (iii) geração de nome, ementa e palavras-chave por IAGen; e (iv) integração das disciplinas criadas ao conjunto de dados original, resultando em uma base expandida.

Primeiro, para identificar grupos de disciplinas com conteúdo similar, constrói-se um grafo não direcionado onde os nós são as disciplinas e as arestas conectam pares de disciplinas $(D_i, D_j)$ cuja similaridade de cosseno atende ou supera o limiar (e $i \neq j$). A partir deste grafo, são extraídos os componentes conectados — que representam os grupos de disciplinas semanticamente relacionadas — por meio de uma Busca em Largura (BFS) [Cormen et al. 2022].

Na segunda etapa, para cada grupo identificado, um modelo de IAGen é instruído a sintetizar os conteúdos. Atuando como um "especialista em currículo acadêmico", o modelo recebe *prompts* no padrão RTF para gerar um novo nome, uma ementa consolidada e um conjunto de palavras-chave que representem a fusão daquelas áreas do conhecimento. As novas disciplinas, compostas por esses três elementos, são então anexadas ao *dataset* original.

**HITL -** Em todas as etapas descritas, um especialista acadêmico deverá revisar a saída antes de prosseguir. Caso identifique inconsistências ou omissões, ele ajusta parâmetros (por exemplo, o limiar de similaridade ou o *prompt*) e solicita a reexecução da etapa correspondente, garantindo que cada artefato — desde os pares filtrados até as disciplinas finais — seja validado e alinhado aos critérios de qualidade definidos.

### 3.5. Seleção de Disciplinas pela IAGen

Nesta etapa, as disciplinas são selecionadas segundo sua aderência ao documento de referência por meio de duas fases: inicialmente, o documento é carregado, fragmentado semanticamente e indexado para consultas precisas; em seguida, a IAGen filtra cada disciplina individualmente, descartando aquelas com alinhamento insuficiente. Por fim, as disciplinas pré-selecionadas são avaliadas em conjunto, considerando critérios do curso — como eixos de formação, conteúdos e competências extraídos do documento de referência — e orientações através do *prompt*, para compor a lista de disciplinas definitiva.

Para montar a matriz curricular, utiliza-se um framework baseado em LLMs que integra modelos de linguagem, documentos de referência e o conjunto de disciplinas. Inicialmente, o documento é carregado, seu conteúdo extraído e segmentado em fragmentos ("*chunks*") para preservar o contexto. Cada fragmento é convertido em *embedding* e indexado em um índice vetorial, permitindo buscas semânticas eficientes (*retriever*) e recuperação dos trechos mais relevantes para as consultas subsequentes.

Paralelamente ao preparo do documento de referência, unificam-se nome, ementa e palavras-chave das disciplinas (originais e geradas) em um único texto por disciplina, convertendo-o em *embeddings*. Com esses vetores, a avaliação de aderência para seleção final ocorre em duas fases.

A primeira fase consiste em filtragem individual. Para cada disciplina, o *retriever* extrai trechos relevantes do documento de referência. A IAGen, então, realiza uma classificação binária de aderência, seguindo um *prompt* estruturado no método RTF com as seguintes orientações: Papel - O modelo atua como um classificador binário, focado em uma decisão objetiva; Tarefa - Avaliar se a disciplina fornecida é aderente ao contexto extraído do documento de referência; Formato - A saída deve ser uma única palavra, "Sim" ou "Não", sem qualquer justificação ou texto adicional, para garantir uma resposta limpa e de fácil processamento.

O *prompt* exato para essa tarefa foi:

> *Para determinar a aderência, seja direto. Se for aderente, responda "Sim"; caso contrário, responda apenas "Não". Não justifique sua resposta.*

Na segunda fase, a lista de disciplinas pré-aprovadas é submetida a uma análise para a seleção final. Para esta tarefa complexa, foi construído um *prompt* detalhado, seguindo a estrutura RTF para guiar o modelo através das orientações de Papel: O modelo assume a persona de um "coordenador de curso de graduação", ativando o conhecimento do modelo sobre planejamento acadêmico; Tarefa: Analisar a lista de disciplinas e o documento de referência para selecionar exatamente 40 disciplinas e organizá-las em uma grade de 8 semestres (5 por semestre), aplicando as restrições de exclusão de TCC e Estágio; Formato: A saída é estritamente definida como uma lista pura contendo apenas os nomes das 40 disciplinas, uma por linha, sem formatação ou justificativas.

