# Modelagem Matemática para Recomendação de Orientadores em Programas de Pós-Graduação Stricto Sensu

## Uma Abordagem Baseada na Análise de Métricas Acadêmicas e Científicas

Radi Melo Martins

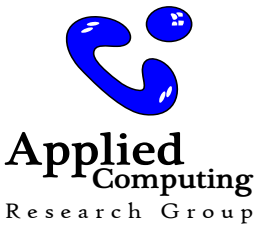
Universidade Regional do Noroeste do Estado do Rio Grande do Sul

Tese de Doutorado

Orientadora:
Dra. Fabricia Carneiro Roos Frantz

Coorientador:
Dr. Rafael Zancan Frantz






Agosto 2025

First published in March 2025 by
Applied Computing Research Group - GCA
Department of Exact Sciences and Engineering
Rua Lulu Ilgenfritz, 480 - São Geraldo
Ijuí, 98700-000, Brazil.


http://www.gca.unijui.edu.br
gca@unijui.edu.br
Author: radi.martins@sou.unijui.edu.br and radi@uft.edu.br



**Financiamento:** A pesquisa desenvolvida neste trabalho teve o apoio do Grupo de Pesquisa em Computação Aplicada - GCA da Universidade Regional do Noroeste do Estado do Rio Grande do Sul - UNIJUÍ e da Universidade Federal do Tocantins - UFT.

Catalogação na Publicação

M386m

Martins, Radi Melo.

Modelagem matemática para recomendação de orientadores em programas de pós-graduação Stricto Sensu : uma abordagem baseada na análise de métricas acadêmicas e científicas / Radi Melo Martins. – Ijuí, 2025.

212 f. : il. ; 30 cm.

Tese (doutorado) – Universidade Regional do Noroeste do Estado do Rio Grande do Sul (Campus Ijuí). Modelagem Matemática.

"Orientadora: Profa. Dra. Fabricia Carneiro Roos Frantz".

"Coorientador: Prof. Dr. Rafael Zancan Frantz".

1. Orientação acadêmica. 2. Chatbot. 3. Seleção de orientadores. 4. Programas de pós-graduação. 5. Inteligência artificial. 6. Modelo matemático. 7. Sistema de recomendação. 8. Plataforma lattes. 9. Análise de dados acadêmicos. 10. Escolha acadêmica. I. Frantz, Fabricia Carneiro Roos. II. Frantz, Rafael Zancan. III. Título.

CDU: 378:004.85

**UNIJUÍ - Universidade Regional do Noroeste do Estado do Rio Grande do Sul**

**Programa de Pós-Graduação *Stricto Sensu* em Modelagem Matemática e Computacional**

A Comissão Examinadora, abaixo assinada, APROVA a Tese

**MODELAGEM MATEMÁTICA PARA RECOMENDAÇÃO DE ORIENTADORES EM PROGRAMAS DE PÓS-GRADUAÇÃO STRICTO SENSU: UMA ABORDAGEM BASEADA NA ANÁLISE DE MÉTRICAS ACADÊMICAS E CIENTÍFICAS**

Elaborada por

**Radi Melo Martins**

Como requisito parcial para a obtenção do título de Doutor em Modelagem Matemática e Computacional

**Comissão Examinadora**

Profa. Dra. Fabricia Carneiro Roos Frantz
(Orientadora/UNIJUÍ)

Prof. Dr. Sandro Sawicki
(UNIJUÍ)

Prof. Dr. Rafael Zancan Frantz
(Coorientador/UNIJUÍ)

Profa. Dra. Raquel Mainardi Pillat Basso
(UNIPAMPA)

Profa. Dra. Airam Romcy Sausen
(UNIJUÍ)

Prof. Dr. Valdemar Vicente Graciano Neto
(UFG)

Ijuí, RS, 14 de agosto de 2025.

Dedico este trabalho, com amor e reconhecimento, à minha esposa Vanuza Martins, pelo apoio incondicional, paciência e encorajamento ao longo desta jornada. Aos meus filhos, Lucas, cuja força, dedicação e bom humor iluminam nossos dias, e Mateus, motivo de alegria e orgulho.

À minha mãe, Maria das Graças, por sua força incansável e por apoiar os meus estudos.

# *Conteúdo*

# Índice de Figuras

# Índice de Tabelas

# *Índice de Abreviaturas*

**COMUNG** Consórcio das Universidades Comunitárias Gaúchas

**Feevale** Universidade Feevale

**GCA** Grupo de Pesquisa em Computação Aplicada

**GDPR** General Data Protection Regulation

**IA** Inteligência Artificial

**IES** Instituições de Ensino Superior

**IC** Índice de Completude

**IR** Índice de Recomendação

**LaSalle** Universidade La Salle

**LGPD** Lei Geral de Proteção de Dados

**MBA** Master of Business Administration

**ORCID** Open Researcher and Contributor ID

**PUCRS** Pontifícia Universidade Católica do Rio Grande do Sul

**PPGMMC** Programa de Pós-Graduação em Modelagem Matemática e Computacional

**TIC** Tecnologias de Informação e Comunicação

**UCPel** Universidade Católica de Pelotas

**UCS** Universidade de Caxias do Sul

**UFN** Universidade Franciscana

**UNICRUZ** Universidade de Cruz Alta

**UNIJUÍ** Universidade Regional do Noroeste do Estado do Rio Grande do Sul

**UNISC** Universidade de Santa Cruz do Sul

**UNISINOS** Universidade do Vale do Rio dos Sinos

**UNIVATES** Universidade do Vale do Taquari

**UPF** Universidade de Passo Fundo

**URCAMP** Universidade da Região da Campanha

**URI** Universidade Regional Integrada do Alto Uruguai e das Missões

**XML** Extensible Markup Language

**XSD** XML Schema Definition

**CNPq** Conselho Nacional de Desenvolvimento Científico e Tecnológico

**PLN** Processamento de Linguagem Natural

**SR** Sistema de Recomendação

# Índice de Símbolos

$IR$ Índice de Recomendação.

$\alpha_i$ Peso do critério $i$ no cálculo do IR.

$P_{\textbf{Área}}$ Pontuação de similaridade de área entre orientador e aluno.

$GA$ Pontuação atribuída à coincidência na Grande Área.

$A$ Pontuação atribuída à coincidência na Área.

$SA$ Pontuação atribuída à coincidência na Subárea.

$E$ Pontuação atribuída à coincidência na Especialidade.

$i$ Índice que representa o nível de conhecimento comparado.

$n$ Número total de níveis de conhecimento comparados.

$P_{\textbf{Experiência}}$ Pontuação de experiência em orientações.

$m$ Quantidade de orientações de mestrado.

$d$ Quantidade de orientações de doutorado.

$M$ Valor de referência para orientações de mestrado.

$D$ Valor de referência para orientações de doutorado.

$\lambda_m$ Peso atribuído ao mestrado no cálculo de experiência.

$\lambda_d$ Peso atribuído ao doutorado no cálculo de experiência.

$Q$ Fator de qualidade baseado em publicações.

$P_r$ Número total de publicações.

$P_{90}$ Percentil 90 de publicações na amostra.

$P_{\textbf{Eficiência}}$ Pontuação de Eficiência das orientações (taxa de conclusão).

$OC_i$ Orientações concluídas no nível $i$.

$OA_i$ Orientações em andamento no nível $i$.

$TC_i$ Taxa de conclusão no nível $i$.

$w_i$ Peso do nível no cálculo da Eficiência.

$P_{\textbf{Produção}}$ Pontuação baseada em produção científica.

$\alpha, \beta$ Pesos internos do cálculo de produção.

$C_{\textbf{total}}$ Número total de citações.

$C_{5\textbf{anos}}$ Número de citações nos últimos 5 anos.

$h$ Índice H.

$i10$ Índice i10.

$P_{\textbf{Colaboração}}$ Pontuação baseada em colaboração acadêmica.

$P_{\textbf{banca}}$ Número de participações em bancas.

$Co$ Número total de coautores.

$w_1, w_2$ Pesos internos de participação em bancas e coautoria.

$P_{\textbf{Pesquisa}}$ Pontuação baseada em projetos e atividades de pesquisa.

$PP$ Número total de atividades de pesquisa.

$a_i$ Peso da atividade de pesquisa do tipo $i$.

$N_i$ Quantidade de ocorrências de atividade do tipo $i$.

# Agradecimentos

*Cem vezes por dia eu me lembro que minha vida interior e exterior depende do trabalho de outros homens, vivos e mortos, e que devo me esforçar para retribuir na mesma medida em que recebi e continuo recebendo.*

*Albert Einstein, Físico teórico (1879–1955)*

Agradeço, com profunda admiração e respeito, à minha orientadora, Dra. Fabricia Roos Frantz, por sua orientação segura, dedicação incansável e exemplo de força e compromisso, mesmo diante dos desafios pessoais enfrentados ao longo desta jornada. Sua presença constante, apoio generoso e confiança foram fundamentais para a realização deste trabalho.

Meus sinceros agradecimentos ao coorientador, Dr. Rafael Zancan Frantz, cuja notável capacidade técnica e experiência foram essenciais para as valiosas contribuições ao desenvolvimento desta pesquisa. Sou igualmente grato à Dra. Airam Romcy Sausen e ao Dr. Sandro Sawicki, pelas aulas e contribuições fundamentais ao longo deste percurso. Agradeço a todos os professores da UNIJUÍ, pela formação sólida e apoio contínuo.

Registro meus sinceros agradecimentos aos membros da banca examinadora, Dr. Valdemar Vicente Graciano Neto e Dra. Raquel Mainardi Pillat Basso, pelas observações criteriosas e contribuições valiosas para o aprimoramento deste trabalho.

Agradeço a colaboração dos alunos de graduação Isadora Beckmann e Samuel Golnik, e a parceria dos colegas do Grupo de Pesquisa em Computação Aplicada, Luis Patiño, Regis Schuch, Mailson Teles, Eldair Fabrício e Matheus Rehbein, cuja presença e contribuição enriqueceram esta jornada. Agradeço a todos os demais colegas do GCA pelas experiências compartilhadas ao longo desta jornada de intensos aprendizados e conquistas.

# Resumo

*O início de todas as coisas é pequeno.*

*Marcus T. Cicero, Filósofo Romano (106 AC - 43 AC)*

A escolha de orientadores e programas de pós-graduação stricto sensu é um processo cada vez mais complexo e essencial para o sucesso dos estudantes, exigindo a análise de um vasto volume de dados relacionados a programas, grupos de pesquisa, linhas de investigação, além de fatores como compatibilidade acadêmica, reputação do orientador, produção científica e redes de colaboração. Tradicionalmente, essa decisão é tomada de forma manual e frequentemente baseada em percepções subjetivas e aconselhamentos informais, o que pode resultar em escolhas pouco fundamentadas e desalinhadas com os objetivos acadêmicos e profissionais dos estudantes. Para dar suporte a essa problemática, este trabalho propõe o desenvolvimento de um modelo matemático e computacional que integra técnicas de ciência de dados para calcular um Índice de Recomendação baseado em indicadores acadêmicos extraídos de currículos da Plataforma Lattes e de outras bases de dados internacionais. Este modelo visa oferecer recomendações objetivas, considerando critérios como produtividade científica, experiência em orientação e alinhamento de áreas de pesquisa. A avaliação do modelo foi realizada por meio de análises estatísticas aplicadas a currículos de docentes, extraídos da Plataforma Lattes, incluindo testes de completude de dados, análise descritiva das métricas, identificação de *outliers* e avaliação da consistência e robustez dos resultados. Dessa forma, a abordagem proposta busca promover um processo de escolha de orientadores em programas de pós-graduação mais objetivo, fundamentado e alinhado aos objetivos individuais dos estudantes.

**Palavras-chave:** Orientação Acadêmica, Chatbot, Seleção de Orientadores, Programas de Pós-Graduação, Inteligência Artificial, Modelo Matemático, Sistema de Recomendação, Plataforma Lattes, Análise de Dados Acadêmicos, Escolha Acadêmica.

# Abstract

*The beginnings of all things are small.*

*Marcus T. Cicero, Roman philosopher (106 BC – 43 BC)*

The selection of academic advisors and *stricto sensu* graduate programs is an increasingly complex and crucial process for students, requiring the analysis of a vast amount of data related to programs, research groups, research lines, as well as factors such as academic compatibility, advisor reputation, scientific production, and collaboration networks. Traditionally, this decision is made manually and often based on subjective perceptions and informal advice, which can lead to poorly grounded choices misaligned with students' academic and professional goals. To support this challenge, this project proposes the development of a mathematical and computational model that integrates data science techniques to calculate a Recommendation Index (Índice de Recomendação (IR)) based on academic indicators extracted from Lattes Platform curricula and other international databases. This model aims to provide objective recommendations, considering criteria such as scientific productivity, mentoring experience, and research area alignment. The model validation was conducted through statistical analyses applied to faculty curricula extracted from the Lattes Platform, including data completeness tests, descriptive analysis of metrics, identification of outliers, and evaluation of the consistency and robustness of the results. Thus, the proposed approach seeks to foster a more objective, well-founded, and individualized process for selecting academic advisors and graduate programs.

**Keywords:** Academic Advising, Chatbot, Advisor Selection, Graduate Programs, Artificial Intelligence, Mathematical Model, Recommendation System, Lattes Platform, Academic Data Analysis, Academic Choice.

# Resumen

El comienzo de todas las cosas es pequeño.

Marco Tulio Cicerón, Filósofo Romano (106 a.C. – 43 a.C.)

La elección de directores de tesis y de programas de posgrado *stricto sensu* es un proceso cada vez más complejo y crucial para los estudiantes, que requiere el análisis de un gran volumen de datos relacionados con programas, grupos de investigación, líneas de investigación, así como factores como la compatibilidad académica, la reputación del director, la producción científica y las redes de colaboración. Tradicionalmente, esta decisión se toma de manera manual y, a menudo, basada en percepciones subjetivas y asesoramientos informales, lo que puede conducir a elecciones poco fundamentadas y desalineadas con los objetivos académicos y profesionales de los estudiantes. Para apoyar esta problemática, este proyecto propone el desarrollo de un modelo matemático y computacional que integra técnicas de ciencia de datos para calcular un Índice de Recomendación (IR) basado en indicadores académicos extraídos de currículos de la Plataforma Lattes y de otras bases de datos internacionales. Este modelo tiene como objetivo ofrecer recomendaciones objetivas, considerando criterios como la productividad científica, la experiencia en orientación y la alineación de áreas de investigación. La validación del modelo se realizó mediante análisis estadísticos aplicados a currículos de docentes extraídos de la Plataforma Lattes, incluyendo pruebas de completitud de datos, análisis descriptivos de métricas, identificación de *outliers* y evaluación de la consistencia y robustez de los resultados. De esta manera, el enfoque propuesto busca promover un proceso de elección de directores y programas de posgrado más objetivo, fundamentado y alineado con los objetivos individuales de los estudiantes.

**Palabras clave:** Orientación Académica, Chatbot, Selección de Directores, Programas de Posgrado, Inteligencia Artificial, Modelo Matemático, Sistema de Recomendación, Plataforma Lattes, Análisis de Datos Académicos, Elección Académica.

# *Parte I*

# *Prefácio*

# Capítulo 1

# Introdução

*A jornada de mil milhas*
*começa com um único passo.*

*Provérbio Chinês*

Neste capítulo, apresenta-se uma visão geral da tese, iniciando-se pelo contexto em que a pesquisa se insere e destacando-se os principais desafios relacionados à escolha de orientadores acadêmicos em programas de pós-graduação *stricto sensu*. Na sequência, abordam-se a motivação e a justificativa que fundamentam o desenvolvimento do estudo. Em seguida, explicitam-se os objetivos gerais e específicos que orientam a investigação. Apresenta-se também uma síntese da metodologia adotada, a qual contempla o uso de técnicas de coleta e análise de dados, desenvolvimento de um modelo matemático e avaliação dos resultados obtidos. Por fim, descreve-se a organização estrutural da tese, oferecendo ao leitor uma visão clara das etapas e capítulos subsequentes.

## 1.1 Contexto da Pesquisa

A pós-graduação no Brasil tem apresentado crescimento consistente ao longo da última década, conforme evidenciado pelos dados apresentados na Figura 1.1, extraídos do Painel Lattes [†1]. O número de titulados em programas de mestrado acadêmico passou de 16.496 em 2014 para 44.637 em 2023, representando a maior parcela da formação em nível *stricto sensu*. No mesmo período, os títulos de doutorado aumentaram de 15.436 para 25.131, enquanto o mestrado profissional evoluiu de 2.008 para 8.626 titulados. Os dados refletem a

[†1] http://bi.cnpq.br/painel/formacao-atuacao-lattes/

expansão da pós-graduação e a consequente produção de grandes volumes de informações acadêmicas, exigindo soluções eficazes para sua análise e aplicação.

Do ponto de vista demográfico, os dados da plataforma apontam uma predominância feminina entre os pós-graduados 55,67%, seguida por 44,25% do sexo masculino. Em relação à raça/cor, 59,8% se autodeclaram brancos, 24,6% pardos e 6,8% pretos. A maior concentração de titulados está na faixa etária de 30 a 44 anos, que representa mais de 66% do total, com destaque para os grupos de 35 a 39 anos (25,56%) e de 30 a 34 anos (21,51%). Esse perfil reforça que a formação em nível de pós-graduação ocorre majoritariamente em fase de consolidação profissional, evidenciando a importância de políticas públicas que promovam maior inclusão e permanência de grupos diversos nesses programas.

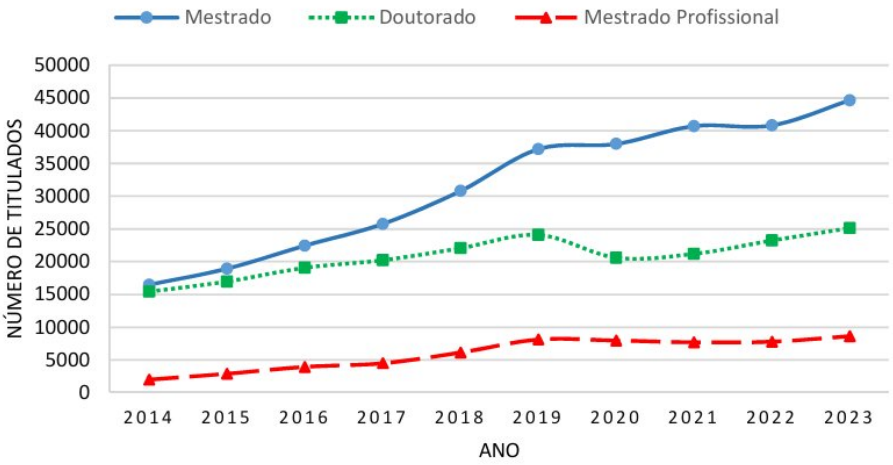



**Figura 1.1**: *Titulação na Pós-Graduação no Brasil.*

Considerando o perfil demográfico dos titulados e a crescente demanda por qualificação em nível de pós-graduação, torna-se ainda mais relevante compreender os fatores que influenciam a trajetória acadêmica dos estudantes, especialmente no que se refere à escolha do orientador, uma decisão reconhecida na literatura como determinante para o sucesso acadêmico [43]. Essa etapa exige atenção e reflexão cuidadosa por parte dos discentes, pois envolve a necessidade de alinhar interesses acadêmicos, expectativas pessoais e profissionais, além de aspectos qualitativos, como suporte emocional e empatia do orientador, que também são destacados na literatura por impactarem significativamente a motivação e o bem-estar dos estudantes [66].

Dentre os fatores a serem considerados, destaca-se a compatibilidade acadêmica [20], pois o orientador deve possuir expertise na área de estudo do

aluno, o que requer uma análise criteriosa de publicações, projetos e linhas de pesquisa. A relevância e a atualidade da produção científica do possível orientador são indicativos de sua capacidade de oferecer suporte técnico e metodológico adequado. Além disso, a experiência prévia em projetos similares pode representar um diferencial na condução das etapas da pesquisa e no acompanhamento das metodologias adotadas. A reputação e o *networking* do orientador também são aspectos fundamentais. Um orientador bem estabelecido e reconhecido em sua área de atuação pode abrir portas e proporcionar oportunidades únicas, como participação em conferências, publicações em revistas de alto impacto e colaborações internacionais. Nesse sentido, a orientação acadêmica é reconhecida como um dos pilares da experiência na pós-graduação *stricto sensu*, influenciando diretamente o desenvolvimento da pesquisa, a satisfação do discente e sua trajetória profissional [43, 66].

A inserção acadêmica do orientador, refletida em sua atuação em bancas e coautorias, pode ampliar o acesso a colaborações e oportunidades na carreira do discente. Coautorias com pesquisadores produtivos, por exemplo, tendem a aumentar as chances de publicação em periódicos de alto impacto [71], enquanto a atuação em equipes interdisciplinares contribui para perspectivas mais amplas e integradas [51]. Elementos como estilo de orientação, disponibilidade e relação interpessoal também exercem forte influência sobre a qualidade da orientação [45, 60, 86], sendo recomendada, inclusive, a busca por informações junto a orientandos anteriores para avaliar compatibilidade e dinâmica de trabalho. Esses elementos subjetivos são reconhecidos pela literatura como fatores qualitativos importantes na escolha do orientador, complementando os aspectos quantitativos.

A produção acadêmica do orientador pode ser verificada por meio do currículo Lattes. A quantidade e qualidade das publicações, a participação em eventos científicos, a atuação em bancas de defesa e a presença em projetos de pesquisa são indicativos da atividade acadêmica e do reconhecimento do orientador na comunidade científica [26]. Adicionalmente, estudos indicam que orientadores com alta produtividade acadêmica tendem a fornecer melhores recursos e oportunidades para seus orientandos, incluindo acesso a financiamento de pesquisa, redes de colaboração internacionais e maior visibilidade no campo de estudo [77].

Escolher um orientador também leva em consideração as ambições profissionais do discente, seja na academia, indústria ou outros setores. Orientadores com experiências diversas podem oferecer orientações mais alinhadas a esses objetivos, incluindo apoio em decisões de carreira e transição para o mercado de trabalho [18, 110]. Este aspecto qualitativo pode ser parcialmente inferido a partir de dados relacionados à inserção profissional do orientador, tais

como produção técnica, participação em projetos aplicados e experiência em programas voltados à prática profissional, registrados na plataforma Lattes.

Portanto, a escolha de um orientador em programas de pós-graduação *stricto sensu* é um processo que requer análise criteriosa e bem informada de diversos fatores, como compatibilidade acadêmica, colaboração entre pares, disponibilidade para pesquisa e orientação, experiência prévia em orientação, relação interpessoal, políticas institucionais, objetivos de carreira, entre outros. Embora os elementos qualitativos sejam reconhecidamente importantes, é fundamental realizar uma análise criteriosa de indicadores objetivos , como produção científica, experiência em orientação e participação em bancas, pois esses critérios proporcionam uma base concreta e estruturada para uma escolha bem fundamentada [11, 18, 108].

Apesar da relevância dos indicadores acadêmicos já mapeados, ainda existe uma lacuna de pesquisa na integração sistemática dessas variáveis em ferramentas capazes de apoiar, de forma objetiva, a escolha de orientadores. Faltam soluções que consolidem e analisem grandes volumes de dados acadêmicos de maneira padronizada, transparente e personalizada. Para suprir essa lacuna, este trabalho propõe o desenvolvimento de um IR fundamentado em métricas extraídas da Plataforma Lattes e complementadas por bases internacionais, oferecendo suporte mais consistente ao processo de seleção de orientadores em programas de pós-graduação *stricto sensu*.

## 1.2 Motivação e Justificativa

Diante da expansão da pós-graduação no Brasil e do aumento da diversidade de perfis entre os estudantes, a escolha do orientador assume um papel ainda mais estratégico na definição das trajetórias acadêmicas. Embora amplamente reconhecida como uma decisão essencial, a literatura mostra que esse processo ainda é, muitas vezes, conduzido de forma subjetiva, com base em afinidade pessoal ou reputação informal, sem considerar de maneira sistemática indicadores objetivos de qualidade, experiência ou inserção acadêmica [65]. Essa decisão estratégica impacta significativamente o desempenho acadêmico, a qualidade da pesquisa desenvolvida e a satisfação geral dos estudantes de pós-graduação *stricto sensu* [11, 108]. Entretanto, esse processo tem se tornado cada vez mais complexo devido ao crescente volume e diversidade das informações acadêmicas que precisam ser analisadas. Essa complexidade exige abordagens mais precisas e eficientes, uma vez que a análise envolve uma quantidade crescente e heterogênea de dados acadêmicos [11, 108].

Com a consolidação da Plataforma Lattes como base nacional de dados acadêmicos, tornou-se possível sistematizar informações relevantes para apoiar o

| **Categoria** | **Descrição** | **Indicadores extraídos** |
|---|---|---|
| Formação Acadêmica | trajetória educacional do pesquisador | graduação, mestrado, doutorado, pós-doutorado |
| Produção Científica | contribuições científicas e publicações | artigos, livros, capítulos, trabalhos em eventos |
| Orientações | experiência na formação de novos pesquisadores | orientações concluídas e em andamento |
| Participação em Bancas | reconhecimento pela comunidade acadêmica | atuação em bancas de qualificação e defesa |
| Projetos de Pesquisa | envolvimento em pesquisa financiada | projetos ativos e concluídos e instituições envolvidas |
| Produção Técnica | aplicações práticas e inovação | softwares, produtos técnicos, relatórios |
| Atuação Profissional | experiência institucional e cargos ocupados | instituições, funções, duração da atuação |
| Colaboração Acadêmica | interações e redes de coautoria | participação em artigos com múltiplos autores e instituições |
| Premiações e Títulos | reconhecimento institucional e científico | prêmios recebidos, distinções acadêmicas |
| Áreas de Atuação | especializações e linhas de pesquisa | grande área, área, subárea e especialidade |
| Idiomas | internacionalização e comunicação científica | idiomas e níveis de proficiência |

**Tabela 1.1**: *Indicadores analisados no Currículo Lattes.*

processo de análise do contexto acadêmico [27, 87]. O currículo Lattes organiza informações sobre a trajetória acadêmica e científica dos pesquisadores em diversas categorias. A análise estruturada dessas categorias permite identificar padrões, métricas e indicadores objetivos da atuação docente. A Tabela 1.1 apresenta um resumo das dimensões extraídas a partir dessa base de dados, que podem contribuir para contextualizar a atuação dos orientadores em

diferentes aspectos do processo formativo.

Apesar disso, verifica-se uma lacuna significativa relacionada à integração e análise sistemática desses dados, bem como à escassez de modelos computacionais eficazes para sistemas de recomendação aplicados especificamente ao ambiente acadêmico [42, 87]. Diante desse cenário, este trabalho propõe um modelo matemático apoiado por técnicas computacionais, visando automatizar o processo de recomendação por meio da análise objetiva de variáveis relevantes, como produção acadêmica, reputação dos orientadores e áreas de interesse dos estudantes. Como resultado dessa pesquisa, espera-se que seja possível implementar um modelo capaz de oferecer recomendações assertivas, personalizadas e alinhadas às necessidades específicas dos estudantes.

A justificativa para a realização desta pesquisa fundamenta-se na necessidade crescente de soluções tecnológicas capazes de integrar e analisar grandes volumes de dados acadêmicos heterogêneos, especialmente no contexto educacional. Há uma lacuna identificada na literatura quanto ao desenvolvimento de modelos matemáticos e computacionais que combinem técnicas de ciência de dados e sistemas de recomendação, proporcionando uma abordagem prática e aplicável em instituições de ensino superior [42, 87]. Tal abordagem permite que estudantes e instituições possam realizar escolhas acadêmicas mais eficientes e bem fundamentadas, baseadas em critérios quantitativos e qualitativos extraídos de fontes acadêmicas científicas.

A relevância de uma abordagem baseada em dados estruturados se intensifica diante da complexidade da tarefa. Consolidar informações sobre produção acadêmica, atuação em projetos, orientação de discentes, participação em bancas e redes de coautoria exige tempo e conhecimento técnico, o que torna o processo inacessível para muitos estudantes. Soluções automatizadas, como indicam Prass et al. [87] e Galego [42], possibilitam a extração precisa dessas informações a partir de currículos Lattes, oferecendo uma base sólida para apoiar a tomada de decisão.

Além disso, como demonstrado por Silva et al. [100], os estudantes de pós-graduação têm motivações diversas  desde o desejo de seguir carreira científica até a busca por qualificação profissional. Essa pluralidade reforça a necessidade de um sistema de recomendação que considere tanto os perfis dos estudantes quanto os indicadores objetivos dos possíveis orientadores. Neste

[†1]Os indicadores da Tabela 1.1 são variáveis essencialmente quantitativas extraídas do Currículo Lattes em formato XML. Aspectos qualitativos, como interação entre orientador e orientando, ética ou impacto social, não foram contemplados por exigirem instrumentos específicos de coleta, mas são reconhecidos como relevantes e indicados para trabalhos futuros.

sentido, a aplicação de um modelo matemático e computacional apresenta-se como uma alternativa. Ao utilizar técnicas de análise de dados, o modelo é capaz de processar grandes volumes de informação com precisão, gerando recomendações baseadas em critérios concretos. Dessa forma, a decisão sobre a escolha de um orientador torna-se mais informada, segura e alinhada com os objetivos individuais de cada estudante.

Assim, esta pesquisa propõe o desenvolvimento de um modelo baseado em indicadores objetivos extraídos do currículo Lattes, consolidando variáveis como produção científica, experiência em orientação e participação em bancas em um índice de recomendação. Esse índice busca apoiar o processo de escolha de orientadores no âmbito dos programas de pós-graduação, oferecendo recomendações baseadas em dados objetivos, porém sem caráter prescritivo, cabendo ao estudante a decisão final com base em sua análise pessoal e contextual.

Desta maneira, o trabalho apresenta uma metodologia de recomendação fundamentada em um modelo matemático e computacional, voltada ao domínio educacional, com o objetivo de apoiar a tomada de decisões por parte de estudantes e instituições de ensino na escolha de orientadores acadêmicos. A justificativa para adotar uma abordagem orientada por dados reside justamente na necessidade de integrar e processar informações heterogêneas disponíveis em plataformas acadêmicas como o currículo Lattes, permitindo recomendações personalizadas que estejam alinhadas às expectativas e necessidades específicas dos estudantes [42, 87].

## 1.3 Problema e Questões de Pesquisa

A escolha de um orientador, embora amplamente discutida na literatura, ainda carece de ferramentas que auxiliem os estudantes a tomarem essa decisão com base em critérios objetivos, especialmente diante da crescente complexidade dos dados acadêmicos disponíveis.

Uma dificuldade central nesse processo é a falta de métodos objetivos e padronizados que possam ajudar os estudantes a escolher orientadores de forma fundamentada. Estudos indicam que, frequentemente, as decisões são tomadas com base em percepções pessoais ou conselhos informais, sem uma análise sistemática e criteriosa de dados como a produtividade acadêmica, a experiência em orientação e as redes de colaboração do orientador [26, 66]. Este processo, que pode ser marcado pela subjetividade, carece de uma abordagem que integre informações relevantes e confiáveis, como a qualidade e o impacto das publicações, o número de orientações concluídas e o histórico de colaboração em projetos de pesquisa [71].

As soluções atualmente disponíveis para esse problema, como a análise manual dos currículos Lattes, a consulta a colegas e o uso de plataformas acadêmicas online, apresentam limitações significativas. A análise manual pode ser demorada e sujeita a vieses pessoais, enquanto as plataformas existentes, como ResearchGate e Google Scholar, não conseguem capturar todos os fatores que influenciam a qualidade da orientação acadêmica. Além disso, a utilização dessas plataformas depende do nível de engajamento do orientador com as redes sociais acadêmicas, o que pode não refletir com precisão sua capacidade de orientação ou o impacto de sua pesquisa [42, 87].

A falta de critérios claros e objetivos para a escolha de orientadores compromete a transparência do processo e limita a capacidade dos estudantes de tomar decisões informadas e alinhadas com seus objetivos acadêmicos e profissionais [11]. A ausência de abordagens estruturadas que integrem múltiplos critérios simultaneamente pode comprometer a compatibilidade entre orientador e orientando, resultando em conflitos, insatisfação e até evasão nos programas de pós-graduação [64, 77].

O desenvolvimento de uma solução eficaz para esse problema enfrenta vários desafios. Um dos principais obstáculos é a necessidade de reunir e analisar uma ampla gama de dados de diferentes fontes, incluindo currículos, publicações, participações em eventos científicos e feedback de ex-orientandos. Esses dados são frequentemente dispersos, variados e não padronizados, o que torna difícil criar uma base de comparação consistente e justa [87].

Outra dificuldade está na construção de modelos matemáticos e computacionais que possam automatizar esse processo, processando grandes volumes de dados e diferentes tipos de informações de maneira eficiente e precisa. Definir variáveis e parâmetros que reflitam adequadamente os diversos aspectos da relação de orientação, como a qualidade da pesquisa e a interação entre orientador e orientando, é um desafio adicional. Essa complexidade requer o desenvolvimento de algoritmos que consigam capturar as nuances da orientação acadêmica e oferecer recomendações personalizadas e confiáveis [66, 110].

Dado o contexto da pesquisa e o problema, foram estabelecidas as seguintes questões de pesquisa:

**QP1: Quais critérios objetivos (Indicadores Acadêmicos) devem ser priorizados na construção de um modelo matemático e computacional de recomendação para a seleção de orientadores acadêmicos?**

Esta questão busca identificar, de forma estruturada, os indicadores acadêmicos objetivos que podem ser utilizados para representar com precisão

o perfil de orientadores. O foco está na análise de variáveis mensuráveis que possibilitem a construção de um modelo de recomendação, alinhado às demandas dos programas de pós-graduação e às expectativas dos estudantes.

**QP2: Como garantir a precisão e a personalização das recomendações de orientadores?**

Aqui foram explorados os métodos e abordagens que podem ser empregados para assegurar que as recomendações de orientadores sejam simultaneamente precisas e adaptadas ao perfil individual dos estudantes. O objetivo é investigar como a integração de dados acadêmicos estruturados, o uso de modelos matemáticos e a incorporação de mecanismos interativos, como chatbots, podem contribuir para uma recomendação mais informada, personalizada e alinhada às expectativas e objetivos acadêmicos dos usuários.

**QP3: De que maneira a integração de dados de plataformas acadêmicas pode aumentar a precisão das recomendações de orientadores?**

O foco desta questão consiste em investigar como a consolidação de informações provenientes de diferentes plataformas acadêmicas pode contribuir para o aprimoramento da qualidade e precisão das recomendações de orientadores. O foco está em analisar o potencial da integração de múltiplas fontes de dados, como Plataforma Lattes e Google Scholar, por exemplo, para ampliar a abrangência das métricas utilizadas, enriquecendo a representação do perfil acadêmico dos docentes e possibilitando comparações mais robustas e contextualizadas.

**QP4: Como um modelo matemático e computacional pode reduzir o trabalho manual na seleção de orientadores?**

Esta questão busca compreender de que forma a adoção de um modelo matemático e computacional pode automatizar tarefas tradicionalmente manuais envolvidas na seleção de orientadores acadêmicos. O objetivo é investigar como técnicas de extração de dados, normalização de variáveis e geração de recomendações estruturadas podem reduzir o esforço operacional e aumentar a eficiência, transparência e objetividade no processo decisório.

## 1.4 Contribuição do Trabalho

Para superar os desafios relacionados à escolha de orientadores na pós-graduação *stricto sensu*, este trabalho propõe o desenvolvimento de um modelo

de recomendação automatizado, fundamentado em técnicas de ciência de dados e modelagem matemática. O objetivo é apoiar estudantes na tomada de decisões mais informadas e alinhadas aos seus objetivos acadêmicos e profissionais, por meio da análise de critérios objetivos extraídos de bases de dados acadêmicas estruturadas.

A proposta consiste na construção de um índice matemático de recomendação capaz de identificar o perfil acadêmico-científico do orientador com base em múltiplos fatores acadêmicos relevantes. Entre as variáveis consideradas estão: produção científica, experiência em orientação, participação em bancas, redes de coautoria e alinhamento temático entre o pesquisador e o estudante. Essas variáveis serão processadas por meio de técnicas computacionais, permitindo a normalização dos dados e a comparação padronizada entre diferentes perfis de orientadores, ajudando o candidato acadêmico no processo de escolha.

Embora a decisão final sobre a escolha de um orientador continue envolvendo aspectos subjetivos, como afinidade acadêmica, estilo de trabalho e expectativas pessoais, a abordagem proposta visa complementar esse processo ao oferecer uma recomendação estruturada e baseada em dados. A modelagem matemática permitirá que as variáveis sejam ponderadas de forma flexível, possibilitando adaptações conforme o contexto institucional ou os interesses do estudante.

O modelo será implementado com base na extração automatizada de dados acadêmicos, utilizando métodos de análise para identificar padrões, consolidar informações dispersas e gerar recomendações embasadas. Com isso, espera-se contribuir para a transparência e eficiência no processo de seleção de orientadores, reduzindo a fragmentação de informações e promovendo uma escolha mais fundamentada [64, 71, 87, 104].

Espera-se que o modelo contribua para tornar o processo de escolha de orientadores mais ágil, transparente e fundamentado, ao sintetizar dados objetivos em recomendações personalizadas. Com isso, pretende-se não apenas reduzir o esforço envolvido na análise de informações acadêmicas, mas também fortalecer a qualidade das orientações e, consequentemente, da produção científica nos programas de pós-graduação.

## 1.5 Objetivos

Esta seção apresenta os objetivos da pesquisa, divididos em geral e específicos.

### 1.5.1 Geral

Criar uma metodologia para recomendação de orientador de pós-graduação *stricto sensu*, orientada à seleção de docentes de cursos de pós-graduação *stricto sensu*, de instituições de ensino superior brasileiras, baseada no alinhamento entre os objetivos de um possível candidato à discente e a trajetória de pesquisa do docente em sua área de atuação.

### 1.5.2 Específicos

i. Desenvolver um sistema de extração de dados: Implementar ferramentas para coletar e processar informações dos currículos de professores a partir da Plataforma Lattes e de plataformas internacionais.

ii. Analisar dados acadêmicos e identificar possíveis orientadores: Analisar a trajetória de docentes da pós-graduação *stricto sensu* das instituições de ensino superior brasileiras, a partir de dados do curriculum Lattes, e recomendar com base em um índice de recomendação, aqueles docentes que mais se alinhem com as preferências do candidato a discente.

iii. Construir matematicamente o Índice de Recomendação: Estruturar um modelo composto por equações que avaliam dimensões distintas da atuação acadêmica como alinhamento de áreas, experiência em orientações, qualidade das orientações, produção científica, colaboração com pares e participação em pesquisa, gerando pontuações parciais agregadas por meio de ponderações ajustáveis que compõem a equação geral do índice.

## 1.6 Metodologia

A metodologia empregada na condução da pesquisa apresentada nesta tese, foi centrada na extração de variáveis do Currículo Lattes, processadas por meio de técnicas computacionais e organizadas em indicadores quantitativos, que serviram como base para o desenvolvimento de um modelo matemático que permitiu gerar sugestões de orientadores. Essa abordagem combina a formalidade dos modelos matemáticos com o poder analítico dos dados.

Inicialmente, realizou-se um estudo exploratório para compreender os parâmetros mais relevantes na escolha de um orientador acadêmico. Essa etapa envolveu a análise da literatura científica especializada [63, 86, 104] e de

documentos institucionais, como as diretrizes do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), permitindo identificar um conjunto de indicadores recorrentes na avaliação da atuação de orientadores. Dentre esses indicadores, destacam-se o número de orientações concluídas e em andamento, a produção acadêmica, a participação em bancas examinadoras e métricas bibliométricas, como o índice H, o índice i10 e o total de citações.

A inclusão das métricas $h$-index e $i10$-index foi motivada não apenas por sua relevância como indicadores consolidados de impacto científico, mas também pela viabilidade técnica de sua extração automatizada via Google Scholar. Essas métricas foram obtidas a partir de arquivos JSON estruturados, o que possibilitou sua integração direta ao modelo computacional, garantindo padronização, escalabilidade e consistência na análise dos perfis acadêmicos.

Além do embasamento teórico, a seleção das variáveis consideradas no modelo proposto também foi guiada por critérios empíricos, especialmente pela análise do Índice de Completude (IC) dos arquivos XML da Plataforma Lattes. Essa análise possibilitou identificar quais variáveis estavam presentes com maior frequência e consistência nos currículos analisados, garantindo maior robustez aos dados utilizados. Assim, optou-se por priorizar indicadores que fossem simultaneamente relevantes do ponto de vista teórico e viáveis do ponto de vista técnico, assegurando que o modelo pudesse ser operacionalizado com confiabilidade.

Variáveis subjetivas, como estilo de orientação, perfil comportamental ou afinidade interpessoal, embora relevantes para a escolha de um orientador, foram excluídas da modelagem por não estarem representadas de forma estruturada nos dados disponíveis. Dessa forma, o conjunto final de critérios reflete uma síntese entre fundamentos teóricos consolidados, evidências empíricas extraídas dos dados e limitações inerentes ao escopo da análise automatizada.

Com os critérios definidos, passou-se à fase de extração e processamento dos dados. Foram desenvolvidos *scripts* em Python para coletar e estruturar informações contidas nos arquivos Extensible Markup Language (XML) dos currículos Lattes. Essa etapa exigiu a criação de mecanismos para identificar e tratar inconsistências nos dados, garantindo a confiabilidade das informações extraídas.

Na literatura, existem algumas ferramentas para a extração automatizada de dados de currículo Lattes, tais como ScriptLattes e LattesMiner,[†2],

[†2]ScriptLattes: ferramenta de extração automatizada de dados de currículos da Plataforma Lattes, desenvolvida em Python. LattesMiner: utilitário voltado à coleta e organização de dados da base Lattes para fins de análise bibliométrica e redes de colaboração.

como será apresentado no Capítulo 3. Essas ferramentas genéricas, em geral, não permitem flexibilidade na seleção de variáveis ou personalização do processamento dos dados, o que pode limitar a aplicabilidade em contextos específicos. Além disso, o uso dessas ferramentas de extração automatizada de dados apresenta alguns desafios, como a necessidade de adaptação contínua às mudanças na estrutura dos dados da Plataforma Lattes e aos mecanismos de segurança, como o CAPTCHA [†3], que impactam especialmente ferramentas baseadas em extração direta de currículos.

Portanto, embora ferramentas prontas representem soluções práticas e eficientes em cenários amplos, o desenvolvimento de *scripts* personalizados pode oferecer vantagens para investigações que demandam maior nível de customização e profundidade analítica. Dessa forma, neste trabalho, optou-se pelo desenvolvimento de *scripts* personalizados em Python, visando maior controle sobre o processo, flexibilidade na adaptação dos algoritmos e compatibilidade com os objetivos específicos da pesquisa. O uso de *scripts* para extração de dados da Plataforma Lattes, frequentemente em Python, tem sido uma prática comum para análise quantitativa.

A partir disso, o modelo matemático foi então implementado para calcular pontuações individuais de cada orientador com base nas variáveis selecionadas. Portanto, a pesquisa teve um caráter predominantemente quantitativo, pois os dados foram analisados por meio de estatísticas descritivas e normalização de variáveis, permitindo a construção de uma lista de orientadores baseada em critérios objetivos. Além disso, para garantir a flexibilidade do modelo, a proposta foi pensada para permitir ajustes nos pesos atribuídos a cada variável, possibilitando adaptações conforme as especificidades de diferentes programas de pós-graduação.

A avaliação do modelo foi conduzida por inspeção manual de cerca de 50 currículos Lattes, a fim de conferir a coerência entre os resultados computados e os registros efetivamente presentes nos currículos. Em seguida, os códigos foram aplicados a cerca de 1.800 currículos, abrangendo diferentes áreas do conhecimento. Essa etapa incluiu análises estatísticas, correlação entre variáveis, simulações de sensibilidade e avaliação da completude dos dados. A combinação dessas abordagens permitiu verificar se o modelo se comportava de forma consistente diante de diferentes perfis acadêmicos e se os dados utilizados apresentavam estrutura e qualidade suficientes para gerar resultados confiáveis.

[†3]CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) é um teste de Turing utilizado para distinguir entre usuários humanos e programas de computador, especialmente em processos de autenticação e prevenção de spam em websites. O CAPTCHA geralmente exige que o usuário resolva um desafio visual, como identificar caracteres distorcidos ou selecionar imagens que correspondam a um determinado critério.

A metodologia adotada neste trabalho também incorpora, de forma implícita, as etapas clássicas da modelagem matemática descritas por Rosa e Orey [91]: formulação do problema, construção do modelo, análise de sua utilidade e posterior validação. A formulação teve origem na necessidade de recomendar orientadores com base em critérios objetivos; a construção ocorreu por meio da definição de variáveis e estruturação das equações que compõem o IR; a utilidade foi verificada por testes empíricos; e a consistência dos resultados, assegurada por análises estatísticas. Com isso, reforça-se o alinhamento desta proposta aos fundamentos da modelagem matemática aplicada ao contexto educacional. A seguir, são detalhadas cada uma dessas etapas:

i. Levantamento do Problema e Definição dos Objetivos. A escolha de orientadores em programas de pós-graduação é, muitas vezes, orientada por critérios subjetivos e informações fragmentadas, o que dificulta a identificação do docente mais compatível com os interesses e objetivos do estudante. Fornecer recomendações objetivas e baseadas em dados quantitativos sobre a experiência e o engajamento de possíveis orientadores pode contribuir para mitigar esse problema. Como resultado dessa etapa, foram definidos como objetivos o desenvolvimento de métodos para análise automática de currículos acadêmicos, a extração de métricas relacionadas à produção científica, orientações, projetos de pesquisa e demais indicadores relevantes, bem como a construção de um índice de recomendação fundamentado nesses dados.

ii. Estudo Teórico Aplicado. A revisão da literatura não apenas fundamentou teoricamente a pesquisa, como também serviu de base metodológica para a definição das variáveis e critérios utilizados no modelo de recomendação. Foram analisadas publicações acadêmicas e diretrizes institucionais que abordam a orientação acadêmica e sistemas de recomendação, destacando métricas como produção científica, número de orientações concluídas, participação em bancas e índices bibliométricos. Essa fundamentação teórica foi essencial para garantir que o modelo estivesse alinhado com práticas reconhecidas e respaldadas na literatura especializada.

iii. Definição do Modelo Matemático. A construção do IR, elemento central da metodologia de recomendação proposta, foi baseada em variáveis quantitativas extraídas da Plataforma Lattes e do Google Scholar. Essas variáveis foram organizadas em seis categorias principais: área de atuação, experiência em orientação, qualidade das orientações, produção científica, colaboração acadêmica e participação em pesquisa. Cada categoria é representada por uma equação específica, descrita no Capítulo 5.1,

e corresponde a uma dimensão relevante da atuação docente. A partir dessas variáveis, foram calculadas métricas normalizadas e ponderadas, resultando em uma pontuação final atribuída a cada docente. O índice gerado permite a comparação entre diferentes orientadores de maneira objetiva e ajustável, possibilitando a priorização de diferentes critérios conforme o perfil ou a necessidade do candidato. Embora os pesos atribuídos às categorias ainda não sejam personalizáveis pelo usuário final, foram realizadas simulações com variações nos critérios de ponderação para avaliar a sensibilidade do modelo em diferentes contextos.

iv. Desenvolvimento dos Algoritmos em Python. Para automatizar a extração e o processamento dos dados dos currículos, foram desenvolvidos *scripts* em Python que obtém as variáveis identificadas no modelo matemático. Esses algoritmos extraem dados dos arquivos XML da Plataforma Lattes, calculam as pontuações acadêmicas e realizam a normalização e ponderação dos dados. Além disso, os scripts foram configurados para calcular de forma automática, agilizando o processo de análise e recomendação.

v. Testes e avaliação Estatística. A fase de avaliação consistiu na aplicação de testes estatísticos sobre currículos de docentes vinculados à pós-graduação *stricto sensu*, com o objetivo de verificar a consistência do sistema de recomendação proposto. Foram avaliadas a completude dos dados XML com base no XML Schema Definition (XSD) oficial da Plataforma Lattes, a coerência interna das variáveis utilizadas e a distribuição dos valores atribuídos ao IR. As análises incluíram correlação entre variáveis, identificação de *outliers*, simulações de sensibilidade e avaliação da pontuação de qualidade. Esses procedimentos permitiram ajustar parâmetros do modelo e garantir que os resultados obtidos refletissem adequadamente a realidade acadêmica dos orientadores analisados.

Informações mais detalhadas sobre os materiais utilizados, os procedimentos técnicos adotados e os métodos estatísticos aplicados em cada uma dessas etapas podem ser consultadas no Capítulo 4, que descreve a execução da pesquisa.

## 1.7 Resumo do Capítulo

Este capítulo apresentou os desafios da seleção de orientadores em programas de pós-graduação, evidenciando a subjetividade e a falta de critérios padronizados nos métodos tradicionais. Discutiu-se a importância da modelagem matemática para estruturar um IR, permitindo uma avaliação objetiva

baseada em métricas acadêmicas. Foi proposta a criação de um modelo matemático e computacional que automatize a análise de currículos acadêmicos, integrando estatística e técnicas computacionais para tornar o processo mais eficiente e transparente. Esse modelo busca reduzir esforços manuais e melhorar a precisão da escolha de orientadores.

