UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

FABIO HENRIQUE ALVES FERNANDES
Orientador: Prof. Dr. Vander Luis de Souza Freitas
Coorientador: Prof. Dr. Eduardo José da Silva Luz

CARACTERIZAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO BRASIL A PARTIR DE DADOS BIBLIOMÉTRICOS DA OPENALEX

Ouro Preto, MG
2024

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

FABIO HENRIQUE ALVES FERNANDES

CARACTERIZAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO BRASIL A PARTIR DE DADOS BIBLIOMÉTRICOS DA OPENALEX

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Vander Luis de Souza Freitas
Coorientador: Prof. Dr. Eduardo José da Silva Luz

Ouro Preto, MG
2024

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Fabio Henrique Alves Fernandes

Caracterização dos Programas de Pós-Graduação em Ciência da Computação no Brasil a partir de dados bibliométricos da OpenAlex

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 10 de Outubro de 2024.

Membros da banca

Vander Luis de Souza Freitas (Orientador) - Doutor - Universidade Federal de Ouro Preto
Eduardo José da Silva Luz (Coorientador) - Doutor - Universidade Federal de Ouro Preto
Gladston Juliano Prates Moreira (Examinador) - Doutor - Universidade Federal de Ouro Preto
Fernando Henrique Oliveira Duarte (Examinador) - Mestre - Universidade Federal de Ouro Preto

Vander Luis de Souza Freitas, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 10/10/2024.

Documento assinado eletronicamente por **Vander Luis de Souza Freitas**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 15/10/2024, às 08:35, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0789765** e o código CRC **52146D4A**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Resumo

Estima-se que até 2025 o Brasil necessitará de mais de 797 mil profissionais qualificados em Computação. Com o aumento significativo no número de profissionais, é natural observar um aumento correspondente na oferta de cursos, destinados tanto à formação de novos talentos quanto à especialização de profissionais, por meio de Programas de Pós-Graduação (PPGs). Para garantir a qualidade do ensino, não apenas nos cursos de Ciência da Computação, mas em todas as áreas oferecidas por universidades e faculdades em todo o país, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) realiza avaliações quadrienais. Essas avaliações atribuem notas de 1 a 7 a todos os PPGs, sendo que as notas 1 e 2 indicam que o curso é considerado inadequado para a emissão de certificados, enquanto as notas 7 denotam excelência no programa. O objetivo deste trabalho é realizar uma análise aprofundada das características que influenciam as notas de cada PPG em Computação. Para isso, utiliza-se dados bibliométricos dos professores credenciados e outras informações públicas das instituições de ensino envolvidas. O intuito é identificar os fatores determinantes que contribuem para a qualidade dos programas. Todos os dados bibliométricos foram retirados do banco de dados da OpenAlex, uma plataforma aberta que abrange trabalhos acadêmicos, autores, locais, instituições e conceitos. Com estes dados bibliométricos, entendemos quais destes dados fazem dos PPGs o que são atualmente, porém, predizer as notas dos PPGs a partir das características dos professores resultou em um F1-Score de 30% a partir da Regressão Logística. Entretanto, foi possível obter uma melhora no F1-Score para 70% para predição de notas dos programas, quando agrupadas em dois grupos, um com Professores de PPGs notas 3, 4 e 5, e outro com as notas 6 e 7. As características que melhor separaram os grupos foram a média de citação em cinco anos, H-Index, I10-Index e o número de publicações.

Palavras-chave: Pós-graduação. OpenAlex. Ciência da Ciência. Capes. Aprendizado de Máquina.

Abstract

It is estimated that by 2025, Brazil will need more than 797,000 qualified professionals in Computer Science. With the significant increase in the number of professionals, it is natural to observe a corresponding increase in the supply of courses, aimed at both training new talent and specializing professionals, through Postgraduate Programs (PPGs). To ensure the quality of education, not only in Computer Science courses, but in all areas offered by universities and colleges across the country, the Coordination for the Improvement of Higher Education Personnel (Capes) conducts four-year evaluations. These evaluations assign grades from 1 to 7 to all PPGs, with grades 1 and 2 indicating that the course is considered inadequate for the issuance of certificates, while grades 7 denote excellence in the program. The objective of this work is to carry out an in-depth analysis of the characteristics that influence the grades of each PPG in Computer Science. To do this, bibliometric data from accredited professors and other public information from the participating educational institutions are used. The intention is to identify the determining factors that contribute to the quality of the programs. All bibliometric data was retrieved from the OpenAlex database, an open platform that encompasses academic works, authors, locations, institutions, and concepts. With this bibliometric data, we understand which of these data make the PPGs what they are today, however, predicting the grades of the PPGs based on the characteristics of the professors resulted in an F1-Score of 30% using Logistic Regression. However, it was possible to obtain an improvement in the F1-Score to 70% for predicting program grades when grouped into two groups, one with PPG professors with grades 3, 4 and 5, and another with grades 6 and 7. The characteristics that best separated the groups were the average citation in five years, H-Index, I10-Index and the number of publications.

Keywords: Postgraduate, OpenAlex, Capes, Science of Science, Machine Learning.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

DECOM	Departamento de Computação
UFOP	Universidade Federal de Ouro Preto
PPG	Programa de Pós-Graduação
ROR	Research Organization Registry
API	Application Programming Interface
REST	Representational State Transfer
ID	Identificador

Sumário

1 Introdução

Segundo Fortunato *et al.* (2018), a Ciência da Ciência, ou do inglês *Science of Science* (SciSci), é um domínio que busca compreender quantitativamente a estrutura e evolução da pesquisa científica, analisando interações entre agentes científicos e as condições que favorecem a criatividade e descobertas. Seu objetivo primordial é desenvolver ferramentas e políticas para impulsionar o avanço científico, proporcionando uma compreensão mais profunda dos fatores que contribuem para o sucesso na ciência. SciSci se dedica a examinar os padrões que emergem no desenvolvimento de novas áreas científicas, os quais são revelados por meio de redes de colaboração, e a investigar as trajetórias das descobertas notáveis por meio da análise das redes de citações. Além disso, a área analisa as escolhas e desafios enfrentados pelos cientistas ao longo de suas carreiras e no cenário científico global, incluindo a preferência por pesquisar tópicos relacionados à sua expertise atual.

Conselho Federal de Educação (1965) nos explica que o sistema de Pós Graduação brasileiro no seu início foi baseado na estrutura da universidade americana, com cursos de graduação conduzindo ao bacharelado e cursos de Pós Graduação focados em estudos avançados para Mestrado ou Doutorado. O Mestrado é visto como um início para aqueles que desejam aprofundar sua educação científica ou profissional recebida em cursos de graduação, mas não têm a inclinação ou a capacidade para as atividades de pesquisa exigidas para um doutorado. Ele é útil para a competência profissional e pode levar a uma melhor remuneração. No entanto, para uma carreira no ensino superior, um doutorado é necessário.

Segundo a Capes (2019), em 2019, a Computação figurava como uma das 20 áreas com mais Programas de Pós Graduação (PPG) do Brasil. No Portal Sucupira, portal de coleta de informações do Sistema Nacional de Pós Graduação, podemos visualizar que a área contém cerca de 89 programas de Pós Graduação inscritos na CAPES, além de 132 cursos na área, todos eles sendo programas acadêmicos ou profissionais.

A Avaliação Quadrienal da CAPES para os programas de Pós Graduação no Brasil é o principal controle de qualidade (Capes, 2021). Cada programa é avaliado com notas que vão de 1 (um) até 7 (sete). Para o mestrado, nota três é o mínimo para se manter em funcionamento, diferente do doutorado, onde a nota de corte é quatro. Programas que tenham notas um ou dois são desqualificados como um programa pela CAPES, não podendo mais emitir certificações. Programas recém admitidos não podem receber notas inferiores a três. Programas com notas seis e sete são considerados programas com excelência internacional, já programas com notas 4 e 5, são considerados programas com excelência nacional. A Figura 1.1 mostra como é feita a avaliação de cada programa, desde a coleta de dados, passando pelo tratamento e análise da Comissão de Área e do Conselho Técnico-Cientifico da CAPES, até a divulgação da nota pela

própria CAPES a partir do Portal Sucupira.

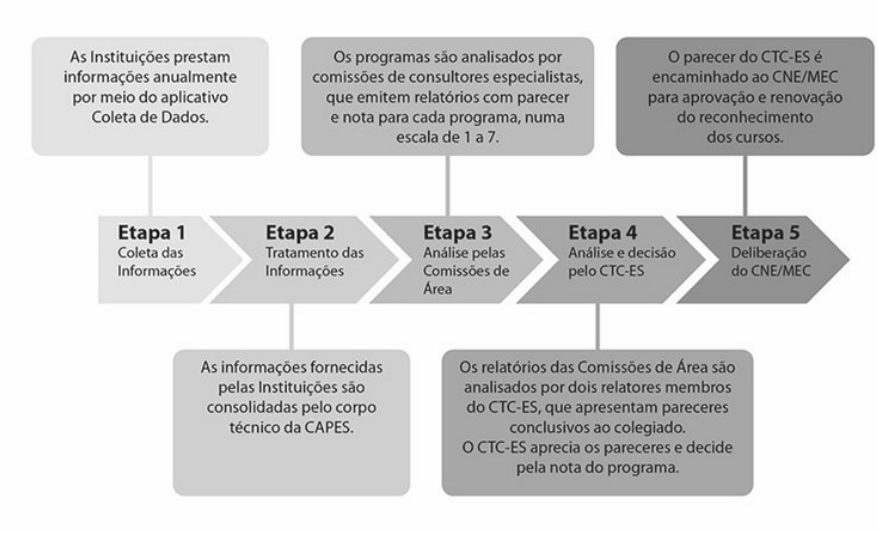

Figura 1.1 – Fluxograma da Avaliação Quadrienal. Fonte: (DAV/CAPES, 2021).

A OpenAlex é uma plataforma aberta que abrange trabalhos acadêmicos, autores, locais, instituições e conceitos. Como dizem Priem, Piwowar e Orr (2022), esta plataforma constitui um grafo direcionado heterogêneo, composto por cinco tipos distintos de entidades acadêmicas e as conexões que existem entre elas. A cada entidade da OpenAlex é atribuída a um ID OpenAlex persistente, que atua como chave primária no conjunto de dados. O conjunto de dados está disponível para acesso através do download de *snapshots* da base, uma API REST ou uma interface gráfica baseada na web. Todos esses recursos são oferecidos gratuitamente, sem a necessidade de registro ou permissão, com um número máximo diário de requisições à API na versão gratuita ou ilimitado na versão paga. O código subjacente à OpenAlex é totalmente aberto, acessível por meio da conta OurResearch no GitHub[1]. A OpenAlex atualmente indexa aproximadamente 109.000 instituições, utilizando o ROR ID como *Canonical External ID* (CEID) para identificar instituições de forma única. A Figura 1.2 nos mostra mais alguns números sobre a plataforma. Para garantir precisão, as afiliações listadas pelos autores são analisadas e normalizadas. Isso é alcançado por meio de um algoritmo de duas etapas que combina estágios baseados em regras e técnicas de aprendizado de máquina. Este processo visa estabelecer conexões sólidas entre instituições e trabalhos acadêmicos.