O *prompt* completo utilizado para executar esta tarefa foi o seguinte:

> *Você é coordenador de curso de graduação e deve selecionar 40 disciplinas que melhor se alinham ao documento de referência. A partir da lista de disciplinas fornecida, monte uma matriz acadêmica para o Bacharelado em Ciência de Dados com 8 semestres de 5 disciplinas cada, excluindo Trabalho de Conclusão de Curso e Estágio. Retorne apenas a lista, com cada disciplina em linha separada, sem justificativas ou formatações extras.*

**HITL -** Em ambas as fases de seleção — avaliação individual e análise consolidada

— um especialista acadêmico valida cada decisão. Caso uma disciplina seja classificada inadequadamente ou não atenda aos critérios, o especialista ajusta o *prompt*, os parâmetros de consulta ou o limiar de similaridade e reexecuta a etapa. Assim, garante-se que a lista final reflita fielmente as diretrizes do documento de referência e os requisitos específicos do usuário.

### 3.6. Organização Curricular via Análise de Grafos

Nesta etapa, construiu-se um grafo não direcionado em que cada nó representa uma disciplina e cada aresta conecta pares cuja similaridade de cosseno (Seção 3.3) atinge ou supera o limiar (média dos coeficientes). A partir desse grafo, calcularam-se as métricas de centralidade de grau (número de conexões diretas), de intermediação (*betweenness*, frequência com que um nó integra caminhos mínimos) e de proximidade (*closeness*, inverso da soma das distâncias ponderadas pelas similaridades) para revelar o papel estrutural de cada disciplina [Otte and Rousseau 2002].

Com base nas combinações desses indicadores, as disciplinas foram classificadas em três grupos pedagógicos a partir das métricas de centralidade: *Básicas* — alta centralidade de grau e proximidade, baixa intermediação (semestres iniciais); *Específicas* — alta intermediação e proximidade média (semestres intermediários); *Complementares* — alta proximidade, baixo grau e baixa intermediação (semestres finais). Esse mapeamento não só identifica pontes interdisciplinares, mas também orienta a distribuição das disciplinas ao longo do curso.

**HITL:** Em cada etapa — da geração do grafo à categorização — um especialista acadêmico revisa os resultados. Se identificar conexões ou classificações inadequadas, ajusta o limiar e reexecuta a fase correspondente, assegurando a aderência aos critérios pedagógicos e curriculares aprovados.

## 4. Estudo de Caso

O presente estudo de caso foi desenhado para validar empiricamente cada etapa do método proposto. A seguir, a eficácia da extração de palavras-chave será avaliada por sua similaridade com as ementas; a qualidade da criação de novas disciplinas será medida pela coesão semântica dos grupos; a seleção final das disciplinas será validada por meio de testes de hipótese contra matrizes de referência e as diretrizes da SBC; e, por fim, a organização curricular em eixos pedagógicos será justificada pela análise estrutural do grafo.

Para a geração de conteúdo, foi empregado o modelo de IAGen *deepseek-chat (v3)* [DeepSeek-AI 2024] e na seleção da lista de disciplinas foi utilizado o modelo *deepseek-reasoner (v1)* [DeepSeek-AI 2025], selecionados por sua alta performance em *benchmarks* e pela relação custo-benefício. A vetorização foi realizada pelo modelo *open-source nomic-embed-text* [Nussbaum et al. 2024], preferido por sua robustez, resultados superiores e contexto de entrada ampliado, sendo executado localmente via *Ollama*[9].

A orquestração de todo o processo foi gerenciada pelo *framework LangChain*[10], utilizado para construir o fluxo que integra a leitura de documentos de referência, a criação de índices vetoriais com *FAISS* e as chamadas aos modelos de linguagem. Por fim, na camada de processamento de dados, a biblioteca *spaCy* (v3.8.4) foi usada para o pré-processamento de textos em português, enquanto a *scikit-learn* (v1.5.0) foi empregada para os cálculos de similaridade de cosseno. Todas as ferramentas utilizadas estão disponibilizadas em repositório público [11].