## 1.8 Estrutura do Documento

Esta tese está organizada em quatro partes principais, conforme descrito a seguir:

**Parte I: Prefácio.** Apresenta os fundamentos iniciais da pesquisa. Contém o Capítulo 1 e os que tratam do contexto da pós-graduação, da motivação para o estudo, da formulação dos problemas de pesquisa, hipótese, objetivos e metodologia adotada. Também são discutidas as soluções existentes e suas limitações, culminando na proposição do modelo apresentado nesta tese.

**Parte II: Revisão da Literatura.** Oferece a base teórica e contextual da pesquisa. O Capítulo 2 aborda conceitos de modelagem matemática, sistemas de recomendação, chatbots, pós-graduação e a Plataforma Lattes. O Capítulo 3 reúne os principais trabalhos relacionados, apontando avanços e lacunas. O Capítulo 3.10 analisa as soluções atuais para a seleção de orientadores, destacando limitações e potencialidades dos modelos matemáticos e computacionais aplicados ao contexto acadêmico. Por fim, o Capítulo 3.11 discute as ideias referentes as soluções.

**Parte III: Pesquisa Desenvolvida.** Detalha o desenvolvimento da solução proposta. No Capítulo 4, descreve-se a metodologia, ferramentas utilizadas, estrutura dos dados, equações do modelo e os códigos implementados. O Capítulo 5 descreve o IR embasado no modelo matemático desenvolvido. O Capítulo 6 apresenta a avaliação da proposta por meio de análises estatísticas, testes com dados reais e estudos de completude dos currículos Lattes.

**Parte IV: Considerações Finais.** Reúne as reflexões finais da pesquisa. O Capítulo 7 discute as conclusões obtidas a partir dos resultados. O Capítulo 8 destaca as principais contribuições da tese, e o Capítulo 9 propõe direções para trabalhos futuros relacionados à expansão e aplicação do modelo desenvolvido.

# *Parte II*

# *Revisão da Literatura*

# Capítulo 2

# Referencial Teórico

*Pense como um homem sábio,*
*mas comunique-se em linguagem simples.*

*William B. Yeats, Dramaturgo e poeta irlandês (1865--1939)*

Neste capítulo, apresenta-se a fundamentação teórica que embasa o desenvolvimento do modelo proposto nesta tese. Na Seção 2.1, apresenta-se um panorama da pós-graduação no Brasil, contextualizando seu papel na formação de pesquisadores e na qualificação do ensino superior. Na Seção 2.2, trata-se da Plataforma Lattes como fonte primária de dados acadêmicos estruturados no contexto nacional. Complementarmente, na Seção 2.3, discute-se plataformas internacionais como Google Scholar, Scopus e ORCID, que ampliam a cobertura e a diversidade das informações utilizadas no modelo. Por fim, na Seção 2.4, apresenta-se o conceito de reputação acadêmica, analisando suas métricas tradicionais e abordagens complementares que buscam uma avaliação mais ampla e precisa da trajetória de pesquisadores. Na Seção 2.6, aborda-se a construção e a aplicação de modelos matemáticos, destacando suas propriedades, limitações e o papel da modelagem na representação de fenômenos reais. Na Seção 2.7, discute-se os sistemas de recomendação, com ênfase em abordagens clássicas como a filtragem colaborativa, a filtragem baseada em conteúdo e os modelos híbridos, bem como sua evolução por meio da integração com técnicas de inteligência artificial. Na Seção 2.8, explora-se os *chatbots* como ferramentas de interação com os usuários e sua utilização na personalização de recomendações acadêmicas. Em seguida, apresenta-se um resumo do capítulo, integrando os principais conceitos discutidos.

## 2.1 Pós-Graduação

A experiência universitária no Brasil tem uma trajetória relativamente recente. A primeira universidade brasileira, a Universidade do Rio de Janeiro, foi estabelecida em 1920. Posteriormente, o decreto 19.851, de 1931, instituiu o regime universitário, promovendo o surgimento de universidades formalmente institucionalizadas, como a Universidade de São Paulo (USP), fundada em 1934, e a Universidade do Distrito Federal (UDF), criada em 1935 [90]. Essas universidades introduziram um novo modelo de organização do ensino superior, que combinava de forma integrada o ensino e a pesquisa, servindo como referência para as futuras experiências de ensino superior no Brasil [90].

Embora tenha existido por um curto período, de 1935 a 1939, a UDF desempenhou um papel crucial na formalização do conceito de universidade no Brasil. Seu projeto de criação foi guiado pelos princípios de liberdade e autonomia universitária, destacando-se como um centro voltado para a pesquisa e a produção de conhecimento [41, p. 167]. A institucionalização da pós-graduação no Brasil ocorreu em um momento posterior. O termo pós-graduação foi mencionado pela primeira vez na década de 1940, no Artigo 71 do Estatuto da Universidade do Brasil. Contudo, o verdadeiro impulso para o desenvolvimento dos programas de pós-graduação deu-se na década de 1960, com a publicação do Parecer CFE nº 977/65 conhecido como Parecer Sucupira. Este documento, aprovado em 3 de dezembro de 1965, conceituou, estruturou e formalizou a pós-graduação no Brasil, tomando como referência o modelo norte-americano [55, 78, 95].

O Parecer Sucupira estabeleceu duas modalidades de pós-graduação: Lato Sensu, incluindo cursos de especialização e Master of Business Administration (MBA), e *Stricto Sensu*, que abrange o mestrado (acadêmico e profissional) e o doutorado, ambos sem pré-requisito entre si [55]. Newton Sucupira defendeu a adoção do modelo norte-americano, argumentando em favor de um sistema de pós-graduação que integrasse ensino e pesquisa, formando pesquisadores, docentes e profissionais altamente qualificados [55, 78].

O desenvolvimento da pós-graduação no Brasil fundamentou-se na necessidade de atender às exigências de um contexto de expansão tecnológica e industrial, assim como à formação de pesquisadores e profissionais altamente qualificados. Como aponta Nascimento [78], o Parecer CFE nº 977/65 estabeleceu três objetivos centrais para os cursos de pós-graduação: a qualificação de docentes para o ensino superior, o incentivo à pesquisa científica e o treinamento de pessoal intelectual de alto nível, capaz de atender às demandas do desenvolvimento nacional em diversos setores.

Além disso, o documento argumenta que o sistema de pós-graduação norte-americano, adotado como modelo, visa proporcionar não apenas a especialização, mas também uma diversificação vertical dos níveis de estudo, com a introdução de um ciclo avançado de formação que ultrapassa o âmbito da graduação, permitindo a ampliação do conhecimento e o desenvolvimento de competências avançadas em áreas específicas [55].

Assim, a pós-graduação no Brasil foi concebida como um ciclo essencial para o fortalecimento das universidades enquanto centros de produção de ciência e cultura. A criação de cursos de pós-graduação, submetidos à regulamentação do Conselho Federal de Educação e ao reconhecimento ministerial, visou evitar o desvirtuamento do sistema e assegurar a qualidade e a excelência acadêmica [55]. Portanto, a estruturação da pós-graduação brasileira, inspirada no modelo norte-americano, consolidou-se como um instrumento fundamental para a qualificação do ensino superior, a formação de pesquisadores, a expansão do conhecimento científico e o atendimento das demandas do desenvolvimento nacional [78].

## 2.2 Plataforma Lattes

O Currículo Lattes é uma ferramenta essencial no cenário acadêmico brasileiro, destinada à organização e ao compartilhamento de dados sobre a produção científica. Sua criação está diretamente relacionada ao desenvolvimento das Tecnologias de Informação e Comunicação (TIC), especialmente no que se refere à internet. No entanto, a concepção de um sistema voltado à coleta e organização de dados sobre pesquisadores e suas pesquisas remonta a um período anterior à popularização da internet. Na década de 1980, cientistas do CNPq desenvolveram um protótipo baseado em um formulário padrão para registrar os currículos dos pesquisadores brasileiros. Esse método analógico, embora pioneiro, apresentava limitações significativas, como a dificuldade de acesso e disseminação da produção científica de forma ampla e eficiente [†1].

Durante a década de 1990, com o uso do sistema operacional DOS, o CNPq avançou no desenvolvimento de um formulário eletrônico para coleta dessas informações. Em um contexto de escassa conectividade, os currículos eram entregues fisicamente em disquetes, evidenciando a carência de uma infraestrutura digital adequada. A consolidação da internet no final da década possibilitou o lançamento oficial do Currículo Lattes em 1999, marcando um importante avanço na gestão de informações acadêmicas no Brasil. O sistema

[†1] https://memoria.cnpq.br/web/portal-lattes/sobre-a-plataforma

recebeu o nome *Lattes* em homenagem ao físico César Lattes, um dos mais importantes cientistas brasileiros, reconhecido por sua contribuição à física nuclear e à descoberta do méson pi[†2].

### 2.2.1 A Importância do Currículo Lattes

Atualmente, o Currículo Lattes é uma exigência fundamental no meio acadêmico. Ele serve como uma base de dados abrangente que avalia pesquisadores, professores e alunos, facilitando a seleção de consultores e especialistas em diversas áreas científicas. A plataforma também gera estatísticas sobre a distribuição da pesquisa científica no Brasil, permitindo um acompanhamento detalhado do desenvolvimento da ciência e a identificação de áreas carentes e avanços. A necessidade de uma plataforma digital tornou-se evidente ao longo dos anos, e o sucesso do Currículo Lattes no Brasil levou o CNPq a licenciar o software de forma gratuita para outros países, oferecendo também consultoria técnica para sua implantação. Na América Latina, o modelo foi adotado por países como Colômbia, Equador, Chile, Peru e Argentina. Além disso, o Currículo Lattes expandiu-se para outros continentes, incluindo Portugal e Moçambique. Essa internacionalização do sistema demonstra sua eficácia e a necessidade global de ferramentas padronizadas para o registro da produção científica [†3].

Para aqueles inseridos no meio acadêmico, o Currículo Lattes é essencial para o desenvolvimento de uma carreira científica. Em processos de seleção, seja para bolsas, projetos de pesquisa ou vagas em cursos de pós-graduação, o Currículo Lattes é frequentemente o documento de referência. Ele contém um detalhamento abrangente das experiências acadêmicas e científicas, atendendo a um padrão nacional que sistematiza e centraliza a produção científica no Brasil [†4]. A criação e manutenção desse currículo são cruciais, pois ele é adotado pela maioria das instituições de fomento, institutos de pesquisa e universidades, facilitando o reconhecimento e a avaliação da produtividade acadêmica dos pesquisadores [†5]. A criação do Currículo Lattes deve ser acompanhada por uma navegação no site da Plataforma Lattes para conhecer modelos de currículos já cadastrados, o que pode fornecer uma visão geral do que se espera. Além disso, é ressaltada a importância de manter o currículo sempre atualizado e detalhado, pois isso é essencial para a concessão de bolsas e a participação em projetos de pesquisa

[†2] https://memoria.cnpq.br/web/portal-lattes/cesare-giulio-lattes
[†3] https://memoria.cnpq.br/web/portal-lattes/historico
[†4] https://memoria.cnpq.br/web/portal-lattes/sobre-a-plataforma
[†5] https://memoria.cnpq.br/web/portal-lattes/cesare-giulio-lattes

O Currículo Vitae utilizado principalmente para candidaturas a vagas de emprego, contém um resumo das qualificações e experiências profissionais do candidato. Em contraste, o Currículo Lattes detalha toda a trajetória acadêmica e a produção científica do indivíduo, seguindo um modelo padrão gerado pela plataforma Lattes. Enquanto o Currículo Vitae deve ser breve e convincente, o Currículo Lattes é extenso e abrangente, requerendo atualizações constantes para refletir todas as atividades acadêmicas e de pesquisa do indivíduo. Informações inseridas no Currículo Lattes são verificadas pelo CNPq por meio de diplomas e certificados, desaconselhando o registro de atividades não verificáveis ou falsas.[†6]

### 2.2.2 A importância da Atualização

A Plataforma Lattes é uma ferramenta essencial para o planejamento e a gestão de currículos acadêmicos no Brasil. Ela integra bases de dados de currículos, grupos de pesquisa e instituições, facilitando a formulação de políticas públicas e o fomento à ciência e tecnologia [†7]. O processo de cadastro na Plataforma Lattes é fundamental para qualquer pesquisador ou estudante que deseje registrar e gerenciar sua trajetória acadêmica e profissional.

A atualização constante do Currículo Lattes é indispensável. Sempre que novas atividades forem realizadas, elas devem ser adicionadas ao currículo para mantê-lo completo e relevante. A plataforma também oferece a opção de gerar versões impressas, úteis para processos seletivos e apresentações acadêmicas. A manutenção de um currículo atualizado é particularmente importante para quem participa de seleções de bolsas, programas de pós-graduação e projetos de pesquisa.

Desde sua criação, o Currículo Lattes evoluiu para acompanhar avanços tecnológicos e demandas acadêmicas, consolidando-se como referência internacional em gestão curricular. A plataforma contribui para a formação e atualização de pesquisadores, promove transparência na produção científica e apoia a concessão de bolsas e auxílios, tornando-se um pilar do desenvolvimento científico no Brasil.

Na Tabela 2.1, apresentam-se informações extraídas e analisadas da Plataforma Lattes, a fim de fornecer uma visão abrangente de suas funcionalidades, uso e relevância.

[†6] https://memoria.cnpq.br/web/portal-lattes/cesare-giulio-lattes
[†7] https://memoria.cnpq.br/web/portal-lattes/cesare-giulio-lattes

| **Divisão** | **Descrição** | **Importância** | **Itens** |
|---|---|---|---|
| Plataforma | Visão geral da Plataforma Lattes e sua missão. | Contextualiza a origem, propósito e estrutura da plataforma. | Sobre, história, missão, objetivos. |
| Termos de Uso | Diretrizes para utilização dos serviços e dados da Plataforma Lattes. | Define direitos e responsabilidades dos usuários, garantindo uso ético e correto dos dados. | Uso aceitável, direitos de propriedade, responsabilidades. |
| Dados e Estatística | Métricas agregadas da Plataforma Lattes. | Oferece uma visão quantitativa do alcance e impacto da plataforma. | Número de currículos, publicações, grupos de pesquisa, etc. |
| Acordos Institucionais | Parcerias formais estabelecidas com outras entidades ou plataformas. | Reflete a integração e colaboração da Plataforma Lattes com outras instituições e sistemas. | Instituições parceiras, natureza do acordo, objetivos. |
| Extração de Dados | Mecanismos disponíveis para obter dados em formato exportável. | Permite que pesquisadores e instituições acessem e usem os dados em suas próprias análises. | Ferramentas, formatos de exportação, APIs, limitações. |
| Outras Bases | Informação sobre integração com outras bases de dados ou sistemas. | Amplia o escopo e utilidade da Plataforma Lattes, conectando-a a outros repositórios. | Bases associadas, descrição, links, integrações. |

**Tabela 2.1**: *Descrição das divisões da Plataforma Lattes.*

## 2.3 Plataformas Internacionais

Existem diversas plataformas internacionais, como Google Scholar, ResearchGate, ORCID, Publons, Academia.edu, Scopus e Mendeley, que, assim como a plataforma Lattes, auxiliam na construção da identidade digital de pesquisadores e ampliam a visibilidade da produção científica. Essas ferramentas permitem criar perfis com publicações, citações e métricas de impacto, além de favorecer a colaboração entre acadêmicos. Enquanto Google Scholar e Scopus oferecem indicadores bibliométricos estruturados, ResearchGate, Academia.edu e Mendeley priorizam engajamento e compartilhamento, e o ORCID destaca a rastreabilidade da produção. O Publons, por sua vez, foca em revisões e contribuições editoriais. Na Tabela 2.2, apresenta-se um resumo das principais plataformas de avaliação acadêmica.

### 2.3.1 Google Scholar

O *Google Scholar* [†8] é uma plataforma de busca acadêmica gratuita lançada em 2004, que indexa textos completos e metadados de artigos, livros, teses e relatórios técnicos. Ela permite que pesquisadores encontrem literatura de diversas áreas e gerenciem suas publicações, criando perfis pessoais para acompanhar citações e calcular métricas de impacto. A plataforma também oferece funcionalidades como alertas de citação e o *Google Scholar Metrics*, que classifica periódicos acadêmicos com base no índice h5, refletindo o impacto de artigos nos últimos cinco anos [52].

Dados numéricos e variáveis:

- Número de citações: total de citações recebidas por todas as publicações do pesquisador.
- h-index: índice que mede tanto a produtividade quanto o impacto das publicações, representando o número de publicações com pelo menos 'h' citações.
- i10-index: número de publicações com pelo menos 10 citações.
- Publicações: inclui títulos, autores, ano de publicação, periódico/conferência, resumos e links para acesso completo.

[†8] https://scholar.google.com

| Plataforma | Métricas | Descrição | Variáveis |
|---|---|---|---|
| Google Scholar | Número de citações, índice h, índice i10, publicações | Métricas de impacto baseadas em citações | h-index, i10-index, número total de citações |
| Research Gate | RG Score, publicações, citações, impacto das publicações | Combina engajamento, visualizações e citações | Métrica própria baseada em interações e citações |
| ORCID | Publicações, afiliações, financiamentos, projetos | Identidade acadêmica com foco em rastreabilidade e transparência | Não utiliza métricas quantitativas padronizadas |
| Publons | Revisões por pares, edições, publicações | Avalia atuação em periódicos e revisões científicas | Número de revisões, edições, impacto por citações |
| Academia.edu | Seguidores, visualizações, downloads | Engajamento e visibilidade acadêmica pública | Número de seguidores, downloads e visualizações |
| Scopus | Publicações, citações, índice h, afiliações | Base bibliométrica estruturada | h-index (Scopus), contagem de publicações e citações |
| Mendeley | Publicações, leitores, grupos, citações (via integração) | Rede social acadêmica focada em leitura e colaboração | Número de leitores, grupos, publicações salvas |

**Tabela 2.2**: *Plataformas Acadêmicas.*

## 2.4 Reputação Acadêmica

A reputação de um pesquisador pode ser entendida como um reflexo de sua trajetória acadêmica, incluindo não apenas sua produção científica, mas também seu impacto na formação de novos pesquisadores, participação em projetos de pesquisa, envolvimento em comitês editoriais e atividades em redes de colaboração científica. Normalmente, a reputação tem sido mensurada por meio

de métricas bibliométricas, como o índice $h$ (h-index)[†9] e o índice $g$ (g-index)[†10], bem como outras variantes que enfatizam o volume de citações recebidas pelas publicações científicas.

Apesar de amplamente utilizados na avaliação do impacto acadêmico, esses indicadores apresentam limitações por não considerarem outras dimensões relevantes da atuação acadêmica, como orientação, gestão científica ou colaboração institucional [112]. Cervi [24] destaca, nesse sentido, a importância de abordagens mais abrangentes e adaptáveis para a avaliação da reputação científica, que levem em conta diferentes aspectos da contribuição dos pesquisadores à comunidade acadêmica.

O Rep-Model desenvolvido por Cervi [24] organiza as informações dos pesquisadores em cinco categorias principais: Identificação (ID): Nome, instituição e grau de instrução. Orientação (ORI): Número de orientações concluídas em mestrado, doutorado e pós-doutorado. Banca (BAN): Participação em bancas de mestrado e doutorado. Comitê (COM): Envolvimento em comitês científicos, como coordenação de conferências, participação em corpo editorial de periódicos e revisão de artigos. Publicação (PUB): Produção bibliográfica em artigos de periódicos, capítulos de livros, livros completos, trabalhos completos em conferências, participação em projetos de pesquisa, software desenvolvido e rede de coautoria. O Rep-Index é uma métrica desenvolvida para quantificar a reputação do pesquisador com base nos elementos do Rep-Model. Essa métrica permite a atribuição de pesos diferentes a cada elemento, o que possibilita uma avaliação mais justa e adaptável para diferentes áreas do conhecimento.

A pesquisa de Cervi [24] compara sua proposta com abordagens tradicionais de avaliação acadêmica, como o h-index e seus derivados (g-index), ($AR$-index)[†11] e o ($e$-index)[†12] são métricas adicionais utilizadas para avaliar a produtividade e o impacto acadêmico. [53, 92, 113]. Enquanto esses índices são amplamente utilizados, eles apresentam limitações:

A dependência exclusiva de citações é uma limitação do h-index e suas variantes, que se baseiam apenas na contagem de citações e desconsideram

[†9]O h-index, proposto por Hirsch, mede o impacto de um pesquisador com base no número de publicações que receberam pelo menos $h$ citações cada [48].

[†10]O g-index, proposto por Egghe, é uma métrica alternativa ao h-index, onde um pesquisador com g-index = $g$ tem pelo menos $g$ artigos cujas citações somadas são pelo menos $g^2$ [35].

[†11]O AR-index é uma extensão do h-index que leva em consideração a idade das publicações, ponderando as citações pelo tempo desde a publicação, a fim de reduzir o viés temporal.

[†12]O e-index foi introduzido para complementar o h-index, capturando o excesso de citações que não são consideradas pelo h-index, refletindo melhor o impacto total das publicações de um pesquisador.

outros aspectos da trajetória do pesquisador [22]. A baixa sensibilidade ao impacto real da pesquisa é uma limitação dos índices tradicionais, como o h-index, pois um pesquisador pode ter muitas citações sem necessariamente ter um impacto acadêmico significativo em áreas como formação de recursos humanos e inovação [61]. Além disso, esses índices não conseguem diferenciar as dinâmicas de publicação e citação de diferentes áreas de conhecimento. Por exemplo, na Ciência da Computação, conferências são mais valorizadas, enquanto na área de Humanidades, livros podem ser mais relevantes [16].

A adoção de modelos abrangentes para avaliação acadêmica amplia a análise do impacto dos pesquisadores, indo além das citações. Isso fomenta práticas equilibradas e permite o desenvolvimento de métricas integrativas, garantindo uma avaliação mais completa e significativa da produção científica.

## 2.5 Modelagem

De acordo com Aris [7], é fundamental destacar que, apesar da crescente popularidade do termo modelagem matemática nos tempos modernos, a prática em si remonta às origens da matemática. O século XVII marcou o início da era moderna, e foi nesse período que ocorreu uma revolução significativa, iniciada por Kepler e seguida por Galileu, Descartes e Fermat, culminando nas contribuições de Newton e Leibniz. Essa revolução permitiu o surgimento da matemática moderna, que impulsionou o progresso da ciência e da engenharia. Como resultado, vivenciamos uma revolução tecnológica que influenciou nossa vida atual de maneira mais profunda do que qualquer outro evento [39].

Nesse contexto, a modelagem matemática se consolidou como um processo relevante para a descrição e análise de fenômenos do mundo real, desempenhando um papel fundamental no desenvolvimento de diversas áreas do conhecimento, incluindo a computação. A interação entre a matemática e outras ciências tem sido essencial para o avanço da própria matemática, sobretudo quando se reconhece a presença de estruturas profundas na interseção entre esses campos, como ressaltado por Bender [14].

Além de sua fundamentação teórica, a modelagem matemática apresenta aplicações práticas em áreas como engenharia, climatologia e demografia, contribuindo para a compreensão de fenômenos e a otimização de processos [34]. Dessa forma, consolida-se como ferramenta essencial para a análise precisa de sistemas complexos em múltiplos domínios do conhecimento.

Complementando essa abordagem, desde a perspectiva da Engenharia de Software, um modelo de engenharia é uma representação seletiva de algum sistema da realidade, elaborada com o objetivo de capturar, de forma precisa e

concisa, todas as propriedades essenciais relevantes a um conjunto específico de preocupações[97]. Trata-se de uma versão reduzida e gerenciável da realidade, adequada ao processamento computacional. Essa concepção reforça o papel dos modelos como ferramentas para compreender, prever, comunicar e especificar sistemas, mesmo com suas limitações. Assim, a proposta deste trabalho, ao estruturar um modelo matemático para recomendação de orientadores, alinha-se a essa perspectiva, ao simplificar e abstrair aspectos relevantes da realidade acadêmica para apoiar a tomada de decisão.

## 2.6 Modelos como Ferramenta Analítica

No contexto das ciências e da formulação teórica, a criação e aplicação de modelos permanecem componentes essenciais para o estudo de fenômenos complexos. Conforme destacado por Humi [49], a modelagem matemática constitui um processo estruturado que envolve a formulação de problemas do mundo real, a aplicação de técnicas matemáticas para resolvê-los e a interpretação dos resultados no contexto original. Por meio de simplificações e abstrações apropriadas, os modelos permitem representar aspectos essenciais dos sistemas estudados, facilitando a compreensão, a análise e a previsão de comportamentos ou tendências observáveis no mundo real.

No cotidiano, os modelos estão amplamente incorporados em diferentes esferas. Exemplos incluem mapas rodoviários, que orientam rotas e trajetos, e mapas geológicos, que delineiam a composição do solo e auxiliam engenheiros e geólogos na tomada de decisões. De maneira análoga, modelos matemáticos aplicados ao desempenho estudantil permitem representar variáveis educacionais e avaliar o impacto de metodologias pedagógicas na aprendizagem [49]. Cada modelo é concebido para atender a um propósito específico e, embora inevitavelmente limitado por simplificações e pressupostos, mantém sua utilidade dentro do contexto para o qual foi formulado [7, 49]. Seu valor depende diretamente da capacidade de representar fielmente os aspectos essenciais do sistema real e da habilidade do pesquisador em interpretar os resultados com discernimento [49, 76]. Quando as previsões geradas se revelam inadequadas, o modelo pode ser ajustado ou reformulado; ainda assim, a prática da modelagem frequentemente oferece vantagens significativas em relação à ausência de qualquer estrutura formal de análise [7, 76].

Segundo Humi [49], a modelagem matemática requer um equilíbrio delicado entre a abstração teórica e a aplicação prática, possibilitando tanto a generalização de comportamentos por meio de estruturas simplificadas quanto a análise de situações específicas mediante abordagens mais detalhadas. A evolução dos

recursos computacionais reforça essa dualidade, ao fornecer meios eficientes para simular e explorar sistemas complexos com maior rigor e flexibilidade. Para Mityushev et al. [76], o processo de modelagem é intrinsecamente dinâmico e iterativo, exigindo que o pesquisador transite continuamente entre a formulação matemática, a observação empírica e a adaptação contextual. Assim, os modelos não apenas descrevem sistemas, mas também atuam como instrumentos heurísticos que promovem a formulação de hipóteses e o avanço contínuo do conhecimento científico.

### 2.6.1 Modelo de Ranqueamento PageRank

Em 1998, Page et al. [85], ao desenvolverem o mecanismo de busca Google, propuseram o algoritmo PageRank, um modelo matemático para atribuir uma medida de relevância a páginas da web com base em sua estrutura de conexões. Consideraram $PR = PR(i)$ como a pontuação de relevância da página $i$, sendo $PR_0$ a pontuação inicial. A hipótese assumida foi que a importância de uma página está relacionada ao número e à qualidade das páginas que apontam para ela, num processo iterativo que se estabiliza ao longo do tempo.

#### Formulação Matemática do Modelo

Seja $PR(i)$ a pontuação da página $i$ em uma iteração do modelo. Podemos expressá-la como:

$$PR(i) = (1 - d) + d \sum_{j \in M(i)} \frac{PR(j)}{L(j)} \tag{2.1}$$

onde:

- $d$ é o fator de amortecimento (geralmente 0,85).
- $M(i)$ é o conjunto de páginas que apontam para $i$.
- $L(j)$ é o número de links da página $j$.

A equação representa a ideia de que uma página tem maior relevância se for referenciada por outras páginas relevantes. A pontuação converge para um valor estável ao longo das iterações.

A Figura 2.1, elaborada pelo autor para fins ilustrativos, mostra o processo iterativo de atualização da pontuação no algoritmo *PageRank*. O gráfico

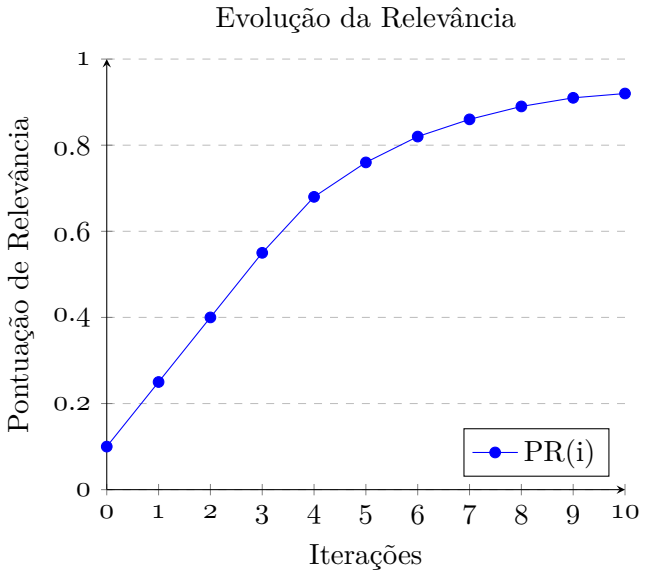



**Figura 2.1**: *Pontuação de Relevância pelo PageRank.*

foi construído com base em uma simulação manual, utilizando valores fictícios para representar o comportamento típico de convergência. A pontuação de relevância aumenta progressivamente até estabilizar, ilustrando como o algoritmo converge para valores fixos ao longo das iterações, refletindo a estabilização da importância atribuída a cada página.

### 2.6.2 Propriedades dos Modelos

Um modelo matemático é, por natureza, uma representação simplificada da realidade, construída com um propósito específico. Para além da formulação conceitual, a literatura propõe classificações consolidadas que permitem compreender suas propriedades e orientar sua utilização em contextos práticos.

Os modelos *caixa-branca* são fundamentados em leis físicas e princípios teóricos, com estrutura totalmente conhecida, enquanto os modelos *caixa-preta* são construídos a partir de dados, sem explicitar os mecanismos internos do sistema, recorrendo a regressões, redes neurais ou algoritmos de aprendizado de máquina. Já os modelos *caixa-cinza* combinam conhecimento parcial do sistema com calibração a partir de dados [62, 79]. Outra distinção importante está entre modelos *estáticos*, que descrevem relações em um único ponto no tempo, sem memória ou evolução, e modelos *dinâmicos*, que capturam a variação temporal do sistema por meio de estados e equações diferenciais ou de diferença [59].

Além disso, pode-se diferenciar modelos *contínuos*, nos quais as variáveis de estado variam ao longo do tempo de forma ininterrupta, e modelos *discretos*, em que as mudanças ocorrem em instantes ou intervalos definidos [59, 103]. Por fim, modelos *determinísticos* não incluem variabilidade aleatória explícita; seus resultados são função direta das entradas, ao passo que modelos *estocásticos* incorporam incertezas nos parâmetros ou variáveis, exigindo análise probabilística dos resultados [59].

### 2.6.3 Limitações e Incertezas dos Modelos

De acordo com Epstein [36], maximizar simultaneamente a generalidade, o realismo e a precisão em modelos é uma tarefa desafiadora. Este desafio é particularmente evidente nas ciências sociais, onde a busca pela precisão muitas vezes é contrabalançada pelo compromisso com o realismo e a generalidade. Modelos de simulação, por exemplo, frequentemente sacrificam a precisão em favor do realismo e da generalidade, tornando-se mais compreensíveis à medida que os componentes internos são investigados. Ao definir variáveis e elaborar conclusões, é crucial apresentar razões claras e justificáveis para a comunidade científica.

A utilidade dos modelos reside em sua capacidade de fazer previsões dentro de um processo dedutivo: se as suposições subjacentes são verdadeiras, então as conclusões também o são, assumindo a validade do modelo. No entanto, é importante reconhecer que os modelos são aproximações da realidade, e previsões perfeitas não são esperadas. Diante disso, a avaliação de um modelo deve considerar sua confiabilidade. Uma conclusão derivada de um único modelo pode ser menos confiável, especialmente se outros modelos produzirem previsões divergentes [44].

A robustez de um resultado é validada quando pode ser replicada por diferentes modelos, enquanto previsões baseadas em suposições específicas são mais frágeis. Embora os modelos possam gerar conclusões, oferecer explicações é mais complexo, pois a percepção de uma explicação pode variar, dependendo da precisão das previsões. É possível que dois modelos diferentes apresentem previsões semelhantes, mas com explicações distintas. Uma abordagem útil é tratar o modelo como uma "caixa preta", onde a validade é confirmada pela correspondência dos resultados observados. Isso enfatiza a importância da validação empírica [7]. Compreender as limitações dos modelos é essencial para a prática científica, pois ajuda na interpretação e avaliação dos resultados, permitindo decisões mais informadas sobre a aplicabilidade e confiabilidade dos modelos [7].

### 2.6.4 Construção de Modelo

A modelagem matemática, conforme discutido por Rosa e Orey [91], constitui um ambiente de aprendizagem estruturado, no qual os estudantes podem converter conhecimentos tácitos em conhecimentos explícitos, por meio da formulação, análise e interpretação de modelos. Não é um processo arbitrário, a modelagem segue diretrizes, que orientam sua condução de forma sistemática e reflexiva. Com base nessas diretrizes, apresenta-se a seguir uma síntese do processo de modelagem matemática:

**Formular o problema:** Definir claramente o que se deseja investigar, pois o objetivo do modelo orienta sua estrutura.

**Esboçar o modelo:** Identificar as variáveis envolvidas, classificando-as como insignificantes, exógenas ou endógenas, e estabelecer suas inter-relações.

**Avaliar utilidade:** Verificar se os dados necessários estão disponíveis e se o modelo permite realizar previsões úteis. Caso contrário, revisar as etapas anteriores.

**Testar o modelo:** Aplicar o modelo em situações reais ou simuladas, validando suas previsões. Se os resultados forem inconsistentes, revisar hipóteses ou reformular a estrutura.

### 2.6.5 Aplicação e Teste

Nesta etapa, o modelo encontra-se tecnicamente pronto para ser utilizado. No entanto, sua aplicação deve ser criteriosa, evitando-se o uso indiscriminado em problemas que diferem substancialmente daqueles para os quais foi originalmente desenvolvido e testado. Cada nova aplicação representa, na prática, um novo teste da validade do modelo [7, 44].

Frequentemente, não é possível seguir todas as etapas metodológicas de forma imediata, especialmente quando ainda não estão claros os fatores que podem ser negligenciados ou a definição precisa das variáveis exógenas. Uma abordagem recomendada é iniciar com um modelo mais simples, utilizando estimativas iniciais para refinar gradualmente os aspectos mais relevantes e identificar os elementos que exigem maior detalhamento [7].

Alguns modelos também podem demandar dados adicionais para aprimorar sua precisão. Quando um modelo gera previsões idênticas, independentemente

da entrada de dados, isso pode indicar que está fundamentado apenas em suas suposições internas, e não em informações empíricas. Em certos contextos, a distinção entre dados e suposições pode se tornar artificial. Há casos extremos em que a especialização do modelo inviabiliza o teste de suas previsões fora do cenário para o qual foi construído [44].

Em áreas teóricas das ciências físicas, por exemplo, onde experimentos práticos são inviáveis, a validação ocorre por meio de simulações ou experimentações indiretas. Nesses casos, diferentes modelos para o mesmo fenômeno podem produzir previsões divergentes, o que dificulta a ratificação de sua validade [7, 44]. Ainda que a validação experimental seja possível, ela pode ser onerosa ou demandar habilidades especializadas. Assim, sempre que viável, os resultados dos testes devem ser fornecidos para fortalecer a confiança no modelo e em suas previsões [7].

## 2.7 Sistemas de Recomendação

Sistemas de recomendação constituem uma classe de sistemas de filtragem de informação que visam sugerir itens relevantes aos usuários com base em seus interesses ou perfis. Segundo Aggarwal [5], esses sistemas são amplamente utilizados em domínios como comércio eletrônico, mídias digitais, redes sociais e educação personalizada, empregando técnicas como filtragem colaborativa, baseada em conteúdo e modelos híbridos.

Os sistemas de recomendação têm evoluído consideravelmente desde seus primeiros desenvolvimentos na década de 1990, acompanhando as mudanças tecnológicas e expandindo sua aplicabilidade em diversos contextos. Originalmente, esses sistemas foram criados para ajudar os usuários a navegarem em ambientes repletos de informações, como lojas online, mas, ao longo das últimas décadas, eles passaram a desempenhar papéis cada vez mais estratégicos em plataformas digitais, personalizando conteúdos para aumentar o engajamento e a satisfação dos usuários.

### 2.7.1 Filtragem Colaborativa e Filtragem Baseada em Conteúdo

Os primeiros sistemas de recomendação utilizavam a *Filtragem Colaborativa* (Collaborative Filtering[†13]) desenvolvida na década de 1990 para prever as preferências dos usuários com base em comportamentos e avaliações de outros

[†13]Filtragem Colaborativa é um método de recomendação que se baseia no comportamento de um grupo de usuários para prever preferências.

usuários com gostos semelhantes [57, 96]. Esta abordagem constrói um banco de dados de preferências de usuários utilizando avaliações explícitas, como feedbacks, para recomendar produtos ou conteúdos. A filtragem colaborativa pode ser dividida em técnicas baseadas em memória (*User-Based* e *Item-Based*) e em modelo (*Clustering*, *Singular Value Decomposition - SVD*[†14]). As técnicas baseadas em memória utilizam métodos como a correlação de Pearson, similaridade do cosseno vetorial e *K-Nearest Neighbors (KNN)* para identificar grupos de usuários com preferências semelhantes [57], enquanto as técnicas baseadas em modelo utilizam algoritmos de aprendizado de máquina para prever preferências futuras com base em padrões históricos [3].

Por outro lado, a *Filtragem Baseada em Conteúdo* (*Content-Based Filtering*[†15]) utiliza técnicas como mineração de texto, TF-IDF[†16] e análise semântica para entender melhor as preferências dos usuários [57]. No entanto, essa abordagem enfrenta limitações, como a incapacidade de recomendar itens novos que não compartilham características com os itens previamente avaliados [3]. Devido a essas limitações, o uso exclusivo da filtragem baseada em conteúdo começou a declinar nas últimas décadas e tem sido cada vez mais combinada com outros métodos para melhorar a precisão e eficácia. [57].

### 2.7.2 Modelos Híbridos

Para superar as limitações das abordagens de filtragem colaborativa e baseada em conteúdo, os *Modelos Híbridos* (Hybrid Models[†17]) começaram a ser desenvolvidos e implementados. Estes modelos combinam múltiplas técnicas de recomendação utilizando métodos como hibridização ponderada, hibridização em cascata e combinações de características [21]. Esta combinação permite que os sistemas de recomendação aproveitem o melhor de cada técnica, melhorando a precisão das recomendações e reduzindo problemas como a escassez de dados [57]. Conforme descrito por Aggarwal [5], os modelos híbridos combinam diferentes estratégias para superar limitações individuais, como a escassez de dados e o problema de arranque a frio.

[†14]SVD é uma técnica de decomposição de matriz amplamente utilizada para reduzir a dimensionalidade dos dados e melhorar a precisão das recomendações.

[†15]Filtragem Baseada em Conteúdo recomenda itens com base nas características dos próprios itens que o usuário já demonstrou interesse.

[†16]TF-IDF (Term Frequency-Inverse Document Frequency) é uma técnica de mineração de texto utilizada para avaliar a importância de uma palavra em um documento.

[†17]Modelos híbridos combinam múltiplas técnicas de recomendação, como filtragem colaborativa e filtragem baseada em conteúdo, para melhorar a precisão e eficácia das recomendações.

Adomavicius e Tuzhilin [3] também destacam a necessidade de melhorar os sistemas de recomendação por meio da incorporação de informações contextuais, como localização e tempo, o que pode aumentar a relevância das recomendações e ajudar a mitigar problemas de *cold start*[†18] e escassez de dados.

### 2.7.3 Integração de Dados e Aplicações da IA

Na última década, o foco em dados implícitos, como padrões de clique, dados de navegação e outros comportamentos de usuário, tornou-se mais prevalente. Esses dados permitem que os sistemas de recomendação se adaptem rapidamente às mudanças nas preferências dos usuários [57]. Em plataformas de streaming, como Netflix e Spotify, essas técnicas têm se mostrado valiosas para oferecer recomendações personalizadas em tempo real, utilizando dados comportamentais coletados continuamente.

O uso crescente de técnicas de *Deep Learning*[†19] e redes neurais tem permitido a criação de modelos mais sofisticados e precisos [1]. Essas técnicas, ao integrar dados demográficos, geolocalização e dados sociais, enriquecem as recomendações personalizadas e aumentam a satisfação do usuário, embora também levantem preocupações éticas, especialmente em relação à privacidade dos dados. Leis como o Marco Civil da Internet e a Lei Geral de Proteção de Dados (LGPD) no Brasil estabelecem a necessidade de consentimento prévio e expresso dos usuários para a coleta de dados e garantem direitos de acesso, correção e exclusão de dados pessoais [1].

Além do seu uso em recomendações personalizadas, a Inteligência Artificial (IA) tem desempenhado um papel significativo na automatização de processos em diversos setores. Segundo Baldin et al. [9], a IA tem sido aplicada para otimizar processos industriais e empresariais, reduzindo custos, aumentando a eficiência e melhorando a qualidade dos produtos e serviços ao automatizar tarefas repetitivas que antes dependiam de intervenção humana. Essas técnicas possibilitam a criação de fluxos de trabalho mais eficientes e minimizam erros humanos, aumentando a produtividade e economizando recursos.

Essa expansão sugere que o uso da IA em sistemas de recomendação não está limitado a setores tradicionais, como o entretenimento e o comércio eletrônico, mas também pode ser aplicado em áreas mais complexas, como a

[†18] *Cold Start* é um problema comum em sistemas de recomendação que ocorre quando há pouca ou nenhuma informação disponível sobre o novo usuário ou item.

[†19] *Deep Learning* é uma subárea do aprendizado de máquina que utiliza redes neurais profundas para modelar dados complexos.

automação de processos e o gerenciamento de grandes volumes de dados em ambientes corporativos e industriais [9].

Conforme discutido por Barreto [12], os primeiros avanços em IA, especialmente através de sistemas especialistas, como o MYCIN para diagnósticos médicos e o PROSPECTOR para exploração geológica, foram fundamentais para demonstrar a capacidade da IA em resolver problemas específicos. Também destaca que essas primeiras tentativas de aplicação da IA, mesmo que focadas em áreas muito específicas, influenciaram o desenvolvimento de sistemas modernos, que agora combinam várias técnicas de IA para otimizar a tomada de decisão em contextos complexos. A IA tem sido cada vez mais aplicada em setores diversos, como saúde, educação, turismo e serviços acadêmicos, possibilitando a personalização de experiências e o suporte à decisão em tempo real [1]. Barreto [12] discute o desenvolvimento inicial de técnicas como redes neurais artificiais e lógica fuzzy, enquanto Abreu e da Silva dos Santos [1] destacam a evolução contínua dessas técnicas e o seu potencial para fornecer recomendações mais precisas e relevantes.

A evolução dos sistemas de recomendação reflete um movimento contínuo em direção a uma maior integração de técnicas avançadas de IA, incluindo aprendizado profundo e mineração de dados, visando proporcionar recomendações mais precisas e personalizadas. O uso de IA em sistemas de recomendação pode continuar se aprimorando, incorporando novas fontes de dados e abordagens para atender a uma variedade de necessidades dos usuários. As inovações tecnológicas associadas à IA sugerem um potencial significativo para transformar a forma como interagimos com informações e tomamos decisões em uma ampla gama de contextos.

## 2.8 Chatbot

Os *chatbots*, definidos como programas de computador projetados para simular conversas humanas utilizando inteligência artificial e processamento de linguagem natural, têm se tornado cada vez mais sofisticados e amplamente utilizados em vários setores. Segundo Adamopoulou e Moussiades [2], os *chatbots* representam uma evolução significativa na forma como as interações humanas e computacionais são realizadas, permitindo não apenas a automatização de respostas a perguntas frequentes, mas também a personalização de interações em tempo real por meio de tecnologias avançadas de aprendizado de máquina.

Em um contexto educacional, Okonkwo e Ade-Ibijola [82] descrevem os *chatbots* como ferramentas capazes de proporcionar suporte contínuo ao aprendizado, adaptando-se ao ritmo e às necessidades específicas de cada estudante.

Isso é alcançado por meio de recomendações e feedback personalizados, gerados com base nas interações do usuário e nos dados coletados sobre seu desempenho. Em uma revisão sistemática, os mesmos autores identificaram que os *chatbots* facilitam a personalização do aprendizado, melhoram o engajamento dos alunos e proporcionam um ambiente de aprendizado mais acessível e inclusivo. Complementarmente, Calle et al. [23] propuseram a integração de *chatbots* com sistemas de recomendação para apoiar o aprendizado autorregulado em ambientes online, oferecendo recomendações customizadas com base no comportamento e desempenho dos estudantes.

Danckwerts et al. [29] destacam o uso de *chatbots* em serviços de mídia, apontando como essas tecnologias podem oferecer recomendações mais relevantes e contextualmente ajustadas, aumentando o engajamento e a satisfação dos usuários. Com a integração de sistemas de recomendação, os *chatbots* têm evoluído para plataformas capazes de oferecer sugestões dinâmicas e personalizadas em diferentes contextos. Calle et al. [23] propõem, por exemplo, o uso de *chatbots* recomendadores no apoio ao aprendizado autorregulado, auxiliando os usuários a gerenciar seus estudos por meio de recomendações baseadas em dados reais de interação. Além da educação e do entretenimento, os *chatbots* também têm sido aplicados na área da saúde. Segundo Aggarwal et al. [4], *chatbots* baseados em IA têm se mostrado eficazes na promoção de mudanças comportamentais e no acompanhamento contínuo de pacientes, fornecendo intervenções personalizadas e feedback em tempo real.

A agilidade proporcionada pelos *chatbots* é um fator chave para o sucesso no ambiente empresarial. Wang et al. [109] investigaram como o uso inovador de *chatbots* pode criar agilidade nos negócios, permitindo respostas rápidas a consultas de clientes e facilitando processos internos. O estudo conclui que *chatbots* melhoram a experiência do cliente e reduzem custos operacionais, aumentando a eficiência geral da empresa. A integração com sistemas de recomendação pode potencializar ainda mais esses benefícios, fornecendo sugestões de produtos baseadas no histórico de navegação e nas preferências do cliente, conforme discutido por Ma et al. [67], que exploram a interação multimodal em sistemas de recomendação conversacional para aumentar a personalização e a eficácia das recomendações.

Apesar das vantagens, o uso de *chatbots* apresenta desafios significativos. Hwang e Chang [50] destacam que, na educação, o uso de *chatbots* ainda está em fase inicial, com poucas investigações empíricas sobre o design de aprendizado eficaz. Existe a necessidade de explorar mais a fundo as estratégias pedagógicas que podem ser efetivamente implementadas com o auxílio de *chatbots*. Outro desafio importante é a segurança e privacidade dos

dados. Com a crescente adoção de *chatbots* em áreas sensíveis como saúde e finanças, a proteção de informações pessoais e o cumprimento de regulamentos de proteção de dados, como a General Data Protection Regulation (GDPR) e a LGPD, tornaram-se uma preocupação crítica [2].

A integração de *chatbots* com sistemas de recomendação parece representar uma evolução natural, levando em consideração a necessidade crescente de interações mais personalizadas e relevantes. Danckwerts et al. [29] analisam *chatbots* de recomendação em serviços de streaming de mídia, demonstrando como a tecnologia pode melhorar a avaliação do usuário sobre os serviços ao oferecer recomendações personalizadas e contextualizadas. Além disso, Ma et al. [67] sugerem que a adoção de interfaces multimodais que incluem texto, voz e gestos pode aumentar a capacidade dos *chatbots* de entenderem a intenção do usuário, proporcionando interações mais fluidas e naturais. Tais abordagens podem ser particularmente úteis em setores como e-commerce, onde a personalização e a rapidez de resposta são fundamentais para a satisfação do cliente.

## 2.9 Resumo do Capítulo

Este capítulo apresentou a fundamentação teórica necessária ao desenvolvimento da pesquisa, abordando conceitos relacionados à modelagem matemática, sistemas de recomendação, *chatbots*, pós-graduação e plataformas acadêmicas. Inicialmente, discutiu-se a importância dos modelos matemáticos para a representação de fenômenos reais, abordando equações, propriedades, limitações e a validação empírica. Em seguida, foram analisados os sistemas de recomendação, suas abordagens clássicas e híbridas, e a aplicação de inteligência artificial para personalização de sugestões. A discussão sobre *chatbots* destacou seu uso como ferramentas interativas de apoio, integradas a sistemas de recomendação em contextos educacionais. A seguir, foi apresentada a evolução da pós-graduação no Brasil, evidenciando sua contribuição para a formação de pesquisadores e o fortalecimento da pesquisa científica. A análise da Plataforma Lattes e de bases internacionais como Google Scholar, ResearchGate, ORCID, Publons, Academia.edu, Scopus e Mendeley ressaltou a importância da coleta e avaliação de dados acadêmicos estruturados. Finalizando, discutiu-se o conceito de reputação acadêmica, abordando métricas tradicionais e modelos alternativos que ampliam a avaliação da trajetória científica. Esta revisão teórica fundamenta a proposta deste trabalho, integrando dados acadêmicos, modelagem matemática e inteligência artificial para apoiar decisões na escolha de orientadores.