[1] <https://github.com/ourresearch/OpenAlex>

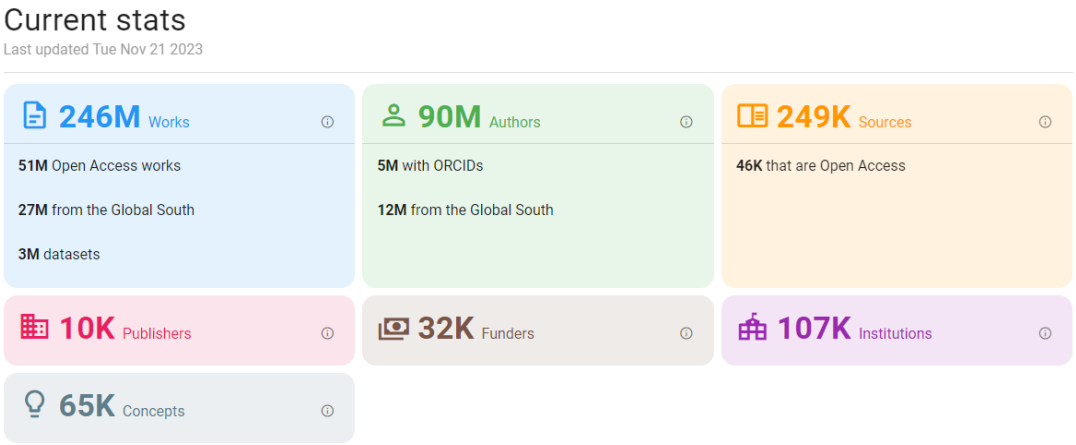

Figura 1.2 – Dados da OpenAlex no dia 21/11/2023 a partir do endereço <https://openalex.org/stats>.

O foco da presente monografia é, a partir de dados bibliométricos, entender quais características de um dado programa acadêmico de Pós Graduação em Ciência da Computação define a nota final na Avaliação Quadrienal da CAPES. A plataforma OpenAlex é usada para a criação da base de dados usada na monografia, agrupando dados de instituições credenciadas pela CAPES e de professores credenciados para lecionar em programas de Pós Graduação. Esta monografia está inserida no contexto de um projeto que visa criar uma base de dados dos professores credenciados nos PPGs em Computação do Brasil, fazendo um cruzamento dos dados disponibilizados pela CAPES, Currículo Lattes e os perfis desses professores e instituições na OpenAlex. A equipe do projeto está construindo a base de dados e analisando os dados sob diferentes perspectivas, incluindo *deep learning*, *embeddings*, ciência das redes e mineração de dados. A base de dados é composta de dados como um *id* único de cada professor, o orientador dele em seu doutorado, se ele possui alguma bolsa de produtividade da CAPES, e outros dados citados na Seção 3.1.2. Ao final, foi possível ter uma ideia de como tais características impactaram na nota final, seja de forma negativa ou positiva.

1.1 Justificativa

Atualmente, as notas são computadas a partir de diversos critérios como número de defesas de Mestrado e Doutorado; tempo médio dos alunos nos respectivos cursos, até a defesa; número de professores credenciados, etc. Busca-se avaliar aqui o quanto os dados bibliométricos, disponibilizados gratuitamente pela OpenAlex, são capazes de explicar as notas dos programas, sem o uso dos demais critérios.

1.2 Objetivos

O objetivo geral desta monografia é estimar a nota da Avaliação Quadrienal da CAPES de programas acadêmicos de Pós Graduação em Ciência da Computação a partir de dados bibliométricos dos professores credenciados nos respectivos programas. Para isso, é necessário definirmos alguns objetivos específicos como:

- Criar uma base de dados com dados dos PPGs em Computação do Brasil, juntamente com dados das respectivas instituições e professores credenciados, a partir da OpenAlex;
- Avaliar as notas de cada programa, encontrando as características que explicam a nota (relevância) do programa, a partir de testes estatísticos;
- Criar um classificador para, a partir das características bibliométricas de um programa, atribuir a ele uma nota entre 3 e 7.

1.3 Organização da Monografia

O Capitulo 2 apresenta a revisão bibliográfica, incluindo trabalhos relacionados e referencial teórico. O Capitulo 3 mostra como foi o processo de construção da metodologia empregada. O Capitulo 4 traz alguns resultados obtidos, junto de uma análise. Por fim o Capitulo 5 apresenta as considerações finais, junto de um cronograma de todas as atividades.

1.3.1 Estrutura da Monografia

Capítulo 1: Introdução.

Capítulo 2: Revisão Bibliográfica/ Embasamento Teórico (com o referencial teórico e trabalhos relacionados).

Capítulo 3: Metodologia.

Capítulo 4: Resultados.

Capítulo 5: Considerações finais.

2 Fundamentação Teórica

Neste capítulo, abordaremos toda a literatura que servirá de base teórica para este trabalho, realizando uma análise de alguns trabalhos relacionados e apresentando um referencial teórico sobre os métodos utilizados.

2.1 Trabalhos Relacionados

Em Goenner e Snaith (2004), os autores buscam prever a relevância de algumas universidades dos Estados Unidos que concedem cursos de doutorado a partir da Classificação de Carnegie (Shulman, 2001), utilizando como atributos características dos alunos de graduação, como idade, sexo, formação, entre outros. Foram analisados cursos de graduação com durações de 4, 5 e 6 anos, e uma análise estatística dos dados foi realizada, juntamente com uma regressão múltipla para prever os resultados. No final, os autores chegaram à conclusão de que compreender as características pessoais dos alunos é a forma ideal para determinar a produtividade individual dentro da universidade.

Em Steiner (2005) discute-se sobre a diversidade e qualidade institucional dos programas de Pós Graduação no Brasil no ano de 2003, quando tínhamos um total de 45 instituições com cursos de doutorado, e 75 de mestrado. O autor também apresenta uma análise sobre a qualidade do ensino de Pós Graduação no país. Além disso, busca-se fazer uma comparação entre a qualidade do ensino em instituições públicas, privadas e filantrópicas, com base na avaliação da Capes.

Pensando na construção da base de dados, Krause e Mongeon (2023) analisam as citações presentes na OpenAlex para entender a relação entre os criadores dos *datasets* e os autores que os citam nos níveis individual, institucional e nacional. A maioria dos países cita conjuntos de dados dos EUA com mais frequência do que os seus, e as instituições sediadas nos EUA costumam usar conjuntos de dados de outros países. Há uma baixa taxa de autocitação e citações correspondentes à instituição para conjuntos de dados. Os autores visam compreender as práticas de citação de dados pesquisando as relações entre autores citantes e criadores de dados.

Soares *et al.* (2016) fazem uma análise bibliométrica da produção científica brasileira na área de Tecnologias da Construção, usando o banco de dados Web of Science, buscando consolidar abordagens e metodologias específicas para indicadores bibliométricos nessa área. A Web of Science foi usada como fonte para avaliar a relação entre autores, instituições, estados, áreas do conhecimento e países dos artigos selecionados, somando um total de 910 artigos científicos publicados em revistas acadêmicas, anais de conferências e periódicos e anais de conferências relacionados na área e cobrindo o período de 1982 a 2014. Ao final, o estudo conclui que houve

um crescimento significativo na produção científica brasileira no campo das Tecnologias da Construção, conforme evidenciado pela análise de 910 artigos publicados entre 1982 e 2014. Também foi possível concluir que a região Sudeste do Brasil deu a maior contribuição para esse desenvolvimento, além de que as principais colaborações de pesquisa são com pesquisadores afiliados a instituições americanas, seguidos por pesquisadores da Inglaterra e da Espanha.

2.2 Fundamentação Teórica

2.2.1 Teste estatístico Mann-Whitney U

O teste estatístico Mann-Whitney U é não paramétrico, usado para comparar medianas entre duas amostras independentes. De acordo com McKnight e Najab (2010), este método é tido como a versão não paramétrica do teste t, já que ambos necessitam de dois grupos de amostra para verificar se ambos diferem em uma única variável contínua. Ainda segundo os autores, o Mann-Whitney U é utilizado para determinar se dois grupos distintos vêm da mesma população ou se possuem diferenças significativas. A formulação do teste considera as seguintes hipóteses:

- **Hipótese Nula (H_0)**: As duas amostras vêm da mesma distribuição, ou seja, não há diferença significativa entre as medianas das populações.
- **Hipótese Alternativa (H_1):** As amostras vêm de distribuições diferentes.

Supondo dois grupos (grupo a e grupo b), o primeiro passo é combinar obervações de ambos os grupos em um único grupo e ranqueá-los com notas de 1 a N, onde N é o tamanho dos grupos a e b unidos ($n_a + n_b = N$). Logo após o processo de ranqueamento, os indivíduos dos grupos são novamente separados a partir dos seus grupos originais e é feita a soma das notas separadamente para cada grupo (T_a e T_b). A equação abaixo, mostra como é feito o calculo da a estatística U:

$$U_a = T_a - \frac{n_a(n_a + 1)}{2}. \tag{2.1}$$

$$U_b = T_b - \frac{n_b(n_b + 1)}{2}. \tag{2.2}$$

Agora, calcula-se os valores de U_a e U_b e considera-se que a estatística será:

$$U = \min\{U_a, U_b\}. \tag{2.3}$$

No caso extremo de as amostras serem claramente separadas, ou seja, os valores de a são sempre menores (ou sempre maiores) que os de b, o valor da estatística U será menor que 0.05. Caso contrário, os valores obtidos podem variar entre 1 (inclusive) e $n_a n_b$ (inclusive). Quando o tamanho dos grupos é pequeno, há tabelas similares às utilizadas no teste t, para avaliar se os valores obtidos pela estatística obedecem a uma dada significância. Para grupos grandes, é

preciso computar o p-valor associado. Bibliotecas como o SciPy[1] possuem métodos[2] para cálculo da estatística que retornam o valor da estatística acompanhado do p-valor associado à hipótese alternativa, ou seja, de as populações serem distintas. Quanto menor o *p-value* retornado, menores as chances de os dois grupos serem iguais, isto é, eles possivelmente são diferentes.

2.2.2 Regressão Logística

Problemas de classificação não podem ser adequadamente abordados por meio de uma regressão linear, pois este método é destinado à tarefa de regressão, que consiste em predizer valores numéricos. Em situações onde a tarefa é definir categorias, como a cor dos olhos ou o tipo de cabelo, torna-se necessário recorrer a classificadores. Conforme mencionado por James *et al.* (2023), um classificador é uma ferramenta utilizada para prever uma resposta qualitativa para uma observação, atribuindo-a a uma categoria específica.