[9] https://github.com/ollama/ollama
[10] https://github.com/langchain-ai/langchain
[11] https://github.com/nefif/SBIE_2025.git

### 4.1. Extração de Palavras-chave

A extração de palavras-chave pela IAGen demonstrou, através da análise quantitativa da similaridade de cosseno entre as palavras-chave geradas e suas respectivas ementas, uma alta similaridade média ($0{,}8862$) e baixa variabilidade (desvio padrão de $0{,}0414$). Por exemplo, para a disciplina *Gerenciamento de Dados e Informações*, cuja ementa era apenas "Algoritmos e estrutura de dados", a IAGen expandiu o vocabulário para: *dados, informação, gerenciamento, algoritmos, estruturas, banco de dados, modelagem, análise, processamento, armazenamento*. Essa capacidade de expansão conceitual fornece uma base de dados mais rica para as etapas subsequentes.

No histograma na Figura 2, a grande maioria dos valores concentrada em um intervalo de alta similaridade e um pico de frequência próximo a $0{,}90$. Essa concentração de dados, evidenciada pelo baixo desvio padrão, mostra que o método produz resultados consistentemente elevados.

### 4.2. Matriz de Similaridade e Criação de Disciplinas

Após a geração da matriz de similaridade, foi realizada uma análise de sensibilidade da modularidade com valores entre 0,75 e 0,95, indicando que a modularidade de $0,90$ mantinha grupos consistentes e semanticamente coesos. Consequentemente, foi adotado o limiar (*threshold*) de $0{,}90$ para agrupar disciplinas com alta similaridade, assegurando fidelidade conceitual e evitando combinações artificiais entre conteúdos pouco relacionados.

Em seguida, aplicou-se o modelo IAGen a cada grupo para sintetizar novas disciplinas com base em seus conteúdos internos. A fusão resultou em elevada fidelidade ao material original, com média geral de similaridade de $0{,}9079$ e baixo desvio padrão ($0{,}0373$), valores calculados pela média e desvio das similaridades de cosseno entre cada disciplina gerada e as demais de seu grupo, seguidos da agregação desses resultados em todos os grupos. Essa consistência foi confirmada visualmente pelo histograma da Figura 3, que evidencia a concentração da maioria dos valores acima de $0{,}80$.

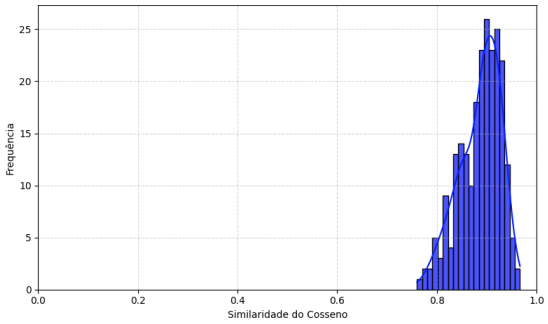


**Figura 2. Distribuição das similaridades de cosseno das palavras-chave geradas com as ementas das disciplinas.**

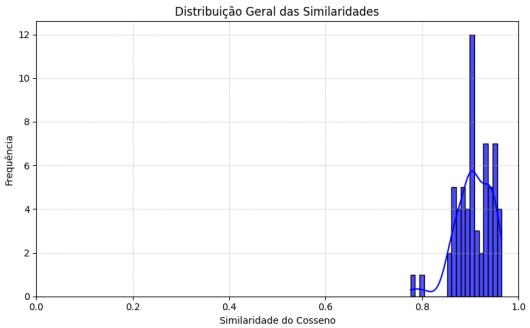


**Figura 3. Distribuição das similaridades entre disciplinas originais e as novas disciplinas geradas por fusão.**

A Tabela 1 apresenta métricas de coesão semântica para algumas disciplinas geradas, destacando que grupos de natureza técnica, como *Banco de Dados I*, alcançaram coesão quase perfeita (média $0{,}9545$, desvio padrão e IC 95% próximos de zero), ao passo que disciplinas mais abstratas, como *Estágio Supervisionado*, exibiram média inferior ($0{,}7890$), maior variabilidade e intervalo de confiança mais amplo, evidenciando maior incerteza. Esse contraste confirma a robustez do método em diferentes contextos.