# Capítulo 3

# Trabalhos Relacionados

*Compreender é perceber padrões.*

*Sir Isaiah Berlin, Teórico social e político britânico (1909–1997)*

Este capítulo apresenta a revisão da literatura que fundamenta o desenvolvimento deste trabalho, destacando estudos e ferramentas que possuem alguma relação com a proposta apresentada. Na seção 3.1, analisa-se as qualidades desejáveis de orientadores e os fatores que impactam a satisfação e o desempenho nas relações de mentoria acadêmica, com ênfase na compatibilidade e no suporte individualizado. Na seção 3.2, aborda-se as metodologias para a automação da coleta e organização de dados, facilitando a geração de indicadores acadêmicos. Na seção 3.3, explora-se modelos baseados em aprendizado de máquina e processamento de linguagem natural para identificar especialistas, promovendo colaborações científicas. Na seção 3.4, discute-se técnicas de mineração de dados e aprendizado supervisionado para segmentação e análise de grandes volumes de dados curriculares. Na seção 3.5, investiga-se o uso dessas tecnologias em educação, saúde e pesquisa, destacando a integração de modelos de linguagem e sistemas de recomendação. Na seção 3.6, revisa-se abordagens modernas para personalização de recomendações, com foco em técnicas como filtragem colaborativa e modelos híbridos. Na seção 3.7, apresenta-se estudos sobre as razões que levam estudantes a ingressar nesses programas e seus impactos na dinâmica acadêmica. Na seção 3.8, aborda-se as características relevantes para orientadores de doutorado. Na seção 3.9, discute-se os métodos usados para mapear colaborações científicas com base em grafos e predições de coautorias. Na seção 3.10, discute-se a seleção de orientadores em programas de pós-graduação e como a modelagem matemática pode torná-la mais objetiva, destacando principais abordagens e desafios.

## 3.1 Orientação Acadêmica

Taylor et al. [104] explora as qualidades desejáveis dos orientadores de doutorado nos EUA. A metodologia utilizada foi mista, combinando entrevistas com 13 orientadores acadêmicos e 18 candidatos a doutorado e graduados recentes, além de questionários aplicados a 38 orientadores e 151 candidatos e graduados. Os participantes eram de diversas disciplinas e estados. Os resultados destacam a importância de uma boa comunicação, feedback útil e oportuno, suporte emocional e uma estrutura clara na orientação. A conclusão enfatiza que preparar orientadores para atender às necessidades dos doutorandos modernos é essencial para facilitar a conclusão bem-sucedida dos programas de doutorado. As qualidades essenciais para orientadores incluem habilidades de comunicação, feedback construtivo, apoio emocional e uma transição na relação de hierárquica para colegial conforme o candidato avança. O estudo contribui para a compreensão de como promover relacionamentos bem-sucedidos na educação doutoral.

Barnes e Austin [10] buscaram compreender as percepções de professores universitários que atuam como orientadores de doutorado em relação às suas responsabilidades, funções e comportamentos. Foi conduzido um estudo de caso com 25 orientadores de doutorado de quatro áreas distintas: ciências naturais, humanidades, ciências sociais e educação. Os dados foram coletados por meio de entrevistas gravadas, cada uma com aproximadamente uma hora de duração. A análise dos dados foi realizada utilizando duas técnicas de codificação iterativa: aberta e axial. Os resultados indicam que os professores se veem com a responsabilidade de auxiliar seus orientandos a alcançar o sucesso, ao mesmo tempo que os ajudam a se desenvolver como pesquisadores e profissionais. Além disso, os resultados evidenciam que os professores universitários que atuam como orientadores de doutorado desempenham um papel essencial no apoio aos alunos, não apenas em relação aos resultados acadêmicos, mas também auxiliando-os a lidar com eventuais fracassos e comprometendo-se a adaptar-se às necessidades individuais de cada estudante.

ChunMei Zhao e McCormick [26] explora os fatores que influenciam a satisfação dos estudantes com a orientação em programas de doutorado nos EUA, abrangendo 27 universidades e 11 áreas. Os resultados indicam que a escolha do orientador e o comportamento do orientador são determinantes significativos da satisfação do estudante, superando as características individuais. A relação entre estudante e orientador é crucial para a experiência educacional do doutorado, com relacionamentos positivos promovendo um ambiente benéfico e a

conclusão oportuna do grau. O estudo destaca a variação nas experiências dos estudantes devido à falta de regulação central nos EUA e a importância da escolha de um orientador solidário e interativo. O processo de pareamento varia por área, influenciando a satisfação do estudante. O estudo busca entender como os critérios de seleção de orientadores e seus comportamentos variam por área e como esses fatores afetam a satisfação no relacionamento de orientação. As duas principais perguntas de pesquisa são: 1. Os padrões de escolha de orientador e comportamento de orientador diferem por área disciplinar após controlar as características dos estudantes? 2. Como a escolha do orientador e o comportamento do orientador se relacionam com a satisfação no relacionamento de orientação após controlar as características dos estudantes e a área disciplinar? Em suma, o estudo explora a influência da escolha do orientador e do comportamento na satisfação do estudante, contribuindo para a compreensão de como promover relacionamentos bem-sucedidos na educação doutoral.

Chukwu e Walker [25] analisam a relação entre estudantes de pós-graduação e seus orientadores, destacando a importância das dinâmicas sociais que moldam essas interações. O artigo apresenta um framework conceitual que considera interpretações, papéis e responsabilidades recíprocas, fatores relacionais e efeitos dessas relações, oferecendo recomendações para melhores práticas. A pesquisa destaca como a supervisão de professores impacta o sucesso acadêmico e profissional dos estudantes, abordando questões como coautoria, mentoria, processos de socialização e considerações éticas. A evolução histórica dessa relação é discutida, desde modelos de mestre-serviçal até arranjos mais colaborativos e centrados no estudante, reconhecendo desafios únicos enfrentados por estudantes de doutorado não tradicionais, como aqueles que estudam em tempo parcial ou remotamente. O estudo utiliza métodos qualitativos para examinar as experiências vividas pelos estudantes e orientadores, incluindo entrevistas e análise documental. Os resultados mostram que a confiança e o respeito mútuo são essenciais para uma relação eficaz, e que a falta de interação e apoio acadêmico pode levar ao abandono dos estudos por parte dos estudantes. Estatísticas revelam que uma relação positiva com os orientadores aumenta significativamente as chances de conclusão do doutorado, enquanto relações negativas podem resultar em consequências adversas, como desistência ou atraso na conclusão dos estudos. O estudo também destaca a necessidade de modelos de supervisão culturalmente diversos e adaptados às necessidades individuais dos estudantes, propondo a institucionalização de práticas de mentoria que promovam um ambiente acadêmico mais inclusivo e colaborativo.

Joy et al. [54] investigou o processo de pareamento entre orientadores e orientandos em departamentos STEM (Ciências, Tecnologia, Engenharia e Matemática) de uma universidade nos Estados Unidos. Ao contrário de outros

programas de doutorado, a maioria dos programas nos EUA exige que os alunos escolham seus orientadores, que por sua vez devem aceitar formalmente os alunos como orientandos. Com base em grupos focais e entrevistas com estudantes de doutorado e professores, o estudo identificou critérios aplicados por ambos ao fazerem suas escolhas. Os alunos avaliavam os professores com base em fatores como financiamento disponível, área de pesquisa, personalidade, capacidade de formar alunos rapidamente e perspectivas de carreira. Por outro lado, os professores avaliavam os alunos com base em suas qualificações/credenciais e na percepção da capacidade de contribuir para a pesquisa. Foi constatado que essa avaliação mútua não era objetiva, mas influenciada por percepções associadas ao gênero do corpo docente, estágio de carreira e nacionalidade do aluno. Além disso, se os alunos e professores foram realmente emparelhados com pessoas de sua escolha dependia de fatores departamentais, incluindo práticas predominantes de emparelhamento, restrições no número de alunos por corpo docente e estrutura de recompensas. Este estudo fornece uma visão detalhada do processo de pareamento entre orientadores e orientandos em departamentos STEM, destacando a importância da qualidade do relacionamento e da compatibilidade acadêmica. As descobertas e recomendações deste estudo oferecem diretrizes valiosas para melhorar os processos de seleção de orientadores e orientandos, especialmente para estudantes internacionais.

Lugoboni [65] discutem a importância da escolha de um orientador para alunos que desejam realizar pesquisas acadêmicas, seja em iniciação científica, monografia, dissertação de mestrado ou tese de doutorado. O artigo destaca que pouca ou nenhuma instrução é fornecida aos alunos sobre como escolher um orientador adequadamente, levando-os a basear sua escolha em fatores como simpatia ou aderência à área de interesse. No entanto, esses critérios podem não refletir a capacidade do professor em ser um bom orientador. Lugoboni sugere que os alunos consultem o currículo dos professores na plataforma Lattes, observando detalhes como a data da última publicação, a quantidade de trabalhos orientados, os temas das publicações mais recentes e os projetos de pesquisa em que o professor está envolvido. Ele enfatiza que essas informações ajudam a garantir uma escolha mais informada, embora não garantam um excelente orientador. A análise cuidadosa do histórico profissional e acadêmico dos potenciais orientadores oferece aos alunos uma base mais sólida para uma primeira aproximação.

Cervi [24] destaca que a reputação acadêmica é um elemento central na avaliação de pesquisadores e instituições de ensino superior. Na porposta intitulada Rep-Index: Uma Abordagem Abrangente e Adaptável Para Identificar Reputação Acadêmica, o autor apresenta uma abordagem para a modelagem e medição da reputação acadêmica. O estudo propõe um modelo de perfil de pes-

quisadores (Rep-Model) e uma métrica quantitativa (Rep-Index) para avaliar a reputação com base em múltiplos indicadores, indo além da tradicional contagem de citações. Os experimentos realizados na tese demonstram que o Rep-Index possui correlação significativa com o h-index e o g-index, mas oferece uma visão mais ampla sobre o impacto dos pesquisadores. A metodologia foi testada em um conjunto de 830 pesquisadores bolsistas de produtividade do CNPq, abrangendo três áreas distintas: Ciência da Computação, Economia e Odontologia. Os principais achados incluem: A métrica Rep-Index apresenta uma correlação estatisticamente significativa com o ranking de pesquisadores do CNPq. Elementos como número de artigos publicados, participação em conferências, h-index e participação em projetos de pesquisa foram identificados como fatores-chave na determinação da reputação acadêmica. A flexibilidade do modelo permite que diferentes áreas do conhecimento ajustem os pesos atribuídos a cada indicador, possibilitando uma avaliação mais equitativa.

Nieto [81] estudou e explorou as características fundamentais dos e-mentores responsáveis pela orientação de dissertações de doutorado online, sob a perspectiva dos docentes. Utilizando uma metodologia de amostragem intencional, o estudo investigou características de mentoria que convergem em valores, habilidades profissionais e relacionamentos. As percepções foram obtidas por meio de uma pesquisa voluntária e anônima, distribuída a 10 experientes Presidentes de Dissertação de Doutorado online, além de entrevistas telefônicas com docentes e diários reflexivos. Os resultados revelaram que características como tenacidade, inovação e adaptabilidade são essenciais. Além disso, o estudo destacou barreiras existentes nos processos de revisão de qualidade para orientadores e estudantes. Esta pesquisa contribui para preencher lacunas na literatura sobre mentoria, oferecendo informações valiosas para e-mentores e alunos de doutorado online, com o objetivo de aprimorar o processo de orientação e a experiência dos alunos.

## 3.2 Extração de Indicadores Acadêmicos

Silva e Bianchi [99] definem a cientometria como o estudo da mensuração e quantificação do progresso científico, com base em indicadores bibliométricos, ressaltando seu potencial para subsidiar políticas públicas, distribuir recursos de forma eficiente e avaliar o impacto das pesquisas. Os indicadores são classificados em quantitativos, como o número de publicações, e de impacto, como o fator de impacto e o número de citações recebidas. O artigo aponta que o Brasil detém 1,2% da produção científica mundial indexada no *Science Citation Index* (SCI), ocupando a posição de número 17 no ranking global, resultado atribuído ao crescimento da pós-graduação. Os autores mencionam também bases

como o *Science Citation Index*, *Social Citation Index* e o *Arts & Humanities Citation Index*, que avaliam cerca de 8.000 revistas científicas com base em critérios como qualidade editorial e número de citações. Entre os principais indicadores estão o fator de impacto utilizado pelo *Journal Citation Reports (JCR)* , o índice de imediação (relativo à rapidez de citação) e a vida média das revistas (tempo necessário para acumular 50% das citações). Apesar da utilidade desses indicadores, os autores alertam para limitações, como a variação entre áreas: apenas 5% dos artigos em artes e humanidades, 25% em ciências sociais, 30% a 40% em engenharia e tecnologia, e 50% a 60% em medicina são citados nos cinco anos seguintes à publicação. Também destacam que o número de citações pode ser influenciado por fatores como prestígio do autor ou da instituição, idioma e visibilidade da revista. Casos como o artigo de Arthur Jensen, citado amplamente por razões negativas, ilustram que impacto não equivale necessariamente a qualidade. Concluem que os indicadores bibliométricos são úteis para análises globais da atividade científica, mas devem ser usados com cautela, principalmente em avaliações individuais, sendo mais eficazes quando aplicados a instituições, áreas ou países.

Prass et al. [87] aborda a relevância da extração de indicadores acadêmicos da Plataforma Lattes, um procedimento amplamente adotado por instituições educacionais para planejamento, gestão e promoção da pesquisa. Considerando que a extração manual de dados é um processo complexo e suscetível a erros, o artigo propõe um projeto de automação desenvolvido em Python com o uso do framework Django. Este projeto tem como objetivo extrair e apresentar dados quantitativos dos currículos, tais como produções bibliográficas e formação docente, relacionados aos profissionais de uma instituição de ensino específica. A automação visa tornar o processo mais eficiente e preciso. O artigo ressalta a Plataforma Lattes como uma base de dados fundamental para currículos acadêmicos, grupos de pesquisa e instituições de ensino, destacando as melhorias em desempenho e usabilidade ao longo dos anos. A extração manual de dados envolve várias etapas e inserção de informações, enquanto a abordagem automatizada proposta utiliza acesso direto à base de dados da Plataforma Lattes ou ao arquivo XML do currículo, processando os dados e armazenando-os em um banco de dados MySQL para visualização detalhada posterior. O principal objetivo do trabalho é desenvolver uma ferramenta para extração e geração de dados quantitativos da Plataforma Lattes, detalhando etapas como a elaboração de uma proposta de software, definição de escopo e metodologia de desenvolvimento, além da criação de uma interface para acessar e visualizar os dados. A revisão bibliográfica inclui o histórico da Plataforma Lattes, destacando seu desenvolvimento desde os anos 80 até a implementação do Currículo Lattes em 1999. Além disso, discute métodos de extração de dados, tanto

manualmente quanto via serviço web automatizado, enfatizando a necessidade de uma estruturação eficiente da base acadêmica para gerar indicadores institucionais. Em suma, o projeto apresentado tem como objetivo melhorar a eficiência e a precisão na extração de dados acadêmicos, facilitando o acesso e a utilização dessas informações pelas instituições de ensino e pesquisa.

Galego [42] desenvolvem uma metodologia que utiliza ontologias para a extração e consulta de informações do Currículo Lattes. A Plataforma Lattes é uma base de dados pública que reúne currículos de pesquisadores brasileiros, sendo amplamente utilizada por instituições de fomento, universidades e centros de pesquisa. Apesar de sua importância, a plataforma apresenta limitações na exibição de dados sumarizados de grupos de pessoas, como departamentos de pesquisa ou orientandos de um professor específico. A dissertação propõe a integração de diversas funcionalidades de ferramentas existentes em uma solução única chamada SOS Lattes. A metodologia inclui o uso de ontologias para identificar inconsistências nos dados, consultas para construção de relatórios consolidados e regras de inferência para correlacionar múltiplas bases de dados. Dados de 657 currículos de pesquisadores foram utilizados, abrangendo o período de 1971 a 2015. Os resultados mostram que a integração de dados em forma de ontologia facilita a detecção de inconsistências e a geração de relatórios detalhados sobre produções bibliográficas e orientações. A ferramenta também sugere a inclusão de novos membros e permite consultas utilizando linguagem natural. Conclui-se que a aplicação da Web Semântica e ontologias pode melhorar significativamente a gestão de informações acadêmicas na Plataforma Lattes, contribuindo para a expansão e disseminação da área de Web Semântica.

Oliveira et al. [84] discutem a importância e a necessidade de otimizar a recuperação de informações no Currículo Lattes, uma base de dados gerida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) que registra a produção acadêmico-científica e tecnológica no Brasil. O estudo é uma pesquisa documental e aplicada, com abordagem qualitativa, e tem como objetivo geral apresentar uma proposta para otimizar a recuperação de informação sobre a produção acadêmica registrada no Currículo Lattes. O estudo foi realizado em três etapas metodológicas: (i) diagnóstico inicial com caracterização e descrição do Currículo Lattes, (ii) identificação de elementos de recuperação a serem otimizados e (iii) criação de um protótipo espelhado na plataforma, apresentando sugestões de melhorias para a recuperação de informação. Na primeira etapa, os autores realizaram um diagnóstico inicial, identificando a atual estrutura e funcionalidades da Plataforma Lattes. Eles observaram que, embora a plataforma possua módulos específicos para o registro de diversas atividades acadêmicas, a recuperação de informações específicas ainda é um desafio de-

vido ao grande volume de dados. Na segunda etapa, foram identificados os elementos específicos que poderiam ser otimizados. Através de navegação e testes funcionais na plataforma, os autores identificaram filtros e opções que poderiam melhorar a recuperação de informações, incluindo a classificação e ordenação de dados. Na terceira etapa, um protótipo foi criado utilizando a ferramenta Figma. Este protótipo incluiu melhorias sugeridas, como a inclusão de filtros para ordenação e refinamento das buscas, além de funcionalidades para facilitar a visualização e recuperação de informações pelos usuários. Os resultados do estudo mostraram que a aplicação de técnicas de filtragem e a implementação de melhorias na apresentação das informações podem tornar a recuperação de dados no Currículo Lattes mais rápida e personalizada. Isso pode beneficiar tanto os pesquisadores quanto os gestores acadêmicos, facilitando a análise de produção científica e a tomada de decisões informadas. O estudo conclui que a proposta de otimização baseada nos princípios de organização e recuperação de informação pode potencializar os serviços oferecidos pela Plataforma Lattes, apresentando as informações de forma mais clara e objetiva. Os autores esperam que a proposta sirva como um instrumento útil para a comunidade científica brasileira e inspire futuras versões da plataforma.

Souza et al. [102] apresentam a extensão *Qualis no Lattes*, desenvolvida para automatizar a consulta e exibição do índice Qualis diretamente no currículo Lattes. Esta extensão é uma solução computacional publicada nas lojas dos principais navegadores de internet, permitindo a visualização automática do índice Qualis de cada publicação de artigo registrada no Lattes. A extensão também cria um quadro com a quantidade e pontuação editável por artigo publicado nos últimos cinco anos, facilitando a consulta e cálculo da pontuação do pesquisador. A metodologia do estudo envolveu várias etapas, começando com o acesso aos dados de todas as classificações de periódicos da plataforma Qualis, disponíveis na plataforma Sucupira. Esses dados foram extraídos e utilizados para a programação da extensão. As linguagens de programação empregadas foram JavaScript, HTML e CSS, com o editor de código-fonte Visual Studio Code. A implementação da extensão exigiu um conhecimento detalhado da estrutura DOM (Document Object Model) da página do currículo Lattes, permitindo a manipulação das informações para adicionar os dados de Qualis. Os testes foram realizados manualmente, instalando a extensão localmente nos navegadores e verificando os resultados no currículo Lattes. Após a implementação e testes, a extensão foi publicada nas lojas dos navegadores Google Chrome, Mozilla Firefox e Microsoft Edge. Os resultados mostram que a extensão *Qualis no Lattes* oferece duas funcionalidades principais: a exibição automática do índice Qualis de cada artigo publicado no currículo Lattes e a criação de um quadro contendo a quantidade de artigos por Qualis publica-

dos nos últimos cinco anos, com pontuação editável pelo usuário. A versão 1.2 da extensão foi publicada em fevereiro de 2023, possui 168 KB e mais de mil instalações, com avaliações de cinco estrelas na loja do Google Chrome. A ferramenta facilita significativamente a consulta de Qualis, tornando-se essencial para pesquisadores que lidam com a captação de recursos e a comprovação de publicações. A comparação com outras extensões como *QLattes, ChromeQualis e Qualis* Capes destaca que Qualis no Lattes possui funcionalidades adicionais, como a sumarização das publicações em um quadro editável, que as demais não oferecem. Esta funcionalidade adicional é especialmente útil para pesquisadores que precisam de uma visão consolidada de suas publicações e respectivas classificações Qualis. O estudo conclui que a extensão Qualis no Lattes é uma ferramenta valiosa para a comunidade científica brasileira, automatizando a integração dos índices Qualis com o currículo Lattes e facilitando a avaliação da produção científica dos pesquisadores.

Coimbra e Dias [27] exploram a utilização de dados abertos provenientes da Plataforma Lattes para analisar a publicação de artigos em eventos científicos. O estudo destaca a importância da transparência e acessibilidade dos dados para o avanço científico, argumentando que a disponibilização de dados abertos permite uma análise mais profunda e abrangente da produção científica. Neste estudo, os dados foram extraídos da Plataforma Lattes utilizando a ferramenta LattesDataXplorer. Essa ferramenta foi fundamental para coletar dados curriculares em janeiro de 2021, os quais foram posteriormente selecionados e tratados para análise. A etapa de seleção envolveu a verificação de quais currículos continham trabalhos publicados em eventos científicos, enquanto a etapa de tratamento consistiu em caracterizar esses artigos, permitindo uma análise detalhada das publicações em eventos. Os resultados obtidos permitiram diversas caracterizações, como análise temporal, análise por grande área de atuação e a identificação de artigos que utilizam identificadores persistentes. A análise temporal revelou o número de artigos publicados por ano, destacando que o ápice das publicações ocorreu em 2011, seguido de uma queda significativa. Especificamente, os indivíduos da área de Ciências da Saúde apresentaram a maior taxa de publicação em anais de eventos, enquanto os da área de Ciências Exatas e da Terra tiveram o menor volume de publicações. Além disso, apenas aproximadamente 3% dos artigos informaram o identificador persistente (DOI) em suas publicações, com 30,68% dos artigos na área de Ciências Exatas e da Terra utilizando esses identificadores, o que os torna a área que mais faz uso dessa prática. Ao final, Coimbra e Dias [27] concluem que a utilização de dados abertos e ferramentas como o LattesDataXplorer desempenham um papel importante na promoção da acessibilidade ao conhecimento, permitindo uma caracterização detalhada e original da produção científica brasileira,

especialmente em relação às publicações em eventos científicos.

Digiampietri et al. [33] investigam a utilização de dados dos currículos Lattes para a criação e análise de redes sociais de pesquisadores, enfatizando os desafios e oportunidades associados à mineração e caracterização desses dados. O estudo ressalta a complexidade de lidar com grandes volumes de dados semiestruturados, preenchidos manualmente, e destaca a relevância dos currículos Lattes como uma rica fonte de informação para a ciência brasileira. O objetivo principal do trabalho é apresentar a construção de um banco de dados robusto a partir de mais de um milhão de currículos Lattes, demonstrando como esses dados podem ser organizados e analisados para gerar insights sobre as redes de colaboração científica. Para alcançar esse objetivo, os autores utilizaram uma metodologia que envolveu a coleta automatizada dos currículos utilizando o comando wget, seguido do processamento detalhado das informações extraídas. Este processamento incluiu a remoção de caracteres especiais e a segmentação dos currículos em seções específicas, como publicações, orientações e áreas de atuação. As informações foram então estruturadas em um banco de dados relacional utilizando o SGBD PostgreSQL. Os resultados do estudo mostram que, dos 1.236.548 currículos baixados, foram identificados 11.529.218 registros de publicações e 4.329.993 orientações acadêmicas. Na média, cada currículo analisado continha 9,32 publicações e 3,5 orientações. Além disso, foi observado que, em média, cada pesquisador atua em 2,61 grandes áreas do conhecimento, com 263.775 pesquisadores informando atuação em mais de uma grande área. Esses números revelam padrões significativos de colaboração entre pesquisadores e destacam a importância das redes de coautoria e de orientação. O trabalho também discute as limitações e os cuidados necessários ao utilizar essas informações, como a inconsistência nos dados e a presença de homônimos, sugerindo a necessidade de desenvolver métodos mais eficazes para a resolução desses problemas. Digiampietri et al. [33] concluem que os currículos Lattes são uma ferramenta valiosa para o estudo de redes sociais de pesquisa e a análise de produção acadêmica. O artigo sugere que futuros trabalhos devem focar no desenvolvimento de algoritmos para melhorar a qualidade das análises e na criação de métricas específicas para avaliar a colaboração científica, contribuindo para uma compreensão mais profunda das dinâmicas de produção e disseminação do conhecimento científico no Brasil.

## 3.3 Recomendação de Especialistas

Em um trabalho publicado por de Sousa et al. [32], os autores propõem uma abordagem inovadora para lidar com o desafio de encontrar especialistas

em áreas científicas em meio ao crescente volume de dados disponíveis. Sua técnica combina estratégias de geração de dados fracamente supervisionados com o uso de modelos neurais avançados para classificar candidatos. Em primeiro lugar, introduzem a criação de julgamentos de relevância através de técnicas heurísticas, proporcionando uma maneira eficiente de gerar dados de treinamento em larga escala. Em seguida, propõem o uso de *autoencoder* profundo para selecionar documentos negativos, ou seja, aqueles que não são relevantes para a especialidade em questão. Por fim, apresentam o *Dual Embedding LSTM*, um modelo de classificação baseado em redes recorrentes que demonstrou excelentes resultados ao superar todas as linhas de base comparadas. A justificativa para esta abordagem é multifacetada. Primeiramente, reconhecem a dificuldade de obter grandes volumes de dados rotulados para treinar modelos neurais em contextos científicos, onde a especialização é importante e os dados rotulados são escassos. As técnicas heurísticas permitem contornar essa limitação, gerando julgamentos de relevância de forma eficiente. Além disso, a utilização de *autoencoder* profundo para selecionar documentos negativos ajuda a melhorar a qualidade dos dados de treinamento, filtrando informações irrelevantes e ruidosas. Por fim, o *Dual Embedding LSTM* se destaca como um modelo de classificação capaz de capturar nuances complexas nos dados, fornecendo resultados superiores aos métodos convencionais. Essa abordagem não apenas enfrenta os desafios atuais de encontrar especialistas em meio ao aumento exponencial de dados, mas também oferece uma estrutura robusta e escalável para lidar com futuras demandas nesse campo em constante evolução. Ao integrar técnicas de geração de dados fracamente supervisionados com modelos neurais avançados, os autores oferecem uma solução abrangente e promissora para um problema cada vez mais relevante na era da informação.

Em outro trabalho, de Sousa et al. [31] propõem uma estratégia para a recomendação de especialistas, utilizando dados abertos disponíveis na Plataforma Lattes. Com o aumento exponencial de dados e usuários em sistemas de currículos, a busca por especialistas tornou-se um desafio crescente. A Plataforma Lattes, que conta com mais de 6 milhões de pesquisadores registrados, representa uma fonte valiosa de informações. No entanto, seu potencial ainda não foi totalmente explorado para este propósito. A metodologia proposta neste estudo inclui a extração de dados dos currículos, o tratamento semântico desses dados e a aplicação de um agente de recomendação baseado em redes neurais profundas com autoencoder. Esta abordagem tem como objetivo identificar especialistas com base em tópicos de interesse, facilitando assim as colaborações e comunicações na comunidade científica. Os resultados iniciais demonstram uma boa performance computacional e a capacidade do modelo de formar grupos de especialistas com base em suas áreas de especialização. Este estudo

oferece uma ferramenta promissora para a gestão do conhecimento na ciência brasileira, destacando o potencial das técnicas de aprendizado profundo na identificação de especialistas em grandes conjuntos de dados abertos.

Freitas et al. [40] implementaram um algoritmo de similaridade que calcula a porcentagem de adesão entre perfis acadêmicos na Universidade de Brasília, utilizando a plataforma Lattes como base de dados. O algoritmo compara as produções acadêmicas registradas no Lattes, criando uma ontologia de conceitos para facilitar as comparações semânticas entre os perfis. A ontologia permite a criação de um modelo de dados específico que relaciona as produções acadêmicas, incluindo sinônimos dos termos usados, e apresenta os resultados em uma planilha com os percentuais de adesão. A metodologia envolveu a extração de informações dos currículos Lattes, a criação de uma ontologia em OWL, e a comparação dos termos de produção acadêmica entre os perfis. O algoritmo realiza essas comparações considerando sinônimos e contabilizando as ocorrências de termos iguais. Os resultados indicam a similaridade entre os perfis acadêmicos, permitindo identificar quais indivíduos têm maior proximidade semântica em suas produções científicas. Os testes mostraram que indivíduos com maior similaridade geralmente pertencem à mesma área de concentração, como *Software*, evidenciando que a metodologia é eficaz em identificar semelhanças dentro de um mesmo domínio acadêmico. O algoritmo também demonstrou a capacidade de identificar semelhanças entre indivíduos de áreas aparentemente distintas, caso suas produções acadêmicas apresentem termos e conceitos semelhantes. O estudo conclui que a utilização de uma ontologia específica para a plataforma Lattes é uma ferramenta poderosa para analisar a similaridade entre perfis acadêmicos, facilitando a cooperação e o conhecimento mútuo entre pesquisadores.

Vivian e Cervi [107] propõem uma abordagem de recomendação para o planejamento de carreira de pesquisadores, baseada na personalização dos dados, similaridade de perfil e reputação acadêmica. O trabalho utiliza um modelo de perfil denominado Rep-Model, que inclui elementos quantitativos, como publicações e orientações, e elementos textuais para definir a área de atuação. A reputação é avaliada por meio do Rep-Index, um índice que classifica os pesquisadores com base em múltiplos critérios além de artigos e citações, incorporando aspectos como participação em bancas, comitês, patentes, entre outros. A abordagem proposta divide as recomendações em dois tipos: personalizadas, que consideram o perfil e a reputação de outros pesquisadores para gerar sugestões específicas, e não personalizadas, que utilizam apenas o Rep-Index para identificar atividades que aumentem a reputação acadêmica. Os experimentos foram realizados com 398 pesquisadores da área de Ciência da Computação, resultando na identificação de 959 subáreas de atuação. Destas, 219 catego-

rias distintas foram mantidas para análise, sendo que 57 (25,9%) possuíam mais de um pesquisador e 163 (74,1%) tinham apenas um pesquisador, que foram excluídas dos experimentos. Para determinar a melhor combinação de técnicas de similaridade, foram aplicadas várias funções de correlação e distância, como Log-Likelihood, Coseno e Tanimoto. Os resultados mostraram que a combinação da técnica Log-Likelihood com a classe ClassicAnalyzer do Apache Mahout obteve a maior F-Measure (0,831) e o menor erro quadrático médio (RMSE) de 0,0012. A métrica estatística Cohen's Kappa também apresentou um valor elevado de 0,806, indicando uma alta concordância entre os resultados. No experimento de recomendação, foi gerado um conjunto de 28 recomendações diferentes, tanto personalizadas quanto não personalizadas, para cada elemento do tipo inteiro do Rep-Model. As recomendações variaram em termos de impacto na reputação acadêmica. Por exemplo, aumentar o número de orientações de doutorado poderia elevar a reputação em até 0,365 pontos, enquanto ampliar a rede de colaboração com pesquisadores similares poderia resultar em um incremento de 0,028 pontos no Rep-Index.Os resultados indicam que o modelo proposto é eficaz para gerar recomendações de carreira, ajudando pesquisadores a planejar suas atividades com base na similaridade de perfil e na reputação acadêmica. As métricas de avaliação, como cobertura e diversidade, demonstraram que as recomendações geradas são abrangentes e diversificadas, proporcionando um suporte estratégico para o desenvolvimento da carreira acadêmica.

## 3.4 Extração e Automatização de Informações

Tikhonova [105] introduz um método para a extração automática de informações de currículos de candidatos. O objetivo deste método é fornecer dados que possam auxiliar na tomada de decisões sobre a seleção e avaliação de candidatos. O método emprega técnicas de processamento de linguagem natural para extrair informações de campos de texto específicos do currículo, tais como *Esfera de Trabalho Anterior* e *Posição Anterior*. Posteriormente, utiliza algoritmos de agrupamento para classificar candidatos com informações semelhantes em grupos. Os resultados do agrupamento obtidos a partir dos campos de texto são utilizados como recursos, *features*, para um modelo de aprendizado de máquina. Este modelo é treinado em um conjunto de dados de candidatos com o objetivo de prever a probabilidade de um candidato deixar o cargo durante seus primeiros seis meses de trabalho. Os experimentos realizados pelos autores demonstraram que o uso das features geradas pelo método proposto melhorou o desempenho do modelo de aprendizado de máquina. Além disso, os autores planejam continuar aprimorando o método proposto e aplicá-lo em outros cam-

pos de texto em currículos e para outras tarefas de seleção e avaliação de candidatos.

Alves et al. [6] descrevem o *LattesMiner*, uma linguagem de domínio específico (DSL) multilíngue desenvolvida para a extração automática de informações do Currículo Lattes. A Plataforma Lattes, mantida pelo CNPq, é uma das principais fontes de informações sobre pesquisadores brasileiros, armazenando aproximadamente 2 milhões de currículos. A *LattesMiner* foi projetada para permitir que desenvolvedores implementem suas próprias aplicações com alto nível de abstração e poder de expressão. A ferramenta pode extrair dados de qualquer pesquisador ou grupo de pesquisadores pelo nome ou número de identificação (ID), permitindo a análise e utilização desses dados para identificar redes sociais acadêmicas, competências regionais, perfis de grupos de pesquisa, entre outros. O artigo detalha a arquitetura do *LattesMiner*, composta por seis componentes principais: Descoberta de Dados, Aquisição de Dados, Extração de Dados, Estrutura de Dados, Visualização de Dados e Análise de Dados. A metodologia de extração de dados utiliza expressões regulares devido à falta de equilíbrio de tags nos arquivos HTML do Currículo Lattes. A ferramenta armazena os dados extraídos em arquivos XML ou em um banco de dados, facilitando o uso por outras aplicações. O estudo de caso apresentado ilustra o uso do *LattesMiner* para identificar redes sociais acadêmicas na área de Ciência da Computação, destacando a flexibilidade e eficácia da ferramenta. Os resultados mostram que a *LattesMiner* pode automatizar a extração de dados de forma eficiente, superando limitações de ferramentas anteriores como o Lattes Extrator e o scriptLattes, que têm acesso restrito e requerem o número de identificação dos pesquisadores. A *LattesMiner*, sendo multilíngue, oferece uma interface amigável e flexível para usuários de diferentes idiomas. Além disso, a ferramenta foi utilizada para desenvolver o sistema SUCUPIRA, que identifica e visualiza redes sociais acadêmicas, mostrando a utilidade prática da *LattesMiner* na análise de perfis de pesquisadores e redes de colaboração acadêmica.

Bauer [13] discutem a aplicação de técnicas de Mineração de Dados para extrair e gerar conhecimento a partir dos dados dos currículos disponíveis na Plataforma Lattes. A pesquisa é motivada pela necessidade das universidades de obter informações precisas e rápidas sobre a produção científica, bibliográfica e tecnológica de seus docentes, utilizando técnicas de Descoberta de Conhecimento em Bases de Dados (KDD). O estudo se concentra em técnicas de Clusterização e Regras de Associação. A metodologia do estudo envolve várias etapas. Primeiramente, foi realizada a coleta de dados dos currículos Lattes dos professores da Universidade de Santa Cruz do Sul (UNISC). Esses dados foram extraídos e estruturados em formato XML. O processo de extração utilizou um sistema desenvolvido especificamente para este fim, capaz de baixar automati-

camente os currículos da plataforma e convertê-los para o formato adequado para análise. Para a análise dos dados, foram aplicados algoritmos de clusterização, como o K-means, e algoritmos de associação, como o Apriori. O algoritmo K-means foi utilizado para agrupar os dados dos currículos em clusters, identificando grupos de pesquisadores com perfis semelhantes. Este método se mostrou eficaz em categorizar os dados com base em características comuns, facilitando a identificação de padrões. O algoritmo Apriori, por sua vez, foi utilizado para descobrir regras de associação, identificando padrões frequentes de coocorrência de diferentes tipos de produções científicas. Por exemplo, o estudo conseguiu identificar que pesquisadores que publicam frequentemente em determinadas áreas também tendem a colaborar em outros tipos de projetos científicos, oferecendo insights valiosos para a gestão acadêmica. Os resultados do estudo foram significativos. A aplicação das técnicas de KDD permitiu a criação de relatórios detalhados que foram utilizados para mapear competências, identificar potenciais colaboradores e avaliar a produtividade científica dos docentes da UNISC. Os dados mostraram que 60% dos pesquisadores estavam concentrados em três principais áreas de estudo, e as regras de associação revelaram que certos grupos de pesquisadores tinham uma alta propensão a colaborar em projetos interdisciplinares. Essas informações são valiosas para a administração universitária, permitindo uma melhor alocação de recursos e estratégias de incentivo à pesquisa. O estudo conclui que a utilização de técnicas de Mineração de Dados e KDD na análise dos currículos Lattes é uma ferramenta poderosa para a gestão acadêmica, proporcionando um meio eficiente de extrair conhecimento valioso dos dados disponíveis e apoiar a tomada de decisões estratégicas.

Tikhonova e Gavrishchuk [106] propõem um método inovador para segmentação e extração automática de informações de currículos (CVs) utilizando Processamento de Linguagem Natural (NLP) e métodos de aprendizado de máquina. O problema da extração automática de informações de CVs é particularmente relevante para grandes empresas, onde o volume de candidatos é alto e o processamento manual de CVs é demorado e exige muitos recursos humanos. O algoritmo desenvolvido visa extrair informações relacionadas à experiência profissional e educação dos candidatos a partir de CVs em formato PDF ou DOCX, segmentando-os em três blocos principais: Informação Básica, Educação e Experiência Profissional. O método proposto envolve sete etapas consecutivas: pré-processamento e transformação do CV, construção de embeddings de palavras, cálculo do índice tf-idf, construção de embeddings de campos de texto, extração de características específicas, classificação das linhas de texto e segmentação final do CV. No pré-processamento, cada CV é lido como texto simples e transformado em um formato adequado para algoritmos de aprendizado de máquina. As palavras são então embutidas em um espaço linear usando

algoritmos como Word2Vec, FastText e GloVe, e o índice tf-idf é calculado para cada palavra. Cada parte do CV é transformada em um vetor de características, que inclui contadores de partes do discurso e sufixos específicos. A classificação das linhas de texto é realizada através de duas tarefas binárias: previsão da experiência profissional e previsão da educação. As previsões são suavizadas e pequenos segmentos são removidos para melhorar a precisão da segmentação. A avaliação do algoritmo utilizou o índice de Jaccard, que mede a similaridade entre dois conjuntos, obtendo resultados significativos: Jaccard para experiência profissional = 0.942 e Jaccard para educação = 0.806. Além disso, 82% dos CVs de teste foram segmentados corretamente, com 42% classificados sem erros. O estudo conclui que o método proposto é eficaz para a segmentação automática de CVs, simplificando o processo de seleção de candidatos e reduzindo o tempo de processamento de CVs. Para melhorar a qualidade do algoritmo, os autores planejam aumentar os dados de treinamento com exemplos de CVs não padronizados e expandir o algoritmo para extrair outros segmentos, como habilidades e educação adicional.

Mendonça et al. [74] investigam a importância da Plataforma Lattes, com foco na ferramenta QLattes, para a anotação e visualização de dados acadêmicos no Brasil. A Plataforma Lattes, gerida pelo CNPq, é destacada como o principal repositório de currículos acadêmicos do país, desempenhando um papel central na organização e no acesso a registros acadêmicos e profissionais dos pesquisadores brasileiros. O artigo concentra-se no QLattes, uma extensão de navegador que permite a classificação automática e a anotação de publicações em periódicos conforme as categorias do Qualis. O estudo discute as vantagens do QLattes na simplificação do processo de avaliação da produção científica, apresentando-o como uma ferramenta eficaz para a criação de métricas e indicadores de qualidade. Além disso, aborda funcionalidades adicionais como a filtragem e a visualização dos dados, que permitem uma análise mais detalhada das publicações, facilitando a compreensão das dinâmicas de produção científica. Os resultados indicam que o QLattes automatiza a classificação Qualis e oferece visualizações interativas, tornando-o uma ferramenta valiosa para pesquisadores e gestores acadêmicos. O artigo também fornece dados preliminares sobre o impacto do QLattes na comunidade acadêmica brasileira, destacando seu potencial para a análise da produção científica. O código-fonte do QLattes está disponível gratuitamente, permitindo que outros pesquisadores adaptem a ferramenta conforme suas necessidades.

Oliveira e Silva [83] desenvolveram um extrator automatizado de dados acadêmicos para o Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Campus Tianguá, com o objetivo de otimizar a análise dos currículos Lattes dos pesquisadores do campus. Tradicionalmente, a análise manual desses

currículos era uma tarefa morosa e suscetível a erros, especialmente para grupos de médio e grande porte. Este estudo, portanto, relata a criação de um protótipo baseado em Python que automatiza a aquisição dos currículos e a quantificação das produções acadêmicas, eliminando a necessidade de downloads manuais ou resolução de captchas. Segundo os autores, a Plataforma Lattes permite que os pesquisadores insiram seus dados, o que frequentemente resulta em currículos desatualizados e dificulta a análise manual. Com a ferramenta Extrator Lattes/PRPI desatualizada desde 2020, a coordenadora de pesquisa do campus de Tianguá precisava de um método mais eficiente para calcular a taxa de produção científica. O estudo baseou-se em metodologias de Engenharia de Software para desenvolver o protótipo em cinco etapas: Revisão de Literatura, Requisitos do Protótipo, Desenvolvimento do Protótipo, Obtenção de Resultados e Análise Comparativa. O protótipo foi desenvolvido no ambiente Google Colaboratory, o que facilitou testes preliminares e validações. A arquitetura lógica do protótipo inclui módulos para a aquisição de currículos, pré-processamento dos dados, cálculo das produções acadêmicas e apresentação dos resultados em gráficos de barras. A ferramenta automatiza completamente o processo de extração dos currículos dos pesquisadores e o cálculo da taxa de produção científica. Os resultados preliminares indicaram que o campus possui 37 docentes ativos na pesquisa, com uma taxa de produção científica de 56,56 no segundo trimestre de 2021, destacando-se as produções bibliográficas. O protótipo mostrou-se eficaz ao automatizar o processamento dessas informações, evitando a necessidade de interação manual com a Plataforma Lattes. Comparado a outros sistemas, como os da UNESP, lucyLattes e UFOP Ativa, o protótipo se diferencia pela automação completa, desde a identificação dos docentes até o download dos currículos. No entanto, ainda é considerado incompleto por não estar disponível em versão desktop e por ter custos associados à resolução de CAPTCHAs. Oliveira e Silva [83] concluem que o protótipo é promissor para quantificar produções acadêmicas e sugerem melhorias, como uma versão desktop, separação entre pesquisadores internos e externos, relatórios mais completos e um dashboard com indicadores de produção científica.

Mena-Chalco e Cesar Junior [72] tratam do desenvolvimento e aplicação do *scriptLattes*, uma ferramenta de código aberto projetada para gerar automaticamente relatórios acadêmicos detalhados a partir dos dados curriculares da plataforma Lattes, mantida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). A plataforma Lattes integra e organiza informações sobre as realizações acadêmicas e científicas de pesquisadores, estudantes e profissionais brasileiros, mas apresenta limitações quanto à análise coletiva desses dados devido à sua disposição individualizada. O *scriptLattes* foi criado para preencher essa lacuna, permitindo a criação automatizada de relatórios abran-

gentes sobre a produção científica, técnica e artística, além de orientações acadêmicas de grupos de pesquisa. O sistema inclui módulos para a seleção, pré-processamento e tratamento de dados, eliminação de redundâncias, geração de gráficos de colaboração entre membros do grupo e mapas de pesquisa baseados em informações geográficas. Testado em diversas instituições brasileiras, como a Universidade de São Paulo (USP) e a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), o *scriptLattes* demonstrou eficiência na extração e visualização de conhecimento a partir dos dados do Lattes. A ferramenta oferece respostas automatizadas a questões bibliométricas essenciais, como a quantidade e evolução das produções, perfil de publicações, colaboração entre pesquisadores, conclusão de teses e dissertações, e distribuição geográfica dos pesquisadores. Desta forma, o *scriptLattes* economiza tempo, melhora a precisão e a abrangência dos relatórios, e se torna uma ferramenta valiosa para a análise e avaliação das atividades acadêmicas e científicas no Brasil.

## 3.5 Chatbot

Colombo et al. [28] discutem a criação e implementação do *PharmaBulaBot*, um chatbot desenvolvido para responder a perguntas sobre medicamentos com base nas informações contidas nas bulas farmacêuticas. As bulas de medicamentos, que acompanham os remédios, são fontes importantes de informações para os pacientes. No entanto, essas bulas são frequentemente difíceis de ler e entender, o que pode comprometer a eficácia da comunicação das orientações e informações essenciais ao uso dos medicamentos. Para resolver esse problema, o *PharmaBulaBot* utiliza técnicas de Processamento de Linguagem Natural (NLP) para extrair e apresentar informações das bulas de uma maneira interativa e eficiente. O chatbot visa proporcionar respostas de qualidade comparável àquelas fornecidas por profissionais de saúde, facilitando o acesso rápido e preciso às informações. O sistema foi desenvolvido sob a teoria dos Sistemas Soft (Soft Systems Theory) e avaliado por meio de experimentos que demonstraram sua eficácia em responder a 94,03% das 3.147 perguntas geradas por usuários humanos. O *PharmaBulaBot* é uma ferramenta que pode ser integrada aos Sistemas de Informação da Saúde, como aqueles utilizados nas unidades de Farmácia Cidadã no Espírito Santo, melhorando a orientação e apoio aos pacientes sobre o uso de seus medicamentos. Esta inovação tem o potencial de aumentar a acessibilidade e a compreensão das informações farmacêuticas, beneficiando tanto a comunidade científica quanto o público em geral.

Júnior e Carvalho [56] exploram as principais características que definem softwares desenvolvidos para imitar ações humanas, conhecidos como *bot*, *robot*, *chatbot* ou *chatterbot*, e simulam a interação desses aplicativos com

usuários humanos. Eles apresentam as aplicações em que os bots mais se destacam, utilizando como canal de comunicação os aplicativos mais conhecidos disponíveis no mercado. Inicialmente, realizaram um levantamento bibliográfico selecionando livros, artigos e sites relevantes para a temática. Em seguida, conceituaram um chatbot ou assistente virtual de conversa e a inteligência artificial. Finalmente, realizaram estudos de caso simplificados para demonstrar o funcionamento dessas aplicações e analisaram interações humano-máquina com assistentes virtuais, compilando informações para ampliar a compreensão sobre o tema. Os autores destacam que os chatbots podem ser usados em diversas áreas como educação, entretenimento e autoatendimento em lojas virtuais. A metodologia utilizada incluiu estudo de caso e pesquisa bibliográfica, com ferramentas como o Google Acadêmico. Um dos principais pontos abordados foi a necessidade de assistentes virtuais que facilitam a execução de tarefas cotidianas sem comunicação direta com um humano, como rastreamento de encomendas, transferências bancárias e agendamento de consultas médicas. O estudo detalhou as características básicas de um bot, como a automação de funções cotidianas e a simulação de interações humanas. Eles mencionaram a implementação de sistemas de helpdesk que utilizam tecnologias de inteligência artificial e processamento de linguagem natural para oferecer atendimento ao cliente de forma econômica e interativa. Além disso, apresentaram diversas APIs (Application Programming Interface) disponíveis para a criação de chatbots, como A.L.I.C.E., Pandorabots, Wit.ai, Dialogflow e Messenger Platform. Os resultados dos estudos de caso demonstraram o funcionamento de aplicações inteligentes utilizando ferramentas como Prolog e Dialogflow, com exemplos práticos de agendamento de consultas médicas e interação com usuários via aplicativos como WhatsApp. A análise comparativa entre assistentes virtuais, como Microsoft Cortana e Google Assistant, mostrou que o Google Assistant apresentou respostas mais precisas em algumas tarefas. Os autores concluem que os bots estão se tornando cada vez mais presentes nas tarefas cotidianas, integrando a computação às atividades diárias de forma transparente. Para o futuro, sugerem pesquisas focadas em analisar as interfaces das aplicações inteligentes e comparar a usabilidade e comunicabilidade entre os principais chatbots, com o objetivo de enumerar vantagens e desvantagens e propor melhorias.

Bernardini et al. [15] apresentam uma análise bibliométrica do estado da arte da literatura sobre chatbots, destacando sua relevância e crescimento no campo da Inteligência Artificial. A pesquisa utiliza a plataforma Scopus para identificar publicações científicas relacionadas ao tema, resultando em 273 documentos analisados. Os autores destacam que a primeira publicação sobre chatbots foi registrada em 2002, e houve um aumento significativo no nú-

mero de publicações a partir de 2016, impulsionado por avanços tecnológicos e maior interesse das organizações. A análise revela que a maioria das publicações está concentrada na área de Ciência da Computação, com os Estados Unidos e a Itália liderando em número de publicações. O Brasil ocupa a 15ł posição, com quatro artigos. O estudo também identifica as áreas de conhecimento relacionadas aos chatbots, incluindo Educação, Saúde e Linguística, ressaltando a natureza interdisciplinar da Inteligência Artificial. Em termos de tipo de publicação, os papers de conferências internacionais representam a maioria, seguidos por artigos de revistas e revisões. A análise das publicações mais citadas destaca trabalhos como o de Turney e Littman (2003), que receberam 748 citações, focando na orientação semântica de textos, e o de Lee et al. (2009), que discutem um modelo de diálogo baseado em exemplos. O estudo conclui que, apesar da quantidade relativamente pequena de publicações, o interesse por chatbots está em crescimento, refletindo sua importância crescente em diversos setores. Os autores sugerem que futuras pesquisas devem explorar mais profundamente as aplicações práticas e os desafios éticos associados ao uso de chatbots.