Dentro da categoria de classificadores, destaca-se a Regressão Logística como uma opção robusta. Esta técnica é aplicável tanto para prever variáveis cujos valores seguem uma ordem natural quanto para variáveis binárias. Ao contrário da regressão linear, a Regressão Logística é especialmente útil em contextos onde a resposta desejada é qualitativa.

É importante ressaltar que, de maneira geral, não existe uma abordagem natural para converter uma variável de resposta qualitativa com mais de dois níveis em uma forma quantitativa adequada para a regressão linear. Nesse cenário, a Regressão Logística se destaca como uma ferramenta valiosa para lidar com problemas de classificação e previsão em diversos contextos.

O modelo logístico é representado por:

$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X, \tag{2.4}$$

onde β_0 e β_1 são os coeficientes e $p(X)$ (ou Y) é o resultado da predição.

Resolvendo a Equação 2.4, o modelo final é descrito por:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}, \tag{2.5}$$

que representa as chances (*odds*) de algo acontecer para determinado valor de x.

Utiliza-se um método chamado *maximum likelihood*, ou máxima similaridade, para estimar os coeficientes β_0 e β_1, que são desconhecidos. Desta forma, diferentemente da Regressão Linear, onde o gráfico das predições é linear, o gráfico da Regressão Logística terá um formato em 'S', conforme ilustra a Figura 2.1.

[1] <scipy.org>
[2] <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html>

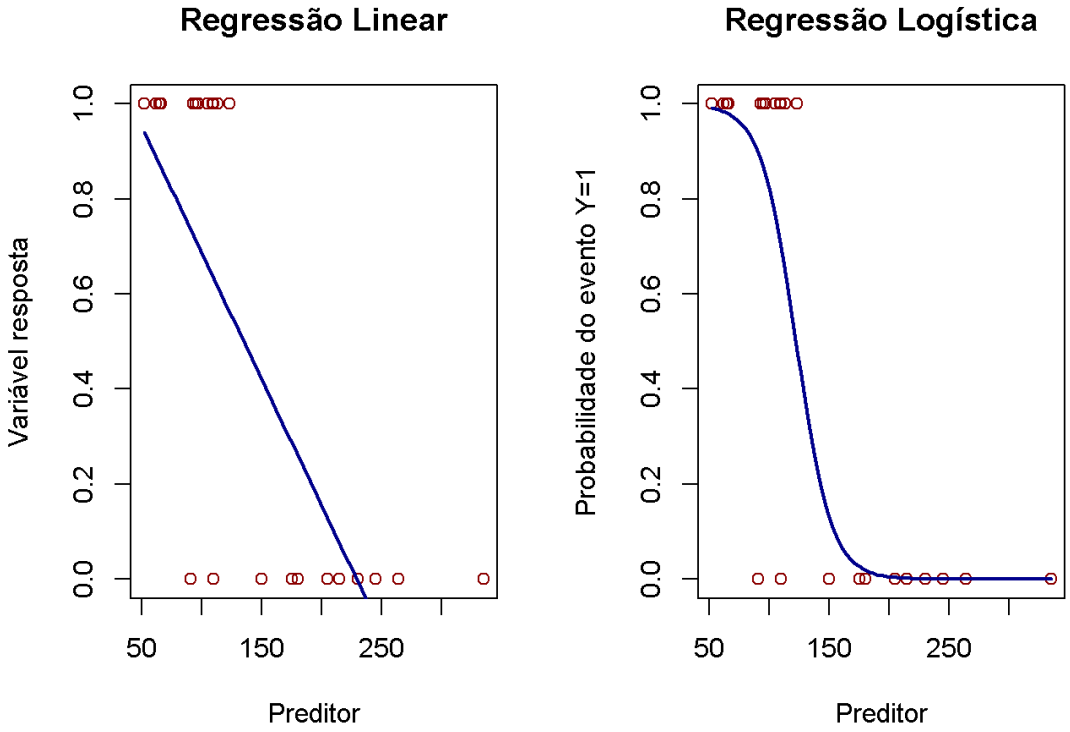

Figura 2.1 – Comparação entre os gráficos da Regressão Linear e a Logística Fonte: (Moreira, 2019).

Para saber se um modelo foi bem treinado, é necessário saber a acurácia. Durante o momento de testes de um modelo, a acurácia é calculada como:

$$\text{Acurácia} = \frac{\text{Soma dos Acertos}}{\text{Total de Previsões Realizadas}} \tag{2.6}$$

O maior dos problemas vêm quando trabalhamos com modelos desbalanceados, já que a predição correta de uma classe dominante pode inflar a métrica sem refletir a verdadeira capacidade do modelo, sendo necessário usarmos a *F1-Score*.

Murphy (2012) nos explica que a F1-Score é uma métrica que combina a Precisão e a Revocação em uma única métrica, e é útil em situações de classes desbalanceadas. Ela é uma média harmônica da precisão e da revocação, equilibrando os dois aspectos:

$$\text{F1-Score} = 2 * \frac{\text{Precisão} * \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{2.7}$$

A precisão é uma métrica que indica, das classificações de uma determinada classe do modelo, quantas foram acertadas. Já a revocação é uma métrica que indica, das amostras existentes de uma classe, quantas o modelo conseguiu classificar corretamente.

3 Desenvolvimento

Neste Capítulo, iremos abordar toda a organização do trabalho, partindo da criação da base de dados, a partir de informações advindas da plataforma OpenAlex, analise estatística das diferenças entre características dos professores de programas diferentes e predição da nota dos programas a partir delas.

3.1 Base de Dados

Inicialmente, criamos a base de dados que será utilizada nas análises durante a monografia. Os dados foram coletados a partir de diferentes fontes, sendo elas:

- Relatório de avaliação quadrienal da CAPES, de 2021 (Capes, 2022);
- Tabela de notas da área de Computação, da CAPES, anexo do relatório supracitado;
- Currículos Lattes[1] dos professores credenciados nos PPGs listados na tabela acima;
- OpenAlex[2].

3.1.1 Programas de Pós-Graduação

Primeiramente é importante saber quais PPGs estão credenciados pela CAPES e suas respectivas instituições. Desta forma é possível entender quais os cursos que tem uma melhor classificação na Avaliação Quadrienal e ter noção de quantos profissionais estão cadastrados. Os dados preenchidos de cada instituição são:

- Codigo_PPG: Um código de identificação do PPG;
- Nome_PPG: Nome do curso de Pós Graduação;
- institution_id: Id da instituição dentro da plataforma OpenAlex;
- institution_name: Sigla da instituição de início que detém o curso;
- Nota_PPG: Nota do curso na Avaliação Quadrienal.

Dados como o código do PPG, nome do PPG, a nota do PPG e o institution_name são retirados do comitê de área, já o institution_id é o ID da instituição dentro da OpenAlex.

[1] <https://lattes.cnpq.br/>
[2] <https://api.openalex.org>

3.1.2 Autores/Professores

Um ponto crucial da monografia foi agrupar dados de todos os professores credenciados pelas instituições, já que os vetores de características foram construídos a partir deles. A tabela contém os seguintes dados de cada professor:

- data_year: Ano da coleta dos dados pela CAPES;
- gp_code: Código do PPG em que o professor está credenciado;
- gp_name: Sigla do PPG em que o professor está credenciado;
- gp_score: Nota do PPG na Avaliação Quadrienal da CAPES;
- institution_id: ID da OpenAlex da instituição em que o professor leciona.
- institution_acr: Sigla da instituição de início que detém o curso;
- author_name: Nome completo do professor;
- phd_year: Ano em que o professor concluiu seu doutorado.
- regime_trabalho: Regime de trabalho em que o professor leciona dentro da instituição;
- carga_horaria: Carga horaria semanal do professor;
- lattes_cv_link: Link do Curriculo Lattes do professor;
- author_id: ID do professor dentro da plataforma OpenAlex;
- productivity_grant: Se o professor recebe ou não bolsa de produtividade pela CAPES;
- productivity_grant_type: Caso o professor receba bolsa de produtividade, esse campo especifica qual o tipo;
- phd_institution_id: ID da instituição em que o foi o autor fez seu doutorado, a partir da OpenAlex;
- phd_institution_name: Nome da instituição em que o foi o autor fez seu doutorado;
- phd_gp_code: Código da CAPES de identificação do PPG;
- phd_supervisor_id: ID do supervisor do doutorado do professor, a partir da OpenAlex;
- phd_supervisor_name: Nome do supervisor do doutorado do professor;

3.1.3 População da base de dados

Para a população da base de dados das instituições, como dito anteriormente, boa parte dos dados foram retirados do comitê de área. Para completar os dados, foi preciso pesquisar pela instituição usando a API da OpenAlex. A Figura 3.1 nos mostra um exemplo de como é feita a requisição e uma parte de como é a resposta, a partir do arquivo JSON retornado pela API.

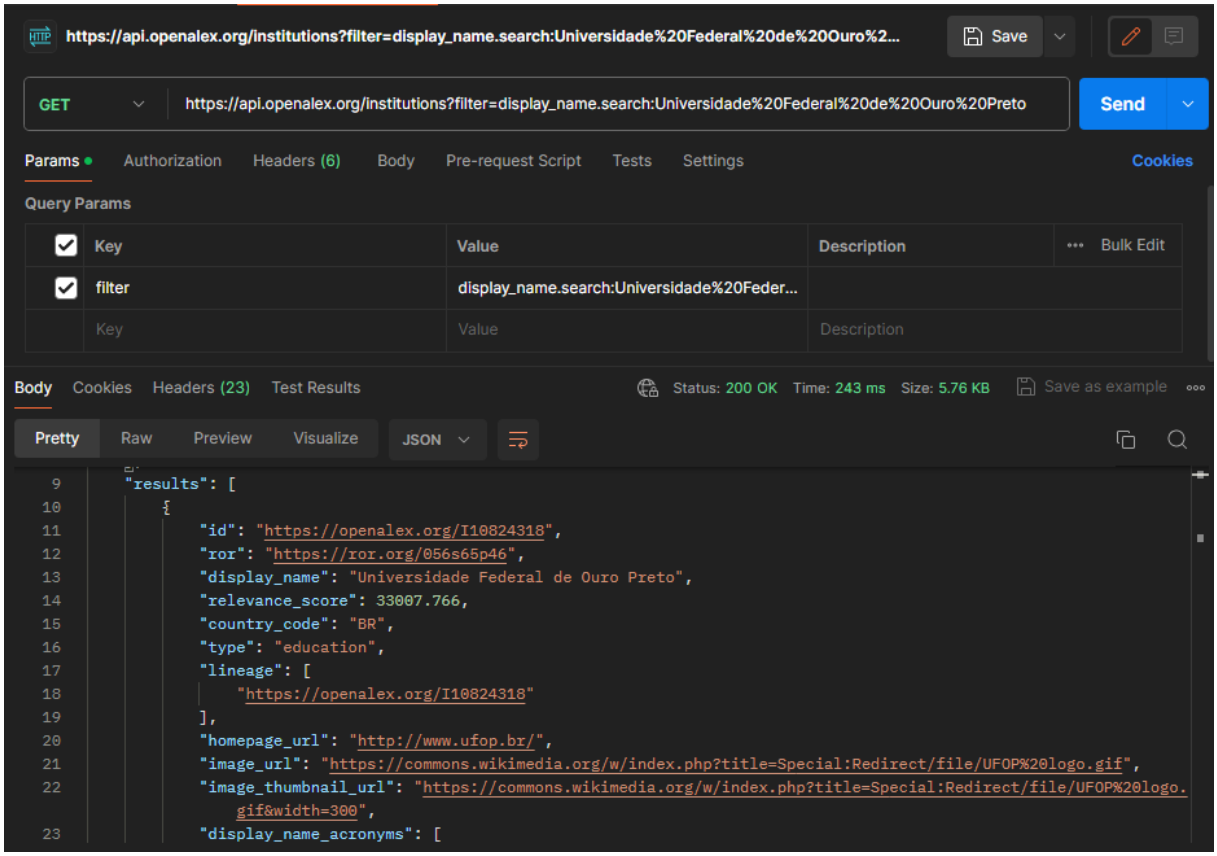

Figura 3.1 – Exemplo de requisição na OpenAlex, procurando as instituições de ensino.