**Tabela 1. Métricas de Coesão Semântica por Grupo de Disciplinas**

| Grupo | Média | Desvio Padrão | IC 95% da Média |
|---|---|---|---|
| Banco de Dados I | 0,9545 | 0,0001 | (0,9542; 0,9547) |
| Engenharia de Software e Sistemas | 0,8961 | 0,0147 | (0,8795; 0,9127) |
| Estágio Supervisionado | 0,7890 | 0,0129 | (0,7638; 0,8143) |
| **Todos os Grupos** | **0,9079** | **0,0373** | **(0,8985; 0,9173)** |

### 4.3. Seleção de Disciplinas pela IAGen

Para validar a qualidade da seleção de disciplinas realizada pela IAGen, foi calculada a similaridade do cosseno entre a lista de disciplinas gerada pela ferramenta e as matrizes curriculares dos cursos de graduação em Ciência de Dados das universidades UFC (23 disciplinas) e UFPB (32 disciplinas). Essa validação buscou verificar se a proposta da IAGen (40 disciplinas) apresenta coesão e alinhamento com estruturas curriculares já consolidadas.

Adicionalmente, foi conduzido um teste de hipótese utilizando Análise de Variância (ANOVA), com o objetivo de avaliar o grau de alinhamento entre a lista de disciplinas selecionadas pela IAGen e as diretrizes presentes no documento de referência para criação de cursos em Ciência de Dados da Sociedade Brasileira de Computação (SBC) [SBC and ABE 2023]. O Teste ANOVA foi escolhido por ser o método estatístico padrão para comparar as médias de três ou mais grupos independentes (IAGen, UFC, UFPB), evitando o erro acumulado que surge ao se realizar múltiplos Testes T [Field 2024].

Na primeira etapa de validação, combinou-se Ementa e palavras-chave de cada disciplina em uma nova coluna e realizou-se sua vetorização via modelo NOMIC, gerando conjuntos de vetores para Ementa, Palavras-chave e a combinação de ambos. Em seguida, compararam-se a similaridade do cosseno dos conjuntos de dados das listas da UFPB e da UFC para estabelecer um *baseline* e, por fim, avaliou-se a similaridade da lista Proposta em relação a essas referências, verificando sua coerência com currículos consolidados. A Tabela 2 apresenta os resultados obtidos na análise de similaridade entre a proposta gerada pela IAGen e as matrizes curriculares dos cursos da UFC e da UFPB. Observa-se que, das 40 disciplinas selecionadas pela IAGen, 10 também estão presentes na matriz da UFC e 13 na da UFPB. As 17 disciplinas restantes não são compartilhadas por nenhuma das duas instituições. Apesar dessa diferença, os resultados de similaridade de conteúdo entre a Proposta e as matrizes da UFC e da UFPB foram superiores ao valor de similaridade observado entre as próprias instituições (UFPB vs. UFC), que serve como *baseline* para comparação. Esses dados reforçam a coerência da proposta construída pela IAGen com relação a cursos já existentes.

**Tabela 2. Análise Comparativa de Similaridade de Conteúdo entre Matrizes Curriculares**

| Métrica de Similaridade | Proposta vs. UFC | Proposta vs. UFPB | UFPB vs. UFC (Baseline) |
|---|---|---|---|
| Ementa | 0,9333 | 0,8482 | 0,8252 |
| Palavras-Chave | 0,8550 | 0,8449 | 0,7314 |
| **Ementa + Palavras-Chave** | **0,8311** | **0,8545** | **0,7741** |
| Disciplinas Compartilhadas | 10 | 13 | - |

A comparação do alinhamento das listas de disciplinas (Proposta, UFC e UFPB) com o

documento de referência da SBC foi realizada por meio do cálculo da similaridade de cosseno entre as informações de cada disciplina (Ementa e Palavras-chave) e o referido documento. Após calcular a similaridade de cada disciplina com o documento de referência, foram obtidas a média e o desvio padrão das similaridades para cada lista. Os resultados demonstram que a lista Proposta obteve uma média de similaridade de 0,73, valor comparável às médias das listas da UFPB (0,72) e da UFC (0,73). No entanto, a lista Proposta apresentou um desvio padrão ligeiramente superior (0,041) em relação ao da UFC (0,036), o que pode indicar uma maior diversidade de conteúdos ou abordagens nas disciplinas geradas pela IAGen.