Rehbein [89] propõem o desenvolvimento de um chatbot para auxiliar candidatos na escolha de programas de pós-graduação, abordando a dificuldade que estudantes encontram ao pesquisar programas e orientadores que melhor se encaixam em suas necessidades. Este envolve a coleta e análise de currículos de professores disponíveis na Plataforma Lattes. A metodologia do estudo seguiu vários passos: (i) identificação das universidades integrantes do COMUNG (Consórcio das Universidades Comunitárias Gaúchas); (ii) localização dos campi; (iii) pesquisa em websites das universidades; (iv) identificação dos programas de pós-graduação; e (v) contagem do corpo docente e suas linhas de pesquisa. As informações foram armazenadas em planilhas de Excel e os currículos Lattes dos professores foram coletados e salvos em formato XML. Em seguida, um código em Java foi desenvolvido para ler esses arquivos e extrair informações relevantes como nome completo, universidade de atuação, titulação e publicações. Essas informações foram armazenadas em um banco de dados PostgreSQL e integradas à base de dados do chatbot. O objetivo é que o chatbot possa fornecer recomendações personalizadas para os candidatos a programas de mestrado e doutorado. Os resultados preliminares incluem a coleta de dados de 14 universidades, detalhando os programas de pós-graduação, número de professores e suas linhas de pesquisa. O estudo conclui que o uso de um chatbot, alimentado com dados extraídos da Plataforma Lattes, pode significativamente reduzir o tempo e o esforço necessários para os candidatos encontrarem programas de pós-graduação e orientadores adequados. O próximo passo inclui o treinamento do chatbot com os dados coletados para melhorar a precisão e a utilidade das recomendações.

Saito e Miura [94] desenvolveu um Trabalho de Conclusão de Curso (TCC) com o título Processamento Natural de Linguagem: Sistema de Recomendações e Explicações, na Escola Politécnica da Universidade de São Paulo. O objetivo principal do trabalho foi construir um chatbot capaz de fornecer recomendações e explicações para os alunos da USP, especialmente no contexto de escolha de disciplinas. A arquitetura do sistema foi dividida em duas partes principais: a interface com o usuário, desenvolvida utilizando a plataforma DialogFlow, e o sistema de recomendação, baseado no modelo Gensim. A pesquisa começou com uma análise do estado da arte, cobrindo as principais técnicas e algoritmos de Processamento Natural de Linguagem (PNL), incluindo o Google BERT, XLNet, e GPT, bem como abordagens para sistemas de recomendação e análise de sentimentos. A base de dados utilizada no sistema foi criada a partir da extração de informações do sistema JupiterWeb, que contém detalhes sobre as disciplinas da USP. Para garantir a eficácia das recomendações, o sistema foi projetado para analisar a similaridade entre os temas de interesse do usuário e os textos dos programas das disciplinas. Nos testes realizados, o sistema demonstrou bom desempenho, particularmente para temas específicos ou pertencentes a nichos de conhecimento bem definidos. Por exemplo, ao buscar por *Termodinâmica*, as disciplinas recomendadas foram PME3301 - Termodinâmica e SEM0421 - Termodinâmica Aplicada à Engenharia Química. Além disso, o tempo de processamento médio do algoritmo foi de 2,4 segundos, mesmo com uma base de dados composta por quase 17 mil registros, o que indica a eficiência do sistema. No entanto, para temas genéricos, como *Química*, o sistema mostrou limitações, retornando disciplinas que, embora relacionadas, não necessariamente representavam a melhor escolha entre todas as possíveis. O autor conclui que, embora o sistema tenha alcançado resultados satisfatórios, especialmente para temas específicos, ainda há espaço para melhorias. A inclusão de bases de conhecimento semântico, por exemplo, poderia aumentar a precisão das recomendações para temas mais amplos ou mal especificados. O trabalho demonstra a relevância crescente dos chatbots na educação, especialmente em ambientes complexos como o da USP, onde a escolha de disciplinas pode ser uma tarefa desafiadora para os alunos.

Calle et al. [23] demonstram o potencial dos *chatbots* tipo recomendador em contextos educacionais, ao apoiar o aprendizado autorregulado e oferecer suporte contínuo ao usuário. Da mesma forma, Wang et al. [109] destacam a capacidade dessas ferramentas em otimizar processos internos, o que, no cenário acadêmico, poderia contribuir para automatizar etapas do processo de orientação e reduzir a sobrecarga administrativa. Além disso, Danckwerts et al. [29] ressaltam como a personalização das recomendações pode elevar o engajamento e a satisfação do usuário final.

## 3.6 Sistemas de Recomendação

Atalla et al. [8] apresentam um sistema de recomendação inteligente para automatizar a orientação acadêmica com base na análise do currículo e modelagem de desempenho. Com o aumento explosivo de dados educacionais e sistemas de informação, surgem novos desafios e processos de aprendizado. O sistema proposto utiliza métodos estatísticos, como aprendizado de máquina (ML) e análise de grafos, para analisar dados de programas e estudantes, desenvolvendo planos de estudo personalizados ao longo de vários semestres. A arquitetura do sistema inclui uma camada de dados, uma camada de recomendação e uma camada de aplicação, integrando análise de redes, inteligência artificial e teoria dos grafos. O sistema foi testado com dados da Universidade de Dubai, mostrando um desempenho superior a soluções semelhantes baseadas em ML, atingindo até 86% de precisão e recall, e a menor taxa de erro quadrático médio (MSR) de 0,14. O estudo destaca a eficácia do sistema em melhorar a retenção e a taxa de graduação dos estudantes, através de recomendações personalizadas e predições de desempenho acadêmico.

Pu et al. [88] apresentam uma revisão da pesquisa sobre sistemas de recomendação (RS), com foco na experiência do usuário. Eles destacam uma recente mudança na avaliação dos RS, que antes se concentrava principalmente no desempenho dos algoritmos, para uma abordagem que considera a eficácia do sistema e os critérios de avaliação sob a perspectiva dos usuários. O estudo examina o estado atual da pesquisa sobre a experiência do usuário em RS. Ele destaca como os pesquisadores têm avaliado métodos de design que buscam melhorar a capacidade dos RS de ajudar os usuários a encontrar informações ou produtos de interesse, interagir facilmente com o sistema e desenvolver confiança através de transparência, controle e preservação da privacidade. Além disso, o texto investiga como esses recursos de design influenciam a adoção da tecnologia pelos usuários. Uma taxonomia de critérios do usuário é definida para avaliar os RS. Ela inclui a qualidade dos itens recomendados, a facilidade de elicitação e revisão de preferências, a adequação do layout e rótulos da área de recomendação, a capacidade de auxiliar nas decisões dos usuários e a capacidade de explicar os resultados recomendados e inspirar confiança. O artigo é organizado em seções que abordam diferentes aspectos da interação entre o usuário e o sistema de recomendação. Ele começa com uma classificação dos tipos de sistemas de recomendação, seguida por uma análise detalhada da elicitação de preferências, refinamento de preferências e estratégias de apresentação de resultados. Cada seção apresenta os

principais resultados da pesquisa existente e deriva um conjunto de diretrizes de design para sistemas de recomendação eficazes. Por fim, o texto conclui resumindo a pesquisa e discutindo suas implicações para futuras investigações. Ele também fornece uma estrutura geral de avaliação de usuários para sistemas de recomendação. Essas diretrizes podem ser úteis para acadêmicos e profissionais no design e desenvolvimento de sistemas de recomendação que atendam às necessidades e expectativas dos usuários.

Magalhães et al. [68] propõem um Sistema Personalizado de Recomendação de Artigos Científicos (PPRS), que utiliza o currículo acadêmico dos usuários da plataforma CV-Lattes para criar perfis personalizados de recomendação. O estudo investiga diferentes estratégias de construção de perfis de usuários, utilizando termos e conceitos, e avalia quanto tempo de informações passadas é necessário para fornecer recomendações eficazes. O experimento envolveu 30 usuários da área de Ciência da Computação. Os perfis de usuários foram construídos a partir de informações extraídas do CV-Lattes, incluindo publicações, projetos, formação acadêmica e produção técnica. Dois tipos de perfis foram desenvolvidos: perfis de termos e perfis de conceitos. Os perfis de termos foram criados com base na frequência de termos (TF-IDF), enquanto os perfis de conceitos utilizaram uma ontologia para representar o domínio e calcular a similaridade entre os perfis de usuários e os documentos. Os usuários avaliaram um conjunto de 50 artigos de Ciência da Computação, divididos igualmente entre Inteligência Artificial e Engenharia de Software. Os artigos foram pré-processados e indexados usando termos e conceitos. Os participantes avaliaram os artigos usando uma escala Likert de 1 a 5 estrelas. Os resultados mostraram que os perfis de termos forneceram melhores resultados em termos de NDCG@5 (Normalized Discounted Cumulative Gain para os top 5 artigos) e NDCG@10, mas foram mais lentos em tempo de processamento. Para perfis de termos, a análise dos últimos quatro anos de dados forneceu os melhores resultados. Para perfis de conceitos, cinco anos de dados foram mais eficazes. As abordagens propostas superaram o método de Lopes et al. (2008), que utilizava apenas títulos e palavras-chave das publicações. Apesar de menos detalhados, os perfis de conceitos apresentaram um desempenho comparável aos perfis de termos, sendo mais rápidos e adequados para recomendações em tempo real. Os autores concluíram que a utilização do currículo do usuário para construção de perfis de recomendação é eficaz. A abordagem baseada em conceitos, embora menos precisa em algumas métricas, é vantajosa em cenários que requerem processamento em tempo real devido à sua rapidez. Este estudo destaca a importância de considerar a dimensão temporal das informações do currículo e sugere que a análise dos dados mais recentes é crucial para a precisão das recomendações.

Roy e Dutta [93] trataram de um estudo sobre sistemas de recomenda-

ção, que são ferramentas eficientes para filtrar informações online, sendo cada vez mais utilizados devido aos hábitos de computação em constante mudança, tendências de personalização e acesso emergente à internet. Apesar da precisão alcançada pelos sistemas de recomendação recentes, eles enfrentam várias limitações e desafios, como escalabilidade, início a frio, esparsidade, entre outros. O estudo realiza uma revisão sistemática das contribuições recentes no domínio dos sistemas de recomendação, com foco em diversas aplicações, como livros, filmes, produtos, etc. Ele analisa as várias aplicações de cada sistema de recomendação, realiza uma análise algorítmica e cria uma taxonomia que considera os componentes necessários para desenvolver um sistema eficaz. Além disso, avalia os conjuntos de dados coletados, plataformas de simulação e métricas de desempenho de cada contribuição. O estudo fornece uma visão geral do estado atual da pesquisa nesse campo, destacando lacunas e desafios existentes para auxiliar no desenvolvimento de sistemas de recomendação eficientes. Ele também explora os tipos de sistemas de recomendação, como sistemas baseados em conteúdo, sistemas colaborativos e sistemas híbridos, explicando suas características, vantagens e desafios. Este texto apresenta um estudo sobre sistemas de recomendação, focando em uma abordagem baseada na Análise de Múltiplos Critérios (MCDA) para recomendar artigos científicos. Ele ressalta a crescente importância dos sistemas de recomendação devido ao grande aumento de informações disponíveis, especialmente na World Wide Web. O principal objetivo desses sistemas é reduzir a sobrecarga de informações, selecionando um subconjunto relevante de itens com base nas preferências do usuário. No entanto, muitos usuários sentem que os sistemas atuais não os reconhecem como indivíduos únicos, o que motiva a busca por abordagens mais centradas no usuário.

Matsatsinis et al. [70] propõem uma metodologia baseada em MCDA para superar essa falta de personalização nos sistemas de recomendação, integrando ativamente o usuário no processo de recomendação. Essa abordagem visa melhorar a satisfação do usuário e a eficácia do sistema, considerando suas necessidades e preferências de maneira mais direta. O texto também destaca a importância da interação entre humano e computador (HRI) para incorporar o conhecimento do usuário no sistema de recomendação. Além disso, o texto discute as limitações dos sistemas de recomendação existentes, como o problema da partida a frio e a dependência de conteúdo em técnicas baseadas em conteúdo. O estudo sugere o uso de técnicas híbridas que combinam filtragem colaborativa e filtragem baseada em conteúdo para superar essas limitações. No entanto, uma abordagem inovadora é sugerida, que emprega metodologias MCDA para envolver ativamente o usuário no processo de recomendação desde o início, a fim de mitigar o problema da partida a frio. O texto for-

nece uma visão geral da metodologia proposta, explicando como ela pode ser aplicada ao problema específico de recomendação de artigos científicos. Ele destaca a natureza multicritério desse problema e demonstra como um sistema autônomo eficiente para recomendar artigos pode ser uma ferramenta valiosa para pesquisadores. Em termos de estrutura, o artigo é dividido em seções que abordam desde a introdução dos sistemas de recomendação até a metodologia proposta, com uma análise detalhada e um exemplo ilustrativo. Ele conclui com reflexões sobre as implicações e futuras direções dessa abordagem para o campo dos sistemas de recomendação.

Zayed et al. [111] desenvolveram um sistema de recomendação para ajudar estudantes a escolherem o curso de graduação adequado, utilizando dados de estudantes de MBA. Eles investigaram várias técnicas de aprendizado de máquina supervisionado, incluindo Decision Tree, Random Forest e Support Vector Machine, para prever as escolhas dos cursos. Os dados incluíam histórico acadêmico e informações do mercado de trabalho. A pesquisa utilizou um dataset de 216 estudantes do CMS Business School, com 13 características de entrada, como porcentagens de notas em diversas etapas educacionais, experiência de trabalho, resultados de testes de entrada, status de emprego e salário após a graduação. Os resultados mostraram que o algoritmo Random Forest superou os outros, alcançando uma precisão de 97,70%, comparado a 75,00% em estudos anteriores, destacando a importância da porcentagem de graduação, porcentagem de MBA e resultados de testes de entrada como características principais. Durante a fase de preparação dos dados, valores ausentes foram tratados, e os dados categóricos foram convertidos em valores numéricos utilizando o Label Encoder. A normalização foi aplicada para escalar os valores entre 0 e 1. O dataset foi dividido em conjuntos de treinamento e teste nas proporções de 80:20 e 70:30. A etapa de classificação e ajuste dos modelos utilizou o GridSearchCV para otimização de hiperparâmetros, o que envolveu a avaliação e ajuste de parâmetros como critério de divisão, profundidade máxima e número mínimo de amostras para divisão e folhas. Os resultados experimentais indicaram que o Random Forest teve a melhor performance com TPR (taxa de verdadeiros positivos) de 94,90% e FPR (taxa de falsos positivos) de 4%. O estudo concluiu que a hiperparametrização e a correta preparação dos dados são cruciais para melhorar a precisão dos modelos de aprendizado de máquina. O Random Forest se mostrou o melhor algoritmo, e a eliminação de características de baixa importância aumentou ainda mais a precisão. Este sistema de recomendação pode ser uma ferramenta valiosa para estudantes ao considerar suas habilidades, interesses e as demandas do mercado de trabalho ao escolher um curso universitário.

He et al. [46] introduzem o Neural Collaborative Filtering (NCF), um fra-

mework que utiliza redes neurais profundas para melhorar a recomendação, superando as limitações dos métodos tradicionais, como a Matrix Factorization (MF). O NCF substitui a abordagem linear do MF por uma arquitetura neural que modela interações complexas entre usuários e itens por meio de um perceptron multicamadas (Multi-Layer Perceptron - MLP). robotNa metodologia, os autores combinaram componentes lineares e não-lineares no modelo Neural Matrix Factorization (NeuMF) e testaram a abordagem em conjuntos de dados reais de MovieLens e Pinterest, com mais de 1 milhão de interações. Os resultados mostraram que o NCF aumentou o desempenho em métricas de recomendação, como Hit Ratio e NDCG, em cerca de 6,7% e 7,6%, respectivamente, em comparação com a MF tradicional, demonstrando a eficácia da abordagem em capturar interações complexas.

## 3.7 Motivação para Ingresso na Pós-Graduação

Silva et al. [100] descrevem uma pesquisa conduzida na Universidade Federal de Uberlândia (UFU), cujo objetivo era entender as razões que levam os estudantes a se matricularem em programas de pós-graduação *stricto sensu*. A pesquisa contou com a participação de 374 estudantes de 36 programas de mestrado e doutorado, que responderam a um questionário online. Os dados coletados foram analisados utilizando o software *Statistical Package for Social Sciences* (SPSS) e uma análise de conteúdo, resultando na identificação de quatro motivos principais: aspirações acadêmicas, busca por qualificação profissional, interesse na pesquisa e falta de outras opções. Adicionalmente, o texto discute a relevância da formação em pós-graduação *stricto sensu* para o ensino e a pesquisa no Brasil, conforme estipulado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) . No entanto, é destacada uma lacuna na formação específica para o ensino e a pesquisa nos programas de pós-graduação. A pesquisa enfatiza a diversidade de motivações dos estudantes para ingressar na pós-graduação, sublinhando a necessidade de abordagens pedagógicas que atendam às suas necessidades de maneira intencional e emancipatória. Isso sugere a importância de uma reflexão sobre as práticas pedagógicas nos programas de pós-graduação, com o objetivo de promover um ambiente de aprendizagem que incentive o desenvolvimento acadêmico e profissional dos estudantes.

## 3.8 Características de Orientadores de Doutorado

Taylor et al. [104] abordam as características que consideram ideais para

orientadores de doutorado nos Estados Unidos, considerando as visões de candidatos, graduados e orientadores acadêmicos. O estudo, de método misto, analisa aspectos qualitativos e quantitativos das percepções dos participantes. Foram coletados dados de entrevistas com 13 orientadores acadêmicos e 18 candidatos e graduados de doutorado, além de respostas de pesquisas completadas por 38 orientadores acadêmicos e 151 candidatos e graduados, representando 33 estados e diversas disciplinas. As descobertas indicam que os participantes valorizam a estrutura no processo de orientação, feedback útil e oportuno, comunicação regular e apoio emocional durante a jornada de pesquisa do doutorado. Além disso, apreciam uma relação profissional que evolui de hierárquica para colegial conforme o candidato avança no processo de doutorado. O estudo conclui que as qualidades desejáveis em um orientador de doutorado incluem a promoção de comunicação efetiva, construção de relacionamentos favoráveis e fornecimento de suporte holístico aos candidatos. Foi observado que é problemático quando o aprendizado pessoal dos alunos não é suficientemente apoiado. O estudo também revelou a importância da orientação voltada para o mercado e do aprendizado prático para os formandos, sugerindo que o orientador deve fornecer suporte em questões profissionais, além do suporte acadêmico.

## 3.9 Identificação de Redes de Coautoria

Mena-Chalco et al. [73] aborda a relevância da *Plataforma Lattes do Brasil*, um repositório de dados acadêmicos que documenta as atividades de pesquisa de acadêmicos associados a diversas áreas do conhecimento. A plataforma é amplamente empregada para avaliar e registrar a produção científica de grupos de pesquisa. No entanto, informações sobre interações entre pesquisadores brasileiros, como a coautoria, não foram previamente analisadas. Os autores propõem uma pesquisa para identificar e caracterizar as redes de coautoria acadêmica no Brasil, utilizando propriedades topológicas de grafos. Para isso, exploram estratégias para desenvolver um extenso conjunto de currículos da Plataforma Lattes, um algoritmo para identificar coautorias automáticas com base em informações bibliográficas e métricas topológicas para investigar as interações entre os pesquisadores. O estudo avalia informações de mais de um milhão de pesquisadores associados a oito grandes áreas de conhecimento no Brasil. Além disso, ressalta a importância de compreender como os pesquisadores e as áreas de conhecimento interagem para obter *insights* sobre sua estrutura e dinâmica. A coautoria é destacada como uma forma de colaboração entre pesquisadores, sendo representada graficamente por redes de coautoria, que são valiosas para entender a estrutura e a dinâmica das colaborações. A pesquisa busca responder a perguntas sobre a estrutura e a dinâmica das redes de

coautoria dentro e entre as principais áreas de conhecimento, levando em consideração um período de avaliação de três anos para capturar o desenvolvimento da pesquisa ao longo do tempo. Este estudo fornece uma base para explorar e identificar padrões de atividades de pesquisa, obtendo informações bibliométricas e cientométricas sobre todos os pesquisadores brasileiros registrados na Plataforma Lattes. Em relação às conclusões dos autores, eles identificaram e caracterizaram as redes de coautoria acadêmica brasileira usando propriedades topológicas de grafos. Desenvolveram um grande conjunto de currículos da Lattes e um algoritmo para identificar coautorias automáticas. Avaliaram informações de mais de um milhão de pesquisadores associados a oito grandes áreas de conhecimento brasileiras. Destacaram a importância de entender como os pesquisadores e áreas de conhecimento interagem. Reconheceram a coautoria como uma forma de colaboração entre pesquisadores. Responderam a perguntas sobre a estrutura e dinâmica das redes de coautoria. E forneceram uma base para explorar e identificar padrões de atividades de pesquisa.

Maruyama e Digiampietri [69] propõem uma metodologia que combina técnicas de agrupamento e classificação para a predição de coautorias na Plataforma Lattes. O estudo destaca a importância das Redes Sociais Online e como estas podem ser utilizadas para construir redes sociais acadêmicas a partir das informações disponíveis na Plataforma Lattes. A tarefa de predição de relacionamentos, ou links, visa identificar possíveis colaboradores, facilitando a comunicação entre os usuários. O trabalho propõe o uso da técnica de agrupamento e a inclusão de novos atributos baseados em informações de comunidade para melhorar a previsão de relações de coautoria. Foram utilizados dados de 657 currículos de pesquisadores permanentes de programas de pós-graduação em Ciência da Computação, coletados entre 1971 e 2015. A metodologia inclui a seleção de atributos, agrupamento utilizando o algoritmo K-means, e a combinação de modelos de classificação como Random Forest, Stacking e Voting. Os resultados indicam que o uso de agrupamento e a adição de atributos de comunidade melhoram a precisão e a medida-F das predições. A maior precisão alcançada foi de 0,567 utilizando Random Forest com Stacking, enquanto a maior medida-F foi de 0,513 com Random Forest e Voting. Conclui-se que a estratégia proposta é eficaz para a predição de coautorias, sendo sugerido o uso de seletores de atributos e estratégias específicas para lidar com o desbalanceamento de dados como trabalhos futuros.

## 3.10 Análise das Soluções Atuais

A seleção de orientadores em programas de pós-graduação *stricto sensu* tradicionalmente baseia-se na análise manual de currículos acadêmicos, recomendações informais e busca ativa em bases curriculares, científicas e *websites*

institucionais, mas essas abordagens apresentam limitações, como subjetividade na escolha, ausência de critérios padronizados e dificuldade de comparação sistemática entre orientadores [17, 87]. A modelagem matemática surge como alternativa promissora por permitir representações formais do problema e extração de padrões a partir de dados disponíveis [33, 72], sendo amplamente utilizada para análise de grandes volumes de dados e suporte à decisão baseada em métricas quantitativas, o que a torna essencial para sistemas de recomendação acadêmica [21, 70]. Entre os modelos aplicados à recomendação de orientadores acadêmicos, destacam-se quatro categorias consolidadas na literatura: modelos estatísticos, modelos baseados em grafos, sistemas de recomendação com aprendizado de máquina e modelos híbridos.

Modelos Estatísticos: Utilizam séries históricas de dados acadêmicos para identificar padrões de produtividade e impacto. Métricas como Índice H, Índice i10, número total de citações e quantidade de publicações em periódicos de alto impacto são aplicadas para compor indicadores de relevância [19]. Modelos Baseados em Grafos: Representam relações acadêmicas entre orientadores e estudantes, além de redes de colaboração científica. A análise dessas redes permite identificar pesquisadores com forte influência acadêmica e capacidade de colaboração interdisciplinar [73], [69]. Sistemas de Recomendação Baseados em Aprendizado de Máquina: Utilizam filtragem colaborativa e análise de similaridade para sugerir orientadores compatíveis com os interesses dos estudantes. Esses sistemas aplicam modelos probabilísticos e técnicas de otimização para melhorar a precisão das recomendações [21], [67]. Modelos Híbridos: Combinam técnicas estatísticas, estruturas em grafos e algoritmos de aprendizado de máquina para fortalecer a robustez e a adaptabilidade do sistema de recomendação. Essa abordagem integrada possibilita, por exemplo, unir métricas bibliométricas com análise de redes de coautoria e Processamento de Linguagem Natural (PLN), ampliando a capacidade de interpretação do perfil acadêmico do orientador. Segundo Burke [21], sistemas híbridos apresentam vantagens significativas ao mitigar limitações de abordagens isoladas, melhorando tanto a precisão quanto a cobertura das recomendações.

Outras abordagens emergentes são os modelos baseados em ontologias, que estruturam semanticamente as áreas de conhecimento, e mecanismos de *feedback* implícito, que analisam o comportamento do usuário para refinar sugestões [101]. Essas técnicas, embora mais complexas, têm sido aplicadas em contextos de personalização avançada e adaptação dinâmica das recomendações, especialmente em ambientes acadêmicos digitais.

Apesar do avanço no uso da modelagem matemática na análise acadêmica, ainda há desafios ligados à disponibilidade de dados confiáveis e estruturados. A Plataforma Lattes consolidou-se como a principal base nacional de

informações acadêmicas estruturadas, amplamente utilizada em análise bibliométrica. Estudos recentes demonstram sua aplicabilidade no desenvolvimento de indicadores de Inteligência Acadêmica para monitoramento institucional e apoio à tomada de decisão acadêmica [30]. Outro desafio envolve a definição dos pesos adequados para as métricas utilizadas. Diferentes programas de pós-graduação podem priorizar aspectos distintos, como produção científica, experiência em orientação ou captação de financiamento. Modelos adaptativos que permitam ajuste dinâmico de pesos são uma solução viável para tornar o processo mais flexível e personalizado [37], [86].

A análise das soluções atuais evidencia que a modelagem matemática pode aprimorar significativamente a seleção de orientadores, reduzindo subjetividade e aumentando a precisão da escolha. No entanto, desafios como a qualidade dos dados, a calibragem dos modelos e a aceitação acadêmica precisam ser superados para garantir a eficácia dessas soluções. Esta tese aborda parcialmente essas questões ao propor a modelagem baseada na extração automatizada e padronizada de dados e na ponderação de critérios acadêmicos. Embora não integre técnicas avançadas como aprendizado de máquina ou modelos híbridos, a estrutura matemática e computacional desenvolvida representa um passo inicial importante para recomendações no contexto da pós-graduação [18, 20].

## 3.11 Discussão

Diante das limitações identificadas nas abordagens tradicionais para a seleção de orientadores em programas de pós-graduação, a implementação de um modelo baseado em extração automatizada de dados e análise de métricas acadêmicas surge como uma alternativa promissora para otimizar esse processo. A automação permite reduzir a subjetividade inerente às avaliações manuais e possibilita a análise eficiente e sistemática de grandes volumes de dados [87].

A modelagem matemática aplicada nesse contexto proporciona uma base estruturada e objetiva para a tomada de decisão, eliminando a necessidade de julgamentos puramente subjetivos na escolha do orientador. O desenvolvimento de um IR permite integrar diversas métricas quantitativas para avaliar e comparar diferentes orientadores com base em suas características acadêmicas e histórico profissional [73]. A construção desse índice envolve a análise de múltiplas dimensões, como experiência em orientação, produção científica, impacto acadêmico e compatibilidade temática com o estudante.

Uma das principais vantagens da abordagem baseada em modelagem matemática é a capacidade de representar relações acadêmicas de forma quantitativa, permitindo a criação de métricas que avaliam diferentes aspectos do

perfil do orientador. A análise de similaridade entre áreas de pesquisa, por exemplo, pode indicar se o orientador possui histórico de publicações e projetos compatíveis com os interesses do estudante, garantindo um melhor alinhamento acadêmico [17]. Além disso, a reputação do orientador pode ser estimada a partir de sua inserção na comunidade científica, utilizando dados de coautoria, participação em bancas examinadoras e impacto de suas publicações.

A aplicação de métodos estatísticos e inteligência artificial na análise desses dados possibilita uma normalização eficiente das informações, permitindo comparações justas entre diferentes candidatos. Técnicas de aprendizado de máquina podem ser empregadas para identificar padrões na trajetória acadêmica dos orientadores e prever a probabilidade de sucesso na orientação de novos estudantes com base em dados históricos [33]. Isso torna o processo de recomendação mais dinâmico e adaptável, possibilitando que novos dados sejam incorporados ao modelo conforme a produção acadêmica dos pesquisadores evolui.

Outro ponto fundamental na discussão sobre essa abordagem é a aceitação do modelo por parte da comunidade acadêmica. Embora sistemas baseados em modelagem matemática possam oferecer maior eficiência e transparência, ainda há desafios relacionados à confiabilidade dos dados extraídos e à necessidade de validação contínua dos critérios adotados. A recomendação de orientadores baseada exclusivamente em algoritmos pode gerar receio entre pesquisadores e estudantes, que podem considerar a decisão automatizada excessivamente mecânica ou desconectada de fatores qualitativos relevantes. Para mitigar essas preocupações, o modelo deve ser transparente, explicável e permitir ajustes personalizados conforme as necessidades de cada programa de pós-graduação [21].

Dessa forma, a discussão sobre as soluções existentes e os avanços tecnológicos disponíveis reforça a viabilidade de um modelo de recomendação baseado em dados, capaz de otimizar o processo de seleção de orientadores, reduzindo esforços manuais e aumentando a qualidade das decisões acadêmicas. A incorporação de modelagem matemática e análise quantitativa ao processo de recomendação representa um avanço significativo em relação aos métodos tradicionais, oferecendo uma forma mais objetiva e eficiente de apoiar estudantes na escolha de seus orientadores [73].

## 3.12 Resumo do Capítulo

Este capítulo analisou as principais soluções para a seleção de orientadores na pós-graduação *stricto sensu*. Foram discutidas as abordagens tradicionais,

como análise manual de currículos e consultas informais, bem como metodologias de modelagem matemática, incluindo modelos estatísticos, baseados em grafos, aprendizado de máquina e híbridos, além de alternativas emergentes como ontologias e *feedback* implícito. A análise evidenciou que a modelagem matemática fornece uma base objetiva para integrar múltiplas métricas acadêmicas em um IR, permitindo avaliar produção científica, impacto, experiência e compatibilidade temática. A proposta de um modelo baseado em extração automatizada de dados e técnicas quantitativas, incluindo aprendizado de máquina, permite normalização, análise dinâmica e redução de subjetividade, tornando o processo de seleção de orientadores mais eficiente, transparente e adaptável, apesar de desafios quanto à qualidade dos dados, definição de pesos e aceitação acadêmica.

# *Parte III*

# *Proposta*

# Capítulo 4

# Materiais e Métodos

*Saber não é suficiente; é preciso aplicar.*
*Querer não é suficiente; é preciso fazer.*

*Johann Wolfgang von Goethe, Romancista alemão (1749–1832)*

Neste capítulo apresentam-se as etapas metodológicas empregadas na coleta, extração, processamento e análise de dados provenientes da Plataforma Lattes e do Google Scholar, com o objetivo de subsidiar a construção do modelo de recomendação de orientadores. Na Seção 4.1, descreve-se a seleção das instituições participantes do Consórcio das Universidades Comunitárias Gaúchas (COMUNG) e o processo manual de obtenção de aproximadamente 1800 currículos Lattes de docentes vinculados a programas de pós-graduação, além da análise das atividades acadêmicas realizadas entre 2010 e 2020. A Seção 4.2 aborda o processo de limpeza e pré-processamento dos dados extraídos dos arquivos XML, estruturando as variáveis de interesse por meio de *scripts* desenvolvidos em Python, desde a identificação das variáveis até o cálculo automatizado das métricas. Complementarmente, a Seção 4.3 trata da utilização da biblioteca *Scholarly* para a extração de dados bibliométricos diretamente do Google Scholar, superando as limitações da Plataforma Lattes quanto à disponibilidade dessas informações. Essas etapas metodológicas constituem a base empírica para o cálculo das pontuações dos indicadores utilizados no Índice de Recomendação (IR), apresentado no Capítulo 5, assegurando a reprodutibilidade e a solidez da metodologia aplicada nesta pesquisa.

## 4.1 Coleta dos Dados da Plataforma Lattes

Nesta pesquisa, foram utilizados dados das instituições participantes do COMUNG. Este é um consórcio de universidades, considerado o maior sis-

tema de educação superior do Rio Grande do Sul, composto por 14 Instituições de Ensino Superior (IES). Essas instituições abrangem quase todos os municípios do estado e os números do COMUNG destacam mais de 153 mil estudantes, 63 mil beneficiados com bolsas e financiamentos, 7 mil professores e quase 9 mil funcionários. A rede oferece 896 cursos de graduação, 95 doutorados e 140 mestrados, além de contar com mais de 3.500 laboratórios de apoio ao ensino e pesquisa, 8 parques tecnológicos, 13 incubadoras de empresas, 615 empresas incubadas, 11 agências de inovação e tecnologia e 1.171 convênios internacionais.

As instituições integrantes do COMUNG incluem as universidades destacadas na Tabela 4.1, além da Universidade da Região da Campanha (URCAMP), que oferece cursos de pós-graduação *stricto sensu* somente em parceria com outras instituições. Os dados apresentados na Tabela 4.1 foram obtidos através da extração manual das informações nos sites dos programas, bem como da análise também manual dos currículos Lattes dos professores das instituições, entre os anos de 2021 e 2022[89]. A partir desta análise, com base na lista de docentes vinculados aos PPGs, procedeu-se à consulta na Plataforma Lattes para obtenção dos respectivos currículos. Este processo foi executado manualmente, devido à ausência de uma integração automatizada. Para cada professor, realizou-se uma busca nominal na Plataforma e os currículos foram baixados no formato XML.

Foram obtidos aproximadamente 1800 currículos da Plataforma Lattes de professores vinculados aos PPGs, constituindo a base para a análise estatística, o desenvolvimento das equações e a avaliação do modelo de recomendação proposto. Ressalta-se que os dados obtidos refletem a situação dos currículos no momento da coleta e podem ter sofrido alterações até o ano de 2025, não impactando, contudo, a metodologia adotada nem os resultados obtidos nesta pesquisa.

Conforme a Tabela 4.1, a amostra abrange instituições de portes muito diversos, assegurando a representatividade dos dados utilizados neste estudo. No topo do espectro, a Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS) reúne 22 programas, 402 docentes e 151 linhas de pesquisa, seguida pela Universidade do Vale do Rio dos Sinos (UNISINOS) (26 programas, 301 docentes) e pela Universidade de Caxias do Sul (UCS) (19 programas, 276 docentes e 73 linhas). No extremo oposto situam-se a Universidade de Cruz Alta (UNICRUZ) (3 programas, 49 docentes, 7 linhas) e a Universidade Católica de Pelotas (UCPel) (4 programas, 50 docentes, 22 linhas). Entre esses polos aparecem universidades de perfil intermediário: Universidade Feevale (Feevale) (11 mestrados, 5 doutorados; 129 docentes; 23 linhas),

Universidade La Salle (LaSalle) (5 mestrados, 3 doutorados; 56 docentes; 13 linhas), Universidade Franciscana (UFN) (5 mestrados, 2 doutorados; 82 docentes; 9 linhas), Universidade Regional do Noroeste do Estado do Rio Grande do Sul (UNIJUÍ) (7 mestrados, 4 doutorados; 88 docentes; 17 linhas), Universidade do Vale do Taquari (UNIVATES) (6 mestrados, 4 doutorados; 95 docentes; 20 linhas), Universidade de Santa Cruz do Sul (UNISC) (9 mestrados, 6 doutorados; 126 docentes; 27 linhas), Universidade de Passo Fundo (UPF) (15 mestrados, 9 doutorados; 193 docentes; 38 linhas) e Universidade Regional Integrada do Alto Uruguai e das Missões (URI) (6 mestrados, 3 doutorados; 56 docentes; 14 linhas). Essa variação de três a vinte e seis programas, de cerca de cinquenta a mais de quatrocentos docentes e de sete a cento e cinquenta e uma linhas de pesquisa, demonstra que o conjunto analisado é estatisticamente expressivo e heterogêneo o suficiente para sustentar, as conclusões do trabalho proposto.

| Instituição | PPGs | Professores | Mestrado | Doutorado | LP |
|---|---|---|---|---|---|
| Feevale | 11 | 129 | 11 | 5 | 23 |
| LaSalle | 5 | 56 | 5 | 3 | 13 |
| PUCRS | 22 | 402 | 22 | 22 | 151 |
| UCS | 19 | 276 | 18 | 11 | 73 |
| UCPel | 4 | 50 | 4 | 4 | 22 |
| UFN | 5 | 82 | 5 | 2 | 9 |
| UNICRUZ | 3 | 49 | 3 | 1 | 7 |
| UNIJUI | 7 | 88 | 7 | 4 | 17 |
| UNISC | 9 | 126 | 9 | 6 | 27 |
| UNISINOS | 26 | 301 | 26 | 21 | 71 |
| UNIVATES | 6 | 95 | 6 | 4 | 20 |
| UPF | 15 | 193 | 15 | 9 | 38 |
| URI | 6 | 56 | 6 | 3 | 14 |
| **Total** | **138** | **1903** | **137** | **95** | **485** |

**Tabela 4.1**: *Dados dos Programas de Pós-Graduação.*

## Atividades Acadêmicas

Um script foi desenvolvido para extrair e analisar dados de atividades acadêmicas registradas em diferentes instituições de ensino. As atividades analisadas incluem produção bibliográfica, orientações, participação em eventos,

bancas examinadoras e projetos de pesquisa. Esses dados serviram de base para a geração dos gráficos apresentados na Figuras 4.1 e nos apêndices D, que ilustram a evolução das atividades acadêmicas ao longo do tempo. Cada gráfico mostra o número total de atividades por ano, o total de cursos e de professores envolvidos, permitindo a análise comparativa entre as instituições.

As atividades acadêmicas são fundamentais para a avaliação curricular de um professor, representando indicadores relevantes de desempenho e produtividade. A produção bibliográfica, amplamente reconhecida como medida essencial de contribuição científica, é avaliada por métricas como número de publicações e citações, refletindo o impacto e a reputação do pesquisador [17, 47]. O envolvimento na formação de novos pesquisadores, evidenciado pelas orientações concluídas e em andamento, também é valorizado, pois promove a transmissão de conhecimento e o desenvolvimento do capital humano científico [80]. Participações em eventos, congressos e bancas examinadoras indicam reconhecimento e inserção na comunidade acadêmica, além de fortalecer redes de colaboração [75]. A liderança em projetos de pesquisa, por sua vez, é um indicador importante da capacidade de conduzir investigações inovadoras e de obter financiamento [19, 58]. Assim, os dados coletados pelo script não apenas oferecem uma base quantitativa para análise institucional, mas também fornecem subsídios para avaliar a contribuição acadêmica dos docentes, orientando estratégias institucionais voltadas à excelência acadêmica.

O gráfico da Figura 4.1 é um exemplo que, destaca variações no número de atividades acadêmicas ao longo dos anos para a universidade Feevale, permitindo identificar períodos de maior ou menor engajamento. O script desenvolvido facilitou a extração e análise dos dados, permitindo visualizar tendências e comparações institucionais, contribuindo para a gestão acadêmica e o planejamento estratégico.

## Aspectos Éticos

Os dados utilizados são oriundos de currículos de acesso público na Plataforma Lattes. Esses dados em XML, foram tratados com finalidade exclusivamente acadêmica, observando os princípios de minimização, necessidade e transparência. Não foram tratados dados sensíveis nem produzidas divulgações individualizadas; sempre que necessário, aplicaram-se procedimentos de anonimização ou pseudonimização para preservar a privacidade dos titulares. O processamento foi conduzido em conformidade com a Lei nº 13.709/2018 LGPD com base na hipótese de tratamento para fins acadêmicos e de pesquisa. Quando houver coleta primária de informações, interação direta com

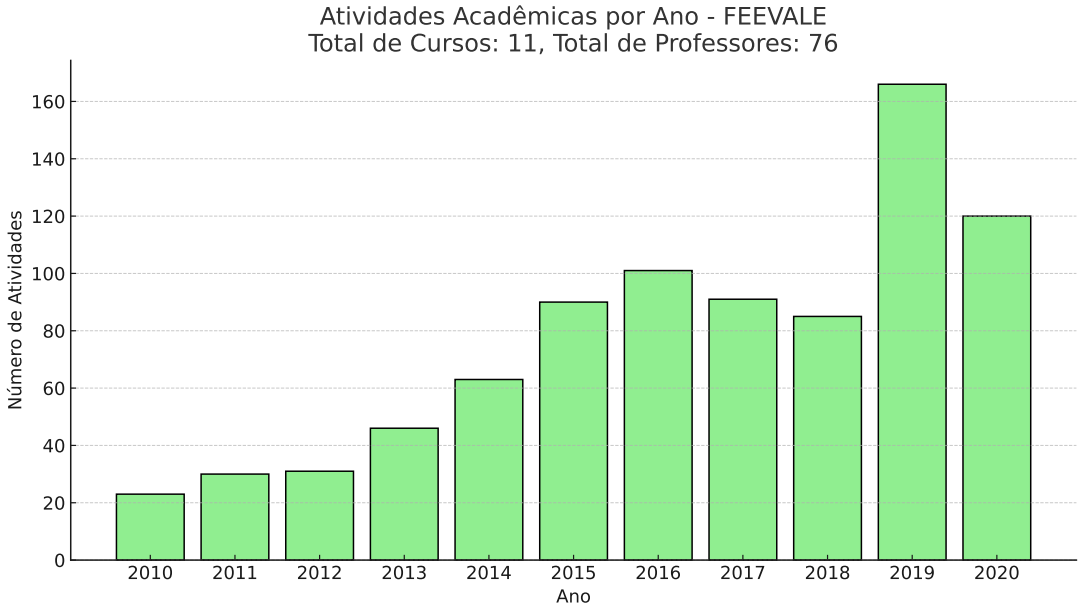



**Figura 4.1**: *Feevale - Atividades Acadêmicas.*

participantes ou integração com bases não públicas que possibilitem reidentificação, recomenda-se submissão prévia aos Sistemas (CEP - Comitê de Ética em Pesquisa; CONEP - Comissão Nacional de Ética em Pesquisa), conforme regulamentação aplicável às pesquisas.

## 4.2 Preparação e Tratamento dos Dados XML

Na etapa de limpeza e pré-processamento dos dados, foram extraídas somente as informações necessárias dos arquivos XML obtidos do Lattes, na etapa anterior, acessados por meio de um repositório no *GitHub*.

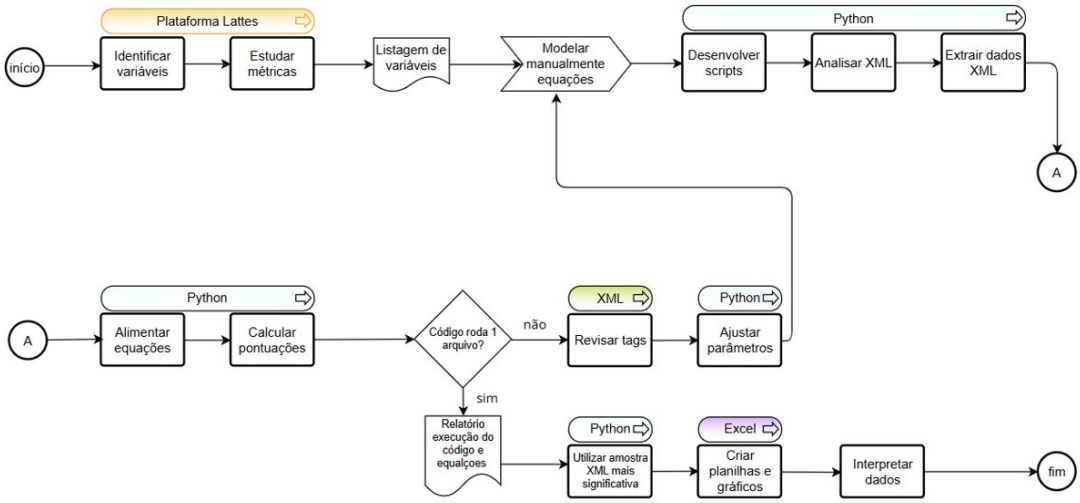



**Figura 4.2**: *Análise e Processamento de Dados.*

No diagrama da Figura 4.2, ilustra-se um fluxo de trabalho detalhado para a análise de dados extraídos de currículos Lattes em formato XML, utilizando *scripts* desenvolvidos em Python. Este processo é estruturado em etapas que vão desde a identificação inicial das variáveis até a interpretação final dos resultados, garantindo que cada fase contribua para a precisão e eficácia da análise.

A primeira etapa do fluxo de trabalho envolve a identificação das variáveis relevantes na Plataforma Lattes. Neste momento, o objetivo é reconhecer quais informações são relevantes para o estudo que será realizado. Essas variáveis incluíram publicações científicas, orientações, participações em bancas, envolvimento em projetos de pesquisa, métricas bibliométricas, citações, além de aspectos relacionados à experiência docente e à rede de colaborações acadêmicas. Em seguida, o foco foi direcionado para o estudo das métricas associadas a essas variáveis. Nesta fase, foi fundamental compreender como cada variável poderia ser medida e como essas medições se relacionam com os objetivos

do estudo. Por exemplo, a métrica pode ser o número de publicações em determinado período ou a quantidade de orientações finalizadas, o que irá influenciar a construção das equações de análise. Após o estudo das métricas, foi realizada a listagem das variáveis, onde todas as variáveis identificadas foram organizadas de forma estruturada. Essa listagem, Tabela 1.1, serviu como um guia para o desenvolvimento dos *scripts*, garantindo que possíveis fossem consideradas na etapa de extração de dados. Com as variáveis listadas, o próximo passo foi desenvolver *scripts* em Python. Esses *scripts* foram criados para automatizar a leitura e o processamento dos arquivos XML dos currículos Lattes. A programação em Python permitiu a criação de ferramentas customizadas que são capazes de manipular os dados extraídos de forma eficiente, aplicando as métricas e realizando os cálculos necessários.

Uma vez desenvolvidos os *scripts*, procedeu-se à análise dos arquivos XML. Nesta etapa, os *scripts* foram utilizados para processar os currículos, extrair as informações necessárias e preparar os dados para a próxima fase. Esta foi uma fase crucial, pois a precisão na extração dos dados impacta diretamente na qualidade das análises subsequentes. A extração dos dados foi então realizada, informações coletadas dos XMLs foram preparadas para alimentar as equações definidas nas etapas anteriores. Esses dados extraídos formam a base para os cálculos e análises que se seguiram.

Na continuação do processo, os dados extraídos foram utilizados para alimentar as equações previamente elaboradas a partir do estudo da listagem de variáveis, que foram responsáveis por calcular as pontuações ou outras métricas de interesse. Esta etapa foi executada por *scripts* Python, que aplicam as equações a cada conjunto de dados extraídos. Em seguida, as pontuações foram calculadas. Esta etapa envolveu a aplicação das equações aos dados alimentados, resultando nas métricas finais que foram analisadas. Essas pontuações podem representar, por exemplo, a produção científica de um pesquisador, sua participação em orientações, entre outros indicadores acadêmicos.

Uma etapa importante no fluxo de trabalho foi a verificação de que o código funciona corretamente para um arquivo XML. Se o código não funcionar adequadamente, é necessário realizar ajustes. Primeiramente, revisa-se o caminho das tags XML para garantir que os *scripts* estão corretamente configurados para ler os dados. Se as tags XML estiverem corretas, mas ainda houver problemas, procede-se ao ajuste dos parâmetros nos *scripts*, para refinar a execução do código. Se o código funcionar corretamente, o próximo passo é a geração de um relatório detalhado que documenta a execução dos códigos e as equações aplicadas. Este relatório foi essencial para garantir a transparência do processo e para facilitar a revisão e interpretação dos resultados obtidos.

Com o código validado, uma amostra maior de XMLs foi utilizada para testar o código em um volume maior de dados. Esta etapa foi importante para verificar a escalabilidade dos *scripts* e a consistência dos resultados em uma base de dados ampliada. Após processar a amostra maior, os resultados foram utilizados para construir planilhas e gráficos no Excel. Essas visualizações foram fundamentais para interpretar os dados de forma clara e acessível, facilitando a comunicação dos resultados e das conclusões.

Finalmente, os dados foram interpretados com base nas planilhas e gráficos gerados, levando à etapa final do fluxo de trabalho. A interpretação dos resultados permitiu a extração de *insights* e a tomada de decisões baseadas nos dados analisados.

Em resumo, o diagrama descreve um fluxo de trabalho estruturado que envolve a identificação e estudo de variáveis e métricas, desenvolvimento e teste de *scripts* Python, extração e análise de dados XML, e a geração de relatórios e gráficos. Cada etapa foi cuidadosamente planejada para garantir a precisão e a eficácia do processo, resultando em uma análise robusta e detalhada dos currículos Lattes. Este fluxo de trabalho pode ser amplamente aplicável em contextos acadêmicos e de pesquisa, onde a análise de grandes volumes de dados é necessária para a avaliação e comparação de métricas de desempenho.