Para popular a base de dados dos autores, buscou-se, inicialmente, dados dos professores na tabela anexa ao Relatório de avaliação quadrienal. Alguns outros dados só foram possíveis a partir de extrações de dados presentes no Currículo Lattes dos autores. Com o nome de cada um, é feita a pesquisa na plataforma Lattes (Figura 3.2), a partir da qual obtêm-se o nome do supervisor e instituição onde o autor fez seu doutorado.

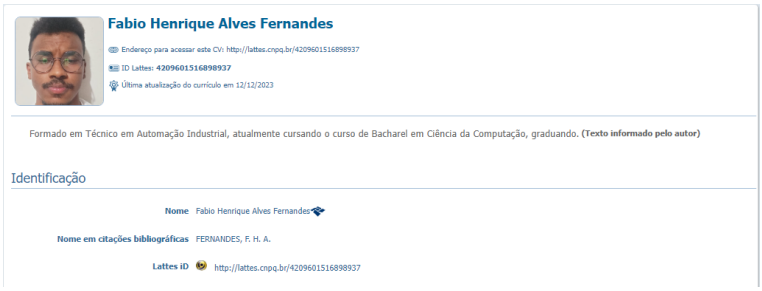

Figura 3.2 – Exemplo de um currículo Lattes.

Para finalizar, faltam alguns dados da plataforma da OpenAlex, que podem ser obtidos a partir da pesquisa pelo nome de cada autor, de cada supervisor, instituições em que cada autor fez seu doutorado que não estão no banco de dados preenchido anteriormente. A Figura 3.3 nos mostra como é feita a requisição de um autor pela OpenAlex.

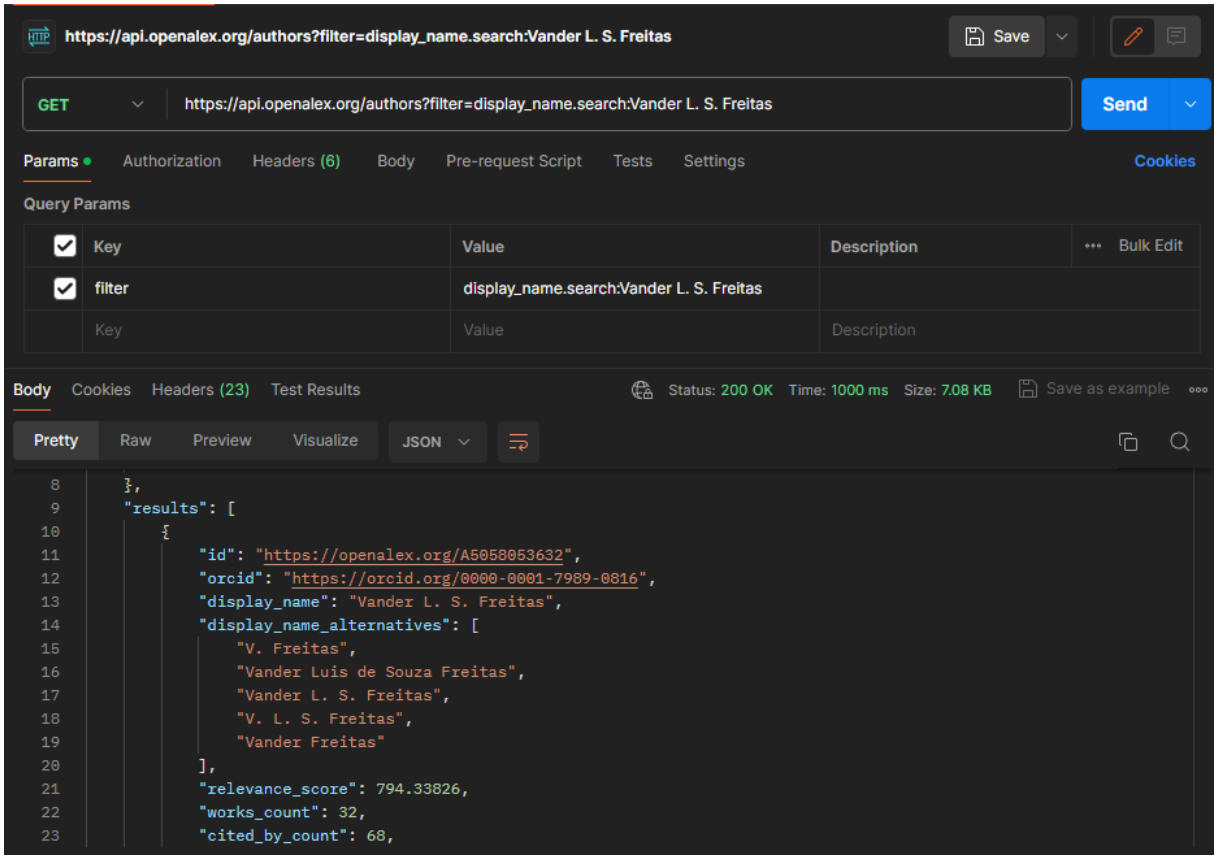

Figura 3.3 – Exemplo de requisição na OpenAlex, procurando autores.

3.2 Metodologia

3.2.1 Análise exploratória dos dados

Nesta seção verifica-se a existência de alguns possíveis padrões, sejam eles a partir das instituições de trabalho, instituição de formação e até dos próprios professores. Foram utilizados dados como a nota do PPG na Avaliação Quadrienal, quantidade de professores em determinada instituição, quantidade de professores que estão inseridos em cursos com uma determinada nota, entre outros. A primeira análise feita foi verificar a quantidade de professores em cada instituição de ensino. A Tabela 3.1 nos mostra essa relação entre cada curso dentro das instituições, já que algumas instituições possuem mais de um Programa de Pós-Graduação, além das notas que cada curso obteve.

Tabela 3.1 – Número médio de professores em cada Programa de Pós-Graduação em Ciência da Computação e as respectivas notas dos PPGs na Avaliação Quadrienal da CAPES de 2021.

Professores por Instituição			
Instituição	Número de Professores	Curso	Nota
UFPE	71	CC	7
UFRGS	57	C	7
PUC-RIO	20	I	7
UFMG	65	CC	7
UNICAMP	41	CC	7
UFRJ	58	ESC	7
USP/SC	61	CCMC	7
PUC/RS	21	CC	7
USP	50	CC	6
UFF	44	C	6
UFPR	34	I	6
UFAM	20	I	6
UNISINOS	11	CA	5
UFSC	27	CC	5
PUC/PR	14	I	5
UFOP	16	CC	5
UFPEL	19	C	5
UFPA	19	CC	5
UFSCAR	38	CC	5
UFC	46	CC	5

Continuação da Tabela			
Instituição	Número de Professores	Curso	Nota
UFRN	27	SC	5
UFU	28	CC	5
UFBA	32	CC	5
UNB	29	I	5
UFLA	14	CC	4
UNESP-SJRP	24	CC	4
UNIRIO	15	I	4
UFV	15	CC	4
UNIFOR	14	IA	4
UFSJ	14	CC	4
UFABC	26	CC	4
UFJF	14	CC	4
UDESC	13	CA	4
PUC/MG	13	I	4
UECE	12	CC	4
UNIVALI	12	C	4
UEL	11	CC	4
FURG	16	C	4
UEM	16	CC	4
IFCE	17	CC	4
UPE	17	EC	4
FUFPI	18	CC	4
UFMS	19	CC	4
UFCG	31	CC	4
UFG	28	CC	4
UFMA	17	CC	4
UFES	21	I	4
UNIFESP	20	CC	4
UFPB-JP	20	I	4
USP/RP	13	CA	3
UFAL	24	I	3
UFRPE	21	IA	3
UNIFACS	11	SC	3
UTFPR	12	CC	3
UFAC	13	CC	3
UFSM	20	CC	3

Continuação da Tabela			
Instituição	Número de Professores	Curso	Nota
UEFS	13	CC	3
FUFSE	19	CC	3
CEFET/RJ	14	CC	3
UNIFEI	19	CTC	3
UNIOESTE	15	CC	3
IME	11	SC	3

É possível notar que os PPGs que estão melhores colocados, nem sempre tem a maior quantidade de professores. Dois exemplos a serem citados são sobre a PUC/RS e a UNB, já que a PUC/RS tem 21 professores e nota máxima, enquanto a UNB tem 29 professores e nota 5.

A Tabela 3.2 entretanto é interessante para nos mostrar quais são as vinte e cinco instituições pelo mundo que mais formaram Doutores que hoje são professores nos PPGs da nossa base de dados.

Tabela 3.2 – Top 25 instituições que mais formaram os Professores da nossa base de dados, ordenadas pelo número de Professores formados por elas.

Instituições de Doutorado	Numero de Professores
Universidade de São Paulo	222
Universidade Estadual de Campinas	122
Universidade Federal de Pernambuco	115
Universidade Federal do Rio de Janeiro	107
Universidade Federal do Rio Grande do Sul	97
Pontifícia Universidade Católica do Rio de Janeiro	82
Universidade Federal de Minas Gerais	76
Universidade Federal de Santa Catarina	36
Universidade Federal do Ceará	32
Universidade Federal Fluminense	27
Universidade Federal de Campina Grande	25
Universidade Federal do Paraná	18
Imperial College London University Of London	16
University Of Kent At Canterbury	14
Universidade Federal do Rio Grande do Norte	13
Instituto Tecnológico de Aeronáutica	11
Institut National Polytechnique de Grenoble	11
Universidade Federal da Bahia	11
Université Pierre et Marie Curie - Paris 6	10
University of Edinburgh	10
University of California	10
Instituto Tecnológico de Aeronáutica	9
Pontifícia Universidade Católica do Rio Grande do Sul	9
Université Joseph Fourier - Grenoble I	9
University of Waterloo	8

Pode-se notar que PPGs bem renomados no Brasil costumam exportar professores para outras instituições, como por exemplo a Universidade de São Paulo, que apareceu com 222 professores que terminaram seus Doutorados na instituição.