Para formalizar a comparação do alinhamento, foi realizada uma Análise de Variância (ANOVA) de uma via, com o objetivo de testar a hipótese de que as médias de similaridade entre as três listas de disciplinas (Proposta: 40, UFPB: 32 e UFC: 23 disciplinas) e o documento de referência da SBC são estatisticamente iguais. Os pressupostos para aplicação do teste foram atendidos: os dados apresentaram distribuição normal (teste de *Shapiro-Wilk*, $p > 0{,}05$) e homogeneidade das variâncias (teste de *Levene*, $p > 0{,}05$), confirmando a adequação da ANOVA. O resultado da análise indicou que não há diferença estatisticamente significativa entre as médias de alinhamento dos grupos ($F(2, 92) = 0{,}31$, $p = 0{,}7341$). Como o valor de $p$ é substancialmente superior ao nível de significância adotado ($\alpha = 0{,}05$). Portanto, a matriz gerada pela presente proposta pode ser considerada similar às matrizes curriculares da UFPB e da UFC conforme os critérios estabelecidos na análise realizada.

### 4.4. Organização Curricular via Análise de Grafos

Foi construído um grafo no qual os nós representam as disciplinas, e as arestas indicam relações de similaridade entre elas. Para estabelecer as conexões, foi adotado um limiar (*threshold*) de $0{,}70$, valor correspondente à média das similaridades de cosseno obtidas a partir da matriz de similaridade entre as disciplinas. Dessa forma, apenas pares de disciplinas com similaridade igual ou superior a esse valor foram conectados no grafo.

Para validar a escolha da média como limiar, foi conduzida uma análise de sensibilidade, técnica que avalia o impacto de diferentes parâmetros nos resultados [Saltelli et al. 2004], contemplando três cenários, cujos resultados são apresentados na Tabela 3. No cenário de limiar baixo ($0{,}60$), a rede tornou-se excessivamente densa (Figura 4), exibindo índice de modularidade negativo, o que inviabiliza a identificação de comunidades. Em contrapartida, um limiar alto ($0{,}80$) fragmentou o currículo em múltiplos componentes, restritos a apenas 26 arestas (Figura 6); embora a modularidade calculada pelo algoritmo Leiden [Traag et al. 2019] apresentasse valor elevado ($0{,}7751$), tal resultado reflete um artefato da fragmentação, tornando-se inadequado para análise de fluxo. O limiar de $0{,}70$, que corresponde aproximadamente à média das similaridades entre pares de disciplinas, ilustrado na Figura 5, mostrou-se o ponto de equilíbrio mais representativo: gera um grafo coeso e conectado, com estrutura modular significativa, preservando conexões interdisciplinares essenciais ao mesmo tempo em que filtra o ruído de similaridades fracas.

As regras de centralidade do método foram aplicadas para classificar as disciplinas de acordo com seu papel pedagógico, conforme detalhado na Tabela 4. As disciplinas de Formação Básica apresentam os maiores valores de Grau. O grupo de Formação Específica contém as disciplinas com maior Intermediação, confirmando sua função de "pontes", enquanto a Formação Complementar é caracterizada por baixa Intermediação, definindo-as como tópicos de alta especialização. É importante notar que a classificação é derivada da topologia do grafo, refletindo o papel estrutural de cada disciplina na rede de similaridade semântica. Disciplinas como "Aprendizado de Máquina", embora fundamentais, podem aparecer como