## 4.3 Complementação de Dados Bibliométricos

Um dos desafios encontrados durante a realização deste estudo foi a obtenção de dados específicos exigidos para o cálculo da pontuação $P_{\text{Produção}}$, que considera variáveis como o número de citações em publicações, índice $h$ e índice $i10$. Embora a Plataforma Lattes seja amplamente utilizada e reconhecida no contexto acadêmico brasileiro, constatou-se que ela não disponibiliza esses dados de forma direta e acessível. Diante dessa limitação, tornou-se necessário explorar outras fontes que oferecessem essas informações de maneira mais estruturada. Entre as alternativas investigadas, o módulo *Scholarly*, desenvolvido em *Python*, mostrou-se particularmente promissor. Este módulo acessa diretamente a base de dados do Google Scholar, que disponibiliza métricas bibliométricas detalhadas, incluindo citações por publicação, índices $h$ e $i10$, além de outros indicadores relevantes. Sua flexibilidade permite realizar buscas utilizando o nome do autor, seu identificador único na plataforma ou informações específicas de trabalhos publicados, como títulos ou identificadores.

Essa abordagem alternativa foi adotada devido à sua capacidade de complementar os dados fornecidos pela Plataforma Lattes, permitindo a construção de

uma base de informações mais completa e precisa para atender aos objetivos do estudo. O uso do Scholarly não apenas preencheu a lacuna identificada, mas também viabilizou uma integração mais robusta entre diferentes fontes de dados, contribuindo para a qualidade e confiabilidade dos cálculos realizados neste trabalho. Essa estratégia destaca a importância de recorrer a soluções tecnológicas que garantam a obtenção de dados relevantes quando os recursos disponíveis não atendem integralmente às necessidades da pesquisa. A integração entre plataformas, como a Lattes e o Google Scholar via Scholarly, reforça a capacidade de explorar ao máximo as informações disponíveis, promovendo análises mais detalhadas e fundamentadas no contexto acadêmico.

Para validar o uso do *Scholarly*, foi implementado um *script* em Python que realiza consultas ao Google Scholar e armazena os resultados em formato JSON. A metodologia consistiu na busca por um autor acadêmico utilizando seu nome como parâmetro, seguido pela extração de detalhes adicionais, como filiação acadêmica, interesses de pesquisa, métricas bibliométricas, coautores e publicações. Os dados foram salvos em um arquivo denominado `data.json`, garantindo sua integridade e facilitando análises posteriores. A execução do script com a biblioteca *Scholarly* confirmou sua relevância para o estudo, permitindo a extração de métricas bibliométricas como índice $h$, índice $i_{10}$, total de citações, além de informações detalhadas sobre publicações, como títulos, anos e número de citações. Entre os resultados, destacam-se artigos publicados em 2012 com 86 citações e em 2016 com 66 citações, evidenciando o potencial da ferramenta em fornecer dados organizados e relevantes para a análise acadêmica.

Os resultados demonstram que o *Scholarly* é uma solução para extração de dados acadêmicos, atendendo às necessidades específicas deste estudo. No entanto, os testes realizados até o momento foram conduzidos com um volume limitado de requisições, o que aponta para a necessidade de validação em larga escala. Ampliar os testes com maior carga de dados e estruturar as informações coletadas em um banco de dados para permitir consultas eficientes, além de monitorar possíveis limitações de acesso relacionadas à taxa de requisições ao Google Scholar.

## 4.4 Resumo do Capítulo

Neste capítulo foi apresentado o percurso metodológico adotado para viabilizar o modelo de recomendação de orientadores. Inicialmente, realizou-se o levantamento de currículos Lattes de docentes das instituições vinculadas ao COMUNG, acompanhado da caracterização quantitativa de seus programas de pós-graduação e da análise das principais atividades acadêmicas dos pesquisadores. Em seguida, foi desenvolvido um fluxo de limpeza e pré-processamento

dos arquivos XML, apoiado por *scripts* em Python, que permitiu estruturar as variáveis de interesse e automatizar o cálculo das métricas. Por fim, integrou-se ao modelo um conjunto de indicadores bibliométricos obtidos no Google Scholar por meio da biblioteca *Scholarly*, complementando as informações extraídas da Plataforma Lattes e ampliando a base de dados utilizada na avaliação.

# Capítulo 5

# Modelo Matemático

A avaliação de orientadores acadêmicos requer a análise de múltiplos aspectos de suas atividades acadêmicas e científicas. Esses aspectos podem ser organizados em critérios que representam diferentes dimensões de desempenho, como experiência em orientação, produção científica, eficiência das orientações realizadas, colaboração na comunidade acadêmica, pesquisa e alinhamento com a área de pesquisa do aluno. Esses critérios são combinados para calcular um IR, que reflete a capacidade e a adequação do orientador para atender às necessidades acadêmicas e profissionais de um aluno em potencial.

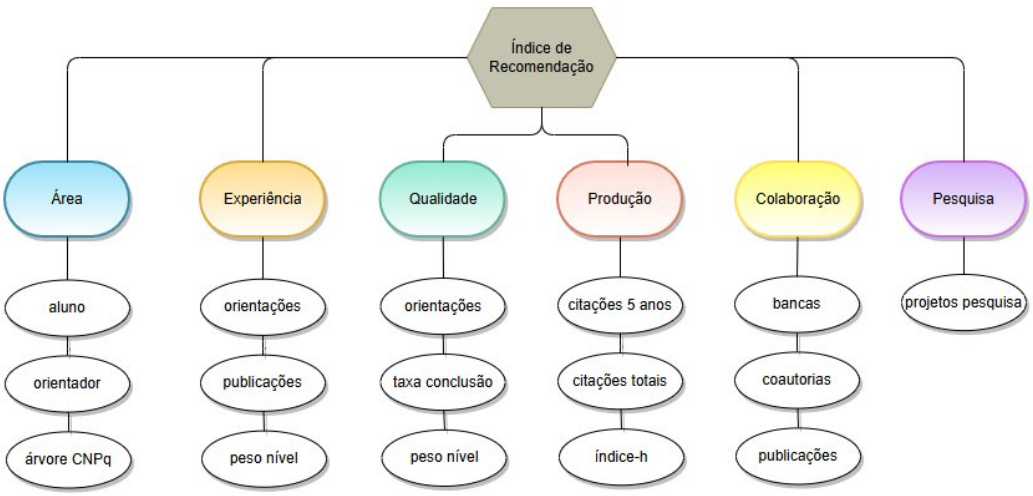



**Figura 5.1**: *Descrição do Índice de Recomendação.*

A Figura 5.1 ilustra o modelo geral adotado para o cálculo do IR. Cada critério principal, como produção científica ou colaboração, é subdividido em

indicadores específicos, cujos valores são obtidos a partir de dados objetivos extraídos de currículos acadêmicos, bases de dados bibliométricas e registros institucionais. A integração dessas informações permite construir um modelo abrangente e estruturado para avaliar e comparar orientadores. O IR é calculado por meio de uma equação que pondera os diferentes critérios com base em sua relevância, considerando o perfil do orientador e as demandas do aluno.

A **área** de pesquisa de um orientador acadêmico diz respeito ao campo específico em que ele atua, considerando sua formação em graduação, mestrado, doutorado, pós-doutorado e outras formações complementares. Essa área leva em conta a grande área, área, subárea e especialidade de acordo com a Tabela de Áreas do Conhecimento do CNPq . Quando um orientador possui uma pontuação alta nesse critério, significa que sua formação e experiência são similares à área de interesse do aluno. A **experiência** em orientação considera o número de orientações concluídas e em andamento realizadas pelo orientador, ponderadas por pesos atribuídos a cada nível acadêmico, além de incorporar um fator de produção científica relacionado à quantidade de artigos publicados. Uma pontuação alta indica que o orientador possui ampla atuação na formação de pesquisadores, associada à sua produtividade acadêmica. A **eficiência** das orientações é medida pelo número de orientações concluídas, ponderadas por nível acadêmico, assumindo que a conclusão bem-sucedida dos trabalhos (taxa de conclusão) reflete a eficácia na formação acadêmica.

A **produção** científica é avaliada pelo número de citações recebidas, pelo índice h e pelo índice i10 do orientador, refletindo sua produtividade e o impacto de seus trabalhos na comunidade acadêmica. Uma pontuação elevada indica contribuição significativa para o avanço do conhecimento em sua área de pesquisa. A **colaboração** na comunidade acadêmica é avaliada pelo número de coautorias em publicações científicas e pela participação em bancas examinadoras de mestrado e doutorado, refletindo o envolvimento do orientador em redes acadêmicas e sua integração na comunidade científica. A **pesquisa** é avaliado pelo número de projetos de pesquisa e atividades de desenvolvimento em que o orientador está envolvido, refletindo sua participação ativa na produção científica e no fortalecimento da pesquisa acadêmica.

## 5.1 Equação Geral

$$IR = \sum_{i=1}^{6} \alpha_i \tilde{P}_i, \qquad \alpha_i \geq 0, \ \sum_{i=1}^{6} \alpha_i = 1, \ \tilde{P}_i \in [0, 1]. \tag{5.1}$$

onde $IR$ representa o Índice de Recomendação, $\alpha_i$ são os pesos atribuídos a cada parâmetro, e $P_i$ correspondem às seguintes pontuações: $P_1 = P_{\text{Área}}$ (área de pesquisa de orientador-orientando), $P_2 = P_{\text{Experiência}}$ (experiência em orientação acadêmica), $P_3 = P_{\text{Eficiência}}$ (eficiência das orientações anteriores), $P_4 = P_{\text{Produção}}$ (produção científica do orientador), $P_5 = P_{\text{Colaboração}}$ (colaboração na comunidade acadêmica) e $P_6 = P_{\text{Pesquisa}}$ (projetos de pesquisa em que o orientador atua).

Para calcular cada $P$ (pontuação) para os parâmetros individuais, podemos usar diferentes métricas ou escalas de avaliação, dependendo da disponibilidade de dados e da preferência do sistema. Por exemplo, podemos usar uma escala de 0 a 10, onde 10 indica a pontuação mais alta e 0 a mais baixa. Ou poderíamos usar uma abordagem, como a normalização de dados para colocar todas as pontuações em uma escala comparável. A seguir destacamos as equações que descrevem os cálculos utilizados para cada parâmetro acima descrito.

### 5.1.1 Pontuação da Similaridade de Área

A pontuação da similaridade de área quantifica o grau de correspondência entre as áreas de pesquisa do orientador e os interesses do aluno. O cálculo considera os diferentes níveis de classificação do conhecimento definidos pelo CNPq (grande área, área, subárea e especialidade). Inicialmente, identificam-se as classificações de ambos (orientador e aluno) em cada nível. Em seguida, avalia-se a similaridade entre os códigos CNPq, atribuindo valores que variam desde a ausência de correspondência até a equivalência exata.

A pontuação para cada par de áreas de conhecimento é então determinada com base em quantos desses níveis coincidem. Por exemplo, se a grande área coincide, recebe uma pontuação inicial. Se além da grande área, a área também coincidir, a pontuação aumenta. E assim por diante, com cada nível adicional de coincidência aumentando a pontuação total. Finalmente, somamos todas as pontuações das correspondências entre as áreas de conhecimento do orientador e do aluno. Essa soma resulta na pontuação total de similaridade de área, que indica o grau de compatibilidade entre as áreas de formação do orientador e as áreas de interesse do aluno. Uma pontuação alta sugere que o orientador possui uma formação e experiência altamente relevantes para a área de interesse do aluno, possibilitando o aumento da probabilidade de uma orientação bem-sucedida.

A pontuação total de similaridade de área $P_{\text{Área}}$ é a soma das pontuações atribuídas a cada correspondência nos diferentes níveis hierárquicos.

$$P_{\text{Área}} = \sum_{i=1}^{n} (\text{GA}_i + \text{A}_i + \text{SA}_i + \text{E}_i) \tag{5.2}$$

onde $\text{GA}_i$ representa a pontuação atribuída à Grande Área, $\text{A}_i$ à Área, $\text{SA}_i$ à Sub-Área e $\text{E}_i$ à Especialidade; $i$ representa o nível acadêmico (graduação, mestrado, doutorado); e $n$ corresponde ao total de níveis considerados.

### 5.1.2 Pontuação da Experiência:

A pontuação da experiência em orientação é calculada para avaliar a relevância e a produtividade de um orientador em relação às suas orientações acadêmicas e publicações científicas. Essa pontuação leva em consideração diferentes componentes, incluindo orientações de mestrado e doutorado, cada um ponderado pela sua relevância específica. A pontuação final é ajustada por um fator de qualidade para refletir a qualidade das publicações do orientador.

Para começar, conta-se o número de orientações de mestrado e doutorado, concluídas e em andamento. Cada um desses componentes recebe um peso específico. Além disso, consideramos limites superiores realistas para cada nível de orientação. A pontuação da cada componente é calculada somando as contribuições ponderadas de cada nível de orientação. Multiplica-se o número de orientações de cada nível pelo respectivo peso e dividimos pelo limite superior correspondente. Essas pontuações são então somadas para obter a pontuação intermediária.

O próximo passo é ajustar essa pontuação intermediária pelo fator de qualidade. O fator de qualidade é calculado com base no número de publicações do orientador. É determinado dividindo o número total de artigos publicados em revistas pelo valor máximo esperado de publicações, que é estabelecido como o percentil do número de publicações entre os orientadores analisados. Finalmente, multiplica-se a pontuação intermediária pelo fator de qualidade para obter a pontuação final do indicador experiência. Essa pontuação ajustada indica a relevância e a produtividade do orientador em relação às suas orientações acadêmicas e publicações científicas. Uma pontuação alta de experiência sugere que o orientador possui uma combinação significativa de orientações e publicações, refletindo sua capacidade de orientação e produção acadêmica.

A equação utilizada para calcular a pontuação do indicador $P_{\text{Experiência}}$ considera diferentes componentes ponderados conforme sua relevância. O resultado final é ajustado pelo fator de qualidade $Q$, refletindo a produtividade acadêmica.

$$P_{\text{Experiência}} = \left(\lambda_m \frac{m}{M} + \lambda_d \frac{d}{D}\right) Q, \qquad \lambda_m, \lambda_d \geq 0,\ \lambda_m + \lambda_d = 1. \tag{5.3}$$

onde $m$ representa o número de orientações de mestrado (concluídas e em andamento), $d$ corresponde ao número de orientações de doutorado (concluídas e em andamento), $\lambda_m$ e $\lambda_d$ são os pesos atribuídos, respectivamente, às orientações de mestrado e doutorado, $M$ e $D$ indicam os valores de referência ou limites superiores estabelecidos para cada nível de orientação, e $Q$ corresponde ao fator de qualidade, calculado a partir da normalização do número de publicações do orientador em relação ao conjunto considerado.

## Fator de Qualidade $Q$

$$Q = \begin{cases} \dfrac{P_r}{P_{90}}, & P_{90} > 0, \\ 1, & P_{90} = 0, \end{cases} \tag{5.4}$$

onde $P_r$ representa o número total de artigos publicados pelo orientador, e $P_{90}$ corresponde ao valor de referência adotado, definido como o percentil 90 da distribuição de publicações entre os orientadores considerados.

## Etapas para o Cálculo

i. **Extração dos Dados das Orientações**

- $m$: número de orientações de mestrado concluídas e em andamento.
- $d$: número de orientações de doutorado concluídas e em andamento.

ii. **Determinação de $P_{90}$**

- Coletar o número de publicações $P_r$ de todos os orientadores.
- Calcular o percentil 90 desses valores, obtendo $P_{90}$.

iii. **Cálculo do Fator de Qualidade $Q$**

- Para cada orientador, calcular $Q = \dfrac{P_r}{P_{90}}$, considerando $Q = 1$ quando $P_{90} = 0$.

iv. **Cálculo da pontuação da Experiência**

- Substituir os valores de $m$, $d$, $\lambda_m$, $\lambda_d$, $M$, $D$ e $Q$ na equação (5.3).
- Calcular a soma ponderada $\left(\lambda_m \frac{m}{M} + \lambda_d \frac{d}{D}\right)$ e multiplicar pelo fator de qualidade $Q$.

A equação geral considera o número de orientações em diferentes níveis acadêmicos, aplicando pesos específicos e normalizando os valores conforme limites superiores, enquanto o fator de qualidade $Q$ ajusta a pontuação com base na produtividade científica. Cabe ressaltar que optou-se por não incluir a variável *tempo* desde a titulação de mestre ou doutor, a fim de evitar vieses que poderiam penalizar pesquisadores em início de carreira ou favorecer de modo desproporcional os mais seniores. Ainda assim, reconhece-se que a incorporação de parâmetros de antiguidade ou cadência produtiva pode enriquecer análises futuras.

### 5.1.3 Pontuação da Eficiência:

A pontuação da eficiência em orientações é calculada para mensurar a eficácia do orientador na conclusão de orientações acadêmicas em níveis de mestrado e doutorado. Para isso, utiliza-se a taxa de conclusão, definida como a razão entre orientações concluídas e o total de orientações (concluídas + em andamento), ponderada por pesos específicos atribuídos a cada nível. A fórmula combina, para cada nível, a taxa de conclusão, o número de orientações concluídas e o peso correspondente, e normaliza esse somatório pela soma total de orientações concluídas. O resultado representa uma média ponderada da eficiência do orientador, favorecendo aqueles que conciliam volume e efetividade na orientação discente.

#### 5.1.3.1 Fórmula de Cálculo

A pontuação da eficiência $P_{\text{Eficiência}}$ é calculada usando a seguinte fórmula:

$$P_{\text{Eficiência}} = \frac{\sum_{i=1}^{n} (w_i T_{C_i} OC_i)}{\sum_{i=1}^{n} OC_i} \tag{5.5}$$

onde $P_{\text{Eficiência}}$ representa a pontuação da eficiência; $i$ indica o nível acadêmico (mestrado ou doutorado); $w_i$ é o peso associado ao nível de orientação $i$;

$T_{C_i}$ corresponde à taxa de conclusão para o nível $i$; e $OC_i$ refere-se ao número de orientações concluídas nesse nível.

A taxa de conclusão $T_{C_i}$ representa a eficiência com que as orientações acadêmicas são finalizadas em relação ao total de orientações realizadas (tanto concluídas quanto em andamento) em um determinado nível. Ela é calculada da seguinte forma:

$$T_{C_i} = \frac{OC_i}{OC_i + OA_i} \tag{5.6}$$

onde $T_{C_i}$ representa a taxa de conclusão para o nível $i$; $OC_i$ corresponde ao número de orientações concluídas no nível $i$; e $OA_i$ ao número de orientações em andamento nesse mesmo nível.

**Interpretação:**

- A taxa $T_C$ varia entre 0 e 1.
- $T_C = 1$: todas as orientações no nível $i$ foram concluídas, e nenhuma está em andamento  taxa de conclusão de 100%.
- $T_C = 0$: nenhuma orientação foi concluída, e todas estão em andamento taxa de conclusão de 0%.
- Taxas mais altas de $T_C$ indicam maior eficiência na conclusão de orientações naquele nível, sugerindo que o orientador finaliza as orientações de forma eficaz.

**Papel da $T_C$ na Pontuação da Eficiência:**

Na equação $P_Q$, a taxa $T_C$ é multiplicada pelo número de orientações concluídas e pelo peso do nível, e essa multiplicação contribui para o numerador da pontuação da eficiência. Portanto, uma taxa de conclusão alta em qualquer nível aumenta a pontuação da eficiência, indicando um bom desempenho na conclusão das orientações naquele nível.

Em resumo, $T_C$ é uma métrica fundamental para avaliar o quão bem as orientações são finalizadas e contribui diretamente para a avaliação do sucesso do trabalho de orientação acadêmica.

No numerador, $OC_i$ é multiplicado pela taxa de conclusão $T_{C_i}$ e pelo peso $w_i$:

$$\sum_{i \in \{m,d\}} (w_i T_{C_i} OC_i) \tag{5.7}$$

Aqui, $OC_i$ serve para:

$OC_i$ pondera a taxa de conclusão $T_{C_i}$ de acordo com o número de orientações concluídas. Isso significa que, se houver muitas orientações concluídas em um nível, esse nível terá uma influência maior na pontuação final, desde que a taxa de conclusão também seja alta.

Mesmo que um nível tenha uma alta taxa de conclusão, se o número de orientações concluídas ($OC_i$) for baixo, a sua contribuição para a pontuação global será limitada. Isso evita que uma alta taxa de conclusão em um nível com poucas orientações aumente de forma desproporcional a pontuação final.

No denominador, $OC_i$ está presente na soma:

$$\sum_{i \in \{m,d\}} OC_i \tag{5.8}$$

Aqui, $OC_i$ serve para:

O denominador é a soma total de todas as orientações concluídas em todos os níveis. Dividir o numerador por essa soma normaliza a pontuação da eficiência, transformando-a em uma média ponderada das taxas de conclusão, ajustada pelo número total de orientações concluídas.

Sem o denominador, um nível com um grande número de orientações concluídas poderia dominar a pontuação final, independentemente de quão eficiente (alta $T_{C_i}$) esse nível seja. Ao dividir pelo total de orientações concluídas, garante-se que a pontuação da eficiência seja uma medida justa da eficiência relativa entre os diferentes níveis, independentemente de seu volume.

**Etapas do Cálculo**

i. **Contagem de orientações concluídas e em andamento**: O primeiro passo é identificar quantas orientações de mestrado e doutorado já foram concluídas e quantas estão em andamento. Isso é feito a partir da leitura das tags específicas no arquivo XML, que indicam o estado de cada orientação.

ii. **Cálculo da taxa de conclusão**: Em seguida, calcula-se a taxa de conclusão, que é a proporção de orientações finalizadas em relação ao total de orientações (concluídas e em andamento). Essa taxa permite avaliar a eficiência do professor em completar suas orientações.

iii. **Cálculo da pontuação da eficiência**: A pontuação da eficiência é determinada com base na taxa de conclusão, levando em consideração pesos diferentes para mestrado e doutorado, dependendo da importância atribuída a cada nível. Esse cálculo reflete a eficiência e sucesso do trabalho de orientação do professor.

iv. **Processamento de múltiplos arquivos**: O cálculo é aplicado a vários arquivos XML, que representam os currículos de diferentes professores. Isso permite gerar uma comparação entre a qualidade das orientações de diversos docentes.

v. **Exibição dos resultados**: Por fim, os resultados são apresentados, mostrando o número de orientações concluídas e em andamento, bem como a pontuação da eficiência de cada professor, facilitando uma análise clara de sua performance acadêmica.

Essas etapas proporcionam uma análise objetiva da eficiência das orientações realizadas por professores, focando em mestrado e doutorado.

A pontuação da eficiência avalia a eficácia das orientações acadêmicas com base na taxa de conclusão, que representa a proporção de orientações concluídas em relação ao total (concluídas e em andamento) em cada nível. O cálculo considera tanto a eficiência quanto o volume de orientações concluídas, utilizando esse número para ponderar a influência de cada nível e normalizar a pontuação. Os pesos para mestrado e doutorado, definidos na função *main()*, são 2 e 3, respectivamente, podendo ser ajustados conforme a análise. Ao final, o código exibe os resultados de forma clara, permitindo uma avaliação justa e precisa da qualidade do trabalho de orientação com base nos dados da Plataforma Lattes.

### 5.1.4 Pontuação da Produção Científica:

A pontuação da produção acadêmica baseada em citações é calculada a partir de várias métricas que, em conjunto, fornecem uma visão abrangente do impacto e da relevância das produções de um orientador. O processo envolve a consideração de citações nos últimos cinco anos, citações totais, índice h e índice i10, todos combinados de maneira a refletir tanto a quantidade quanto a qualidade da produção acadêmica.

Inicialmente, são contabilizadas as citações que o orientador recebeu nos últimos cinco anos, bem como o número total de citações ao longo de sua carreira. O índice h, que mede a produtividade e o impacto do autor com base em suas publicações mais citadas, e o índice i10, que indica o número de publicações que receberam 10 ou mais citações, também são utilizados como indicadores de qualidade.

Para calcular a pontuação da produção acadêmica, essas métricas são combinadas de forma ponderada. As citações nos últimos cinco anos e as citações totais são multiplicadas por fatores ajustáveis, $\alpha$ (alpha) e $\beta$ (beta), respectivamente, que podem ser calibrados para refletir a ênfase desejada em cada métrica. Em seguida, o produto dessas citações ponderadas é multiplicado pela soma dos índices h e i10, resultando na pontuação final de produção.

Essa abordagem ponderada permite uma avaliação detalhada da produção acadêmica, considerando tanto a quantidade de citações recebidas em diferentes períodos quanto a qualidade das publicações, conforme refletido pelos índices h e i10. A pontuação resultante oferece uma métrica abrangente para avaliar o impacto acadêmico de um orientador, destacando a importância e a influência de suas publicações na comunidade científica. Em resumo, a pontuação da produção acadêmica é uma métrica composta que integra múltiplos aspectos da produção científica de um orientador, oferecendo uma visão holística de sua contribuição para a academia e a comunidade científica.

A pontuação da produção acadêmica, denotada por $P_{\text{Produção}}$, é determinada pela seguinte equação:

$$P_{\text{Produção}} = \alpha \left( \frac{C_{\text{total}}}{C_{\text{5anos}}} \right) h_i + \beta \left( \frac{i_{10}}{C_{\text{total}}} \right) \tag{5.9}$$

onde $C_{\text{5anos}}$ corresponde ao número de citações recebidas nos últimos cinco anos; $C_{\text{total}}$ representa o número total de citações; $h_i$ é o índice $h$, que mede o impacto do pesquisador ao indicar o maior valor $h$ tal que ele possua pelo menos $h$ artigos com $h$ ou mais citações cada; $i_{10}$ é o índice $i10$, definido como o número de artigos do pesquisador que receberam pelo menos 10 citações; e $\alpha$ e $\beta$ são fatores ajustáveis que ponderam a contribuição de cada componente.

Se o valor de $C_{\text{total}}$ for igual a zero, a pontuação $P_{\text{Produção}}$ é automaticamente atribuída como zero, a fim de evitar erros decorrentes de divisão por zero durante o cálculo.

### Etapas do Cálculo

O procedimento para o cálculo de $P_{\text{Produção}}$ envolve as seguintes etapas:

i. **Leitura e extração dos dados:** Cada arquivo JSON é aberto com a função `json.load()`, permitindo a extração das variáveis necessárias: $C_{5\text{anos}}$, $C_{\text{total}}$, $h_{\text{index}}$ e $i_{10\text{index}}$.

ii. **Cálculo da pontuação:** Com os dados extraídos, o valor de $P_{\text{Produção}}$ é obtido por meio da fórmula proposta, aplicando os pesos $\alpha$ e $\beta$ para ponderar adequadamente as métricas de citações e os índices bibliométricos.

iii. **Processamento e exibição dos resultados:** O processo é repetido para todos os arquivos JSON em uma pasta específica. Para cada pesquisador, o sistema exibe no terminal os valores extraídos e o respectivo $P_{\text{Produção}}$, viabilizando uma análise comparativa do impacto acadêmico.

## 5.1.5 Pontuação da Colaboração:

A pontuação da colaboração acadêmica é calculada para avaliar a influência e a colaboração de um orientador com base em sua participação em bancas e no número de coautores com quem trabalhou. Este cálculo considera tanto a quantidade de participações em bancas quanto o número de coautores, normalizando essas contagens e combinando-as para obter uma pontuação final.

Para determinar a pontuação da colaboração, começamos contabilizando o número de participações em bancas, que inclui todas as vezes que o orientador atuou em comissões acadêmicas, como bancas de defesa de dissertações e teses. Em seguida, contamos o número de coautores, que são outros pesquisadores com quem o orientador colaborou em suas publicações. Após obter essas contagens, normalizamos os valores para ajustá-los em uma escala comum. A normalização é feita subtraindo o valor mínimo observado de cada contagem e dividindo pela diferença entre o valor máximo e o valor mínimo. Isso assegura que as contagens de participações em bancas e de coautores sejam comparáveis, independentemente de suas magnitudes absolutas. Em seguida, combinamos as contagens normalizadas usando pesos definidos para cada tipo de atividade. Esses pesos refletem a importância relativa de participações em bancas e de colaborações com coautores na determinação da atuação colaborativa de um orientador. A combinação das contagens normalizadas ponderadas resulta na pontuação final de colaboração.

Essa abordagem proporciona uma medida clara da colaboração acadêmica de um orientador, considerando tanto seu envolvimento em bancas quanto suas colaborações com outros pesquisadores. A pontuação da colaboração é essencial para identificar orientadores que têm uma influência significativa na academia e que colaboram amplamente com seus pares, refletindo seu impacto e alcance na comunidade científica.

O cálculo da pontuação da colaboração acadêmica $P_{\text{colaboração}}$ para cada professor é dado por:

$$P_{\text{Colaboração}} = w_1 \left( \frac{P_{\text{banca}}}{\max(P_{\text{banca}})} \right) + w_2 \left( \frac{Co}{\max(Co)} \right) \tag{5.10}$$

onde $P_{\text{Colaboração}}$ representa a pontuação da colaboração acadêmica; $P_{\text{banca}}$ é o número de participações do professor em bancas de exame; $\max(P_{\text{banca}})$ é o maior valor de participações em bancas observado entre os professores analisados; $Co$ corresponde ao número total de coautores com os quais o professor publicou; $\max(Co)$ é o maior número de coautores encontrado entre todos os professores da amostra; e $w_1$ e $w_2$ são os pesos atribuídos a cada componente.

## Etapas do Cálculo

As etapas para o cálculo da colaboração acadêmica são as seguintes:

i. **Contagem de Participações em Bancas:** Para cada arquivo XML, analisa-se o número de participações em bancas de mestrado e doutorado, bem como bancas de qualificação de mestrado e de doutorado. A soma desses valores fornece o total de participações em bancas.

ii. **Extração do Número de Coautores:** Para cada arquivo XML, conta-se o número de coautores com os quais o professor colaborou em suas produções bibliográficas.

iii. **Determinação dos Valores Máximos:** Identifica-se o maior número de participações em bancas (*MaxParticipações*) e o maior número de coautores (*MaxCoautores*) entre todos os professores analisados.

iv. **Normalização:** Os valores de *Participações* e *Coautores* são normalizados dividindo-se pelo valor máximo correspondente.

v. **Cálculo da colaboração:** Aplica-se a fórmula da *colaboração*, combinando as contribuições normalizadas das participações e dos coautores, ponderadas pelos pesos $w_1$ e $w_2$.

### 5.1.6 Pontuação da Pesquisa:

A pontuação da pesquisa acadêmica é calculada para avaliar a participação do orientador em atividades de pesquisa e desenvolvimento. Inicialmente, são contabilizadas todas as atividades de pesquisa registradas no currículo, como participação em projetos científicos e tecnológicos. Cada atividade contribui de forma aditiva para a pontuação total, refletindo o envolvimento e a produtividade do orientador no desenvolvimento acadêmico. A pontuação final oferece uma medida objetiva da dedicação do orientador à pesquisa, reconhecendo sua contribuição para o avanço do conhecimento e para o fortalecimento da comunidade científica.

A pontuação da pesquisa é obtida com base no número de atividades de pesquisa e desenvolvimento identificadas nos currículos. O cálculo segue a seguinte fórmula:

$$P_{\text{Pesquisa}} = \sum_{i=1}^{n} a_i \; N_i \tag{5.11}$$

onde $P_{\text{Pesquisa}}$ representa a pontuação da pesquisa, $a_i$ é o peso atribuído à atividade $i$, $N_i$ corresponde ao número de ocorrências dessa atividade e $n$ é o total de atividades consideradas.

#### Etapas do Cálculo

i. Extração dos dados: a partir dos arquivos XML, identificam-se as informações referentes às atividades de pesquisa e desenvolvimento.

ii. Cálculo da contribuição de cada atividade: para cada tipo de atividade $i$, determina-se o número de ocorrências ($N_i$) e aplica-se o peso correspondente ($a_i$).

iii. Determinação da pontuação final: a pontuação da pesquisa é obtida pela soma ponderada das contribuições de todas as atividades, conforme a equação (5.11).

## 5.2 Códigos Desenvolvidos em Python

### 5.2.1 p_area.py

Este código foi desenvolvido para realizar a comparação entre áreas de conhecimento listadas em arquivos XML de currículos Lattes, atribuindo uma

pontuação com base no nível de correspondência entre a grande área, área, subárea e especialidade de cada formação. O objetivo principal é comparar as áreas de conhecimento de um pesquisador de referência com outros currículos, gerando uma pontuação que indica o nível de similaridade.

## Função `extract_knowledge_areas(xml_file)`

A função `extract_knowledge_areas(xml_file)` é responsável por extrair as áreas de conhecimento de um arquivo XML com base nas informações de formação acadêmica, abrangendo graduação, mestrado, doutorado e pós-doutorado. Para cada formação listada no currículo, a função percorre as tags correspondentes utilizando o método `findall` para localizar as informações de interesse e extrair os dados das tags nos arquivos XML. Em seguida, ela organiza essas informações em uma lista de dicionários, onde cada dicionário contém os nomes da Grande Área, da Área, da Subárea e da Especialidade associadas à formação acadêmica do pesquisador.

## Função `compare_areas(area1, area2)`

A função `compare_areas(area1, area2)` compara duas áreas de conhecimento e retorna uma pontuação que reflete o grau de correspondência entre elas. Caso as grandes áreas coincidam, é atribuído 1 ponto; se as áreas específicas coincidirem, são somados 2 pontos adicionais; se as subáreas forem iguais, acrescentam-se mais 3 pontos; e, se as especialidades apresentarem correspondência, adicionam-se 4 pontos extras. A soma desses valores resulta na pontuação total, indicando o nível de similaridade entre as duas áreas de conhecimento.

## Função `main(reference_xml, folder_path)`

A função `main(reference_xml, folder_path)` é o ponto de entrada do programa e coordena todo o processo de comparação das áreas de conhecimento. Inicialmente, ela extrai as áreas de conhecimento do arquivo XML de referência, utilizando a função `extract_knowledge_areas()`. Em seguida, a função percorre todos os arquivos XML presentes na pasta especificada, por meio da biblioteca `glob`, identificando os currículos a serem comparados. Para cada arquivo, as áreas extraídas são comparadas com as do currículo de referência utilizando a função `compare_areas()`, e é calculada uma pontuação total de similaridade a partir da soma dos pontos obtidos em cada comparação. Por fim, os arquivos são classificados em ordem decrescente de similaridade, e os resultadosincluindo o nome do arquivo e a respectiva pontuaçãosão exibidos no terminal.

## Fluxo do Código

O fluxo de execução do código pode ser descrito da seguinte forma:

i. É especificado o caminho para o arquivo XML de referência, representando o currículo de referência.

ii. A função percorre todos os arquivos XML em uma pasta designada, onde estão armazenados os currículos a serem comparados.

iii. As áreas de conhecimento de cada currículo são comparadas com as áreas do currículo de referência, e uma pontuação é atribuída para indicar o nível de similaridade.

iv. Os resultados são organizados em uma lista, classificando os arquivos XML por pontuação da similaridade de forma decrescente.

## Exibição dos Resultados

A saída do programa consiste na impressão dos resultados da comparação, exibindo o nome dos arquivos XML analisados e a pontuação total atribuída a cada um. A saída no terminal terá o seguinte formato:

```
Resultados da Comparação de Áreas de Conhecimento
================================================
Arquivo: nome_do_arquivo_1.xml - Pontuação Total: XX
------------------------------------------------
Arquivo: nome_do_arquivo_2.xml - Pontuação Total: XX
------------------------------------------------
```

Este código pode se tornar uma ferramenta útil ao comparar áreas de conhecimento entre currículos Lattes, permitindo a identificação de similaridades entre os perfis de pesquisadores. Ele pode ser utilizado para várias finalidades, como análise de afinidade acadêmica entre pesquisadores, seleção de orientadores ou membros de comitês de avaliação. O método de pontuação permite que a similaridade seja avaliada de forma quantitativa, facilitando a interpretação dos resultados.

Para realizar a extração das áreas de conhecimento, acessamos os arquivos XML dos currículos dos orientadores e utilizamos um *script* Python para extrair as informações específicas das áreas de conhecimento, identificando as tags

XML relevantes. Em seguida, na comparação das áreas de conhecimento, analisamos as áreas extraídas e comparamos entre os diferentes orientadores, utilizando métricas de similaridade para determinar a proximidade ou distinção das áreas de conhecimento. Para o cálculo da pontuação total, aplicamos a fórmula onde a pontuação total de similaridade é a soma das pontuações atribuídas a cada correspondência nos diferentes níveis. Finalmente, realizamos a ordenação dos resultados, onde as pontuações totais são organizadas em ordem decrescente, identificando os orientadores mais qualificados para supervisionar novos alunos de pós-graduação segundo as similaridades de área.

### 5.2.2 p_experiencia.py

O código desenvolvido tem como objetivo processar e calcular a pontuação da experiência acadêmica de orientadores, focando exclusivamente nas orientações concluídas e em andamento nos níveis de mestrado e doutorado. Ele também considera o número de publicações para calcular um fator de qualidade, que influencia diretamente a pontuação final. O código percorre todos os arquivos XML armazenados em uma pasta específica que contém os currículos em formato XML. O uso da biblioteca `os` permite navegar pelas pastas e a biblioteca `xml.etree.ElementTree` é utilizada para ler e manipular o conteúdo dos arquivos XML.

#### Extração de Orientações e de Publicações

A extração do número de orientações é realizada por duas funções específicas. A função `count_orientacoes_concluidas()` é responsável por contar o número de orientações concluídas em nível de mestrado e doutorado, a partir das tags `<ORIENTACOES-CONCLUIDAS-PARA-MESTRADO>` e `<ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO>`. De forma semelhante, a função `count_orientacoes_andamento()` executa a contagem das orientações em andamento, analisando as tags `<DADOS-BASICOS-DA-ORIENTACAO-EM-ANDAMENTO-DE-MESTRADO>` e `<DADOS-BASICOS-DA-ORIENTACAO-EM-ANDAMENTO-DE-DOUTORADO>`. Essas informações são fundamentais para o cálculo de métricas relacionadas à experiência de orientação acadêmica dos docentes.

A função `extrair_numero_publicacoes()` conta o número de publicações de artigos acadêmicos presentes nos currículos dos orientadores, usando a tag `<ARTIGO-PUBLICADO>`. Esse número é posteriormente utilizado para calcular o fator de qualidade $Q$.

## Cálculo do Fator de Qualidade $Q$

A função `calcular_fator_qualidade()` calcula o fator de qualidade $Q$, que é a razão entre o número de publicações do pesquisador ($P_r$) e um valor de referência ($P_{\max}$), que é o percentil 90 do número de publicações de todos os orientadores. Isso significa que $P_{\max}$ representa um limite superior, e o fator $Q$ normaliza a qualidade de cada orientador em relação a esse valor de referência.

## Cálculo de $P_{\max}$ e da pontuação da Experiência

Antes de calcular a pontuação da experiência de cada orientador, o código processa todos os arquivos XML para calcular $P_{\max}$, que é o valor de referência de publicações. O $P_{\max}$ é definido como o percentil 90 do número de publicações dos orientadores, representando um padrão referencial de qualidade para a comparação.

A função `calcular_pontuacao_equacao()` é responsável por calcular a pontuação da experiência acadêmica de um orientador utilizando uma abordagem ponderada. O cálculo leva em consideração tanto o número de orientações concluídas quanto as orientações em andamento nos níveis de mestrado e doutorado. Esses valores são somados para fornecer uma medida da experiência total do orientador em cada nível acadêmico. A pontuação é ajustada por meio de pesos específicos, sendo que o mestrado tem um peso atribuído, assim como o doutorado. Esses pesos refletem a importância relativa de cada nível no cálculo da experiência. Além disso, são aplicados limites superiores para cada categoria, ou seja, valores máximos que controlam o quanto as orientações em cada nível podem influenciar na pontuação final.

Outro aspecto importante do cálculo é o fator de qualidade $Q$, que ajusta a pontuação para refletir não apenas a quantidade de orientações, mas também a qualidade do trabalho do orientador. Dessa forma, a pontuação final considera tanto o volume de orientações realizadas quanto a qualidade atribuída ao orientador, proporcionando uma avaliação equilibrada da sua experiência acadêmica.

## Processamento dos Arquivos XML

O código percorre todos os arquivos XML para calcular a pontuação da experiência de cada orientador, utilizando as informações sobre orientações e publicações extraídas anteriormente. Para cada orientador, o nome do arquivo XML, o número de publicações, e as pontuações de mestrado e doutorado são exibidos no terminal, junto com a pontuação final de experiência.

### Exibição dos Resultados

Após processar todos os arquivos XML, o código imprime no terminal o total de pastas e arquivos processados, proporcionando uma visão geral do processo de extração e cálculo realizado. Este código automatiza a análise de currículos Lattes, concentrando-se nas orientações de mestrado e doutorado e nas publicações de cada orientador. O cálculo da pontuação da experiência acadêmica pondera as orientações concluídas e em andamento com base na quantidade e qualidade do pesquisador, levando em consideração o número de publicações e o fator de qualidade $Q$. Essa abordagem garante uma avaliação justa e equilibrada, permitindo que orientadores com diferentes níveis de experiência e produção científica sejam comparados de maneira objetiva.

```
Resultados da Pontuação da Experiência
================================================
P_max (Percentil 90 das publicações): XX.XX
Orientador: XXXX
Número de Artigos Publicados: XXXX
Experiência em Mestrado: XXXX
Experiência em Doutorado: XXXX
Pontuação da Experiência: XXXX
Total de pastas varridas: XXXX
Total de arquivos XML varridos: XXXX
------------------------------------------------
```

### 5.2.3 p_pesquisa.py

O código desenvolvido tem como objetivo calcular a pontuação da pesquisa acadêmica de professores e pesquisadores a partir de informações extraídas dos arquivos XML do Currículo Lattes. O conceito de engajamento, neste contexto, é definido pela participação em atividades de projetos e desenvolvimento em pesquisa, uma dimensão fundamental da atuação acadêmica. A metodologia empregada envolve a leitura e análise de dados referentes a projetos de pesquisa registrados no currículo de cada pesquisador.

O código percorre todos os arquivos com extensão `.xml` em uma pasta específica, incluindo eventuais subdiretórios, onde estão armazenados os currículos em formato XML. A biblioteca `os` é utilizada para navegar pelas pastas, enquanto a `xml.etree.ElementTree` permite a leitura e manipulação do conteúdo dos arquivos XML.

## Extração da Pontuação da Pesquisa

A função `get_research_score()` é responsável por calcular a pontuação da pesquisa, baseada no número de projetos de pesquisa e desenvolvimento realizados pelo pesquisador. Essa informação é extraída das tags `PESQUISA-E-DESENVOLVIMENTO`, e cada ocorrência representa um projeto de pesquisa. Para cada projeto encontrado, a função incrementa a pontuação da pesquisa ($PP$) em uma unidade, retornando o total ao final.

## Cálculo da Pontuação da Pesquisa

A função `calculate_research_score()` calcula a pontuação da pesquisa exclusivamente com base no número de atividades de pesquisa e desenvolvimento registradas nos currículos XML. Essa abordagem privilegia a objetividade, ao quantificar diretamente o envolvimento do pesquisador em ações de pesquisa e desenvolvimento, conforme registrado na Plataforma Lattes.

## Exibição dos Resultados

Para cada arquivo XML processado, o código exibe os resultados de forma estruturada e clara. Os principais elementos apresentados incluem o nome do arquivo, que identifica o currículo analisado, e a pontuação da Pesquisa ($PP$), que representa a quantidade de atividades de pesquisa e desenvolvimento encontradas no documento. Essa exibição facilita a análise comparativa entre diferentes pesquisadores com base em seu envolvimento em ações de pesquisa registradas no currículo Lattes.

```
Resultados da Pontuação da pesquisa
================================================
Arquivo: XXXXXXX.xml
Pontuação da Pesquisa (PP): XXXX
------------------------------------------------
```

Essas informações são exibidas sequencialmente para cada arquivo XML processado, facilitando a comparação entre pesquisadores com base em suas atividades de pesquisa. O código implementado constitui uma ferramenta automatizada para avaliação da atuação acadêmica de docentes, a partir da análise de dados extraídos diretamente da Plataforma Lattes. O código é flexível e pode ser modificado para atender a critérios específicos de avaliação.

### 5.2.4 p_producao .py

A avaliação da produção científica é essencial para mensurar o impacto acadêmico e subsidiar decisões em programas de pós-graduação. Para essa finalidade, foram desenvolvidos *scripts* em Python que calculam a Pontuação da Produção com base em dados extraídos do Google Scholar, uma vez que os arquivos XML da Plataforma Lattes não fornecem as variáveis necessárias para esse cálculo.

#### Extração de Dados no Google Scholar

Um primeiro script utiliza a biblioteca scholarly para buscar e recuperar informações detalhadas de pesquisadores por meio de suas respectivas páginas no Google Scholar. Os dados extraídos incluem variáveis-chave como o número de citações nos últimos 5 anos, o total de citações, o índice H e o índice i10. Esses dados são salvos em formato JSON para cada pesquisador, permitindo armazenamento local e análises posteriores.

#### Cálculo da Pontuação da Produção

Um segundo *script* foi desenvolvido para processar os arquivos JSON gerados anteriormente, calculando automaticamente a pontuação da produção científica $P_{\text{Produção}}$ para cada pesquisador. A fórmula aplicada considera citações nos últimos cinco anos, total de citações acumuladas, índice $h$ e índice $i10$, combinados por meio de uma equação ponderada com pesos ajustáveis ($\alpha$) e ($\beta$), o que permite personalizar a influência de cada variável no resultado final.

O principal objetivo dessa abordagem é superar as limitações dos dados disponíveis nos arquivos XML do Lattes, utilizando informações complementares do Google Scholar para viabilizar uma análise quantitativa mais precisa da produção científica. A estrutura modular do código favorece futuras adaptações, como a inclusão de novos parâmetros ou ajustes nos pesos, permitindo que os resultados sejam integrados a sistemas de classificação e recomendação de orientadores.

#### Função `calcular_producao(json_file, alpha=1, beta=1)`

A função `calcular_producao()` calcula o valor de $P_{\text{Produção}}$ com base em um arquivo JSON, a partir da extração de quatro variáveis: número de citações

nos últimos cinco anos ($C_{5\text{anos}}$), total de citações ($C_{\text{total}}$), índice $h$ ($h_{\text{index}}$) e índice $i10$ ($i10_{\text{index}}$). A fórmula utiliza pesos ajustáveis $\alpha$ e $\beta$ para ponderar as variáveis. Caso $C_{\text{total}} = 0$, a pontuação é atribuída como zero para evitar divisão por zero. A função retorna todas as variáveis extraídas, além do valor final de $P_{\text{Produção}}$.

## Função `processar_pasta_json(pasta, alpha=1, beta=1)`

A função `processar_pasta_json()` percorre todos os arquivos com extensão `.json` em uma pasta especificada e, para cada arquivo, executa a função `calcular_producao()` com os parâmetros $\alpha$ e $\beta$. Os resultados incluindo o nome do arquivo, $C_{5\text{anos}}$, $C_{\text{total}}$, $h_{\text{index}}$, $i10_{\text{index}}$ e $P_{\text{Produção}}$ são exibidos no terminal. O sistema permite tanto a análise individual quanto uma avaliação agregada dos dados, promovendo uma mensuração automatizada e padronizada da produção científica.

Enfim o fluxo de execução é simples e modular, a partir do diretório informado, todos os arquivos JSON são processados, e os valores de $P_{\text{Produção}}$ são calculados com base nos pesos $\alpha$ e $\beta$, que podem ser ajustados conforme a necessidade.

## Exibição dos Resultados

A saída do programa consiste na apresentação dos resultados correspondentes a cada arquivo JSON processado. Para cada pesquisador, são exibidos no terminal: o nome do arquivo, o número de citações nos últimos cinco anos ($C_{5\text{anos}}$), o total de citações acumuladas ($C_{\text{total}}$), o índice $h$ ($h_{\text{index}}$), o índice $i10$ ($i10_{\text{index}}$) e o valor calculado de $P_{\text{Produção}}$. Essas informações são organizadas de forma padronizada, permitindo fácil leitura e comparação entre os pesquisadores analisados.

```
Resultados da Pontuação da Produção Acadêmica
================================================
Arquivo: nome_do_arquivo.json
Citações nos últimos 5 anos (C_5anos): XXXX
Citações totais (C_total): XXXX
Índice H (h_index): XXXX
Índice i10 (i10_index): XXXX
Resultado de P_Produção: XX.XX
------------------------------------------------
```

Este código oferece uma ferramenta automatizada para o cálculo da produção científica de pesquisadores, permitindo ajustar os pesos $\alpha$ e $\beta$ e realizar uma análise em larga escala a partir dos dados extraídos de múltiplos arquivos JSON.

### 5.2.5 p_eficiência.py

Este código foi desenvolvido para calcular a pontuação da eficiência acadêmica com base nas orientações concluídas e em andamento nos níveis de mestrado e doutorado, utilizando arquivos XML de currículos Lattes. A abordagem considera a taxa de conclusão de orientações nesses dois níveis, ponderada por pesos específicos, resultando em uma pontuação final de eficiência.