A Tabela 3.3 mostra a quantidade de professores presentes em PPGs, porém separados pelas notas da Avaliação. Ela corrobora como que foi discutido sobre a independência entre o número de professores e a nota dos programas, como visto na Tabela 3.1, já que podemos notar que a quantidade de professores em PPGs de nota 6 é menor do que os de nota quatro.

Tabela 3.3 – Número de professores presentes nos PPGs com relação às notas da Avaliação Quadrienal.

Nota	Autores
7	369
6	129
5	279
4	481
3	232

A Tabela 3.4 traz um calculo da média de anos em que os professores obtiveram seus doutorados, estratificada pela média das notas dos PPGs. Com ela é possível notar que PPGs com notas superiores possuem Professores que defenderam seus Doutorados em geral antes dos Professores de PPGs de notas menores. Veja os PPGs nota sete, onde a média de formação é de 2003, diferentemente dos cursos nota três por exemplo, onde a média é do ano de 2010.

Tabela 3.4 – Média de ano em que cada professor obteve seu doutorado para cada nota da Avaliação Quadrienal

Nota	Media
7	2003
6	2004
5	2008
4	2008
3	2010

3.2.2 Caracterização dos professores dos PPGs

O próximo passo é construir o vetor de características/atributos dos professores. Dentro dele, teremos os dados mencionados na Seção 3.1.2 e mais alguns dados provenientes da OpenAlex[3], como:

- Afiliações;
- Número total de citações;
- Número de citações divididos por ano;
- Data de criação do perfil dentro da plataforma;
- IDs de outras plataformas que o perfil possa ter, como ORCID, SCOPUS, Twitter e Wikipedia;
- Últimas instituições conhecidas;

[3] Mais informações sobre os dados que a OpenAlex disponibiliza sobre os autores podem ser encontrados em <https://docs.openalex.org/api-entities/authors/author-object>.

- Métricas de citação do autor;
- Data da última modificação
- Um link que contém todas as publicações do autor;
- A quantidade de publicações do autor;
- Uma lista que contém todas as áreas de interesse do professor.

Com este vetor preenchido, é possível comparar os valores dos atributos dos professores que estão inseridos em PPGs com notas diferentes, sendo possível avaliar quais atributos explicam a diferença entre essas notas. Usando o Mann-Whitney U, queremos saber quais atributos obtiveram valores de U mais próximos a 0 (zero), o que mostra que fazem parte de grupos distintos. Desta forma podemos verificar quais atributos melhor separam os grupos, isto é, PPGs de notas diferentes.

3.2.3 Mann Whitney U

Após a extração das características de cada autor, é necessário separar os autores em bases de dados separadas, agrupados pelas notas de seus respectivos PPGs. Desta forma é possível realizar os teste estatístico para cada característica, considerando-se a estratificação dos Professores a partir das notas dos PPGs. O objetivo é verificar quais atributos melhor separam os grupos e em quais os grupos são semelhantes.

3.2.4 Modelo de Regressão Logística

Com relação ao modelo de Regressão Logística, o dataset foi separado em conjunto e treinamento e teste, na proporção 80% e 20%, respectivamente. Também foi utilizada validação cruzada, para ajudar a avaliar o desempenho do modelo. A validação cruzada divide a base de dados em novos conjuntos chamados *folds*, treinando o modelo várias vezes com vários outros conjuntos de treino e teste, sendo esses parte dos 80% dos dados que foram separados para treinamento.

Todo o trabalho foi feito na linguagem Python. Na Seção 3.2.3 foi utilizada a biblioteca *SciPy* (Jones; Oliphant; Peterson, 2001), que contém métodos capazes de realizar o teste estatístico. Já na Seção 3.2.4, utilizaremos a biblioteca *Scikit-Learn* (Geron, 2021), que contém implementação de vários modelos de aprendizado de máquina, incluindo a Regressão Logística.

4 Resultados

Este Capítulo apresenta os resultados obtidos, tanto nos testes estatísticos com o Mann Whitney U quando na predição dos dados usando a Regressão Logística. Inicialmente, discute-se os resultados dos testes estatísticos para cada característica e logo após sobre os resultados das predições como um todo.

4.1 Mann Whitney U

4.1.1 Número de citações

Seguindo a definição do Mann Whitney U, vemos que na Tabela 4.1, podemos assumir que os grupos referentes aos autores que pertencem a PPGs de nota 5 e 6 são estatisticamente parecidos, em se tratando do atributo relacionado a número de citações. O valor obtido de p-value, associado à hipótese alternativa (as variáveis são distintas), entre os dois foi maior do que o limite de 0.05, diferentemente dos outros grupos, mostrando que eles são completamente distintos uns dos outros.

Tabela 4.1 – Valores de *p-value* para o teste Mann Whitney U, relacionados à hipótese alternativa, para a variável número de citações de cada autor, estratificada pelas notas dos PPGs

	Nota 3	Nota 4	Nota 5	Nota 6	Nota 7
Nota 3	——	0.000000	0.000000	0.000000	0.000000
Nota 4		——	0.000000	0.000000	0.000000
Nota 5			——	**0.060628**	0.000000
Nota 6				——	0.000132
Nota 7					——

O gráfico presente na Figura 4.1 ajuda a ilustrar discrepâncias entre a quantidade de citações entre os programas mais bem ranqueados em relação aos menos ranqueados. Podemos notar que os programas de nota 7 tem uma frequência muito baixa de autores com nenhuma citação, ao contrário de programas nota 3, onde acontece o completo oposto.

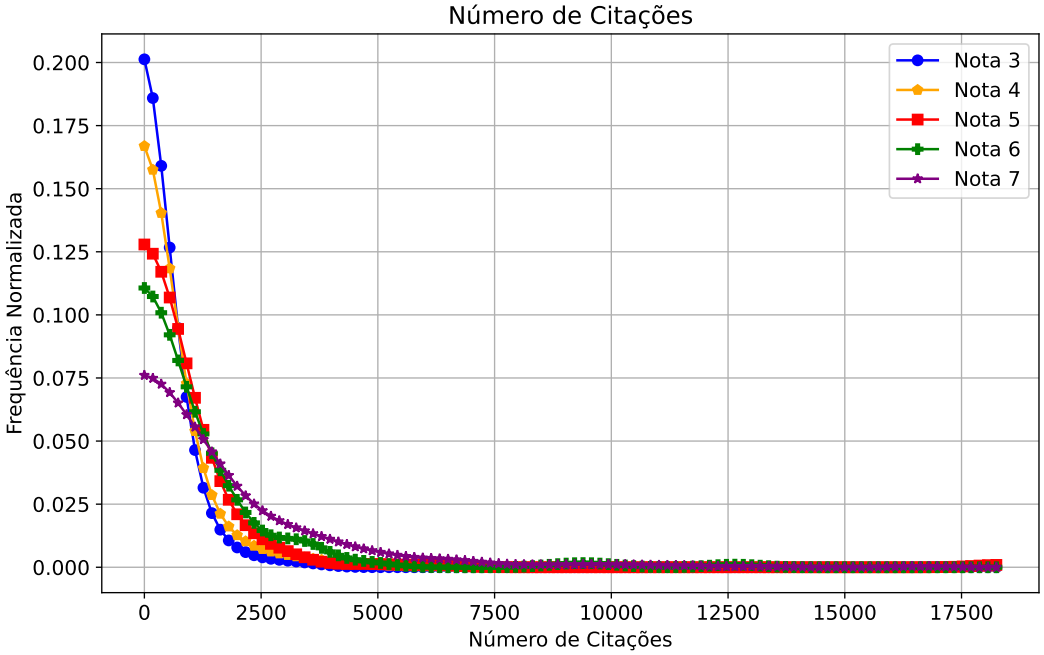

Figura 4.1 – Histograma do número de citações separados pelas notas dos PPGs na área de Computação.

4.1.2 Número de artigos publicados

Em se tratando do atributo número de artigos publicados, a Tabela 4.2 apresenta valor de *p-value* muito próximo do limiar pré-definido, mas menor do que ele entre os grupos que representam as notas 5 e 6. Ou seja, apesar de o *p-value* ser próximo de 0,05, considera-se que os grupos são distintos. Nos outros casos, os p-values são baixos, indicando também que são todos distintos no atributo em questão.

Tabela 4.2 – Valores de *p-value* para o teste Mann Whitney U, relacionados à hipótese alternativa, para a variável número de artigos publicados de cada autor, estratificada pelas notas dos PPGs.

	Nota 3	Nota 4	Nota 5	Nota 6	Nota 7
Nota 3	——	0.000000	0.000000	0.000000	0.000000
Nota 4		——	0.000027	0.000000	0.000000
Nota 5			——	0.049183	0.000000
Nota 6				——	0.005927
Nota 7					——

No gráfico da Figura 4.2 podemos ver que para nesta característica acontece o mesmo comportamento do anterior, onde é possível perceber que os autores de cursos nota 7 tem mais publicações em relação aos outros. É possível ver também uma discrepância muito alta entre

a frequência de autores com nenhuma publicação nos grupos dos autores de cursos nota 3 em relação aos outros.

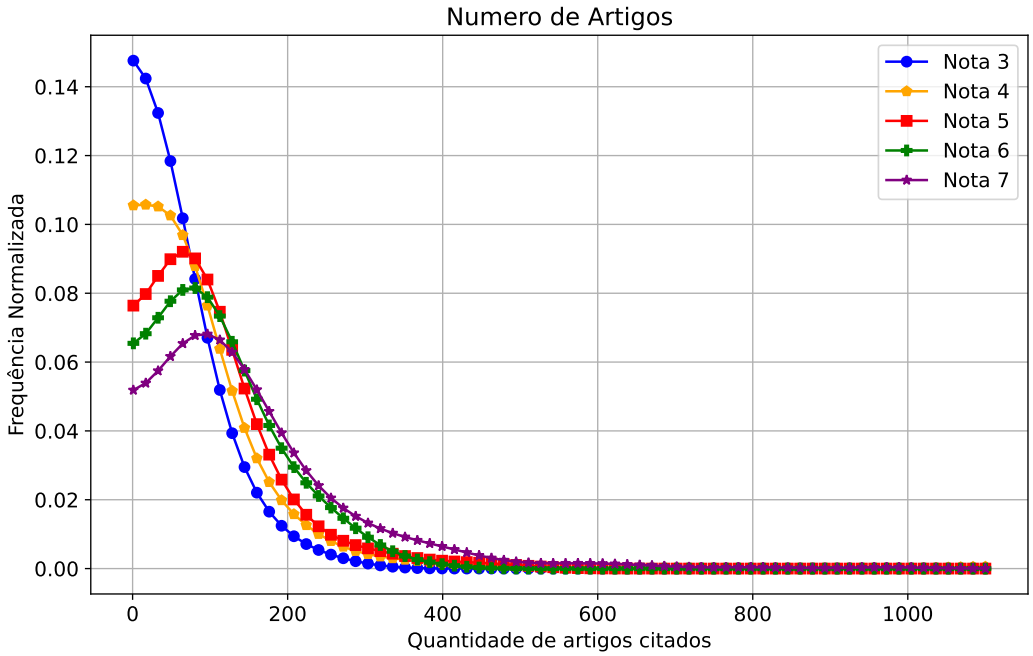

Figura 4.2 – Histograma do número de artigos publicados, separados pelas notas dos PPGs na área de Computação.