**Tabela 3. Análise de Sensibilidade do Grafo com Diferentes Limiares de Similaridade**

| Limiar | Arestas | Média de Centralidade | | | Modularidade |
|---|---|---|---|---|---|
| | | Grau | Intermediação | Proximidade | (Leiden) |
| 0.60 (Baixo) | 774 | 0,9923 | 0,0002 | 0,6962 | -0,0090 |
| **0.70 (Padrão)** | **311** | **0,3987** | **0,0164** | **0,4715** | **0,1787** |
| 0.80 (Alto) | 26 | 0,0333 | 0,0003 | 0,0305 | 0,7751 |

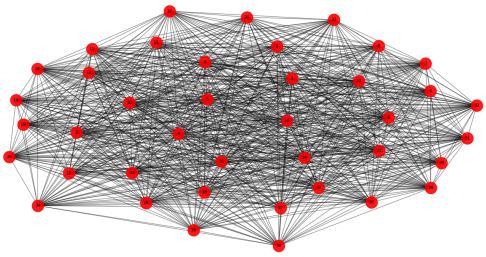

**Figura 4. Limiar Baixo (0.60) com modularidade de -0.0090.**

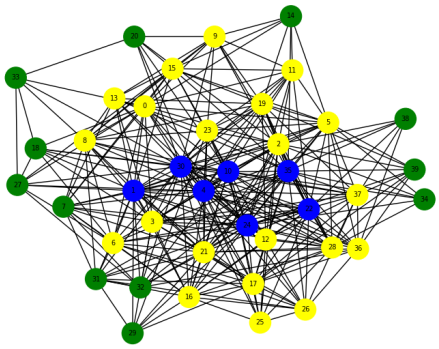

**Figura 5. Limiar Padrão (0.70) com modularidade de 0.1787.**

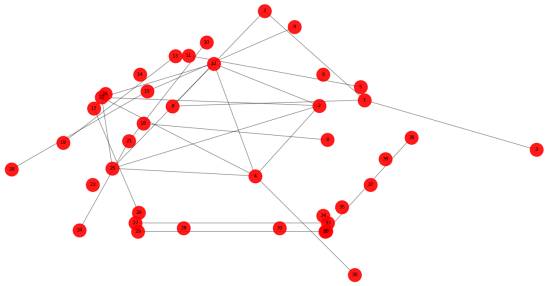

**Figura 6. Limiar Alto (0.80) com modularidade de 0.7751.**

"Complementares" se possuírem baixa intermediação, indicando que são tópicos com menor conexão semântica a múltiplos eixos do curso. Esta classificação é, portanto, um suporte à decisão; a validação final do especialista (HITL) é o que confirmaria ou ajustaria a posição dessas disciplinas na grade curricular. A estrutura final do currículo segue um padrão em "funil", ilustrado na Figura 5. As disciplinas de Formação Básica (Verde) ocupam a borda do grafo. As de Formação Complementar (Azul), de baixa intermediação, compõem o núcleo do grafo. Entre elas, as de Formação Específica (Amarelo), com maior intermediação, atuam como pontes, articulando o fluxo de conhecimento do básico para o avançado.

### 4.4.1. Discussão dos Resultados

Os resultados do estudo de caso validam de forma clara as perguntas de pesquisa e evidenciam as contribuições centrais do método proposto. No que diz respeito à PP 01, a combinação de IA-Gen e similaridade semântica mostrou-se eficaz em dois estágios complementares: a extração de palavras-chave padronizou e enriqueceu as ementas, resultando em média de similaridade de $0{,}8862$ com baixo desvio-padrão; em seguida, a aplicação de um limiar de $0{,}90$ para a fusão