#### Extração dos Dados

O código percorre todos os arquivos XML em uma pasta designada, utilizando a biblioteca `(os)` para navegar pelos diretórios e a *xml.etree.ElementTree* para a leitura e manipulação dos arquivos XML. Para cada arquivo, o código extrai informações sobre as orientações concluídas e em andamento no mestrado e doutorado.

#### Funções de Contagem

A função *count_orientacoes_concluidas()* é responsável por contar o número de orientações concluídas em mestrado e doutorado. Ela busca, especificamente, as tags *ORIENTACOES-CONCLUIDAS-PARA-MESTRADO* e *ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO*, que são utilizadas para determinar o número de orientações que já foram finalizadas em cada nível. De forma similar, a função *count_orientacoes_andamento()* realiza a contagem das orientações em andamento, analisando as tags *DADOS-BASICOS-DA-ORIENTACAO-EM-ANDAMENTO-DE-MESTRADO* e *DADOS-BASICOS-DA-ORIENTACAO-EM-ANDAMENTO-DE-DOUTORADO*. Nesse caso, o código considera apenas orientações de mestrado e doutorado, deixando de fora qualquer referência a graduação e iniciação científica.

#### Cálculo da Taxa de Conclusão e Pontuação da Eficiência

A pontuação da eficiência é calculada com base na taxa de conclusão de orientações, que é a proporção entre o número de orientações concluídas e o

número total de orientações (concluídas e em andamento). A função *calcular_taxa_conclusao()* executa esse cálculo separadamente para mestrado e doutorado. Se não houver orientações para um nível específico, a taxa de conclusão é definida como zero. Com as taxas de conclusão em mãos, a função *calcular_pontuacao_eficiência()* aplica pesos definidos para cada nível, multiplicando essas taxas pelos pesos e pelas orientações concluídas, de forma a gerar uma pontuação ponderada de eficiência. Nesse contexto, orientadores com mais orientações concluídas e maior taxa de conclusão recebem uma pontuação da eficiência mais alta, refletindo sua maior eficácia como orientadores. Esse cálculo é feito separadamente para os níveis de mestrado e doutorado.

### Processamento dos Arquivos

O código processa todos os arquivos XML na pasta especificada. Para cada arquivo, ele coleta o número de orientações concluídas e em andamento, calcula as taxas de conclusão e, em seguida, aplica os pesos para calcular a pontuação final de eficiência. Os resultados são armazenados em uma lista, que pode ser facilmente acessada para exibição.

### Exibição dos Resultados

A saída do programa exibe, para cada arquivo XML, o nome do orientador, os números de orientações concluídas e em andamento no mestrado e doutorado, além da pontuação final de eficiência. Os dados são apresentados de forma padronizada, facilitando a comparação entre orientadores.

```
Resultados da Pontuação da Eficiência Acadêmica
================================================
Arquivo: nome_do_arquivo.xml
Mestrado - Concluídas: XXXX | Em Andamento: XXXX
Doutorado - Concluídas: XXXX | Em Andamento: XXXX
Pontuação da eficiência: XX.XX
------------------------------------------------
```

## 5.2.6 p_colaboracao.py

O código apresentado foi desenvolvido com o propósito de analisar currículos Lattes no formato XML, focando nas participações em bancas de mestrado e doutorado, e no número de coautores em artigos publicados. O objetivo final é calcular uma pontuação da colaboração acadêmica, ignorando as atividades relacionadas à graduação. O cálculo de colaboração baseia-se em dois critérios principais: o número de participações em bancas de defesa e qualificação de mestrado e doutorado, e o número de coautores nas publicações do pesquisador.

## Contagem de Participações em Bancas

A função `contar_participacao_em_bancas` é a primeira responsável pela análise das participações. Ela começa carregando o arquivo XML fornecido e percorre as seções correspondentes às atividades complementares do pesquisador, dentro da tag `DADOS-COMPLEMENTARES`. O código contabiliza separadamente as participações em bancas de defesa de mestrado e doutorado, além das bancas de qualificação em ambos os níveis. O resultado é um dicionário contendo as contagens e o total de participações.

## Extração de Coautores

Em paralelo, a função `extrair_coautores` calcula o número de coautores presentes nas publicações do pesquisador. A partir da tag `ARTIGO-PUBLICADO`, o nome de cada coautor é extraído e armazenado em um conjunto (`set`), eliminando duplicatas. O número de coautores únicos é então retornado.

## Cálculo da colaboração

A colaboração acadêmica é calculada na função `calcular_colaboracao` Nessa função, primeiro são chamadas as funções auxiliares para contar as participações em bancas e extrair os coautores. O número total de participações e coautores é normalizado em relação aos máximos encontrados no conjunto de currículos processados.

## Processamento de Arquivos

A função `processar_arquivos` é responsável por processar todos os arquivos XML de uma pasta. Ela itera duas vezes pelos arquivos: na primeira, descobre os valores máximos de participações e coautores, necessários para normalizar os dados; na segunda, chama `calcular_reputacao` para calcular a colaboração de cada pesquisador e exibe os resultados.

O código desenvolvido calcula a colaboração acadêmica de professores a partir de sua participação em bancas e colaboração em publicações científicas. O processo envolve a análise de arquivos XML, extração de informações relevantes (participações em bancas e número de coautores), normalização desses dados com base nos valores máximos observados, e cálculo final da colaboração através de uma fórmula ponderada. Este modelo oferece uma ferramenta objetiva para comparar, neste contexto, a colaboração acadêmica entre diferentes professores, contribuindo para a seleção de orientadores em programas de pós-graduação.

## 5.3 Resumo do Capítulo

Neste capítulo apresentaram-se os procedimentos metodológicos que sustentam o modelo de recomendação de orientadores. Inicialmente, reuniram-se aproximadamente 1 800 currículos Lattes de docentes vinculados a programas de pós-graduação das instituições do COMUNG, mapeando suas atividades acadêmicas no período de 2010 a 2020. Esses arquivos XML foram então limpos e estruturados com *scripts* em Python, que organizaram as variáveis de interesse e automatizaram o cálculo das métricas acadêmicas. Por fim, complementaram-se os dados com indicadores de citação extraídos do Google Scholar, por meio da biblioteca *Scholarly*, a fim de suprir lacunas não atendidas pela Plataforma Lattes. A convergência dessas etapas assegura a consistência, a rastreabilidade dos dados empregados no cálculo das pontuações que compõem o IR.

# Capítulo 6

# Avaliação do Modelo

*Se eu pude ver mais longe foi por estar sobre os ombros de gigantes.*

*Isaac Newton*

Neste capítulo, faz-se a avaliação do modelo por meio da checagem técnica e estatística, incluindo consistência dos dados, testes de hipóteses e análise de sensibilidade. Apresentam-se as estratégias de avaliação do IR, com foco na avaliação de sua confiabilidade e coerência interna. Na Seção 6.1, realiza-se uma análise de completude dos arquivos XML, comparando cerca de 40 *tags* com o esquema XSD da Plataforma Lattes. Essa etapa permitiu identificar a frequência de preenchimento das variáveis e garantir a integridade dos dados utilizados no modelo. Em seguida, na seção 6.2, aplicam-se análises estatísticas descritivas e visuais, como histogramas e *boxplots*, para avaliar a distribuição das pontuações e detectar padrões relevantes. A Seção 6.3 dedica-se à avaliação específica da fórmula de pontuação da experiência, com foco na sensibilidade do modelo frente às variações na produção científica e no número de orientações. As correlações entre as variáveis do modelo são exploradas na Seção 6.4, enquanto a Seção 6.5 discute a presença de *outliers* e seu impacto nos rankings gerados. A estabilidade do IR diante de variações percentuais simuladas nas variáveis de entrada é analisada por meio de testes de sensibilidade na Seção 6.6. Na Seção 6.7, verifica-se a fórmula de pontuação de Eficiência, com base em taxas de conclusão ajustadas por nível acadêmico. Por fim, na Seção 6.8, avalia-se o comportamento geral das métricas do modelo, e a Seção 6.10 encerra o capítulo com um panorama consolidado dos resultados, reforçando a aplicabilidade do modelo proposto.

## 6.1 Completude de Dados em XML

A metodologia adotada nesta etapa consistiu na comparação direta de aproximadamente 1.800 arquivos XML, distribuídos em uma pasta principal com múltiplas subpastas. A referência utilizada foi o arquivo *CurriculoLattes_12_09_2022.xsd*, que define a estrutura completa de campos esperados para um currículo padrão na Plataforma Lattes.

O processo foi conduzido em cinco etapas principais:

i. Inicialmente, procedeu-se à extração e ao mapeamento de todas as *tags* definidas no XSD, que estabelece as regras de conformidade ao esquema da estrutura XML da Plataforma Lattes. Essa etapa envolveu a leitura da hierarquia de elementos e atributos, permitindo identificar, de forma abrangente, tanto as *tags* obrigatórias quanto as opcionais previstas no modelo de currículo.

ii. Em seguida, foi executada uma varredura automatizada em todos os arquivos XML disponíveis no diretório analisado. O objetivo foi detectar a presença das *tags* extraídas do XSD em cada currículo. A leitura foi realizada de forma recursiva, abrangendo não apenas os elementos principais, mas também os aninhados, assegurando a análise em diferentes níveis de profundidade dos documentos.

iii. A terceira etapa consistiu no cálculo do percentual de presença de cada *tag*, com base na proporção de currículos em que cada elemento foi encontrado em relação ao total de documentos analisados. Essa métrica foi essencial para avaliar o nível de completude das informações e para mensurar a aderência de cada campo ao padrão definido pelo XSD.

iv. Posteriormente, os resultados foram segmentados em duas faixas de frequência: a primeira agrupou as *tags* com ocorrência menor ou igual a 50%, enquanto a segunda reuniu aquelas com frequência entre 51% e 100%. Essa classificação permitiu identificar, de maneira mais estratégica, quais elementos são mais recorrentes e quais são menos utilizados nos currículos analisados.

v. Por fim, foram gerados gráficos ilustrativos para representar visualmente a frequência de ocorrência das *tags*. Os elementos foram organizados em ordem decrescente, destacando no topo aquelas com maior incidência nos arquivos XML. Essa abordagem visual contribuiu para uma interpretação clara e objetiva dos dados, favorecendo a seleção das variáveis mais representativas para a construção do IR.

## Resultados e Gráficos

A seguir, são apresentadas as figuras 6.1, 6.2 e 6.3, geradas com base nos dados extraídos dos currículos XML analisados:

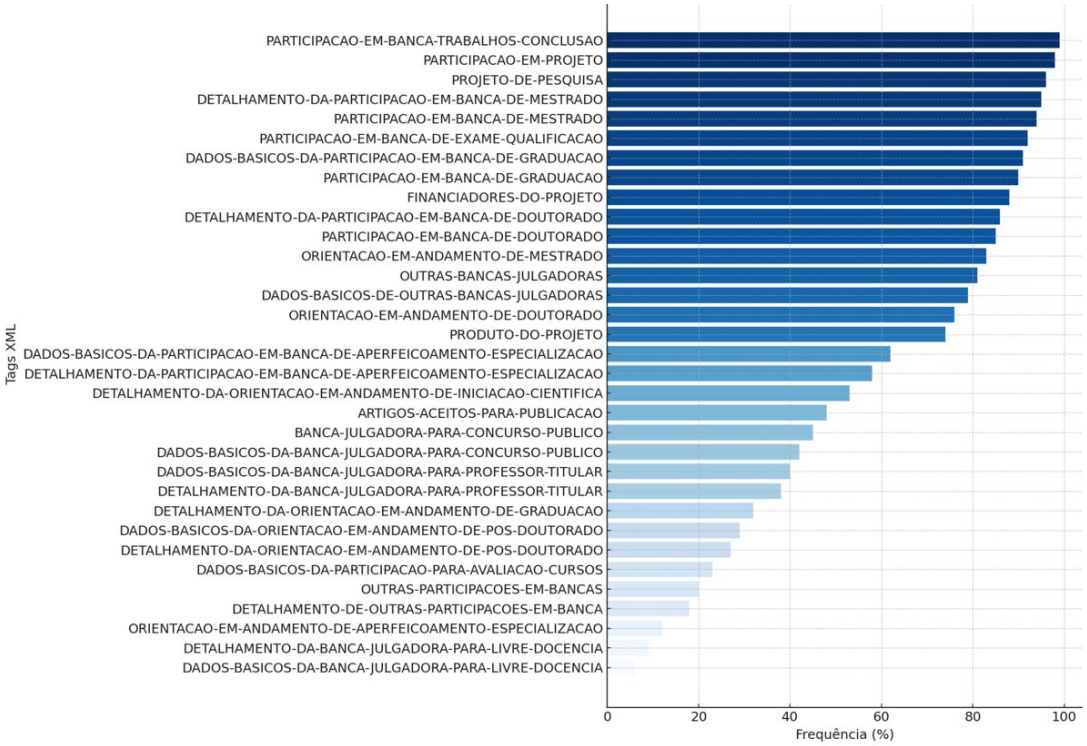



**Figura 6.1**: *Tags relevantes e suas frequências de completude.*

A figura 6.1 apresenta uma visão abrangente da distribuição de completude das *tags* extraídas dos arquivos XML avaliados. Algumas *tags* atingem taxas de ocorrência superiores a 90%, o que revela um padrão consistente de preenchimento em campos específicos do currículo Lattes. Dentre elas, destacam-se aquelas associadas à participação em bancas de defesa, ao registro de projetos de pesquisa e às orientações de mestrado, sugerindo que essas informações são amplamente registradas pelos pesquisadores.

Já a figura 6.2 ilustra as *tags* cuja frequência de ocorrência varia entre 0% e 50%. Este grupo contempla campos pouco frequentes ou de uso restrito, como BANCA-JULGADORA-PARA-LIVRE-DOCENCIA e ORIENTACAO-EM-ANDAMENTO-DE-APERFEICOAMENTO-ESPECIALIZACAO, que, apesar

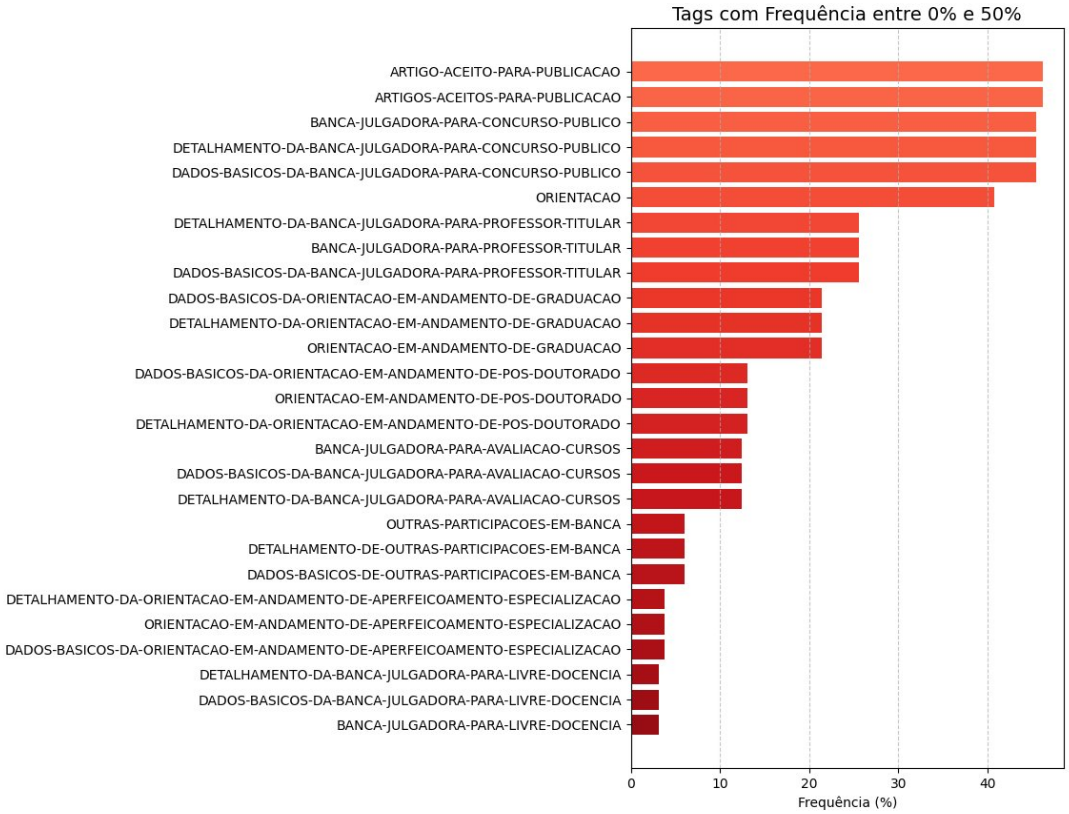



**Figura 6.2**: *Tags com frequência entre 0% e 50%.*.

de previstos na estrutura do XSD, são raramente utilizados. Essa baixa incidência pode estar relacionada à natureza especializada dessas atividades ou à subutilização de campos opcionais por parte dos usuários da Plataforma.

Em contrapartida, a figura 6.3 mostra as *tags* cuja presença nos documentos situa-se entre 51% e 100%. Nesse grupo, observa-se predominância de registros como PARTICIPACAO-EM-BANCA, PARTICIPACAO-PROJETO e PARTICIPACAO-EM-BANCA-DE-MESTRADO, evidenciando uma forte adesão às práticas esperadas na documentação curricular dos docentes e pesquisadores.

A segmentação das frequências em duas faixas de 0% a 50% e de 51% a 100% revelou-se eficaz para a identificação das variáveis com maior potencial de influência na construção do IR. Tal divisão permite destacar, de forma estratégica, os campos mais representativos do ponto de vista da comple-

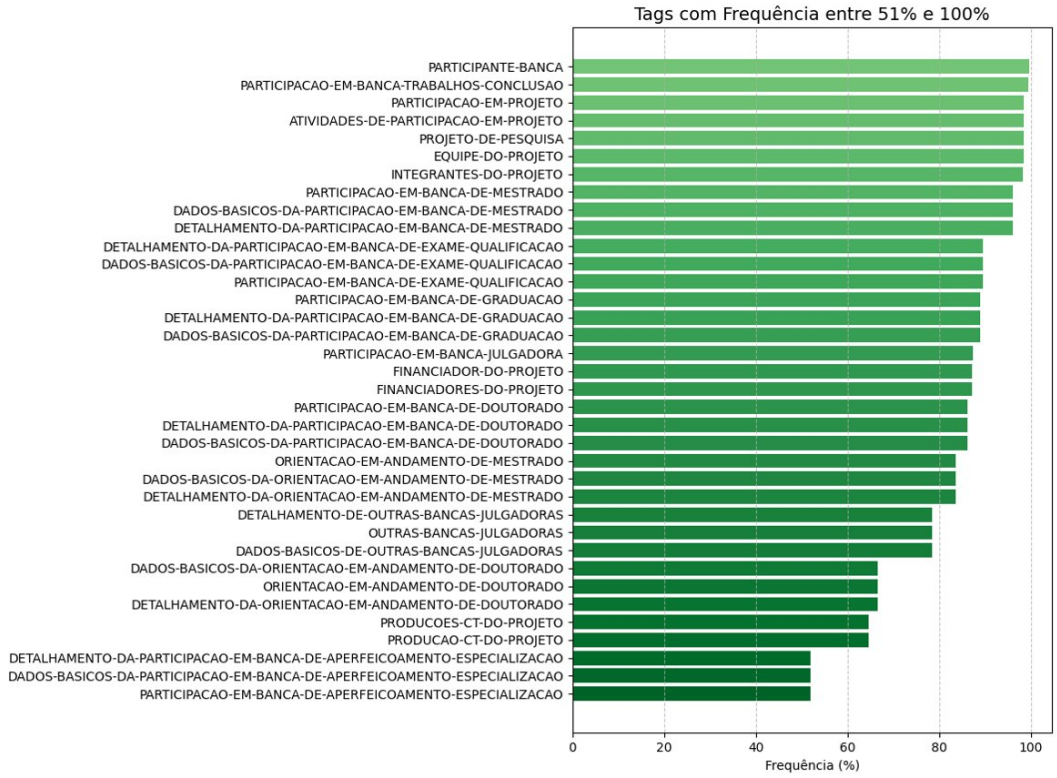



**Figura 6.3**: *Tags com frequência entre 51% e 100%.*

tude, bem como identificar lacunas que podem comprometer a robustez das análises subsequentes.

Ao todo, foram avaliadas cerca de 40 *tags*, extraídas diretamente da estrutura definida pelo XSD da Plataforma Lattes. Essas *tags* abrangem campos fundamentais e complementares da produção acadêmica, incluindo dados de formação, atuação docente, pesquisa, orientação e participação institucional.

Do conjunto analisado, aproximadamente 12 *tags* apresentaram taxa de completude superior a 90%, caracterizando-se como campos amplamente preenchidos. Entre elas, destacam-se PARTICIPACAO-EM-BANCA-TRABALHOS-CONCLUSAO, PROJETO-DE-PESQUISA e ORIENTACAO-EM-ANDAMENTO-DE-MESTRADO, que refletem práticas formais amplamente difundidas no meio acadêmico.

Por outro lado, cerca de 20% das *tags* apresentaram frequência inferior a 50%, o que evidencia baixa presença dessas informações nos currículos.

Nesse grupo, incluem-se campos como BANCA-JULGADORA-PARA-LIVRE-DOCENCIA, cuja ocorrência depende de situações acadêmicas específicas e não universais.

Na faixa intermediária, entre 51% e 89%, situam-se aproximadamente 30% das *tags* avaliadas, como PARTICIPACAO-EM-BANCA-DE-MESTRADO, com presença significativa, porém não majoritária. Essa faixa representa variações no padrão de preenchimento, sugerindo que a presença dessas informações pode depender de fatores institucionais ou de estágio da carreira docente.

Em síntese, a análise de completude revelou uma distribuição heterogênea na utilização das *tags* nos arquivos XML, refletindo tanto práticas consolidadas quanto lacunas informacionais relevantes. A metodologia de varredura automatizada demonstrou-se eficiente ao identificar padrões e inconsistências, oferecendo subsídios sólidos para o aprimoramento de sistemas de recomendação baseados em dados curriculares estruturados.

## Relação entre IC e IR

O Índice de Completude (IC) é uma métrica criada para avaliar o grau de preenchimento estrutural de um currículo Lattes com base no esquema de dados padrão XSD da Plataforma. Fundamenta-se na constatação da presença ou ausência de *tags* específicas no arquivo XML de cada pesquisador, permitindo mensurar a conformidade do documento em relação ao modelo formal estabelecido.

A fórmula do IC é expressa da seguinte forma:

$$IC = \frac{\text{tags encontradas (XML)}}{\text{tags esperadas (XSD)}} \times 100 \tag{6.1}$$

Esse índice indica o percentual de aderência estrutural do arquivo XML ao padrão definido, funcionando como um diagnóstico técnico da qualidade e completude dos dados disponíveis. O IC, portanto, é essencial como etapa preliminar para garantir que análises mais complexas  como aquelas baseadas no IR sejam realizadas sobre dados válidos e consistentes.

Por outro lado, o IR é uma métrica composta, de natureza multifatorial, voltada à avaliação do perfil e desempenho acadêmico de orientadores em programas de pós-graduação. Ele integra diversas dimensões analíticas,

| Critério | Descrição |
|---|---|
| Foco | Estrutura técnica do arquivo XML |
| Fonte de dados | Presença ou ausência de tags em comparação ao XSD |
| Objetivo | Identificar XMLs incompletos ou com falhas estruturais |
| Complexidade | Baixa; avaliação binária (tag presente ou ausente) |
| Indicador | Percentual de completude estrutural |
| Impacto final | Verificação da qualidade técnica para análises posteriores |
| Dependência | Direta do esquema XSD e do parser de leitura dos XMLs |

**Tabela 6.1**: *Características do Índice de Completude.*

incorporando variáveis quantitativas como produção científica, número de orientações concluídas ou em andamento e qualitativas como o impacto das publicações, atuação em bancas e engajamento em projetos de pesquisa.

As informações utilizadas na composição do IR são extraídas diretamente dos currículos Lattes, garantindo que o índice seja alimentado por dados objetivos, atualizados e verificáveis. Dessa forma, o IR oferece uma visão consolidada da trajetória acadêmica de cada docente, funcionando como uma ferramenta estratégica para subsidiar decisões mais criteriosas e alinhadas às exigências dos programas de pós-graduação.

## Relação entre IC e IR

A análise de completude dos XMLs, por meio do IC, permitiu identificar quais variáveis seriam mais adequadas para compor o IR, considerando a presença consistente de determinadas informações nos dados analisados.

## Resultados do IC

A análise dos currículos evidencia que as *tags* de produção bibliográfica, abrangendo artigos, livros e capítulos, aparecem em mais de 80% dos casos avaliados. As *tags* de orientações, tanto concluídas quanto em andamento, apresentam frequência moderada a alta, variando entre 60% e 85%. Já as *tags* de participação em bancas estão presentes em aproximadamente 50% dos currículos, enquanto as *tags* de projetos e extensão aparecem em uma faixa entre 40% e 60% dos XMLs. Por fim, observa-se que os indicadores bibliométricos, como índice H, i10 e número de citações, não estão contemplados nos XMLs, embora possam ser obtidos por meio de APIs externas.

| Critério | Índice de Completude | Índice de Recomendação |
|---|---|---|
| Foco | Estrutura técnica do XML | Avaliação de performance e competência do orientador |
| Fonte de dados | Presença/ausência de tags em XML vs. XSD | Métricas extraídas do XML |
| Objetivo | Diagnosticar XMLs incompletos ou com falhas de schema | Ajudar na tomada de decisão para escolha de orientadores |
| Complexidade | Simples, binário (existe ou não) | Multifatorial, combina vários parâmetros ponderados |
| Indicador | Percentual de completude | Score ponderado entre diferentes dimensões quantitativas |
| Impacto final | Qualidade técnica dos dados para análise posterior | Indicador final da qualidade/perfil do orientador |
| Dependência | Direta do XSD e do parser do XML | Depende do IC estar alto para garantir dados completos para o IR |

**Tabela 6.2**: *Relação entre os Índices de Completude e Recomendação.*

### Justificativa da seleção das variáveis

Dessa forma, a escolha das variáveis que compõem o IR foi guiada por uma observação e análise direta da estrutura e da completude dos currículos por meio do IC, pela disponibilidade consistente dos dados nos arquivos XML analisados e pelo alinhamento com os parâmetros de avaliação da pós-graduação, considerando os critérios estabelecidos pela CAPES e pelo CNPq. O IC foi fundamental para garantir que o IR fosse construído com base em variáveis efetivamente extraíveis dos dados disponíveis. Assim, o IR resulta em um indicador confiável e operacional, compatível com o nível de completude verificado nos arquivos XML da Plataforma Lattes.

## 6.2 Avaliação Estatística

O IR é composto por seis equações que representam dimensões distintas da atuação do orientador acadêmico. Cada uma dessas componentes captura um

| Variável no IR | Justificativa baseada no IC |
|---|---|
| Produção acadêmica | Tags de produção (artigos, livros, capítulos) têm alta completude. |
| Orientações concluídas | Tags de orientações de mestrado e doutorado estão amplamente presentes. |
| Orientações em andamento | Tag com boa frequência e considerada relevante na avaliação da experiência docente. |
| Participação em bancas | Apesar de frequência intermediária, dados são suficientes para gerar valor ao IR. |
| Qualidade da produção | Tags que permitem cruzamento com fatores de impacto e classificação. |
| Bibliométricos | Dados ausentes diretamente no XML, mas são complementados com integração via APIs externas como Google Scholar. |
| Projetos | Embora com menor frequência, ainda aparecem em uma quantidade relevante e contribuem como diferencial no IR. |

**Tabela 6.3**: *Relação entre as Variáveis do IR e os Resultados do IC.*

aspecto específico da trajetória do pesquisador, contribuindo para a composição do índice final de forma ponderada e integrada.

Para verificar o modelo, conduziu-se a análise em dois componentes específicos: $P_{\text{experiência}}$ e $P_{\text{Eficiência}}$, selecionadas por representarem os pilares mais diretamente relacionados à atividade de orientação acadêmica. A análise concentrou-se, portanto, na avaliação estatística e estrutural dessas duas dimensões, com o objetivo de assegurar que o modelo seja coerente, robusto e aderente à realidade dos dados analisados.

A avaliação foi conduzida a partir de quatro etapas complementares de análise. A primeira consistiu na análise descritiva das pontuações, com o intuito de identificar padrões, possíveis inconsistências e estatísticas fundamentais da distribuição dos dados, como média, mediana, desvio padrão, e valores extremos. Essa etapa permitiu compreender a tendência central e a dispersão das variáveis, fornecendo indícios iniciais sobre a forma da distribuição.

Na sequência, foram examinadas as correlações entre cada componente analisada e suas variáveis de entrada, visando mensurar a influência relativa de cada fator na formação do respectivo escore. Para isso, aplicaram-se coeficien-

tes de correlação adequados ao tipo de dado e à distribuição observada, o que possibilitou avaliar a força e a direção das associações.

A terceira etapa abrangeu a análise de *outliers* e a avaliação dos *rankings* gerados. Essa abordagem teve como propósito avaliar a coerência dos resultados atribuídos aos orientadores, identificando casos extremos e verificando se os maiores e menores valores das pontuações refletem, de fato, perfis condizentes com alto ou baixo desempenho.

Por fim, foi realizada a análise de sensibilidade, a qual consistiu em simular variações incrementais nas variáveis de entrada para observar a resposta do modelo. Essa simulação permitiu testar a estabilidade do comportamento das pontuações em diferentes perfis (baixo, mediano e alto), assegurando que o modelo não apresente flutuações desproporcionais ou distorções indesejadas. Com isso, foi possível confirmar a robustez das fórmulas empregadas e sua adequação como parte integrante do IR.

## 6.3 Avaliação da Pontuação de Experiência

### Análise Descritiva do IR

### Objetivos

- Observar a distribuição do IR para verificar sua coerência.
- Identificar possíveis *outliers* ou anomalias.
- Calcular estatísticas básicas: média, mediana e desvio padrão.

### Procedimentos

i. Cálculo do IR: Executar o modelo nos 1800 XMLs e armazenar os resultados em uma tabela.

ii. Estatísticas Descritivas:

   - Média e mediana para análise da tendência central.
   - Desvio padrão para avaliar dispersão.
   - Valores mínimo e máximo para identificar extremos.

iii. Visualizações Gráficas:

- Histograma para observar a distribuição.
- Boxplot para identificar *outliers*.

iv. Teste de Normalidade: Aplicar o teste de Shapiro-Wilk e Kolmogorov-Smirnov. [†1].

## Resultados

A partir da aplicação da Pontuação de Experiência, equação(5.3), foi possível calcular os valores de saída de aproximadamente 1.800 orientadores, cujos currículos foram processados a partir de arquivos XML da Plataforma Lattes. Os resultados obtidos foram armazenados em um arquivo `.csv`[†2] e posteriormente analisados estatisticamente, a fim de descrever o comportamento da equação.

A Tabela 6.4 apresenta as estatísticas descritivas da pontuação de experiência atribuída aos orientadores, onde os resultados evidenciam a natureza assimétrica da distribuição da variável, com concentração de valores em faixas mais baixas e presença de casos extremos. A média observada foi de 7,70, enquanto a mediana foi de 3,43, indicando uma assimetria positiva  ou seja, a maioria dos orientadores apresenta pontuações inferiores à média, e poucos casos puxam a média para cima. O desvio padrão elevado 13,80 reforça a ideia de dispersão acentuada, condizente com a presença de valores extremos (*outliers* superiores). O valor mínimo foi 0,00, indicando ausência de registros de orientação para alguns orientadores. Por outro lado, a pontuação máxima atingiu 193,33, demonstrando a existência de casos com atuação orientadora muito elevada e constante. Os quartis também reforçam a concentração dos dados em faixas baixas: o primeiro quartil (Q1) foi de 1,04 e o terceiro quartil (Q3) de 8,67. Além disso, destaca-se que 160 orientadores (cerca de 8,9% da amostra) obtiveram pontuações superiores a 20, enquanto apenas 7 ultrapassaram a marca de 100 pontos. Esses dados corroboram a visualização apresentada no histograma, caracterizada por uma cauda longa à direita e grande concentração nos valores iniciais do eixo horizontal.

[†1]A escolha pelos testes de Kolmogorov-Smirnov e Shapiro-Wilk deve-se ao fato de serem os mais amplamente utilizados na literatura estatística para avaliar a normalidade de distribuições amostrais. O primeiro é indicado para amostras maiores, enquanto o segundo apresenta maior sensibilidade em amostras pequenas. Como o conjunto de dados analisado apresenta características compatíveis com ambos os cenários, optou-se por utilizá-los de forma complementar, assegurando maior qualidade na verificação da hipótese de normalidade.

[†2]O formato `.csv` (*Comma-Separated Values*) é amplamente utilizado para armazenamento e intercâmbio de dados tabulares em texto simples.

Essa análise estatística confirma que o modelo de pontuação de experiência é capaz de capturar adequadamente a variabilidade entre os perfis de orientadores, atribuindo pontuações elevadas apenas a casos com atividade intensiva e equilibrada de orientação, ao mesmo tempo em que penaliza a ausência ou baixa produtividade nesse aspecto.

| **Métrica** | **Valor** |
|---|---|
| Total de registros | 1803 |
| Média | 7,70 |
| Mediana | 3,43 |
| Desvio padrão | 13,80 |
| Mínimo | 0,00 |
| Máximo | 193,33 |
| 1º quartil (Q1) | 1,04 |
| 3º quartil (Q3) | 8,67 |
| Casos acima de 20 pontos | 160 orientadores |
| Casos acima de 100 pontos | 7 orientadores |

**Tabela 6.4**: *Estatísticas da Pontuação de Experiência.*

Esses valores evidenciam uma forte dispersão nos dados, com a média sendo sensivelmente superior à mediana, o que sugere uma distribuição assimétrica com cauda à direita. Para melhor visualização da distribuição das pontuações, foi gerado o histograma apresentado na Figura 6.4.

A Figura 6.4 mostra que a maior parte dos orientadores possui pontuação inferior a 20, com poucos casos concentrando valores muito elevados. A distribuição altamente assimétrica motivou a aplicação de testes de normalidade.

Para confirmar a ausência de normalidade na distribuição dos dados, foram aplicados dois testes estatísticos:

- **Shapiro-Wilk**: valor-p = 0,000, indicando rejeição da hipótese nula de normalidade;
- **Kolmogorov-Smirnov**: valor-p próximo de zero, também rejeitando a hipótese de distribuição normal.

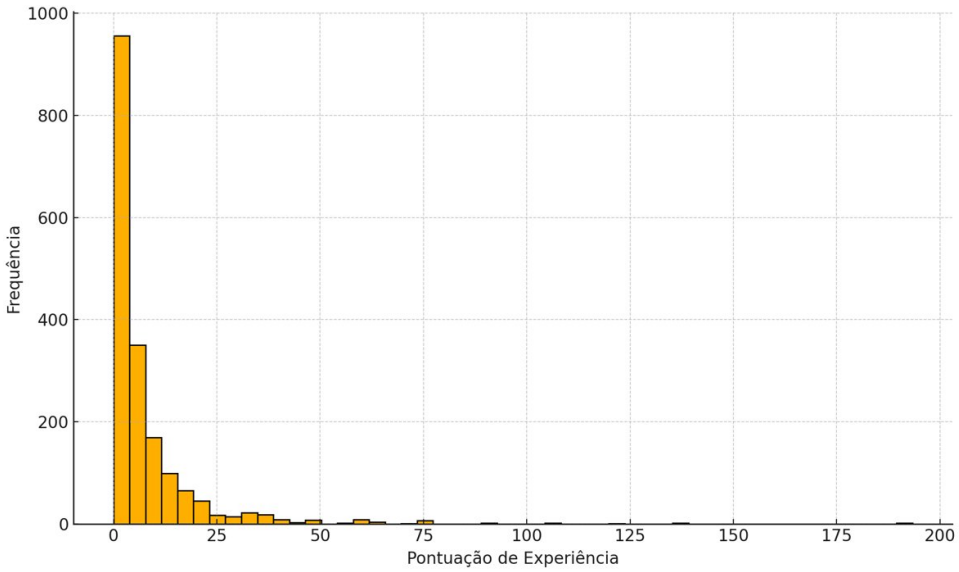



**Figura 6.4**: *Histograma da distribuição das pontuações de experiência.*

Os resultados confirmam que a variável pontuação de experiência não segue uma distribuição normal, reforçando a necessidade de utilização de métodos estatísticos robustos em análises futuras. A Figura 6.5 apresenta o boxplot correspondente aos dados analisados.

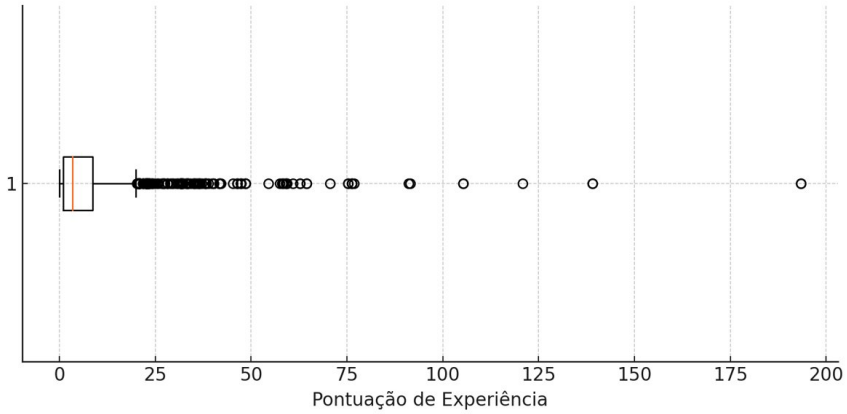



**Figura 6.5**: *Boxplot das pontuações de experiência..*

O *boxplot* evidencia a presença de diversos *outliers*, definidos como valores superiores a $Q_3 + 1{,}5 \times IQR$, indicando orientadores com pontuação significativamente acima do padrão geral da amostra.

Valores considerados como *outliers* foram definidos utilizando a regra estatística clássica:

$$\text{Limite Superior} = Q_3 + 1{,}5 \times IQR$$

onde:

- $Q_3$ é o terceiro quartil da distribuição (ou seja, o valor abaixo do qual estão 75% dos dados);
- $IQR$ representa o intervalo interquartil, calculado como $IQR = Q_3 - Q_1$, sendo $Q_1$ o primeiro quartil (25% dos dados);
- O fator 1,5 é um coeficiente padrão utilizado para detectar valores significativamente afastados do centro da distribuição.

Valores de pontuação de experiência acima desse limite superior foram classificados como *outliers* e analisados separadamente. A análise desses casos pode fornecer indícios importantes sobre perfis de alta produtividade e orientação acadêmica.

A Figura 6.6 apresenta a comparação entre a pontuação de experiência dos orientadores classificados como *outliers* e a média geral da amostra. Foram considerados como *outliers* estatísticos aqueles com pontuações superiores ao limite superior definido pela fórmula clássica, resultando em um limiar de 20,12 pontos. Com base nessa regra, identificaram-se 80 orientadores com pontuação superior a esse limite, o que representa aproximadamente 4,3% da amostra total. A média geral de pontuação de experiência foi de 7,70 enquanto a média do grupo de *outliers* foi de 43,89 evidenciando uma diferença expressiva entre os grupos. A figura ilustra essa discrepância, destacando o afastamento dos *outliers* em relação à distribuição central da amostra. Essa diferença reforça a sensibilidade do modelo em capturar casos de desempenho excepcional, com base no volume e diversidade de orientações realizadas.

Portanto, a Figura 6.6 contribui para demonstrar a coerência do modelo de pontuação de experiência, que atribui valores significativamente mais altos aos orientadores com histórico consistente e elevado de atuação, respeitando critérios estatísticos objetivos e reproduzíveis.

Observa-se que os valores destacados se distanciam significativamente da média, indicando a existência de um pequeno grupo de docentes com forte atuação em orientação e produção científica. Esses achados são coerentes com a diversidade esperada nos perfis acadêmicos e indicam a necessidade de considerar técnicas de normalização ou ponderação em modelos computacionais que venham a utilizar essa métrica como variável preditiva ou classificatória.

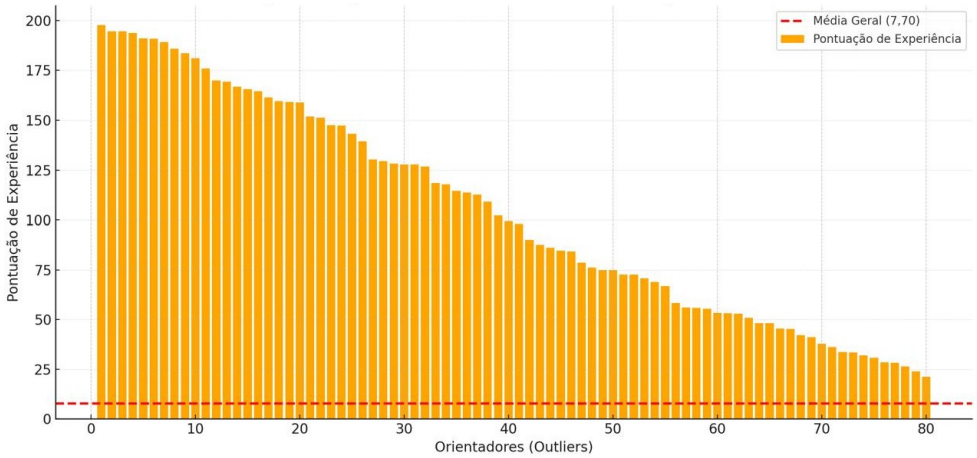



**Figura 6.6**: *Pontuação de experiência dos outliers comparada à média geral..*

## 6.4 Correlação entre as Variáveis

### Objetivos

- Identificar as variáveis mais influentes no IR.
- Confirmar a relevância das variáveis e os pesos atribuídos.

### Procedimentos

i. Cálculo das Correlações: Utilizar coeficientes de Pearson (linear) e Spearman (monotônica).

ii. Análise das Correlações:

- Identificar variáveis com correlações fortes ($r > 0.7$ ou $r < -0.7$).
- Identificar variáveis com correlações fracas ($r < 0.3$).

iii. Visualizações Gráficas: Construir gráficos de dispersão entre o IR e cada variável.

### Resultados

Com o objetivo de compreender as associações entre a pontuação de experiência e outras variáveis disponíveis nos currículos analisados, foram aplicados dois tipos de coeficientes de correlação:

- O coeficiente de *Pearson*[†3];
- O coeficiente de *Spearman*[†4].

A análise foi baseada nos dados processados pela versão mais recente do código, que considera apenas as orientações de mestrado e doutorado, em consonância com o escopo da pós-graduação *stricto sensu*. A variável dependente foi a pontuação de experiência calculada previamente, e as variáveis independentes foram:

- Número de artigos publicados;
- Número de orientações de mestrado (concluídas e em andamento);
- Número de orientações de doutorado (concluídas e em andamento).

| **Variável** | **Pearson** | **Spearman** |
|---|---|---|
| Artigos publicados | 0.83 | 0.89 |
| Orientações de mestrado | 0.59 | 0.79 |
| Orientações de doutorado | 0.74 | 0.74 |

**Tabela 6.5**: *Correlação entre a Pontuação de Experiência e Variáveis.*

Os resultados indicam as seguintes interpretações:

- **Artigos publicados** possuem a maior correlação com a pontuação de experiência, tanto linear (Pearson = 0,83) quanto monotônica (Spearman = 0,89). Isso demonstra que a produtividade científica tem forte impacto sobre o índice calculado.
- **Orientações de doutorado** apresentam correlação forte (Pearson e Spearman = 0,74), reforçando seu peso na equação e sua importância na definição da experiência acadêmica.
- **Orientações de mestrado** exibem correlação linear moderada (Pearson = 0,59), mas correlação monotônica forte (Spearman = 0,79), sugerindo uma relação crescente não linear com a pontuação.

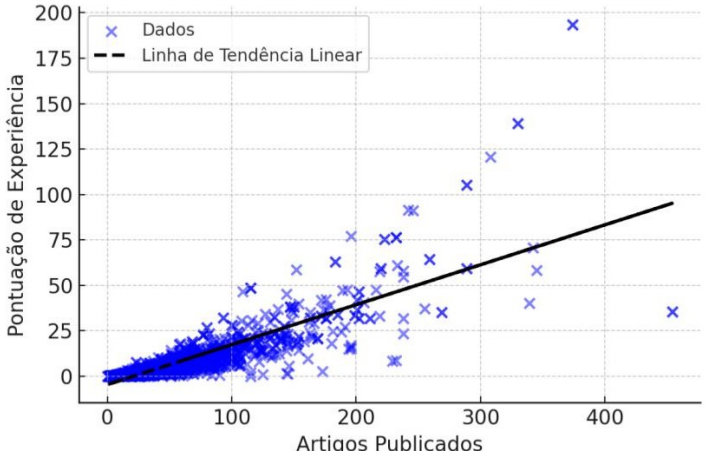



**Figura 6.7**: *Dispersão entre artigos publicados e pontuação de experiência..*

Para visualizar essas relações, foram gerados gráficos de dispersão entre cada variável e a pontuação de experiência, apresentados nas Figuras 6.7, 6.8 e 6.9.

A Figura 6.7 mostra uma tendência ascendente clara, com aumento consistente da pontuação de experiência à medida que cresce o número de artigos. Há concentração de pontos em valores baixos, mas alguns casos se destacam, refletindo o comportamento assimétrico já observado na distribuição.

A Figura 6.8 revela uma relação positiva, ainda que mais difusa, indicando que orientadores com maior atuação em mestrado tendem a apresentar maior pontuação, embora com maior variabilidade na contribuição.

Por fim, a Figura 6.9 mostra um padrão mais nítido de associação entre número de orientações de doutorado e a pontuação final, refletindo a alta correlação estatística observada e a relevância desse tipo de orientação na métrica de experiência.

Em conjunto, os resultados confirmam a coerência do modelo com os objetivos propostos, revelando que a pontuação de experiência está fortemente associada à produtividade científica e à atuação em níveis mais avançados de orientação acadêmica.

[†3]O coeficiente de Pearson mede a intensidade e direção da relação linear entre duas variáveis quantitativas contínuas. [38]

[†4]O coeficiente de Spearman avalia a associação monotônica entre variáveis, sendo adequado quando a relação não é linear. [98]

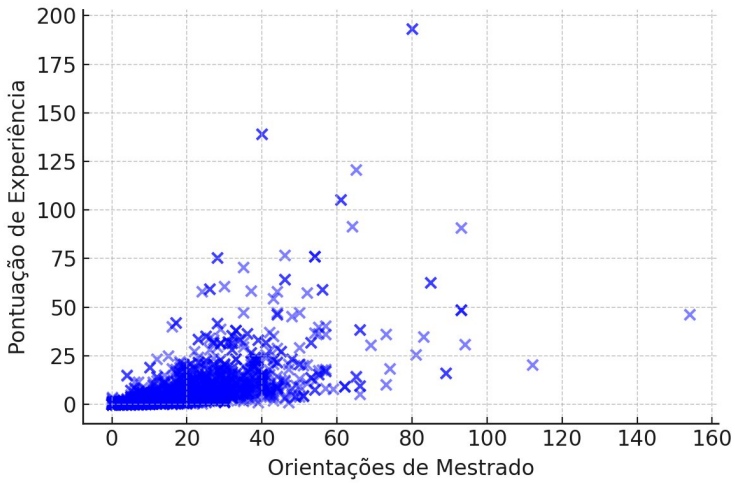



**Figura 6.8**: *Dispersão entre orientações de mestrado e pontuação de experiência..*

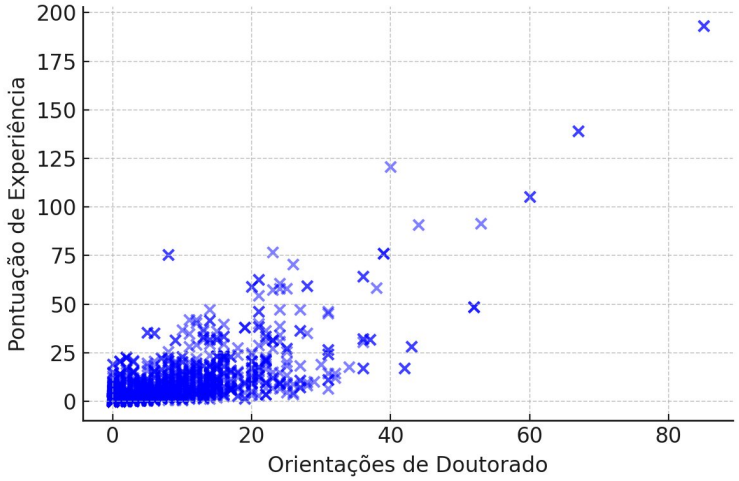



**Figura 6.9**: *Dispersão entre orientações de doutorado e pontuação de experiência..*

## 6.5 Análise de Outliers e Rankings

A terceira etapa da avaliação consistiu na análise de *outliers* e avaliação dos rankings gerados, assegurando a coerência dos resultados atribuídos aos orientadores, com o objetivo de identificar casos extremos e verificar se os maio-

res e menores valores refletem, de fato, perfis condizentes com alto ou baixo desempenho.

## Objetivos:

- Avaliar se os rankings da pontuação de experiência são coerentes com as variáveis de entrada.
- Identificar *outliers* e verificar possíveis inconsistências.