4.1.3 Média de citações em dois anos

Na Tabela 4.3, relacionada ao teste Mann Whitney U para a variável *média de citações em dois anos*, conseguimos ver uma certa semelhança entre vários grupos, como nos grupos que representam as notas 3 e 6, os grupos que representam as notas 4 e 5, notas 4 e 6, notas 4 e 7, notas 5 e 6, notas 5 e 7 e notas 6 e 7, mostrando que nessa característica em específico não há tantas discrepâncias entre os grupos.

Tabela 4.3 – Valores de *p-value* para o teste Mann Whitney U, relacionados à hipótese alternativa, para a variável média de citações em dois anos de cada autor, estratificada pelas notas dos PPGs.

	Nota 3	Nota 4	Nota 5	Nota 6	Nota 7
Nota 3	——	0.004044	0.000164	0.037536	0.000000
Nota 4		——	**0.182020**	**0.894468**	0.000560
Nota 5			——	**0.252964**	**0.074526**
Nota 6				——	0.007581
Nota 7					——

O gráfico representado pela Figura 4.3 ilustra bem toda a semelhança, mostrando bem

como os grupos estão bem parecidos, mesmo com a contínua discrepância entre os grupos que representam as notas 3 e 7.

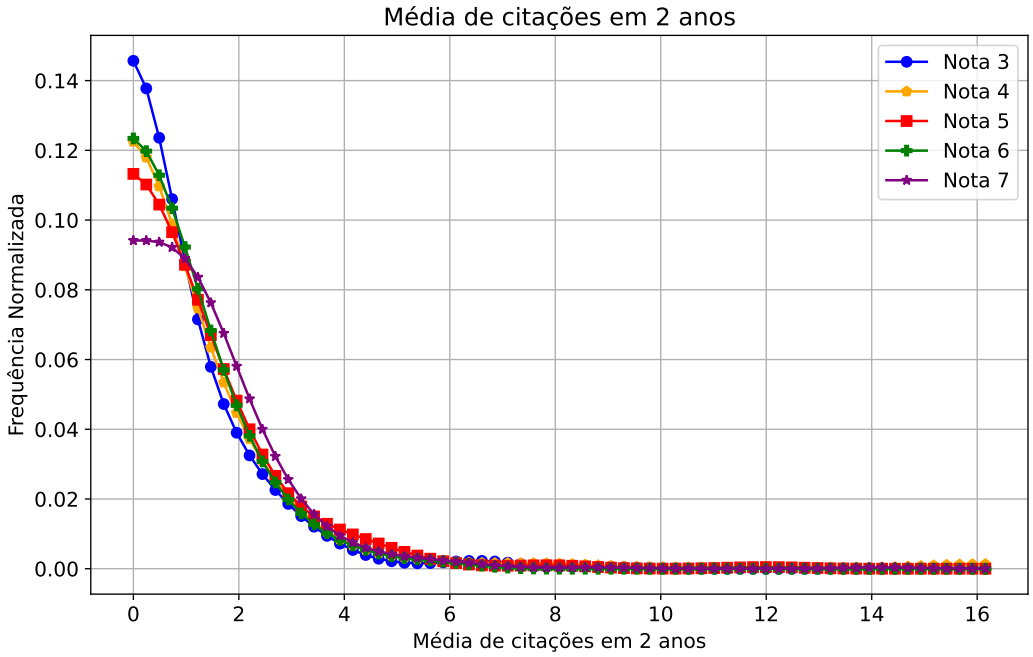

Figura 4.3 – Histograma da média de citações em 2 anos, separados pelas notas dos PPGs na área de Computação.

4.1.4 Bolsa produtividade

Na Tabela 4.4 é possível notar que somente os grupos que representam as notas 5 e 6 tem um *p-value* que corroboram com a hipótese nula de que as variáveis possuem mesma distribuição, e os grupos que representam as notas 4 e 5 estão bem próximos.

Tabela 4.4 – Valores de *p-value* para o teste Mann Whitney U, relacionados à hipótese alternativa, para a variável que diz respeito a existência ou não de bolsa produtividade para cada autor, estratificada pelas notas dos PPGs.

	Nota 3	Nota 4	Nota 5	Nota 6	Nota 7
Nota 3	——	0.000174	0.000000	0.000000	0.000000
Nota 4		——	0.028095	0.000715	0.000000
Nota 5			——	**0.125640**	0.000000
Nota 6				——	0.003008
Nota 7					——

O gráfico representado pela Figura 4.4 é diferente dos demais pois representa um valor do tipo *Boolean*, então a melhor opção é um gráfico de barras. Com isso, podemos perceber que o grupo que representa a nota 3 tem poucos ou quase nenhum autor que recebe bolsa de

produtividade, diferentemente dos grupos que representam os programas notas 6 e 7, que tem uma presença maior.

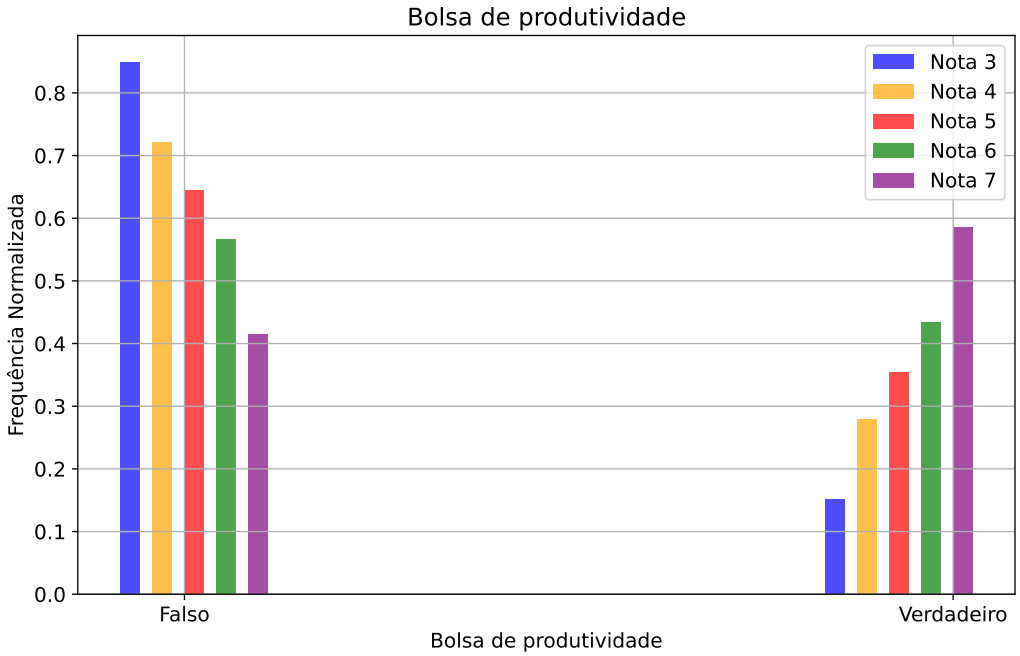

Figura 4.4 – Histograma da Existência ou não de Bolsa Produtividade, Separados pelas notas dos PPGs na área de Computação.

4.1.5 Ano de doutorado

Na Tabela 4.5, podemos perceber que os grupos que tem um valor de *p-value* aceitável são os grupos que representam as notas 4 e 5 e os grupos que representam as notas 6 e 7.

Tabela 4.5 – Valores de *p-value* para o teste Mann Whitney U, relacionados à hipótese alternativa, para a variável ano de doutorado de cada autor, estratificada pelas notas dos PPGs.

	Nota 3	Nota 4	Nota 5	Nota 6	Nota 7
Nota 3	——	0.000001	0.000168	0.000000	0.000000
Nota 4		——	**0.512028**	0.000006	0.000000
Nota 5			——	0.000009	0.000000
Nota 6				——	**0.338054**
Nota 7					——

Já olhando para o gráfico da Figura 4.5, podemos notar que os autores com maior idade acadêmica estão presentes nos grupos que representam os programas notas 6 e 7. O contrário também ocorre para os cursos representam os programas notas 3 e 4, já que são neles que estão concentrados os autores com a titulação de doutor mais recente.

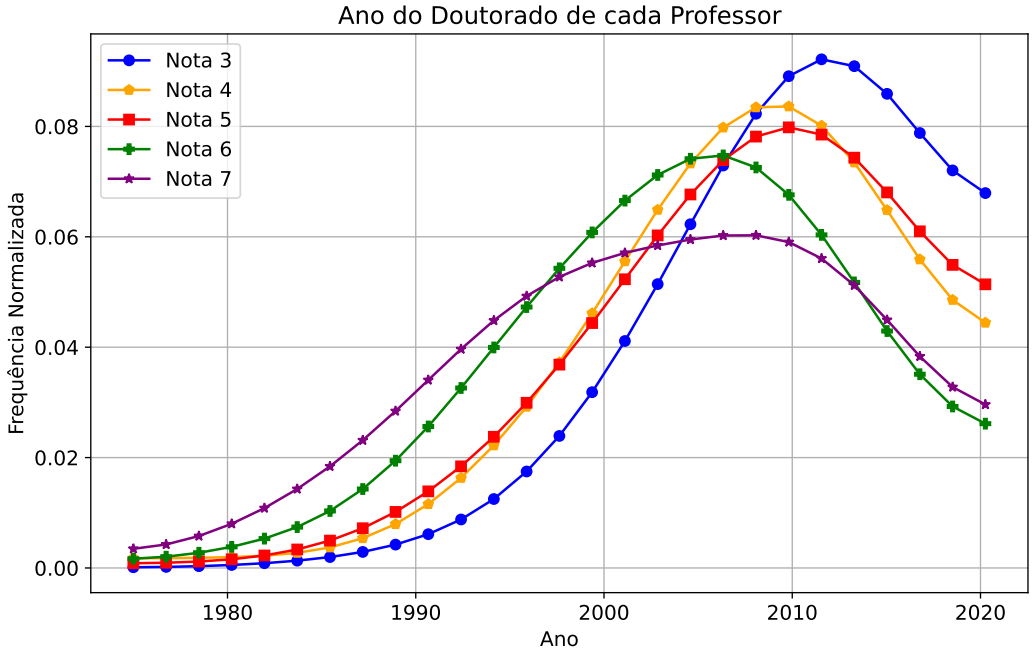

Figura 4.5 – Histograma do ano de doutorado, separados pelas notas dos PPGs na área de Computação.