**Tabela 4. Disciplinas organizadas por papel pedagógico, definido pelas métricas de centralidade.**

| Formação Básica | Formação Específica | Formação Complementar |
|---|---|---|
| Probabilidade e Estatística | Sistemas de Apoio à Decisão | Aprendizado de Máquina |
| Introdução à Ciência de Dados | Otimização Contínua | Aprendizado Profundo |
| Algoritmos e Estruturas de Dados | Otimização Não-Linear | Redes Neurais Artificiais |
| Teoria das Probabilidades | Estatística e Probabilidade Computacional | Análise Multivariada e Aprendizado Não |
| Matemática Discreta Aplicada | Big Data | Supervisionado |
| Álgebra Linear e Geometria Analítica | Análise e Projeto de Algoritmos | Visualização de Dados |
| Cálculo Diferencial e Integral I | Inteligência Artificial | Banco de Dados Integrado |
| Cálculo Vetorial e Geometria Analítica | Mineração de Dados | Administração de Banco de Dados |
| Equações Diferenciais Ordinárias | Métodos Numéricos I | Engenharia de Software e Sistemas |
| | Lógica Computacional Aplicada | Organização de Computadores e Sistemas |
| | Análise de Regressão I | Operacionais |
| | Teoria dos Grafos Aplicada | Paradigmas de Linguagens Computacionais |
| | Pesquisa Operacional | Laboratório de Ciência de Dados |
| | Metodologia Científica e Filosofia da | Projeto Integrador Avançado |
| | Ciência | Gestão da Informação e do Conhecimento |
| | Estatística Aplicada à Administração | Introdução à Programação |

de disciplinas semanticamente próximas produziu grupos com coesão média de $0{,}9079$, comprovando a capacidade do método não apenas de identificar equivalências, mas também de sintetizar disciplinas inéditas a partir de fontes heterogêneas.

Quanto à PP 02, a integração de IAGen com Análise de Grafos permitiu construir uma matriz curricular interdisciplinar e normativamente adequada. A lista final de 40 disciplinas alcançou similaridade média de $0{,}73$ em relação ao documento da SBC, valor estatisticamente equivalente aos currículos da UFC $(0{,}73)$ e da UFPB $(0{,}72)$. A estruturação via grafos — com a classificação das disciplinas em papéis pedagógicos (Básica, Específica e Complementar) realizada através da análise das métricas de centralidade do grafo — resultou em um formato em "funil" que reforça a progressão conceitual.

## 5. Considerações Finais

Este trabalho apresentou uma proposta que integra IAGen, Processamento de Linguagem Natural e Análise de Grafos a fim de elaborar matrizes curriculares e criar novas disciplinas através da fusão de conteúdos de disciplinas de outras matrizes e de conteúdos de normas e diretrizes curriculares. A abordagem se mostrou eficaz, indicando uma alta similaridade semântica entre disciplinas geradas e originais, além de boa aderência às diretrizes de formação, mesmo com ementas heterogêneas. A Análise de Grafos, por sua vez, organizou a matriz curricular com base em métricas de centralidade, revelando o papel de cada disciplina na trajetória acadêmica. Assim, o método combina inovação tecnológica, rigor pedagógico e transparência.

Como perspectivas de trabalhos futuros, propõe-se, inicialmente, explorar modelos de linguagem mais recentes e ampliar a base de dados para conferir maior robustez aos resultados, ao mesmo tempo em que se refinam os critérios de similaridade semântica para adaptá-los a contextos disciplinares específicos. Adicionalmente, a ponderação das arestas no grafo poderá ser aprimorada ao incorporar variáveis como nível de dificuldade, carga horária e pré-requisitos, tornando o mapeamento mais preciso. Pretende-se também estender o método à geração de componentes curriculares complementares — planos de ensino, trilhas de aprendizagem e sugestões de atividades avaliativas — e aprimorar a definição dos semestres, de modo a equilibrar a progressão de complexidade com a coesão e similaridade dos conteúdos ofertados em cada período.

## Agradecimentos

Os autores agradecem ao Observatório de Dados da Graduação (ODG) e à Gestão Institucional da UFRPE pelo essencial suporte e pelos recursos disponibilizados para a realização deste trabalho, desenvolvido no âmbito das iniciativas do ODG para fomentar a gestão universitária baseada em evidências. Ferramentas de Inteligência Artificial Generativa foram utilizadas como apoio na elaboração deste documento, incluindo o auxílio na redação, revisão textual e organização de conteúdo. Os autores revisaram e editaram criticamente todo o conteúdo gerado e assumem total responsabilidade pela obra final.