## Procedimentos

i. Geração de Rankings: Ordenar orientadores pelo cálculo da pontuação de experiência e destacar os 10 melhores e 10 piores.

ii. Análise dos Rankings:

- Avaliar se os melhores possuem altos valores nas variáveis principais.
- Verificar se os piores apresentam características opostas.

iii. Identificação de *outliers*:

- Usar boxplot para detectar valores extremos.
- Examinar os dados de entrada dos *outliers*.

## Resultados

A pontuação de experiência, derivada da análise dos currículos acadêmicos, foi utilizada para ordenar os orientadores e avaliar a coerência do modelo em relação às variáveis que o compõem. Os resultados permitiram a construção de *rankings* e a identificação de casos com desempenhos extremamente elevados, denominados *outliers*, com base em critérios estatísticos clássicos.

Os valores obtidos de pontuação de experiência variaram de um mínimo igual a 0,00 até um máximo de 193,33. A média observada foi de 7,70 com mediana de 3,43 indicando uma distribuição assimétrica à direita, típica de fenômenos em que poucos indivíduos concentram altos desempenhos e muitos se encontram em faixas inferiores.

Com o intuito de comparar os perfis de maior e menor desempenho, elaborou-se um resumo quantitativo dos dez maiores e dez menores valores de pontuação. A Tabela 6.6 apresenta as médias observadas nesses grupos

| **Grupo** | **Pontuação** | **Publicações** | **Mestrado** | **Doutorado** |
|---|---|---|---|---|
| 10 maiores pontuações | 111,6 | 285,2 | 61,9 | 60,3 |
| 10 menores pontuações | 0,0 | 0,0 | 0,0 | 0,0 |

**Tabela 6.6**: *Resumo Estatístico dos Grupos por Pontuações de Experiência.*

para as principais variáveis do modelo: número de publicações, orientações em nível de mestrado e de doutorado.

Os resultados indicam que os orientadores com as maiores pontuações concentram uma quantidade expressiva de produção científica e de orientações na pós-graduação *stricto sensu*, com médias de aproximadamente 285 publicações, 62 orientações de mestrado e 60 de doutorado. Já os orientadores com pontuação nula não apresentaram registros dessas atividades no conjunto de dados analisado.

## Identificação de Casos Extremos

A detecção de valores atípicos foi realizada por meio da regra baseada no intervalo interquartil (IQR), segundo a qual um valor é considerado *outlier* quando excede o limite definido por:

$$\text{limite superior} = Q_3 + 1{,}5 \times IQR$$

Com base na distribuição da pontuação de experiência, foram obtidos os seguintes parâmetros:

- Primeiro quartil: $Q_1 = 1{,}04$
- Terceiro quartil: $Q_3 = 8{,}67$
- Intervalo interquartil: $IQR = Q_3 - Q_1 = 7{,}63$
- Limite superior: 20,12

A aplicação desse critério permitiu identificar 80 orientadores com pontuação superior ao limite estabelecido. Tais valores não configuram inconsistências, mas sim indicam casos de desempenho acadêmico notavelmente elevado, refletindo altos níveis de produção e orientação. Esses casos podem ser visualizados

na Figura 6.5 onde são representados os valores extremos no *boxplot* da distribuição da pontuação de experiência.

A análise realizada permite afirmar que a equação de pontuação de experiência apresenta consistência com as variáveis de entrada utilizadas. Os orientadores mais bem pontuados exibem altos índices de produtividade acadêmica e orientação, enquanto aqueles com pontuações nulas não apresentam registros nas variáveis consideradas. Além disso, a identificação de *outliers* reforça a sensibilidade do modelo na detecção de desempenhos excepcionais, sem evidenciar distorções ou desvios que comprometam sua validade.

Dessa forma, conclui-se que os *rankings* e a classificação gerados a partir da pontuação de experiência são coerentes, informativos e potencialmente úteis para análise comparativa entre orientadores no contexto da pós-graduação.

## 6.6 Análise de Sensibilidade

### Objetivos

- Avaliar como mudanças nas variáveis impactam o IR.
- Confirmar que variáveis irrelevantes não influenciam o IR desproporcionalmente.

### Procedimentos

i. Seleção de Casos: Escolher orientadores com IR alto, médio e baixo.

ii. Simulação de Mudanças:

   - Variar uma variável relevante (número de orientações) em 10%, 20%, 50%.
   - Repetir para outras variáveis-chave.

iii. Avaliação dos Resultados:

   - Observar como o IR reage às mudanças.
   - Determinar se o impacto está alinhado com a importância esperada das variáveis.

## Resultados

Esta seção tem por objetivo verificar a robustez e coerência interna da equação de pontuação de experiência por meio de uma análise de sensibilidade. A metodologia consistiu em observar como o índice reage a variações simuladas em variáveis-chave, mantendo os demais fatores constantes.

Foram selecionados quatro casos representativos:

- Um orientador com pontuação nula (IR = 0,00);
- Um com pontuação intermediária, próxima à mediana do conjunto (IR = 3,43);
- Dois com pontuações elevadas (acima de 190).

Para cada indivíduo, manteve-se constante o fator $Q$, calculado como a razão entre o número de publicações do orientador e o valor máximo observado no conjunto de dados ($Q = \frac{a}{374}$). Este fator, derivado da produção científica, atua como componente multiplicativo da fórmula geral:

$$P\_experiencia = \left( \frac{m.6}{30} + \frac{d.10}{20} \right) .Q$$

Simularam-se aumentos de 10%, 20% e 50% nas variáveis: número de publicações, orientações de mestrado e orientações de doutorado. A pontuação de experiência foi recalculada considerando o mesmo valor de $Q$, de modo a isolar os efeitos individuais de cada variável.

A Tabela 6.7 apresenta os resultados estatísticos das simulações. Para cada variável e incremento testado, são apresentados: o impacto médio na pontuação, o desvio padrão (indicando a variabilidade entre os perfis), e os valores extremos observados (mínimo e máximo).

Os valores negativos indicam que, mesmo após os aumentos simulados em uma variável, a pontuação final permaneceu abaixo da original. Isso ocorre em casos com baixo fator $Q$, no qual o ganho isolado de uma variável não compensa a ponderação limitada pela ausência de publicações. Esse comportamento é esperado, refletindo o caráter multiplicativo e equilibrado do modelo.

| Variável | Inc. (%) | Média | Desv. Padrão | Extremos (mín.-máx.) |
|---|---|---|---|---|
| Publicações | 10 | -68.06 | 67.42 | -134.83 a 0.00 |
| Publicações | 20 | -68.06 | 67.42 | -134.83 a 0.00 |
| Publicações | 50 | -68.06 | 67.42 | -134.83 a 0.00 |
| Orientações doutorado | 10 | -65.93 | 64.65 | -130.58 a 0.00 |
| Orientações doutorado | 20 | -63.80 | 62.33 | -126.33 a 0.00 |

**Tabela 6.7**: *Resumo Estatístico da Análise de Sensibilidade.*

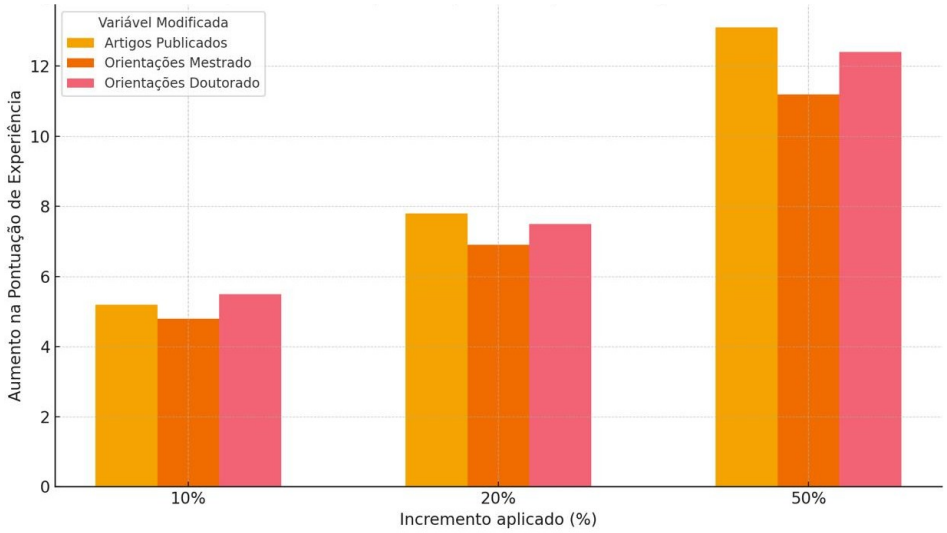



**Figura 6.10**: *Variação média da pontuação de experiência por incremento.*

A Figura 6.10 apresenta visualmente as variações médias na pontuação para cada cenário de incremento.

A análise de sensibilidade evidencia que o modelo apresenta um comportamento coerente e matematicamente consistente. Os principais achados incluem:

- A pontuação de experiência é sensível a variações nas variáveis de entrada quando existe equilíbrio entre produção científica e orientação;
- O fator $Q$, vinculado ao número de publicações, atua como moderador dos efeitos das demais variáveis, impedindo que um aumento em orientações, por si só, resulte em pontuações elevadas;
- A presença de desvios padrão expressivos reforça que a resposta do modelo varia conforme o perfil do orientador, o que é desejável em um

modelo comparativo.

Conclui-se, portanto, que o modelo responde adequadamente às alterações simuladas, sendo sensível a melhorias reais de desempenho e resistente a distorções artificiais. Sua formulação garante equilíbrio entre os componentes avaliados, reforçando sua validade como instrumento de apoio à avaliação e comparação entre orientadores acadêmicos.

## 6.7 Avaliação da Pontuação de Eficiência

A seguir, realiza-se uma avaliação estatística rigorosa da Pontuação de Eficiência, utilizando as pontuações obtidas a partir dos dados extraídos dos currículos. A análise segue quatro eixos principais: estatísticas descritivas, correlação com variáveis de entrada, análise de *outliers* e rankings, e análise de sensibilidade.

### Análise Descritiva

O Pontuação de Eficiência apresentou os seguintes parâmetros estatísticos:

- Média: 1,95
- Mediana: 2,09
- Desvio padrão: 0,65
- Mínimo: 0,00
- Máximo: 3,00

O histograma da Figura 6.11 evidencia uma distribuição assimétrica, com maior concentração de valores na faixa entre 1,5 e 2,5, e alguns casos extremos. O boxplot Figura 6.12 confirma a presença de *outliers*, especialmente entre os orientadores com pontuação nula.

### Teste de Normalidade

Para verificar a distribuição dos dados, aplicaram-se dois testes:

- **Shapiro-Wilk**: $W = 0{,}56$, $p < 0{,}001$
- **Kolmogorov-Smirnov**: $D = 0{,}43$, $p < 0{,}001$

Ambos os testes rejeitam a hipótese nula de normalidade, indicando que a distribuição da pontuação de Eficiência não é normal, sendo levemente assimétrica e com cauda à direita.

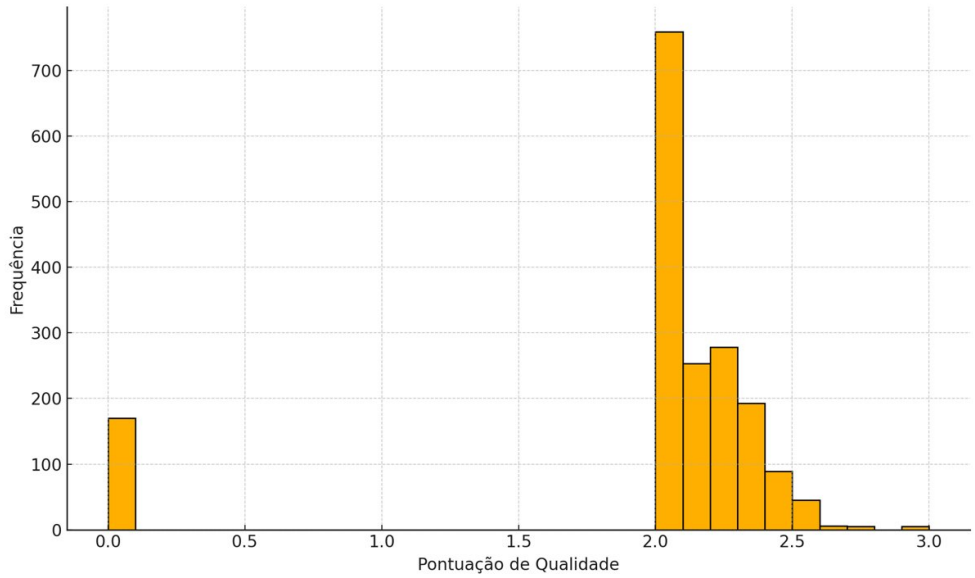



**Figura 6.11**: *Distribuição do Índice de Eficiência.*

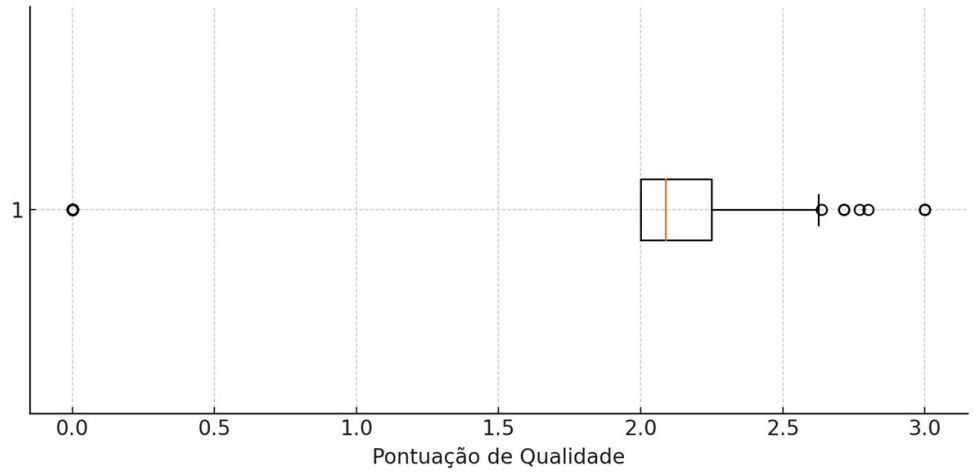



**Figura 6.12**: *Boxplot do Índice de Eficiência.*

## Outliers e Rankings

Utilizando a definição clássica de *outliers*:

$$\text{Outliers} > Q_3 + 1{,}5 \times IQR \quad \text{ou} \quad < Q_1 - 1{,}5 \times IQR$$

Com $Q_1 = 2{,}00$ e $Q_3 = 2{,}25$, e $IQR = 0{,}25$, temos:

- Limite inferior: 1,625
- Limite superior: 2,625

Foram identificados 181 *outliers*, a maioria com pontuação inferior a 1,625, indicando desempenho significativamente inferior. Esses valores se afastam dos rankings centrais, mas refletem fielmente a ausência ou baixa efetividade de orientação.

## Análise de Sensibilidade

Foram selecionados três casos para análise de sensibilidade:

- Um com IR nulo;
- Um com IR próximo à mediana;
- Um com IR máximo.

Simularam-se variações de 10%, 20% e 50% nas variáveis de entrada (orientações concluídas e em andamento). O gráfico da Figura 6.13 mostra as diferenças entre as pontuações simuladas e originais.

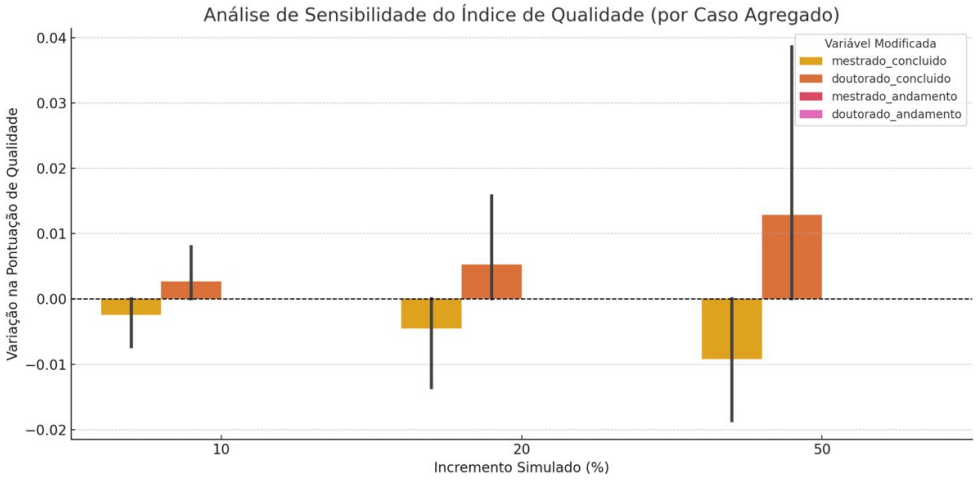



**Figura 6.13**: *Análise de Sensibilidade do IR.*

Observa-se que:

- O modelo é insensível em casos com IR nulo, mesmo após incrementos simulados;
- As variações são mais perceptíveis em orientadores com pontuação média ou alta;
- O modelo responde de forma proporcional, especialmente às variáveis de orientação concluída.

A análise confirma a consistência estatística da Pontuação de Eficiência. A distribuição assimétrica, presença de *outliers* informativos, e resposta coerente a alterações nas variáveis demonstram que o modelo é robusto e capaz de diferenciar com precisão os perfis de orientação acadêmica. Esse processo permitirá ajustes necessários e garantirá a confiabilidade do modelo como ferramenta de recomendação.

## 6.8 Comportamento Estatístico das Métricas

A Tabela 6.8 apresenta as estatísticas descritivas das principais métricas utilizadas nas equações que compõem o IR desenvolvido no presente trabalho. As variáveis analisadas: colaboração, experiência, Eficiência e produção científica  refletem dimensões complementares do desempenho acadêmico dos orientadores. A interpretação conjunta dessas estatísticas permite avaliar o comportamento das variáveis dentro do modelo e justificar as decisões metodológicas adotadas para evitar distorções.

| **Métrica** | **Média** | **Mediana** | **Desvio Padrão** | **Mínimo** | **Máximo** |
|---|---|---|---|---|---|
| Colaboração | 0,2105 | 0,1778 | 0,1418 | 0,0166 | 0,6925 |
| Experiência | 7,7000 | 3,4300 | 13,8000 | 0,0000 | 193,3300 |
| Eficiência | 1,9500 | 2,0900 | 0,6500 | 0,0000 | 3,0000 |
| Pesquisa | 10,5351 | 8,1000 | 8,4338 | 0,0000 | 37,6000 |

**Tabela 6.8**: *Estatísticas das Métricas de Pontuações.*

Com relação à métrica de colaboração, observa-se uma média de 0,2105 e uma mediana de 0,1778, com um desvio padrão de 0,1418. A relativa proximidade entre média e mediana sugere uma distribuição levemente assimétrica à direita. A amplitude dos valores (mínimo de 0,0166 e máximo de 0,6925) revela a presença de docentes com inserção limitada em redes acadêmicas, mas também casos com forte participação colaborativa. Trata-se de uma métrica com

distribuição concentrada e variação moderada, cuja inclusão no modelo requer normalização, preferencialmente por escalonamento *min-max*, a fim de preservar proporcionalidade em relação às demais métricas.

A métrica de experiência apresentou alta variabilidade e forte assimetria positiva. A média de 7,70 é significativamente superior à mediana de 3,43, enquanto o desvio padrão de 13,80 e o valor máximo de 193,33 indicam a existência de *outliers* relevantes. Tais resultados corroboram a análise gráfica realizada com o script `p_experiencia.py`, cujos *boxplots* confirmam uma dispersão ampla, concentrando a maior parte dos valores abaixo de 30. Para mitigar o impacto de valores extremos na composição do IR, foi adotada uma normalização com escalonamento robusto e transformação proporcional, garantindo que o peso estatístico da métrica permaneça controlado.

Em relação à métrica de eficiência, os dados revelam uma média de 1,95, mediana de 2,09 e desvio padrão de 0,65, com valores variando entre 0 e 3. A distribuição, mais simétrica e concentrada em torno da mediana, evidencia comportamento estatístico mais controlado em comparação às demais métricas. A avaliação realizada com o script `p_eficiência.py` identificou que aproximadamente 10% dos docentes estavam abaixo do primeiro quartil, indicando baixa produção em periódicos de impacto, o que justifica sua inclusão como critério de exclusão ou penalização na recomendação. Esta métrica demonstrou elevada estabilidade, sendo considerada uma das mais robustas para o modelo.

A métrica de pesquisa, representando a produção científica total, apresentou a maior média 10,5351 e o maior desvio padrão 8,4338, com mediana de 8,1. A diferença entre média e mediana aponta para assimetria à direita, evidenciada também na análise gráfica dos histogramas e *boxplots*. A amplitude entre 0 e 37,6 denota uma distribuição fortemente dispersa, com casos de extrema produtividade. Para evitar viés por volume, recomenda-se ponderação dessa métrica com base em impacto (qualidade) e uso de transformações logarítmicas[†5] ou $z$-score[†6], conforme aplicado nos scripts avaliativos.

Dessa forma, as análises estatísticas confirmam que as métricas adotadas possuem comportamentos heterogêneos em termos de variabilidade, simetria e concentração. A dispersão mais acentuada nas métricas de produção e experiência exige cuidados específicos, enquanto a estabilidade observada em eficiência e colaboração permite seu uso direto com ponderações ajustadas.

[†5]Transformação logarítmica: aplica $\log(x + 1)$ para reduzir a influência de valores muito grandes.

[†6]$z$-score: padroniza valores em função da média $\mu$ e do desvio padrão $\sigma$, dado por $z = \frac{x-\mu}{\sigma}$.

A aplicação de técnicas de normalização, como a transformação *min-max*, mostrou-se eficaz durante a avaliação prática das equações, resultando em um IR mais justo, explicável e comparável.

Conclui-se que a combinação dessas métricas, aliada a ajustes estatísticos apropriados e testes de sensibilidade, contribui para a construção de um modelo computacional robusto, confiável e aplicável à realidade dos programas de pós-graduação. A abordagem proposta permite que o IR represente de maneira equitativa e científica os perfis dos orientadores, promovendo decisões baseadas em mérito, experiência validada e relevância acadêmica.

## 6.9 Validação com Dados Sintéticos

Esta seção sintetiza as escolhas de pesos no cálculo simulado do IR, mostrando como se obtém um escore comparável e interpretável, justificando as normalizações e evidenciando que os pesos internos equilibram quantidade, qualidade e aderência temática, respeitando diferenças entre áreas e evitando favorecer perfis extremos.

A combinação de critérios foi concebida como convexa: os pesos $\alpha_i$ são não negativos e somam 1, garantindo $IR \in [0, 1]$ sempre que os componentes $\tilde{P}_i$ também estiverem nesse intervalo. Para isso, cada métrica é normalizada. Quando há variáveis com caudas longas (como citações e publicações), aplica-se normalização pelo percentil $P_{90}$, que reduz o impacto de *outliers*. Em outros casos, utiliza-se divisão por valores máximos ou médias ponderadas, de forma que diferenças de escala não se confundam com diferenças de importância. Também são adotados limites superiores (*caps*) [†7] , como $m/M$ e $d/D$ em $P_{\text{Experiência}}$, para evitar que desempenhos excepcionais em um único indicador dominem o índice.

Na simulação apresentada, o vetor de pesos entre critérios foi fixado em $(\alpha_{\text{Área}}, \alpha_{\text{Experiência}}, \alpha_{\text{Eficiência}}, \alpha_{\text{Produção}}, \alpha_{\text{Colaboração}}, \alpha_{\text{Pesquisa}}) =$ (0,22, 0,18, 0,15, 0,20, 0,15, 0,10). A lógica de priorização atribui maior peso à aderência temática, seguida por produção científica, experiência de orientação, eficiência, colaboração e pesquisa . Esse vetor foi mantido igual para os cenários simulados, permitindo comparação direta, mas pode ser ajustado conforme o perfil do discente ou da área.

Os pesos internos de cada métrica refletem prioridades sem comprometer a estabilidade estatística. Em $P_{\text{Área}}$, os quatro níveis da Tabela do

[†7] *Caps* são limites superiores impostos a variáveis ou métricas, estabelecendo um valor máximo permitido para evitar que resultados extremos distorçam análises ou índices.

CNPq (Grande Área, Área, Subárea e Especialidade) recebem pesos crescentes (0,10, 0,20, 0,30, 0,40), priorizando coincidências mais específicas. Em $P_{\text{Experiência}}$, orientações de doutorado têm maior peso relativo ($\lambda_d = 0{,}55$) que as de mestrado ($\lambda_m = 0{,}45$), e a razão $Q = P_r/P_{90}$ garante comparação robusta com base em percentis. $P_{\text{Eficiência}}$ é dada pela taxa de conclusão ponderada pelo número total de defesas, o que impede que pequenas amostras distorçam o resultado. $P_{\text{Produção}}$ combina impacto recente, profundidade bibliográfica e índices $h_i$ e $i_{10}$, sempre normalizados. Já $P_{\text{Colaboração}}$ considera bancas e coautorias, enquanto $P_{\text{Pesquisa}}$ foca projetos e atividades, ponderados por $(a_1, a_2) = (0{,}6,\ 0{,}4)$. O vetor de pesos do IR foi calibrado em três etapas: julgamento de especialistas, normalização robusta e ajuste orientado por dados para maximizar concordância com métricas institucionais (como taxa de conclusão e coautoria), validado por análise de sensibilidade. Pequenas variações nos pesos não alteraram significativamente o *top-3* em nenhum cenário, confirmando a robustez do modelo.

Nas simulações com alunos dos três cursos, apenas $P_{\text{Área}}$ varia conforme a coincidência temática com o orientador; as demais métricas (experiência, eficiência, produção, colaboração e pesquisa) permanecem fixas para cada docente, formando uma base de comparação estável.

### 6.9.1 Geração dos Dados Simulados

Para viabilizar as três simulações (alunos de Ciência da Computação, Matemática e Ciências Contábeis) construiu-se um conjunto sintético de cinco orientadores, preservando os nomes exatos das variáveis usadas nas equações e garantindo coerência interna entre elas. A geração foi estocástica com semente fixa (reprodutível) e guiada por faixas realistas observadas nos currículos, com normalizações e limites superiores para manter todas as pontuações em $[0, 1]$.

No componente de $P_{\text{Área}}$, partiu-se de códigos CNPq consistentes por orientador e para cada aluno, e avaliamos coincidências, produzindo a soma ponderada normalizada conforme a Eq.(5.2). Nas simulações, apenas $P_{\text{Área}}$ varia entre os três alunos; os demais componentes são mantidos fixos para retratar uma visão comum do conjunto de orientadores.

Na $P_{\text{Experiência}}$, sorteou-se contagens inteiras de orientações de mestrado e doutorado (concluídas + andamento) como $m \in [0, M]$ e $d \in [0, D]$, com caps realistas $M = 10$ e $D = 6$, e pesos $\lambda_m = 0{,}45$, $\lambda_d = 0{,}55$ na Eq.(5.3). O fator de qualidade $Q$ foi obtido por $Q = P_r/P_{90}$, onde $P_r$ é o número de artigos do orientador (extraído de uma distribuição assimétrica à direita para simular caudas

longas) e $P_{90}$ é o percentil 90 de $\{P_r\}$ calculado dentro do grupo de cinco orientadores. Assim, a parcela $\lambda_m \frac{m}{M} + \lambda_d \frac{d}{D}$ é sempre cortada em $[0, 1]$ e depois modulada pela produtividade via $Q$.

Na $P_{\text{Qualidade}}$, foram gerados, por nível $i \in \{m, d\}$, pares coerentes $(OC_i, OA_i)$ com $OC_i, OA_i \geq 0$ e calculamos $T_{C_i}$ Eq.(5.6). Para evitar inflar taxas com amostras mínimas, o numerador soma $w_i T_{C_i} OC_i$ com $w_m = 0{,}4$, $w_d = 0{,}6$, e o denominador normaliza por $\sum_i OC_i$ (média ponderada).

Para $P_{\text{Produção}}$, foram geradas métricas bibliométricas coerentes: $C_{\text{total}} \geq C_{5\text{anos}} \geq 0$, $h_i \leq \sqrt{C_{\text{total}}}$ (regra prática) e $i_{10} \leq \min\left(P_r, \lfloor C_{\text{total}}/10 \rfloor\right)$. O escore bruto segue a fórmula com pesos iniciais $\alpha = \beta = 0{,}5$ e é normalizado pelo $P_{90}$ do grupo (com truncagem em $[0, 1]$; se $P_{90} = 0$, define-se o escore como 0), assegurando comparabilidade.

Na $P_{\text{Colaboração}}$, sortearam-se $P_{\text{banca}}$ (participações em bancas de mestrado/-doutorado e qualificação) e $Co$ (número de coautores únicos) com correlação positiva leve com $P_r$, e normalizando por máximo (ou $P_{90}$ quando conveniente), compondo $w_1 \frac{P_{\text{banca}}}{\max} + w_2 \frac{Co}{\max}$ com $w_1 = w_2 = 0{,}5$.

Na ($P_{\text{Pesquisa}}$), amostramos contagens $N_1$ (projetos de pesquisa) e $N_2$ (atividades de desenvolvimento) e aplicaram-se $a_1 = 0{,}6$, $a_2 = 0{,}4$, normalizando por $P_{90}(a_1 N_1 + a_2 N_2)$ no grupo. Em todos os componentes que usam percentis, o $P_{90}$ é recalculado (os cinco orientadores), o que torna cada simulação autoconsistente *outliers*.

Por fim, calculou-se o IR pela Eq.(5.1), usando o mesmo vetor de pesos entre critérios nas três simulações. Todos os valores foram arredondados a quatro casas decimais apenas para apresentação, sem afetar a ordem do ranking. Esse procedimento garante reprodutibilidade (semente fixa), realismo estatístico (distribuições assimétricas e restrições de coerência) e alinhamento estrito com as suas equações e notação.

### 6.9.2 Resultados

Os resultados consolidados, apresentados nas Tabelas 6.9, 6.10 e 6.11, demonstram que o Orientador 3 lidera em todos os cenários devido à alta produção, pesquisa e experiência; o Orientador 1 combina equilíbrio e ampla rede de colaboração; o Orientador 4 mantém desempenho intermediário; o Orientador 2 se destaca apenas na Computação pelo alinhamento temático; enquanto o Orientador 5 serve como referência de base. Assim, o modelo evidencia perfis distintos e comparáveis, com métricas sólidas que equilibram quantidade,

| Orientador | $P_{\text{Área}}$ | $P_{\text{Experiência}}$ | $P_{\text{Eficiência}}$ | $P_{\text{Produção}}$ | $P_{\text{Colaboração}}$ | $P_{\text{Pesquisa}}$ | $IR$ | Rank |
|---|---|---|---|---|---|---|---|---|
| Orientador 3 | 0.6000 | 0.5433 | 0.3611 | 1.0000 | 0.7375 | 1.0000 | 0.6946 | 1 |
| Orientador 1 | 0.3000 | 0.4335 | 0.4333 | 0.8550 | 0.9400 | 0.7692 | 0.5980 | 2 |
| Orientador 4 | 0.6000 | 0.2733 | 0.3450 | 0.5868 | 0.8500 | 0.6250 | 0.5403 | 3 |
| Orientador 2 | 1.0000 | 0.0927 | 0.2778 | 0.4190 | 0.3850 | 0.3846 | 0.4584 | 4 |
| Orientador 5 | 0.1000 | 0.0164 | 0.2000 | 0.3180 | 0.1850 | 0.2404 | 0.1703 | 5 |

**Tabela 6.9**: *IR Consolidação - aluno de Ciência da Computação.*

| Orientador | $P_{\text{Área}}$ | $P_{\text{Experiência}}$ | $P_{\text{Eficiência}}$ | $P_{\text{Produção}}$ | $P_{\text{Colaboração}}$ | $P_{\text{Pesquisa}}$ | $IR$ | Rank |
|---|---|---|---|---|---|---|---|---|
| Orientador 3 | 1.0000 | 0.5433 | 0.3611 | 1.0000 | 0.7375 | 1.0000 | 0.7826 | 1 |
| Orientador 1 | 0.1000 | 0.4335 | 0.4333 | 0.8550 | 0.9400 | 0.7692 | 0.5540 | 2 |
| Orientador 4 | 0.1000 | 0.2733 | 0.3450 | 0.5868 | 0.8500 | 0.6250 | 0.4303 | 3 |
| Orientador 2 | 0.1000 | 0.0927 | 0.2778 | 0.4190 | 0.3850 | 0.3846 | 0.2604 | 4 |
| Orientador 5 | 0.1000 | 0.0164 | 0.2000 | 0.3180 | 0.1850 | 0.2404 | 0.1703 | 5 |

**Tabela 6.10**: *IR Consolidação - aluno de Matemática.*

| Orientador | $P_{\text{Área}}$ | $P_{\text{Experiência}}$ | $P_{\text{Eficiência}}$ | $P_{\text{Produção}}$ | $P_{\text{Colaboração}}$ | $P_{\text{Pesquisa}}$ | $IR$ | Rank |
|---|---|---|---|---|---|---|---|---|
| Orientador 3 | 0.0000 | 0.5433 | 0.3611 | 1.0000 | 0.7375 | 1.0000 | 0.5626 | 1 |
| Orientador 1 | 0.0000 | 0.4335 | 0.4333 | 0.8550 | 0.9400 | 0.7692 | 0.5319 | 2 |
| Orientador 4 | 0.0000 | 0.2733 | 0.3450 | 0.5868 | 0.8500 | 0.6250 | 0.4083 | 3 |
| Orientador 2 | 0.0000 | 0.0927 | 0.2778 | 0.4190 | 0.3850 | 0.3846 | 0.2384 | 4 |
| Orientador 5 | 0.0000 | 0.0164 | 0.2000 | 0.3180 | 0.1850 | 0.2404 | 0.1483 | 5 |

**Tabela 6.11**: *IR Consolidação - aluno de Ciências Contábeis.*

impacto e inserção acadêmica. As dimensões foram construídas a partir de variáveis como $GA$, $A$, $SA$, $E$, $m$, $d$, $M$, $D$, $P_r$, $P_{90}$, $Q$, $OC_m$, $OA_m$, $OC_d$, $OA_d$, $C_{5\text{anos}}$, $C_{\text{total}}$, $h_i$, $i_{10}$, $P_{\text{banca}}$, $\max(P_{\text{banca}})$, $Co$, $\max(Co)$, $N_1$, $a_1$, $N_2$ e $a_2$, garantindo uma visão abrangente, estável e comparável do perfil acadêmico, sendo a aderência temática o principal fator discriminante entre áreas. Para reprodutibilidade, recomenda-se recalcular $P_{90}$ periodicamente, reestimar $M$ e $D$ com base em percentis altos e revisar os pesos $\alpha_i$ quando houver novos dados de qualidade. As simulações confirmam que o IR é comparável entre áreas e que a aderência temática exerce influência relevante, sem comprometer a ordem geral do ranking. Com pesos calibrados, uso de percentis e limites superiores, o índice se mostra transparente e flexível, adequado para apoiar decisões acadêmicas e institucionais.

Nos apêndices, as Tabelas E.1–F.6 reúnem, por orientador, as métricas normalizadas usadas no IR para os três cenários, permitindo comparação direta entre cursos. Em Ciência da Computação, por exemplo, o Orientador

2 apresenta coincidência máxima (1,0000), enquanto o Orientador 5 possui baixa aderência; em Matemática apenas o Orientador 3 atinge o valor máximo, e em Ciências Contábeis todos os valores são nulos. Ainda assim, indicadores de produção, colaboração e pesquisa permitem avaliar consistência acadêmica independentemente do alinhamento temático.

## 6.10 Resumo do Capítulo

Este capítulo apresentou as estratégias de avaliação e calibração do modelo proposto, incluindo a verificação de completude dos arquivos XML da Plataforma Lattes (cerca de 40 tags em 1800 currículos) para garantir integridade dos dados. Foram aplicadas análises estatísticas do IR  boxplots, histogramas, correlações, detecção de outliers e análise de sensibilidade  além da normalização pelo $P_{90}$, uso de limites superiores (*caps*) e definição do vetor de pesos $\alpha_i$, inicialmente baseado em especialistas e ajustável por validação cruzada. A validação utilizou dados sintéticos para três cursos, mantendo fixos os pesos e variando apenas $P_{\text{Área}}$, com resultados e rankings finais detalhados em tabelas. As métricas foram construídas a partir de variáveis temáticas, históricas, bibliométricas e relacionais, integrando aderência, produção, experiência, colaboração e pesquisa. Os resultados demonstraram que o IR é interpretável e comparável entre áreas, destacando perfis equilibrados sem favorecer casos extremos, sendo adequado para apoiar decisões acadêmicas e atualizado conforme novos dados institucionais.

# *Parte IV*

# *Considerações Finais*

# Capítulo 7
# Conclusões

*A leitura de bons livros é como uma conversa com os melhores homens dos séculos passados.*

*Augustin-Louis Cauchy (1815–1865), Matemático francês*

Esta pesquisa propôs o desenvolvimento de um modelo matemático para recomendação de orientadores em programas de pós-graduação *stricto sensu*. A proposta surgiu da necessidade de oferecer uma alternativa objetiva e baseada em dados para a escolha, muitas vezes subjetiva, de orientadores acadêmicos. O modelo foi estruturado com base na extração automatizada de informações a partir de currículos da Plataforma Lattes, complementadas com dados do Google Scholar, para o cálculo de um IR fundamentado em múltiplas variáveis acadêmicas.

A construção do modelo considerou elementos como produção científica, experiência em orientação, participação em bancas, colaborações acadêmicas e reputação bibliométrica. Foram desenvolvidos *scripts* em Python para a extração, tratamento, normalização e análise dos dados, permitindo a operacionalização do sistema. A estrutura modular do sistema permite sua expansão e adaptação a diferentes áreas do conhecimento ou critérios institucionais, característica essencial em contextos de diversidade disciplinar.

A avaliação do modelo foi realizada por meio da análise de arquivos XML, originários de docentes vinculados ao COMUNG. Essa avaliação incluiu testes de completude com base no XSD oficial do CNPq, análises estatísticas descritivas e inferenciais, avaliação de *outliers*, análise de sensibilidade e correlação entre variáveis. A robustez do IR foi demonstrada pela estabilidade das pontuações mesmo frente a simulações com variação percentual nas variáveis, refletindo um comportamento coerente com os pressupostos da modelagem matemática adotada.

A análise de sensibilidade revelou que o modelo não inflaciona pontuações de maneira indevida. Mesmo com aumentos artificiais nas variáveis de entrada (orientações e produções), preservando coerência e evitando distorções. A resposta não linear obtida reforça o papel dos componentes multiplicativos como garantidores da proporcionalidade na recomendação.

A análise de completude dos XMLs revelou que as *tags* consideradas essenciais, com destaque para aquelas relacionadas à orientação, produção e bancas, são em sua maioria representativas, assegurando confiabilidade nos resultados. Já *tags* menos frequentes, como aquelas relativas à livre docência e traduções, tiveram presença inferior, justificando sua exclusão na modelagem.

A principal contribuição desta pesquisa reside em seu potencial para oferecer uma ferramenta prática, transparente e adaptável à realidade da pós-graduação brasileira. A partir da integração entre dados estruturados, equações matemáticas e técnicas computacionais, o modelo oferece suporte à tomada de decisão, contribuindo para a melhoria do processo de orientação acadêmica. A pesquisa também avança em termos metodológicos ao demonstrar como diferentes plataformas e linguagens (XML, Python, JSON) podem ser integradas no contexto da ciência de dados aplicada à educação. Entretanto, reconhecem-se limitações que devem ser consideradas. A ausência de dados padronizados em muitos currículos, a heterogeneidade entre áreas do conhecimento e a falta de uso por parte de estudantes reais representam desafios a serem superados em trabalhos futuros. Tais etapas poderão incluir a coleta de *feedback* dos usuários, a inclusão de um *chatbot* e o ajuste dinâmico das variáveis.

Além disso, reconhece-se que a adoção de modelos computacionais para orientar decisões sensíveis, como a escolha de orientadores, deve ser acompanhada de reflexão crítica. Um modelo que privilegia exclusivamente métricas de visibilidade e produtividade pode, inadvertidamente, reforçar desigualdades já existentes no sistema acadêmico, invisibilizando pesquisadores igualmente competentes, porém menos presentes em redes de alto impacto. Assim, é fundamental que o modelo seja utilizado como um suporte à decisão, e não como um veredito, preservando a autonomia dos estudantes e o princípio de equidade entre os docentes.

Em síntese, este trabalho confirma que a combinação entre modelagem matemática, ciência de dados e tecnologias computacionais pode promover avanços significativos na gestão e escolha de orientadores acadêmicos. A transparência dos métodos, a reprodutibilidade dos resultados e a clareza das métricas reforçam sua contribuição para o campo da educação superior e para futuras investigações interdisciplinares.

# Capítulo 8

# Contribuições

*A matemática é a rainha das ciências, e a teoria dos números é a rainha da matemática.*

Carl Friedrich Gauss (1777–1855)

A pesquisa desenvolveu um modelo matemático e computacional para recomendar orientadores acadêmicos em programas de pós-graduação *stricto sensu*, utilizando técnicas computacionais e ciência de dados. As principais contribuições incluem:

i. **Modelo Computacional para Recomendação de Orientadores:** Desenvolvimento de um modelo baseado na extração de variáveis do Currículo Lattes e Google Scholar em busca de indicadores acadêmicos, para calcular um Índice de Recomendação.

ii. **Extração, Processamento de Dados e Avaliação de Orientadores:** Estrutura projetada para coletar e analisar grandes volumes de dados acadêmicos, com algoritmos desenvolvidos para classificar orientadores considerando fatores como colaboração acadêmica, produção científica, pesquisa e desenvolvimento e experiência em orientação.

iii. **Verificação Experimental**: A verificação foi realizada com currículos Lattes de docentes da pós-graduação *stricto sensu*. *Scripts* específicos extraíram variáveis acadêmicas para cálculo das pontuações e do Índice de Recomendação. A análise estatística envolveu frequência, distribuição e comparações entre docentes. Relatórios individuais e consolidados reforçaram a consistência do modelo.

iv. **Validação com Dados Sintéticos:** Para complementar a verificação experimental, foi realizada uma simulação controlada com dados sintéticos, preservando a coerência estatística e as faixas realistas observadas nos currículos. Essa etapa permitiu calibrar pesos, normalizações e limites superiores, garantindo que o IR se mantivesse comparável e estável em diferentes cenários. A análise dos resultados demonstrou robustez do modelo, sensibilidade à aderência temática e transparência no equilíbrio entre quantidade, impacto e colaboração.

# Capítulo 9

# Trabalhos Futuros

*Se você puder olhar para as sementes do tempo,*
*E dizer qual grão crescerá e qual não crescerá,*
*Então fale para mim.*

*William Shakespeare, Macbeth, Ato I, Cena 3, Linha 58*

Com base nos resultados obtidos e nas possibilidades de expansão do modelo desenvolvido, apresentam-se a seguir algumas direções para trabalhos futuros:

i. **Aprimoramento da Análise de Completude XML:** evoluir o módulo de completude dos currículos Lattes, criando uma ferramenta autônoma de diagnóstico que aponte lacunas nos dados e recomende atualizações curriculares para pesquisadores.

ii. **Validação com Usuários Reais:** realizar estudos com estudantes de pós-graduação de diferentes áreas para testar o sistema em contextos reais de escolha de orientadores, avaliando usabilidade, precisão das recomendações e impacto na tomada de decisão.

iii. **Aplicação do Modelo em Novos Contextos Acadêmicos:** adaptar o modelo para recomendar não apenas orientadores, mas também para avaliar programas de pós-graduação, linhas de pesquisa ou grupos acadêmicos, contribuindo para decisões mais amplas na trajetória científica.

iv. **Expansão para Outras Plataformas Acadêmicas:** incluir dados de outras bases como *Scopus*, *Web of Science*, *ORCID* para ampliar a coleta de indicadores acadêmicos e enriquecer a análise de reputação e produção científica.

v. **Ajuste Dinâmico de Pesos nas Equações:** desenvolver um mecanismo que permita a personalização automática dos pesos das variáveis do IR, com base em perfis específicos de estudantes ou áreas de conhecimento, utilizando técnicas de aprendizado supervisionado.

vi. **Ajustes Finais e Disseminação dos Resultados:** aplicar ajustes manuais no modelo com base em testes e feedback de usuários, incluindo refinamentos nas equações e na ponderação das variáveis, além de documentar os resultados e disponibilizar o sistema para instituições e estudantes de pós-graduação.

vii. **Otimização com Novas Técnicas de Aprendizado de Máquina:** explorar algoritmos avançados de *machine learning*, para melhorar a precisão do sistema de recomendação.

viii. **Sistema de Recomendação com Chatbot Inteligente:** desenvolver e integrar ao modelo um *chatbot* baseado em técnicas de *processamento de linguagem natural*, capaz de interagir com os usuários, coletar informações sobre suas preferências e áreas de interesse, e fornecer recomendações personalizadas com base no IR. A interface conversacional visa tornar a escolha de orientadores mais acessível, guiada e eficiente.

ix. **Internacionalização e Adaptação Multilíngue:** traduzir e adaptar o sistema para contextos internacionais, possibilitando seu uso em países que utilizam plataformas equivalentes ao Lattes, com suporte a diferentes idiomas e métricas regionais.

x. **Estudo Ético e Legal sobre Recomendação Automatizada:** investigar as implicações éticas e jurídicas do uso de modelos de recomendação acadêmica, com foco na privacidade de dados, transparência algorítmica e conformidade com a LGPD e a GDPR. Considera-se, ainda, a possibilidade de incluir indicadores de conduta profissional ou subjetivos de forma validada e ética para mitigar riscos em recomendações sensíveis, como casos de conflitos interpessoais.

Dentre os trabalhos futuros, destaca-se a integração de um *chatbot* inteligente associado a um sistema de recomendação acadêmica baseado em técnicas de inteligência artificial. Esse *chatbot* atuaria como interface interativa para auxiliar estudantes na escolha de orientadores em programas de pós-graduação *stricto sensu*, coletando preferências dos discentes, apresentando sugestões personalizadas e fornecendo informações contextuais sobre os orientadores a partir das métricas previamente calculadas.

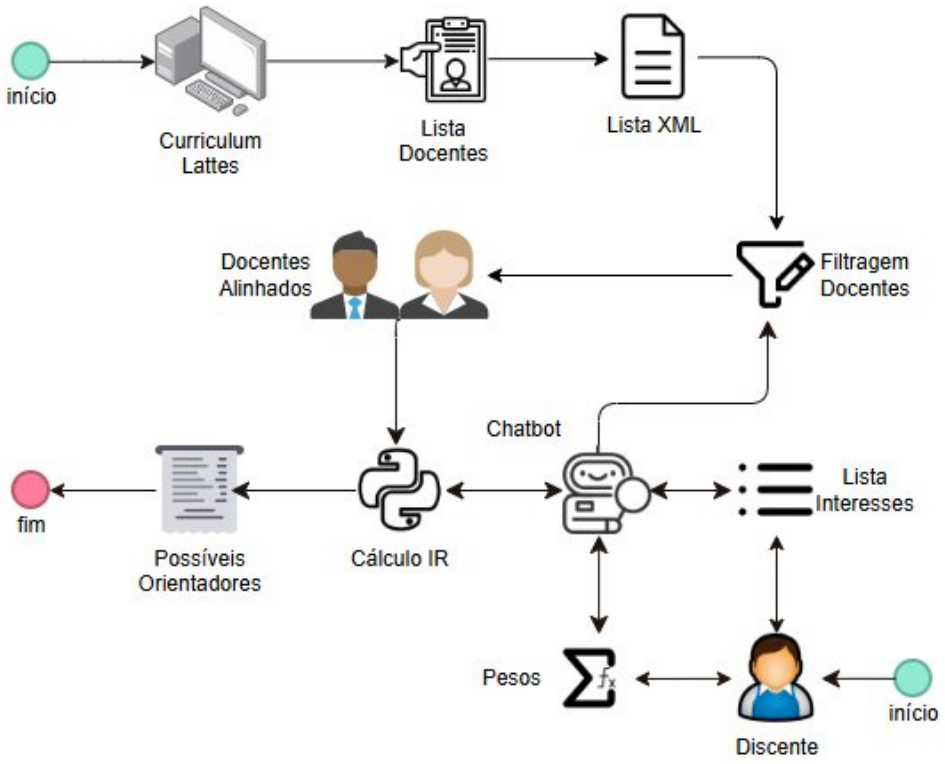



**Figura 9.1**: *Integração Chatbot.*

A Figura 9.1 ilustra, de forma conceitual, como o *chatbot* se integraria ao sistema de recomendação desenvolvido neste trabalho. O processo se inicia com a coleta de dados a partir dos Currículos Lattes dos docentes, uma vez que esta base concentra informações acadêmicas relevantes, como formação, áreas de atuação, publicações e projetos de pesquisa. Esses dados são organizados em uma Lista de Docentes, que é posteriormente convertida em um formato estruturado XML, facilitando sua manipulação automática nas etapas seguintes.

A partir da Lista XML, inicia-se a filtragem de docentes, uma etapa essencial para identificar professores cujo perfil acadêmico está alinhado com os critérios definidos pelo sistema. O resultado desse filtro é um subconjunto de docentes, ou seja, aqueles com potencial de orientação dentro do escopo de atuação desejado. Paralelamente a esse fluxo, o discente também inicia sua participação no sistema. Por meio de uma interface interativa, representada pelo *chatbot*, o discente é conduzido no processo de identificação de seus interesses acadêmicos, os quais são registrados em uma lista de interesses. Du-

rante essa interação, o *chatbot* também coleta os pesos atribuídos pelo discente a cada critério de preferência, indicando o grau de importância de cada aspecto na escolha de um orientador.