4.1.6 H-Index

Na Tabela 4.6 podemos notar que nenhum dos grupos, dois a dois, podem ser considerados advindos da mesma distribuição, ou seja, são distintos.

Tabela 4.6 – Valores de *p-value* para o teste Mann Whitney U, relacionados à hipótese alternativa, para a variável H-Index de cada autor, estratificada pelas notas dos PPGs.

	Nota 3	Nota 4	Nota 5	Nota 6	Nota 7
Nota 3	——	0.000000	0.000000	0.000000	0.000000
Nota 4		——	0.000000	0.000000	0.000000
Nota 5			——	0.039905	0.000000
Nota 6				——	0.000276
Nota 7					——

No gráfico da Figura 4.6 conseguimos notar que a maior concentração de professores com uma boa métrica H-Index continua sendo de professores que fazem parte do grupo dos programas de nota 7, onde poucos dos autores tem essa métrica com valor 0. Já os grupos que contém os cursos de nota 3 ainda é totalmente o contrário, tendo vários autores com esta métrica com valor 0, e uma frequência muito baixa onde as métricas são altas.

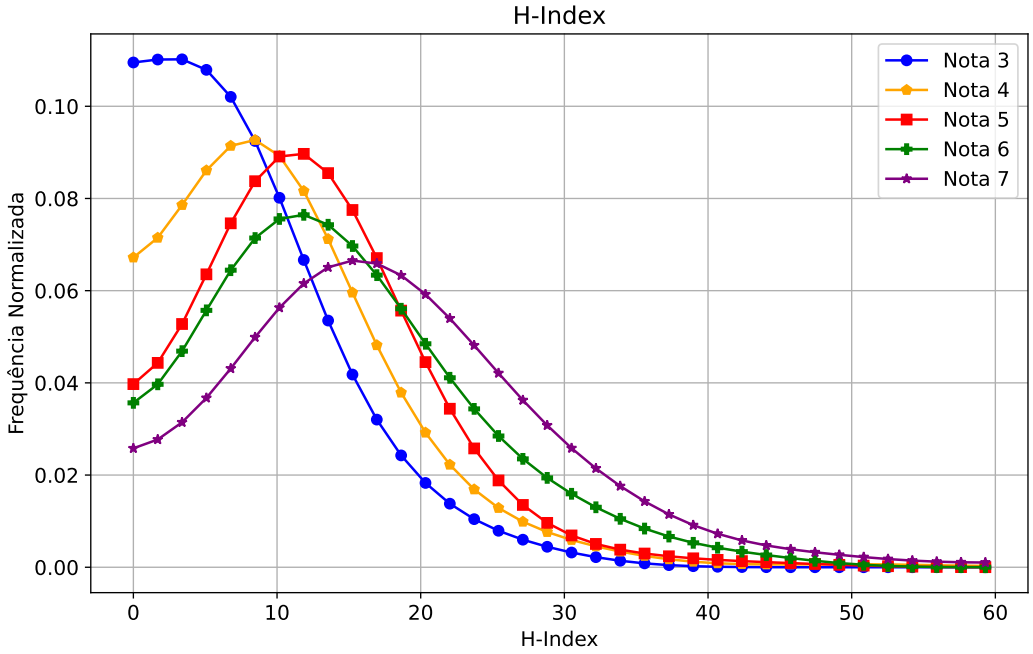

Figura 4.6 – Histograma do H-Index, separados pelas notas dos PPGs na área de Computação.

4.1.7 I10-Index

Na Tabela 4.7 acontece o mesmo do caso anterior, onde podemos considerar que todos os grupos são distintos.

Tabela 4.7 – Valores de *p-value* para o teste Mann Whitney U, relacionados à hipótese alternativa, para a variável I10-Index de cada autor, estratificada pelas notas dos PPGs.

	Nota 3	Nota 4	Nota 5	Nota 6	Nota 7
Nota 3	——	0.000000	0.000000	0.000000	0.000000
Nota 4		——	0.000000	0.000000	0.000000
Nota 5			——	0.042954	0.000000
Nota 6				——	0.000095
Nota 7					——

A Figura 4.7 nos mostra exatamente o mesmo ocorrido no gráfico anterior, com o grupo que contém os programas de nota 7 tendo menos autores com a métrica I10-Index com valor 0 do que os outros, e sendo o grupo em que a frequência de boas métricas é maior.

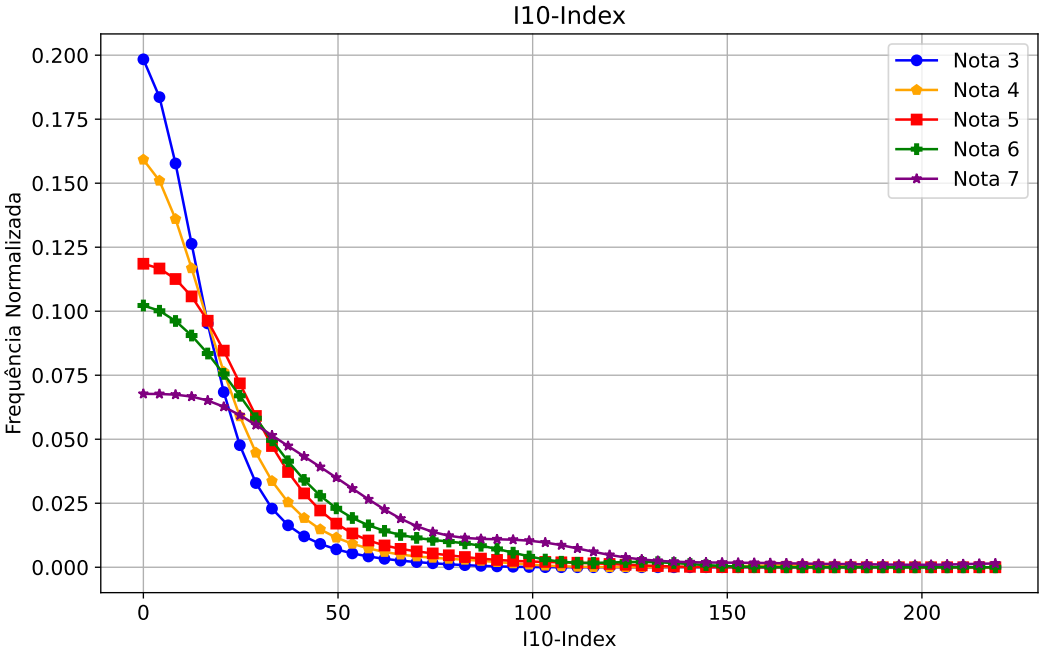

Figura 4.7 – Histograma do I10-Index, separados pelas notas dos PPGs na área de Computação.

4.1.8 Número de trabalhos publicados nos últimos cinco anos

Na Tabela 4.8 os grupos que contém as notas 5 e 6, as notas 5 e 7 e as notas 6 e 7 são os únicos que, par a par, tem seu valor de *p-value* aceitáveis.

Tabela 4.8 – Valores de *p-value* para o teste Mann Whitney U, relacionados à hipótese alternativa, para a variável número de trabalhos publicados nos últimos cinco anos de cada autor, estratificada pelas notas dos PPGs.

	Nota 3	Nota 4	Nota 5	Nota 6	Nota 7
Nota 3	——	0.000002	0.000000	0.000000	0.000000
Nota 4		——	0.001710	0.004311	0.000019
Nota 5			——	**0.643816**	**0.307942**
Nota 6				——	**0.727065**
Nota 7					——

Já graficamente, a Figura 4.8 traduz um pouco da similaridade estatística mencionada. Desta vez, o grupo que contém os autores pertencentes a programas nota 7 estão um pouco mais próximo dos grupos que contém os de programa nota 5 e 6.

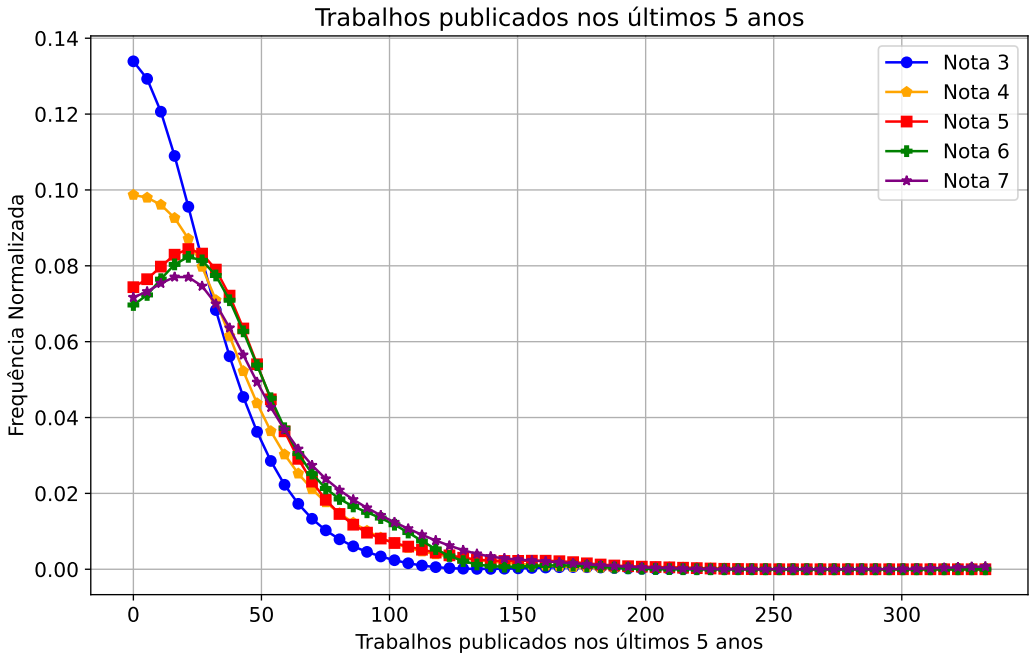

Figura 4.8 – Histograma do número de trabalhos publicados nos últimos cinco anos, separados pelas notas dos PPGs na área de Computação.

4.1.9 Número de citações nos últimos cinco anos

E na Tabela 4.9 mostra que somente os grupos que contém os autores de programas de notas 5 e 6 podem ser considerados advindos da mesma distribuição.

Tabela 4.9 – Valores de *p-value* para o teste Mann Whitney U, relacionados à hipótese alternativa, para a variável número de citações nos últimos cinco anos de cada autor, estratificada pelas notas dos PPGs.