## Considerações Éticas

Este estudo utiliza exclusivamente dados de domínio público, consistindo em matrizes curriculares e ementas de disciplinas disponibilizadas oficialmente por instituições de ensino superior. Por não envolver a participação de seres humanos ou o uso de dados privados, a pesquisa dispensa a submissão a um Comitê de Ética em Pesquisa.

## Disponibilidade de artefatos

Os artefatos deste estudo serão disponibilizados pelos autores correspondentes[12].

## Referências

Al-Omari, M., Al-Hmouz, A., and Al-Khanjaf, Z. (2020). An nlp-based approach for course similarity detection in curriculum analysis. *International Journal of Advanced Computer Science and Applications*, 11(6).

Bahroun, Z., Anane, C., Ahmed, V., and Zacca, A. (2023). Transforming education: A comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. *Sustainability*, 15(17):12983.

Balakrishnan, V. and Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances. *Lecture notes on software engineering*, 2(3):262.

Bianchessi, C., editor (2021). *Educação, Currículo, Cultura Digital: reflexões para a escola na atualidade*. Editora Bagai, Bagé, RS.

Chris, E. and Sheriffdeen, K. (2024). Shaping ai-driven curriculum development: How generative ai is modern educational content.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2022). *Introduction to algorithms*. MIT press.

da Silveira, A. C., de Faria, A. V., de Azevedo, D. S., de Souza, D. S., Vieira, E. A. O., de Assis, F. A., Andrade, J. A., Braz, J. E., Pinto, K. E. V., Alves, R. L., et al. (2024). *PESQUISAS EM EDUCAÇÃO MEDIADA POR TECNOLOGIAS DIGITAIS DE INFORMAÇÃO E COMUNICAÇÃO*. Editora BAGAI.

DeepSeek-AI (2024). Deepseek-v3 technical report.

DeepSeek-AI (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Field, A. (2024). *Discovering statistics using IBM SPSS statistics*. Sage publications limited.

[12] `https://github.com/nefif/SBIE_2025.git`

Holmes, W., Miao, F., et al. (2023). *Guia para a IA generativa na educação e na pesquisa*. UNESCO Publishing.

Kumar, J. A., Zhuang, M., and Thomas, S. (2024). Chatgpt for natural sciences course design: Insights from a strengths, weaknesses, opportunities, and threats analysis. *Natural Sciences Education*, 53(2):e70003.

Lu, W. and Zoghi, B. B. (2024). Using generative ai for a graduate level capstone course design—a case study. In *2024 ASEE Annual Conference & Exposition*.

Mehta, D., Guruprasad, K. R., Wang, Y., Zhu, Y., and Hager, G. D. (2023). Curricullm: A framework for llm-based curriculum generation and task expansion in reinforcement learning. *arXiv preprint arXiv:2310.03153*.

Nussbaum, Z., Morris, J. X., Duderstadt, B., and Mulyar, A. (2024). Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

Otte, E. and Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, 28(6):441–453.

Padovano, A. and Cardamone, M. (2024). Towards human-ai collaboration in the competency-based curriculum development process: The case of industrial engineering and management education. *Computers and Education: Artificial Intelligence*, 7:100256.

Qu, K., Li, K. C., Wong, B. T., Wu, M. M., and Liu, M. (2024). A survey of knowledge graph approaches and applications in education. *Electronics*, 13(13):2537.

Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M., et al. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models*, volume 1. Wiley Online Library.

SBC and ABE (2023). *Referenciais de Formação para o Curso de Bacharelado em Ciência de Dados*. Porto Alegre: Sociedade Brasileira de Computação (SBC).

Stavrinides, P. and Zuev, K. M. (2023). Course-prerequisite networks for analyzing and understanding academic curricula. *Applied Network Science*, 8(1):19.

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12.

Tuzón, P., Salvà, A. S., and Fernández-Gracia, J. (2024). Complex networks approach to curriculum analysis and subject integration: a case study on physics and mathematics. *arXiv preprint arXiv:2412.15929*.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt.

Yamagiwa, H., Oyama, M., and Shimodaira, H. (2024). Revisiting cosine similarity via normalized ica-transformed embeddings. *arXiv preprint arXiv:2406.10984*.