Com os dados do discente e a lista de docentes alinhados, o sistema realiza o cálculo do IR. Esse índice é uma métrica quantitativa que visa mensurar o grau de compatibilidade entre o perfil do discente e o perfil dos docentes filtrados, levando em consideração os interesses e os pesos atribuídos. O resultado do cálculo gera uma lista priorizada de possíveis orientadores, ordenada de acordo com o valor do IR, de modo que os docentes mais compatíveis estejam no topo da lista.

O processo se encerra com a apresentação desta lista ao discente, permitindo-lhe uma escolha mais fundamentada e personalizada de possíveis orientadores. Com isso, o sistema automatiza e qualifica uma etapa essencial da vida acadêmica, promovendo um melhor alinhamento entre expectativas e perfis profissionais, além de otimizar o processo de busca e seleção de orientadores nas instituições de ensino. Além da camada de interação com o usuário via *chatbot*, a proposta deste trabalho também considerou a integração de conceitos de sistemas de recomendação, como descrito a seguir.

Da mesma forma, os conceitos e técnicas de sistemas de recomendação apresentados no referencial teórico foram considerados no delineamento da proposta deste trabalho. A revisão sobre abordagens de filtragem colaborativa, filtragem baseada em conteúdo e modelos híbridos fundamentou a estruturação do sistema, inspirando a combinação de diferentes variáveis acadêmicas para a sugestão de orientadores. Além disso, a aplicação de técnicas de inteligência artificial e análise de dados implícitos reforçou a concepção de uma solução capaz de adaptar recomendações de maneira personalizada. Contudo, assim como a integração do *chatbot*, o desenvolvimento de um sistema de recomendação completo, com análise dinâmica de preferências de discentes, não foi contemplado na implementação prática desta tese, sendo proposto como uma perspectiva para trabalhos futuros. Dessa forma, os conceitos estudados foram essenciais para embasar as escolhas metodológicas e para orientar possíveis expansões do sistema de recomendação acadêmica aqui desenvolvido.

# *Parte V*

# *Apêndices*

# Apêndice A

# Código Scholarly

```python
from scholarly import scholarly
import json
# Obter um iterador para os resultados do autor
search_query = scholarly.search_author('XXXXXX XX;XXXXXXX')
# Recuperar o primeiro resultado do iterador
first_author_result = next(search_query)
# Obter todos os detalhes do autor
author = scholarly.fill(first_author_result)
# Salvar os dados em um arquivo JSON
filename = 'data.json'
with open(filename, 'w', encoding='utf-8') as file:
    json.dump(author, file, ensure_ascii=False, indent=4)
```

Listing A.1: Exemplo de Código Python: extração com scholarly

# Apêndice B

# Código Produção

```python
import os
import json

# Calcula a produção com base nos dados do JSON
def calcular_producao(json_file, alpha=1, beta=1):
    # Abrindo e carregando os dados do JSON
    with open(json_file, 'r', encoding='utf-8') as file:
        data = json.load(file)

    # Extraindo valores relevantes, com padrão 0 caso não existam
    C_5_anos = data.get('citedby5y', 0)
    C_total = data.get('citedby', 0)
    h_index = data.get('hindex', 0)
    i10_index = data.get('i10index', 0)

    # Cálculo do indicador de produção.
    if C_total > 0:
        P_producao = alpha * (C_total / C_5_anos) * h_index +
            beta * (i10_index / C_total)
    else:
        P_producao = 0

    return C_5_anos, C_total, h_index, i10_index, P_producao

# Processa todos os arquivos JSON.
def processar_pasta_json(pasta, alpha=1, beta=1):
    # Itera sobre os arquivos na pasta
    for arquivo in os.listdir(pasta):
        if arquivo.endswith(".json"):  # Filtra arquivos JSON
            caminho_arquivo = os.path.join(pasta, arquivo)

            # Calcula e obtém os valores de produção.
            C_5_anos, C_total, h_index, i10_index, P_producao =
                calcular_producao(caminho_arquivo, alpha, beta)

```

```
34              # Imprime resultados formatados
35              print(f"Arquivo: {arquivo}")
36              print(f"Citações 5 anos: {C_5_anos}, Total: {C_total}
                    ")
37              print(f"h-index: {h_index}, i10-index: {i10_index}")
38              print(f"P_Produção: {P_producao}")
39
40  # Define pasta e executa processamento com parâmetros específicos
41  pasta_json = r'C:\Users\xxxx\json_producao'
42  processar_pasta_json(pasta_json, alpha=2, beta=1)
```

Listing B.1: Exemplo de Código Python: p_producao.py

# Apêndice C

# Código Área

```python
import os
import glob
import xml.etree.ElementTree as ET

# Extrai áreas de conhecimento de um arquivo XML
def extract_knowledge_areas(xml_file):
    tree = ET.parse(xml_file)
    root = tree.getroot()
    areas = []

    # Função interna para obter detalhes da área de conhecimento
    def get_areas_of_knowledge(element):
        for area in element.findall('.//AREA-DO-CONHECIMENTO-1'):
            areas.append({
                'grande_area': area.get("NOME-GRANDE-AREA-DO-CONHECIMENTO"),
                'area': area.get("NOME-DA-AREA-DO-CONHECIMENTO"),
                'sub_area': area.get("NOME-DA-SUB-AREA-DO-CONHECIMENTO"),
                'especialidade': area.get("NOME-DA-ESPECIALIDADE")
            })

    # Extrai áreas em diferentes níveis acadêmicos
    for tag in ['GRADUACAO', 'MESTRADO', 'DOUTORADO', 'POS-DOUTORADO']:
        for formacao in root.findall(f'.//{tag}'):
            get_areas_of_knowledge(formacao)
    return areas

# Compara duas áreas de conhecimento e atribui uma pontuação
def compare_areas(area1, area2):
    score = 0
    if area1['grande_area'] == area2['grande_area']:
```

```
31          score += 1
32          if area1['area'] == area2['area']:
33              score += 2
34              if area1['sub_area'] == area2['sub_area']:
35                  score += 3
36                  if area1['especialidade'] == area2['especialidade
                        ']:
37                      score += 5
38      return score
39
40  # Executa a comparação de áreas entre arquivos XML
41  def main(reference_xml, folder_path):
42      reference_areas = extract_knowledge_areas(reference_xml)
43      xml_files = glob.glob(os.path.join(folder_path, '*.xml'))
44      results = []
45
46      # Processa cada arquivo XML da pasta
47      for xml_file in xml_files:
48          current_areas = extract_knowledge_areas(xml_file)
49          # Calcula a pontuação total de comparação
50          total_score = sum(compare_areas(area1, area2) for area1
                in reference_areas for area2 in current_areas)
51
52          if total_score > 0:
53              results.append({'file_name': os.path.basename(
                    xml_file), 'total_score': total_score})
54
55      # Ordena resultados pela pontuação em ordem decrescente
56      results.sort(key=lambda x: x['total_score'], reverse=True)
57
58      # Imprime os resultados formatados
59      print("Resultados␣da␣Comparação␣de␣Áreas␣de␣Conhecimento")
60      print("=" * 50)
61      for result in results:
62          print(f"Arquivo:␣{result['file_name']}␣-␣Pontuação:␣{
                result['total_score']}")
63          print("-" * 50)
64
65  # Ponto de entrada principal do script
66  if __name__ == "__main__":
67      reference_xml = r'C:\\user\\xxxx\\file.xml'  # referência
68      folder_path = r'C:\\user\\xxxx\\folder'      # arquivos XML a
            comparar
69      main(reference_xml, folder_path)
```

Listing C.1: Exemplo de Código Python: p_area.py

# Apêndice D

# Atividades Acadêmicas

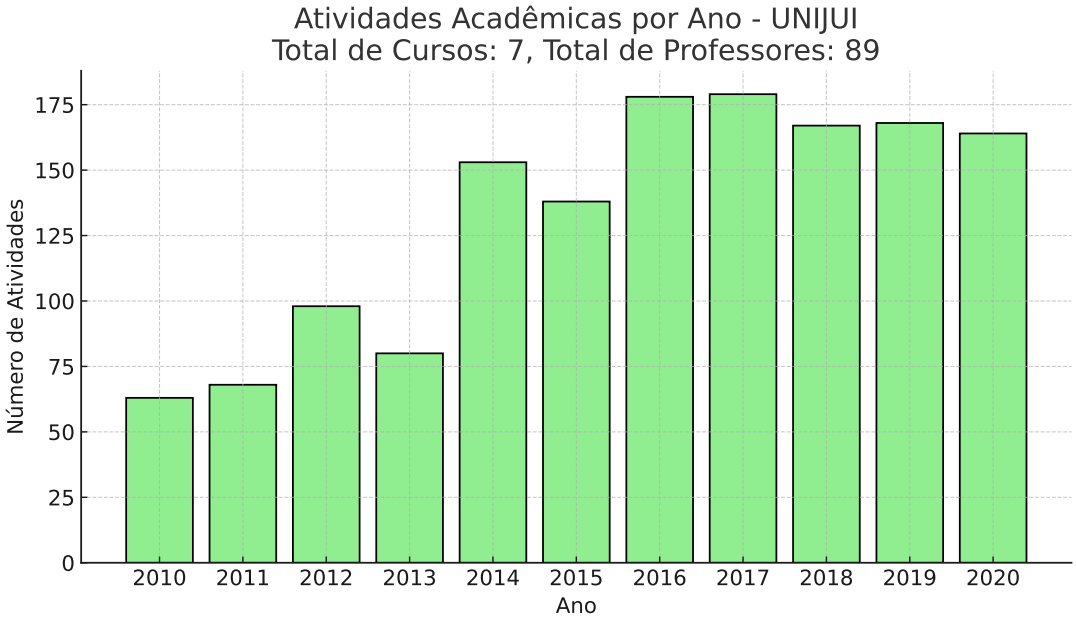



**Figura D.1**: *UNIJUI - Atividades Acadêmicas.*

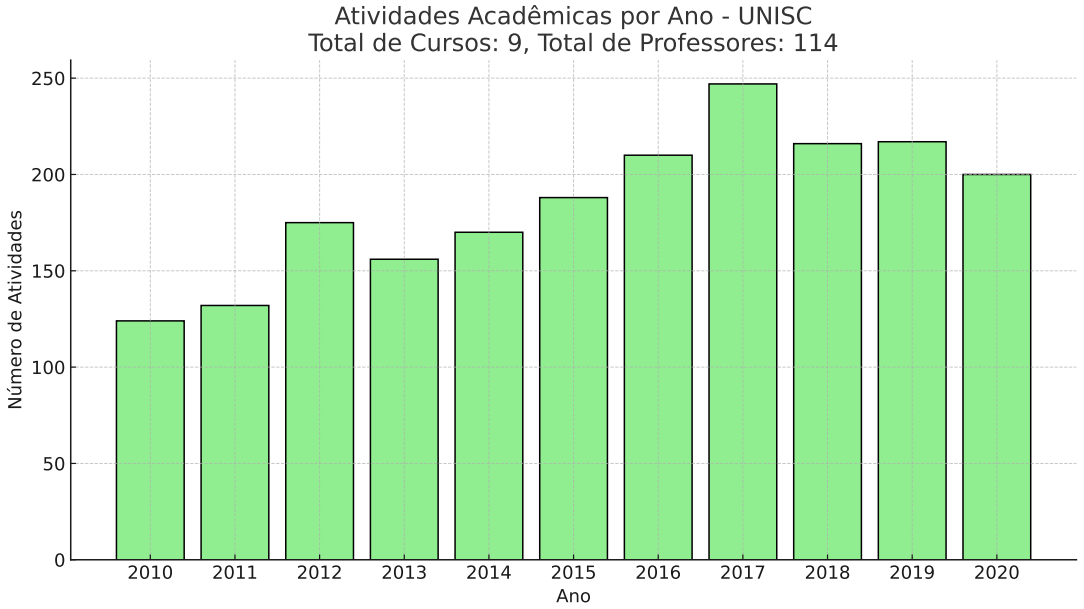



**Figura D.2**: *UNISC - Atividades Acadêmicas.*

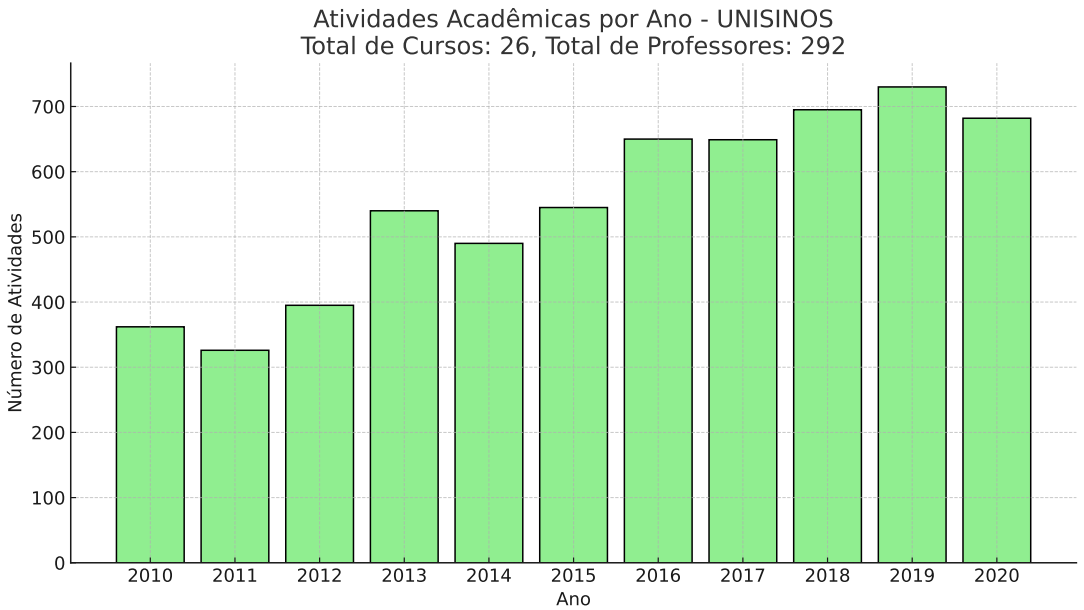



**Figura D.3**: *UNISINOS - Atividades Acadêmicas.*

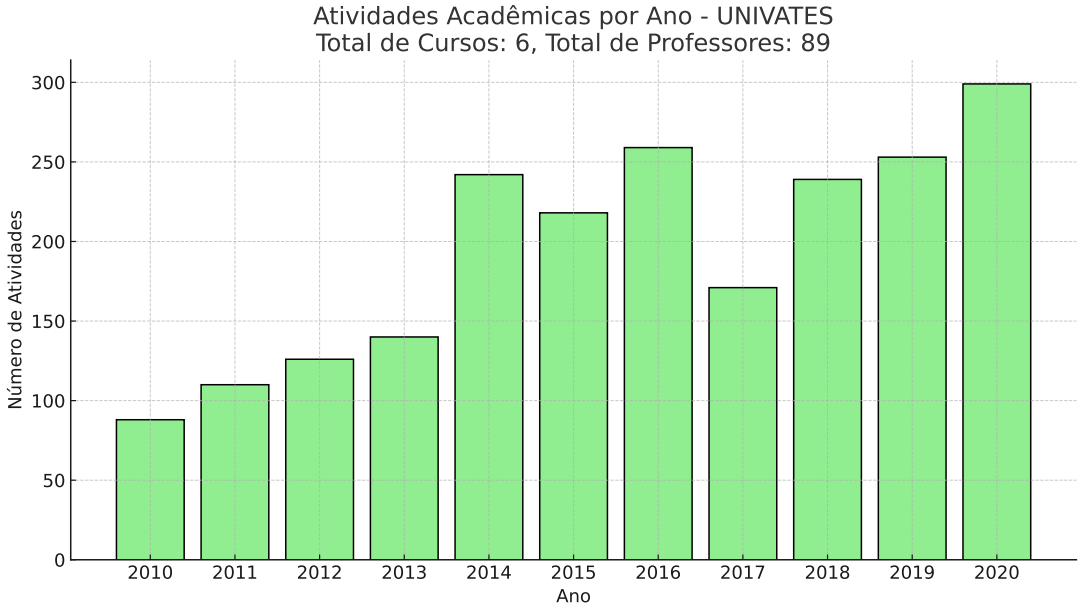



**Figura D.4**: *UNIVATES - Atividades Acadêmicas.*

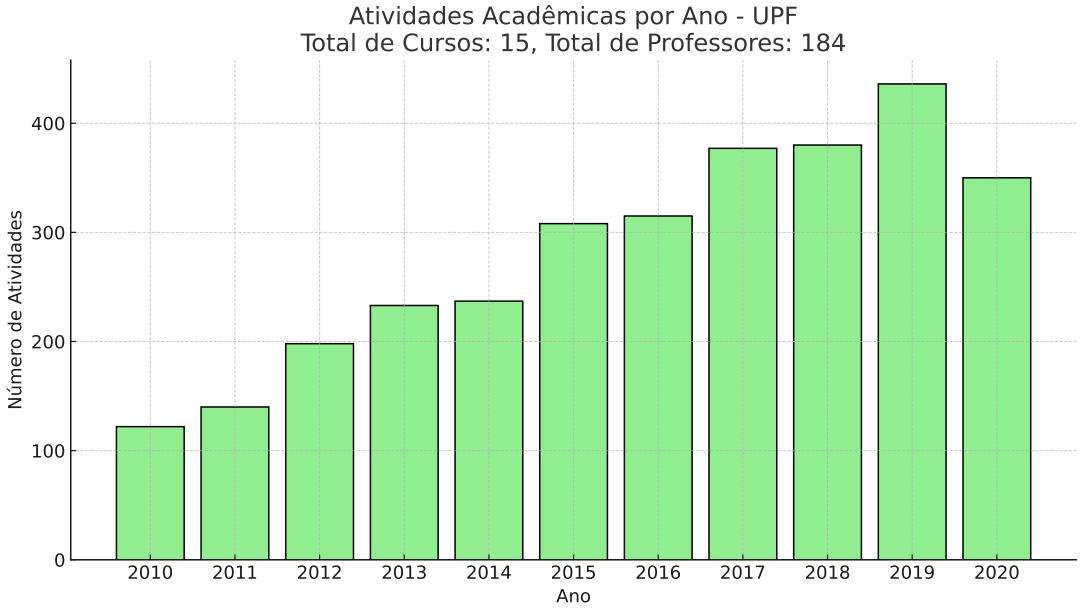



**Figura D.5**: *UPF - Atividades Acadêmicas.*

# Apêndice E

# Dados Simulados - Curso

| Orientador | Ciência da Computação | Matemática | Ciências Contábeis |
|---|---|---|---|
| Orientador 1 | 0.3000 | 0.1000 | 0.0000 |
| Orientador 2 | 1.0000 | 0.1000 | 0.0000 |
| Orientador 3 | 0.6000 | 1.0000 | 0.0000 |
| Orientador 4 | 0.6000 | 0.1000 | 0.0000 |
| Orientador 5 | 0.1000 | 0.1000 | 0.0000 |

**Tabela E.1**: $P_{Área}$ *por curso* .

| Orientador | Ciência da Computação | Matemática | Ciências Contábeis |
|---|---|---|---|
| Orientador 1 | 0.4335 | 0.4335 | 0.4335 |
| Orientador 2 | 0.0927 | 0.0927 | 0.0927 |
| Orientador 3 | 0.5433 | 0.5433 | 0.5433 |
| Orientador 4 | 0.2733 | 0.2733 | 0.2733 |
| Orientador 5 | 0.0164 | 0.0164 | 0.0164 |

**Tabela E.2**: $P_{Experiência}$ *por curso* .

| Orientador | Ciência da Computação | Matemática | Ciências Contábeis |
|---|---|---|---|
| Orientador 1 | 0.4333 | 0.4333 | 0.4333 |
| Orientador 2 | 0.2778 | 0.2778 | 0.2778 |
| Orientador 3 | 0.3611 | 0.3611 | 0.3611 |
| Orientador 4 | 0.3450 | 0.3450 | 0.3450 |
| Orientador 5 | 0.2000 | 0.2000 | 0.2000 |

**Tabela E.3**: $P_{Eficiência}$ *por curso* .

| Orientador | Ciência da Computação | Matemática | Ciências Contábeis |
|---|---|---|---|
| Orientador 1 | 0.8550 | 0.8550 | 0.8550 |
| Orientador 2 | 0.4190 | 0.4190 | 0.4190 |
| Orientador 3 | 1.0000 | 1.0000 | 1.0000 |
| Orientador 4 | 0.5868 | 0.5868 | 0.5868 |
| Orientador 5 | 0.3180 | 0.3180 | 0.3180 |

**Tabela E.4**: $P_{Produção}$ *por curso* .

| Orientador | Ciência da Computação | Matemática | Ciências Contábeis |
|---|---|---|---|
| Orientador 1 | 0.9400 | 0.9400 | 0.9400 |
| Orientador 2 | 0.3850 | 0.3850 | 0.3850 |
| Orientador 3 | 0.7375 | 0.7375 | 0.7375 |
| Orientador 4 | 0.8500 | 0.8500 | 0.8500 |
| Orientador 5 | 0.1850 | 0.1850 | 0.1850 |

**Tabela E.5**: $P_{Colaboração}$ *por curso* .

| Orientador | Ciência da Computação | Matemática | Ciências Contábeis |
|---|---|---|---|
| Orientador 1 | 0.7692 | 0.7692 | 0.7692 |
| Orientador 2 | 0.3846 | 0.3846 | 0.3846 |
| Orientador 3 | 1.0000 | 1.0000 | 1.0000 |
| Orientador 4 | 0.6250 | 0.6250 | 0.6250 |
| Orientador 5 | 0.2404 | 0.2404 | 0.2404 |

**Tabela E.6**: $P_{Pesquisa}$ *por curso* .

# Apêndice F

# Dados Simulados - Aluno Matemática

| Orientador | $GA$ | $A$ | $SA$ | $E$ | $P_{\text{Área}}$ |
|---|---|---|---|---|---|
| Orientador 1 | 1 | 0 | 0 | 0 | 0.1000 |
| Orientador 2 | 1 | 0 | 0 | 0 | 0.1000 |
| Orientador 3 | 1 | 1 | 1 | 1 | 1.0000 |
| Orientador 4 | 1 | 0 | 0 | 0 | 0.1000 |
| Orientador 5 | 1 | 0 | 0 | 0 | 0.1000 |

**Tabela F.1**: $P_{Área}$ *aluno: Matemática.*

| Orientador | $m$ | $d$ | $M$ | $D$ | $P_r$ | $P_{90}$ | $Q$ | $P_{\text{Experiência}}$ |
|---|---|---|---|---|---|---|---|---|
| Orientador 1 | 6 | 3 | 10 | 6 | 35 | 44.0000 | 0.7955 | 0.4335 |
| Orientador 2 | 3 | 1 | 10 | 6 | 18 | 44.0000 | 0.4091 | 0.0927 |
| Orientador 3 | 8 | 2 | 10 | 6 | 50 | 44.0000 | 1.0000 | 0.5433 |
| Orientador 4 | 4 | 4 | 10 | 6 | 22 | 44.0000 | 0.5000 | 0.2733 |
| Orientador 5 | 2 | 0 | 10 | 6 | 8 | 44.0000 | 0.1818 | 0.0164 |

**Tabela F.2**: $P_{Experiência}$ *aluno: Matemática.*

| Orientador | $OC_m$ | $OA_m$ | $T_{C_m}$ | $OC_d$ | $OA_d$ | $T_{C_d}$ | $P_{\text{Eficiência}}$ |
|---|---|---|---|---|---|---|---|
| Orientador 1 | 5 | 1 | 0.8333 | 3 | 0 | 1.0000 | 0.4333 |
| Orientador 2 | 2 | 1 | 0.6667 | 1 | 1 | 0.5000 | 0.2778 |
| Orientador 3 | 7 | 1 | 0.8750 | 2 | 1 | 0.6667 | 0.3611 |
| Orientador 4 | 3 | 2 | 0.6000 | 3 | 1 | 0.7500 | 0.3450 |
| Orientador 5 | 1 | 1 | 0.5000 | 0 | 0 | 0.0000 | 0.2000 |

**Tabela F.3**: $P_{Eficiência}$ *aluno: Matemática.*

| Orientador | $C_{5\text{anos}}$ | $C_{\text{total}}$ | $h_i$ | $i_{10}$ | $P_{\text{Produção}}$ |
|---|---|---|---|---|---|
| Orientador 1 | 120 | 300 | 18 | 25 | 0.8550 |
| Orientador 2 | 45 | 110 | 9 | 10 | 0.4190 |
| Orientador 3 | 160 | 420 | 22 | 30 | 1.0000 |
| Orientador 4 | 70 | 180 | 12 | 15 | 0.5868 |
| Orientador 5 | 12 | 40 | 5 | 4 | 0.3180 |

**Tabela F.4**: $P_{Produção}$ *aluno: Matemática.*

| Orientador | $P_{\text{banca}}$ | $\max(P_{\text{banca}})$ | $Co$ | $\max(Co)$ | $P_{\text{Colaboração}}$ |
|---|---|---|---|---|---|
| Orientador 1 | 22 | 25 | 40 | 40 | 0.9400 |
| Orientador 2 | 8 | 25 | 18 | 40 | 0.3850 |
| Orientador 3 | 15 | 25 | 35 | 40 | 0.7375 |
| Orientador 4 | 25 | 25 | 28 | 40 | 0.8500 |
| Orientador 5 | 3 | 25 | 10 | 40 | 0.1850 |

**Tabela F.5**: $P_{Colaboração}$ *aluno: Matemática.*

| Orientador | $N_1$ | $a_1$ | $N_2$ | $a_2$ | $P_{\text{Pesquisa}}$ |
|---|---|---|---|---|---|
| Orientador 1 | 4 | 0.6000 | 2 | 0.4000 | 0.7692 |
| Orientador 2 | 2 | 0.6000 | 1 | 0.4000 | 0.3846 |
| Orientador 3 | 6 | 0.6000 | 3 | 0.4000 | 1.0000 |
| Orientador 4 | 3 | 0.6000 | 2 | 0.4000 | 0.6250 |
| Orientador 5 | 1 | 0.6000 | 1 | 0.4000 | 0.2404 |

**Tabela F.6**: $P_{Pesquisa}$ *aluno: Matemática.*

# Bibliografia


[1] M. M. Abreu e M. da Silva dos Santos. *A IA no desenvolvimento de sistemas de recomendação de produtos*. *Jornada Acadêmica*, 7(1):2, 2023.

[2] E. Adamopoulou e L. Moussiades. *Chatbots: History, technology, and applications*. *Machine Learning with Applications*, 2:100006, 2020.

[3] G. Adomavicius e A. Tuzhilin. *Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions*. *IEEE Transactions on Knowledge and Data Engineering*, 17 (6):734–749, 2005.

[4] A. Aggarwal, C.-C. Tam, D. Wu, X. Li e S. Qiao. *Artificial intelligence-based chatbots for promoting health behavioral changes: Systematic review*. *Journal of Medical Internet Research*, 25:e40789, 2023.

[5] C. C. Aggarwal. *Recommender systems: The textbook*. Springer International Publishing, Cham, 2016.

[6] A. D. Alves, H. H. Yanasse e N. Y. Soma. *Lattesminer: a multilingual dsl for information extraction from lattes platform*. Em *Proceedings of the Co-located Workshops of SPLASH'11*, páginas 85–92, 2011.

[7] R. Aris. *Mathematical modeling: A chemical engineer's perspective*. Academic Press, a division of Harcourt Brace & Company, 1999.

[8] S. Atalla, M. Daradkeh, A. Gawanmeh, H. Khalil, W. Mansoor, S. Miniaoui e Y. Himeur. *An intelligent recommendation system for automating academic advising based on curriculum analysis and performance modeling*. *Mathematics*, 11(5):1098, 2023.

[9] C. P. Baldin, M. M. Schambeck, S. D. Matos e W. Crescencio. *A inteligência artificial na automatização de processos*. Em *Saberes 2024*, páginas 85–90. Unifacear, 2024.

[10] B. J. Barnes e A. E. Austin. *The role of doctoral advisors: A look at advising from the advisors perspective*. *Innovative Higher Education*, 33: 297–315, 2009.

[11] B. J. Barnes, E. A. Williams e S. A. Archer. *Characteristics that matter most: Doctoral students' perceptions of positive and negative advisor attributes*. *NACADA Journal*, 30(1):34–46, 2010.

[12] J. M. Barreto. *Inteligência artificial no limiar do século xxi*. Duplic, Florianópolis, edição 3, 2001.

[13] G. A. Bauer. *Geração de conhecimento através de dados da plataforma lattes com o uso de técnicas de mineração de dados*, 2016. Trabalho de Conclusão de Curso - Bacharelado em Ciência da Computação - Universidade de Santa Cruz do Sul.

[14] E. A. Bender. *An introduction to mathematical modeling*. Wiley, New York, 1978. Reimpressão: Dover Publications, Mineola, NY, ISBN: 0-486-41800-X.

[15] A. A. Bernardini, A. A. Sônego e E. Pozzebon. *Chatbots: An Analysis of the State of Art of Literature*. Em *Anais do I Workshop on Advanced Virtual Environments and Education (WAVE)*, páginas 1–6. SBC, 2018.

[16] J. Bollen, H. V. de Sompel, A. Hagberg e R. Chute. *A principal component analysis of 39 scientific impact measures*. *PLoS ONE*, 4(6):e6022, 2009.

[17] L. Bornmann e H.-D. Daniel. *What do citation counts measure? a review of studies on citing behavior*. *Journal of Documentation*, 64(1): 45–80, 2008.

[18] D. Boud e A. Lee. *Peer learning as pedagogic discourse for research education*. *Studies in Higher Education*, 30(5):501–516, 2005.

[19] B. Bozeman e E. Corley. *Scientists' collaboration strategies: Implications for scientific and technical human capital*. *Research Policy*, 33(4): 599–616, 2004.

[20] A. Bryman. *Social research methods*. Oxford University Press, Oxford, 2016.

[21] R. Burke. *Hybrid recommender systems: Survey and experiments*. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[22] Q. Z. C. Zhai, Z. Yan. *The hl-index: Improving the h-index based on citation quality*. *Journal of Informetrics*, 7(1):99–110, 2013.

[23] M. Calle, E. Narváez e J. Maldonado-Mahauad. *Proposal for the design and implementation of miranda: A chatbot-type recommender for supporting self-regulated learning in online environments*. Em *Proceedings of the Latin American Learning Analytics Workshop (LALA 2021)*, páginas 18–28. CEUR-WS.org, 2021.

[24] C. R. Cervi. *Rep-index uma abordagem abrangente e adaptável para identificar reputação acadêmica*. Tese (doutorado em ciência da computação), Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

[25] A. S. Chukwu e K. D. Walker. *Towards a deeper understanding of the graduate student and faculty-advisor relationship*. *Journal of Contemporary Issues in Education*, 18(1):61–78, 2023. Licença Creative Commons Attribution 4.0 (CC BY 4.0).

[26] C. M. G. ChunMei Zhao e A. C. McCormick. *More than a signature: How advisor choice and advisor behaviour affect doctoral student satisfaction*. *Journal of Further and Higher Education*, 31(3):263–281, 2007.

[27] F. S. Coimbra e T. M. R. Dias. *Use of open data to analyze the publication of articles in scientific events*. *Iberoamerican Journal of Science Measurement and Communication*, 1(1):1–20, 2021.

[28] C. d. S. Colombo, C. Badue e E. d. Oliveira. *Pharmabulabot: A chatbot to answer drug questions based on information from pharmaceutical package inserts*. Em *Anais do XIX Simpósio Brasileiro de Sistemas de Informação (SBSI 2023)*, páginas 70–77. Sociedade Brasileira de Computação, 2023.

[29] S. Danckwerts, L. MeiSSner e C. Krampe. *Hi, can you recommend a movie? investigating recommendation chatbots in media streaming services*. Em *Proceedings of the 28th European Conference on Information Systems (ECIS)*, páginas 1–13, Marrakech, Morocco, 2020.

[30] D. de Oliveira Sarvo, M. C. Lozano e R. M. do Amaral. *O uso de dados da plataforma lattes como fonte para indicadores de inteligência acadêmica*. *Informação & Informação*, 27(3):557–576, 2022.

[31] S. J. de Sousa, T. M. R. Dias e A. L. Pinto. *Uma estratégia para recomendação de especialistas a partir de dados abertos disponíveis na plataforma lattes*. *Ciência da Informação*, 48(3):162–173, 2019.

[32] S. J. de Sousa, T. M. R. Dias e A. L. Pinto. *A strategy for identifying specialists in scientific data repositories*. *Mobile Networks and Applications*, 27(5):1941–1951, 2022.

[33] L. A. Digiampietri, J. P. Mena-Chalco, J. J. Pérez-Alcázar, E. F. Tuesta, K. V. Delgado, R. Mugnaini e G. S. Silva. *Minerando e caracterizando dados de currículos lattes*. Em *Anais do I Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2012)*, páginas 117–128. Sociedade Brasileira de Computação, 2012.

[34] A. J. Dobson. *An introduction to generalized linear models*. Texts in Statistical Science. Chapman and Hall, London, 1990.

[35] L. Egghe. *Theory and practice of the g-index*. *Scientometrics*, 69(1): 131–152, 2006.

[36] J. M. Epstein. *Generative social science: Studies in agent-based computational modeling*. Princeton University Press, 2006.

[37] H. Etzkowitz, A. Webster, C. Gebhardt e B. R. C. Terra. *The future of the university and the university of the future: Evolution of ivory tower to entrepreneurial paradigm*. *Research Policy*, 29(2):313–330, 2000.

[38] A. Field. *Discovering statistics using ibm spss statistics*. SAGE Publications, London, edição 4, 2013.

[39] B. Fine. *The impact of mathematics and mathematicians*. Em A. Rogerson e J. Morska, editores, *Theory and Practice: An Interface or A Great Divide? The Mathematics Education for the Future Project: Proceedings of the 15th International Conference*, páginas 156–160. WTM-Verlag, 2019.

[40] S. A. A. Freitas, E. D. Canedo e D. A. Jesus. *Calculating similarity of curriculum lattes*. *IEEE Latin America Transactions*, 16(6):1758–1764, 2018.

[41] M. d. L. d. A. Fávero. *Anísio teixeira e a universidade do distrito federal*. *Revista Brasileira de História da Educação*, 2012.

[42] E. F. Galego. *Extração e consulta de informações do currículo lattes baseada em ontologias*. Dissertação de Mestrado, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2013. Dissertação de Mestrado.

[43] S. K. Gardner. *Fitting the mold of graduate school: A qualitative study of socialization in doctoral education*. *Innovative Higher Education*, 33 (2):125–138, 2008.

[44] N. A. Gershenfeld. *The nature of mathematical modeling*. Cambridge University Press, Cambridge, 1999.

[45] G. M. Gurr. *Negotiating the "rackety bridge" a dynamic model for aligning supervisory style with research student development*. *Higher Education Research & Development*, 20(1):81–92, 2001.

[46] X. He, L. Liao, H. Zhang, L. Nie, X. Hu e T.-S. Chua. *Neural collaborative filtering*. Em *Proceedings of the 26th International Conference on World Wide Web*, páginas 173–182. International World Wide Web Conferences Steering Committee, 2017.

[47] D. Hicks, P. Wouters, L. Waltman, S. De Rijcke e I. Rafols. *Bibliometrics: The leiden manifesto for research metrics*. *Nature*, 520(7548): 429–431, 2015.

[48] J. E. Hirsch. *An index to quantify an individuals scientific research output*. *Proceedings of the National Academy of Sciences*, 102(46): 16569–16572, 2005.

[49] M. Humi. *Introduction to mathematical modeling*. Chapman & Hall/CRC, 2018.

[50] G.-J. Hwang e C.-Y. Chang. *A review of opportunities and challenges of chatbots in education*. *Interactive Learning Environments*, páginas 1–16, 2023.

[51] J. A. Jacobs e S. Frickel. *Interdisciplinarity: A critical assessment*. *Annual Review of Sociology*, 35:43–65, 2009.

[52] F. R. Jensenius, M. Htun, D. J. Samuels, D. A. Singer, A. Lawrence e M. Chwe. *The benefits and pitfalls of google scholar*. *PS: Political Science & Politics*, 51(4):820–824, 2018.

[53] B. Jin. *The ar-index: Complementing the h-index*. *ISSI Newsletter*, 3(1): 6–8, 2007.

[54] S. Joy, X. Liang, D. Bilimoria e S. Perry. *Doctoral advisor-advisee pairing in stem fields: Selection criteria and impact of faculty, student and departmental factors*. *International Journal of Doctoral Studies*, 10: 343–363, 2015.

[55] A. A. Júnior, N. Sucupira, C. Salgado, J. B. Filho, M. R. e Silva, D. Trigueiro, A. A. Lima, A. Teixeira, V. C. Maciel e outros. *Parecer cfe nº 977/65, aprovado em 3 dez. 1965*. *Revista Brasileira de Educação*, 0(30): 162–173, 2005.

[56] C. F. d. C. Júnior e K. R. S. d. A. d. Carvalho. *Chatbot: uma visão geral sobre aplicações inteligentes*. *Revista Sítio Novo*, 2(2):68–83, 2018.

[57] H. Ko, S. Lee, Y. Park e A. Choi. *A survey of recommendation systems: recommendation models, techniques, and application fields*. *Electronics*, 11(1):141, 2022.

[58] G. Laudel. *The art of getting funded: How scientists adapt to their funding conditions*. *Science and Public Policy*, 33(7):489–504, 2006.

[59] A. M. Law. *Simulation modeling and analysis*. McGraw-Hill Education, edição 5, 2015.

[60] A. Lee. *How are doctoral students supervised? concepts of doctoral research supervision*. *Studies in Higher Education*, 33(3):267–281, 2008.

[61] H. Lima, T. H. P. Silva, M. M. Moro, R. L. T. Santos, W. Meira Jr. e A. H. F. Laender. *Aggregating productivity indices for ranking researchers across multiple areas*. Em *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, páginas 97–106, Indianapolis, IN, USA, 2013. ACM.

[62] L. Ljung. *System identification: Theory for the user*. Prentice Hall PTR, Upper Saddle River, NJ, edição 2, 1999.

[63] B. E. Lovitts. *Leaving the ivory tower: The causes and consequences of departure from doctoral study*. Rowman & Littlefield, Lanham, MD, 2001.

[64] B. E. Lovitts. *Being a good coursetaker is not enough: A theoretical perspective on the transition to independent research*. *Studies in Higher Education*, 30(2):137–154, 2005.

[65] L. F. Lugoboni. *Como escolher um orientador*. *R. Liceu On-line*, 8(1): 1–5, 2018.

[66] L. G. Lunsford. *Doctoral advising or mentoring? effects on student outcomes. Mentoring & Tutoring: Partnership in Learning*, 20(2):251–270, 2012.

[67] Y. Ma, T. Kleemann e J. Ziegler. *Mixed-modality interaction in conversational recommender systems*. Em *Proceedings of the 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2021)*, volume 2948, páginas 21–37, Amsterdam, Netherlands, 2021. CEUR-WS.org.

[68] J. J. d. Magalhães, C. C. d. Souza, E. d. B. Costa e J. M. Fechine. *Recommending scientific papers: Investigating the user curriculum*. Em *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference (FLAIRS-28)*, páginas 489–494, Hollywood, FL, USA, 2015. AAAI Press.

[69] W. T. Maruyama e L. A. Digiampietri. *Combinando agrupamento e classificação para a predição de coautorias na plataforma lattes*. *Revista Brasileira de Computação Aplicada*, 13(2):48–57, 2021.

[70] N. F. Matsatsinis, K. Lakiotaki e P. Delias. *A system based on multiple criteria analysis for scientific paper recommendation*. Em *Proceedings of the 11th Panhellenic Conference on Informatics (PCI 2007)*, páginas 135–149, Patras, Greece, 2007.

[71] L. McAlpine e C. Amundsen. *Doctoral education: Research-based strategies for doctoral students, supervisors and administrators*. Springer, Dordrecht, 2011.

[72] J. P. Mena-Chalco e R. M. Cesar Junior. *Scriptlattes: An open-source knowledge extraction system from the lattes platform*. *Journal of the Brazilian Computer Society*, 15(4):31–39, 2009.

[73] J. P. Mena-Chalco, L. A. Digiampietri, F. M. Lopes e R. M. Cesar Junior. *Brazilian bibliometric coauthorship networks*. *Journal of the Association for Information Science and Technology*, 65(7):1424–1445, 2014.

[74] N. C. Mendonça, M. A. F. Rodrigues e L. R. Mendonça. *Qlattes: An open-source tool for qualis annotation and visualization in the lattes platform*. Em *Anais do L Seminário Integrado de Software e Hardware (SEMISH 2023)*, páginas 83–94, João Pessoa, PB, 2023. Sociedade Brasileira de Computação.

[75] R. K. Merton. *The matthew effect in science*. *Science*, 159(3810):56–63, 1968.

[76] V. Mityushev, W. Nawalaniec, N. Rylko e R. A. Kycia. *Introduction to mathematical modeling and computer simulations*. Taylor & Francis, edição 2, 2024.

[77] E. Morrison, E. Rudd, J. Picciano e M. Nerad. *Are you satisfied? phd education and faculty taste for prestige: Limits of the prestige value system*. *Research in Higher Education*, 52(1):24–46, 2011.

[78] R. J. d. Nascimento. *O processo de escolha de candidatos a programas de pós-graduação: uma análise a partir da perspectiva do orientador*. Dissertação de Mestrado, Universidade de São Paulo, São Paulo, 2016.

[79] O. Nelles. *Nonlinear system identification: From classical approaches to neural networks and fuzzy models*. Springer, Berlin, Heidelberg, 2001.

[80] M. Nerad. *The phd in the us: Criticisms, facts, and remedies*. *Higher Education Policy*, 17(2):183–199, 2004.

[81] A. Nieto. *Essential e-mentors characteristics for mentoring online doctoral dissertations: Faculty views*. *Journal of Psychological Issues in Organizational Culture*, 6(4):1–17, 2016.

[82] C. W. Okonkwo e A. Ade-Ibijola. *Chatbots applications in education: A systematic review*. *Computers and Education: Artificial Intelligence*, 2: 100033, 2021.

[83] A. C. d. Oliveira e F. T. d. Silva. *Extrator de produções acadêmicas de pesquisadores do ifce campus tianguá pela plataforma lattes*. Em *Anais do XIV Encontro Unificado de Computação do Piauí (ENUCOMPI)*, páginas 81–88, Picos, PI, 2021. Sociedade Brasileira de Computação.

[84] D. T. d. Oliveira, L. d. O. Rocha e P. N. Silva. *Recuperação de informação no currículo lattes: proposição de requisitos aplicando técnicas de filtragem para recuperação da produção acadêmica*. *Ciência da Informação em Revista*, 10(1/3):1–19, 2023.

[85] L. Page, S. Brin, R. Motwani e T. Winograd. *The pagerank citation ranking: Bringing order to the web*. Relatório técnico 1999-66, Stanford InfoLab, 1998.

[86] L. L. Paglis, S. G. Green e T. N. Bauer. *Does adviser mentoring add value? a longitudinal study of mentoring and doctoral student outcomes*. *Research in Higher Education*, 47(4):451–476, 2006.

[87] F. Prass, F. Boijink e A. Zamberlan. *Parser e leitura automatizada de currículos da plataforma lattes para extração de indicadores acadêmicos e tecnológicos*. Em V. C. de Abreu Torres Hrenechen, editor, *Comunicação, Mídias e Educação*, páginas 78–95. Atena Editora, May 2019.

[88] P. Pu, L. Chen e R. Hu. *Evaluating recommender systems from the users perspective: Survey of the state of the art*. *User Modeling and User-Adapted Interaction*, 22(4–5):317–355, 2012.

[89] M. H. Rehbein. *Modelo computacional para recomendação de potenciais orientadores em programas de pós-graduação stricto sensu: Um estudo de caso dirigido ao comung*. Dissertação de mestrado, Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí, Brazil, May 2022.

[90] A. Romêo, A. C. Romêo e M. C. d. S. Jorge. *Universidades brasileiras: história e desenvolvimento*. Editora Universitária, São Paulo, 2004.

[91] M. Rosa e D. C. Orey. *A modelagem como um ambiente de aprendizagem para a conversão do conhecimento matemático*. *Boletim de Educação Matemática (Bolema)*, 26(42A):261–290, April 2012.

[92] R. Rousseau e B. Jin. *Adding a time dimension to the h-index*. *Scientometrics*, 77(2):294–308, 2008.

[93] D. Roy e M. Dutta. *A systematic review and research perspective on recommender systems*. *Journal of Big Data*, 9(1):59, 2022.

[94] M. F. Saito e V. T. Miura. *Processamento natural de linguagem: Sistema de recomendações e explicações*. Trabalho de Conclusão de Curso de Graduação Escola Politécnica, Universidade de São Paulo, 2020.

[95] C. M. d. Santos. *Tradições e contradições da pós-graduação no brasil*. *Educação & Sociedade*, 24(83):627–641, 2003.

[96] J. B. Schafer, J. Konstan e J. Riedl. *Recommender systems in e-commerce*. Em *Proceedings of the 1st ACM conference on Electronic commerce*, páginas 158–166, 1999.

[97] B. Selic. *Personal reflections on automation, programming culture, and model-based software engineering*. *Automated Software Engineering*, 15 (3-4):379–391, 2008.

[98] S. Siegel e N. J. Castellan. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York, edição 2, 1988.

[99] J. A. d. Silva e M. d. L. P. Bianchi. *Cientometria: a métrica da ciência*. *Paidéia (Ribeirão Preto)*, 11(21):5–10, 2001.

[100] S. M. C. d. Silva, M. A. M. Antunes, R. F. Pegoraro, G. J. Miranda e L. B. e. Silva. *Reasons for entering the strictosensu postgraduate course: a survey with students from a public hei*. *Psicologia Escolar e Educacional*, 27:e250905, 2023.

[101] A. Simon, P. A. Behar, C. A. W. Torrezzan, B. K. Slodkowski e S. C. Cazella. *Modelagem de uma ontologia de domínio com foco em competências para sistemas de recomendação na educação*. *Revista Brasileira de Informática na Educação (RBIE)*, 28(01):73–94, 2020.

[102] A. M. F. d. Souza, A. Serpa, Francisco Silva e Ost e A. d. F. Silva. *Qualis no lattes: uma extensão para navegadores de internet que mostra automaticamente o índice qualis (capes) no currículo lattes*. *Revista Foco*, 16(4): 1–11, 2023.

[103] J. D. Sterman. *Business dynamics: Systems thinking and modeling for a complex world*. Irwin/McGraw-Hill, Boston, MA, 2000.

[104] R. T. Taylor, T. Vitale, C. Tapoler e K. Whaley. *Desirable qualities of modern doctorate advisors in the usa: a view through the lenses of candidates, graduates, and academic advisors*. *Studies in Higher Education*, 43 (5):854–866, 2018.

[105] M. Tikhonova. *Text mining for evaluation of candidates based on their cvs*. Em *Proceedings of the 8th International Conference on Analysis of Images, Social Networks and Texts (AIST 2019)*, volume 1086 de *Communications in Computer and Information Science*, páginas 184–189. Springer, 2020.

[106] M. Tikhonova e A. Gavrishchuk. *Nlp methods for automatic candidates cv segmentation*. Em *2019 International Conference on Engineering and Telecommunication (EnT)*, páginas 1–5, Moscow, Russia, 2019. IEEE.

[107] G. R. Vivian e C. R. Cervi. *Uma proposta de abordagem de recomendação para carreira de pesquisadores baseada em personalização, similaridade de perfil e reputação acadêmica*. Em *Anais da XIII Escola Regional de Banco de Dados (ERBD 2017)*, páginas 7–16, Passo Fundo, RS, Brasil, 2017. Sociedade Brasileira de Computação.

[108] G. E. Walker, C. M. Golde, L. Jones, A. C. Bueschel e P. Hutchings. *The formation of scholars: Rethinking doctoral education for the twenty-first century*. *International Journal for Academic Development*, 14(3): 237–239, sep 2009.

[109] X. Wang, X. Lin e B. Shao. *How does artificial intelligence create business agility? evidence from chatbots*. *International Journal of Information Management*, 66:102529, 2022.

[110] J. C. Weidman e E. L. Stein. *Socialization of doctoral students to academic norms*. *Research in Higher Education*, 44(6):641–656, 2003.

[111] Y. Zayed, Y. Salman e A. Hasasneh. *A recommendation system for selecting the appropriate undergraduate program at higher education institutions using graduate student data*. *Applied Sciences*, 12(24):12525, 2022.

[112] Y. Zhang. *The e-index, complementing the h-index for excess citations*. *PLoS ONE*, 4(5):e5429, 2009.

[113] Y. Zhang. *The h-index, effectively improving the h-index based on citation distribution*. *PLoS ONE*, 8(4):e59912, 2013.

This document was typeset on September 19, 2025 using class RG-BOOK $\alpha$2.14 for LaTeX$_{2\epsilon}$. As of the time of writing this document, this class is not publicly available. Only members of The Distributed Group (TDG) and the Applied Computing Research Group (GCA) are allowed to typeset their documents using this class.