	Nota 3	Nota 4	Nota 5	Nota 6	Nota 7
Nota 3	——	0.000000	0.000000	0.000000	0.000000
Nota 4		——	0.000003	0.000013	0.000000
Nota 5			——	**0.393281**	0.000000
Nota 6				——	0.000272
Nota 7					——

Graficamente, na Figura 4.9, conseguimos ver a similaridade entre os grupos que contém os autores de programas de notas 5 e 6, e a mesma discrepância entre o grupo dos autores de nota 7 com os demais, assim como boa parte das outras características já mencionadas.

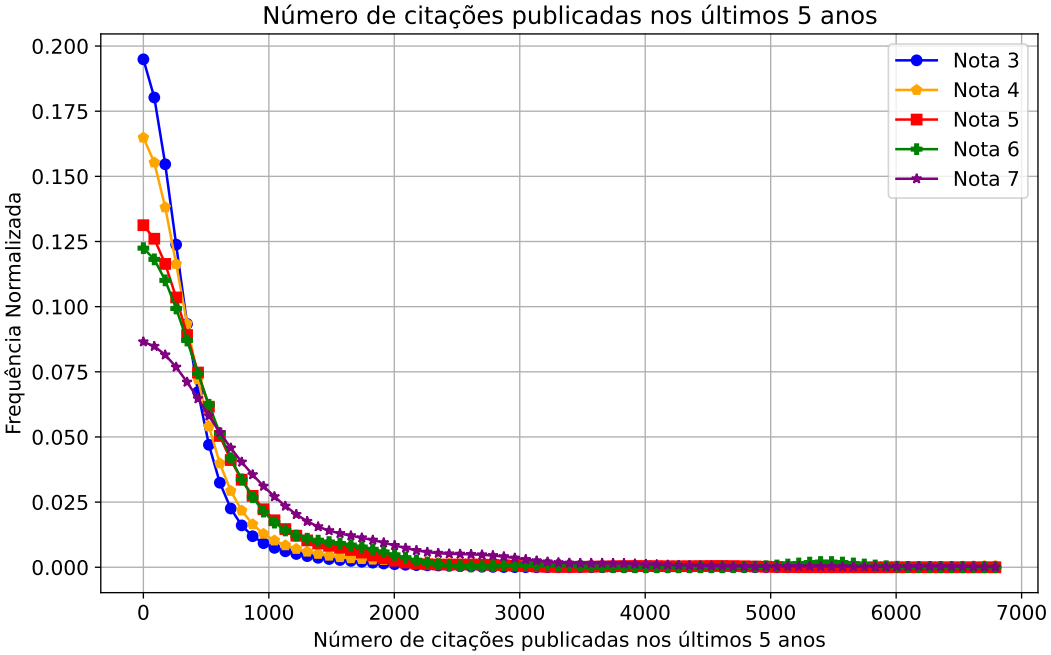

Figura 4.9 – Histograma do número de citações nos últimos cinco anos, separados pelas notas dos PPGs na área de Computação.

4.2 Regressão Logística

Para a base de dados atual, foram usadas todas as características mencionadas anteriormente, atingindo uma F1-Score de apenas 32%.

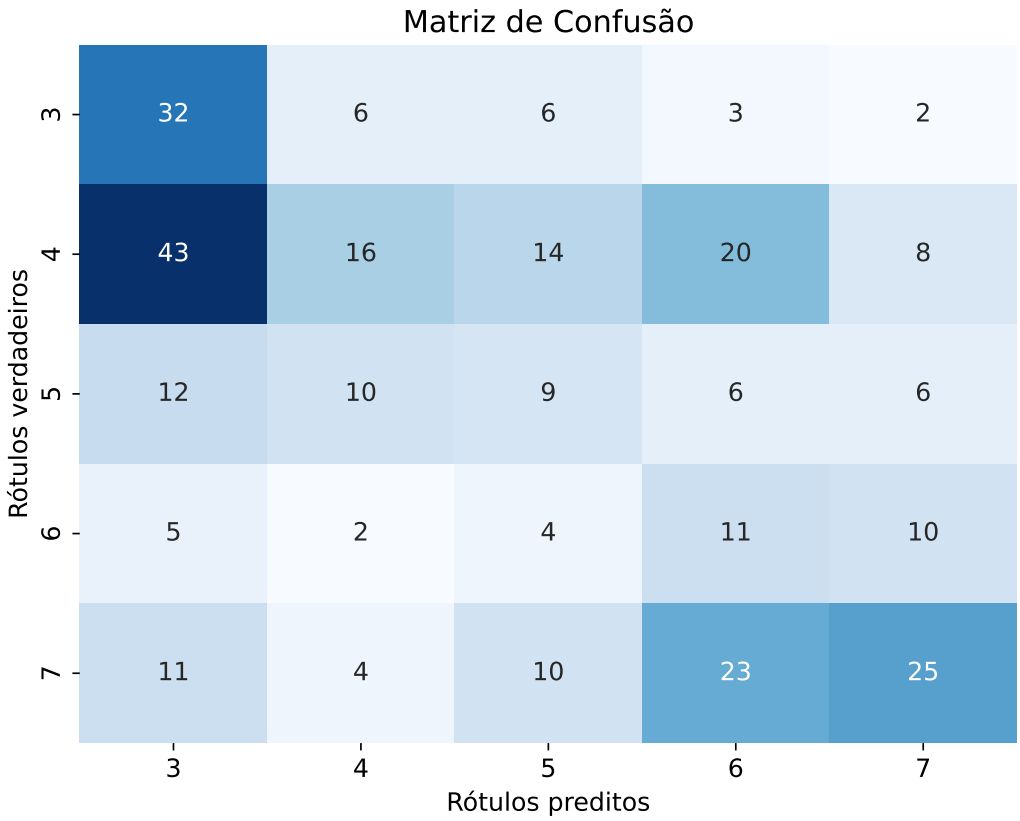

Figura 4.10 – Matriz de confusão da primeira tentativa de predição

A Figura 4.10 é uma Matriz de Confusão que mostra os erros e acertos da predição para o conjunto de teste que foi separado. Como dito, é notável que os resultados não são satisfatórios. Por conta destes resultados, foi necessário uma reclassificação das classes dos autores. Optou-se por transformar o problema de classificação de multi classes para o caso binário, onde os autores de notas que vão de 3 até 5 fazem parte da classe '0' e os autores de notas que vão de 6 a 7 fazem parte da classe '1'. Dessa forma, todos os testes estatísticos e o modelo de Regressão Logístico precisaram ser refeitos.

4.3 Testes estatísticos após alterar os problema para classificação binária

alterar os rótulos das classes (problema binário), os valores de *p-value* para os pares entre os grupos que integram os grupos das novas notas 0 e 1 negaram a hipótese nula para todas as características. A característica com maior estatística calculada foi a média de citação em dois anos, como podemos ver na Tabela 4.10

Tabela 4.10 – Valores de *p-value* para o teste Mann Whitney U, relacionados à hipótese alternativa, para a variável número de citações nos últimos cinco anos de cada autor, estratificada pelas notas dos PPGs após a reclassificação.

	Classe 0	Classe 1
Classe 0	——	0.000263
Classe 1	0.000263	——

Como podemos notar pela Figura 4.11, é possível notar nos novos histogramas uma grande semelhança com os histogramas antes da reclassificação.

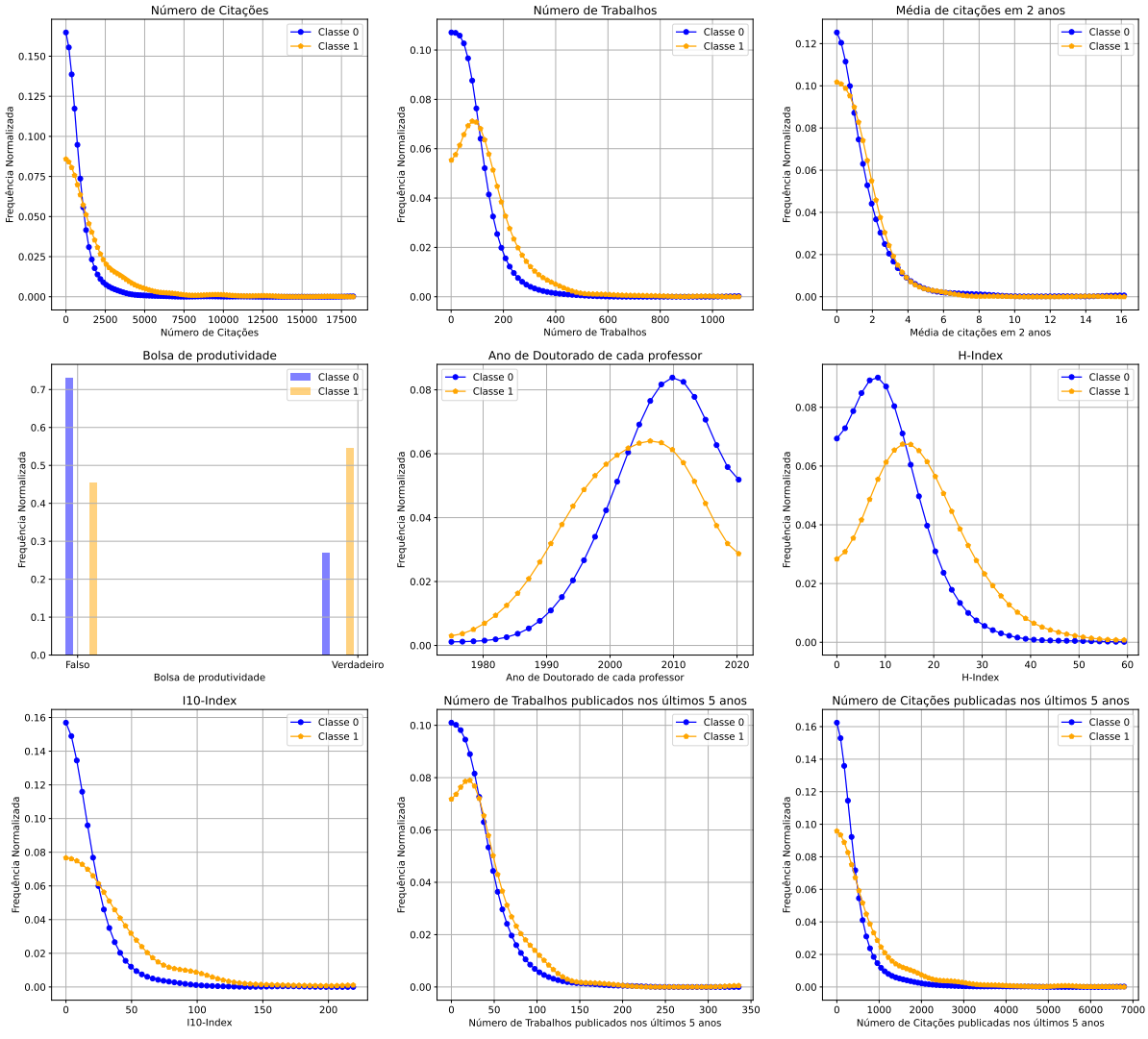

Figura 4.11 – Histogramas das características após a reclassificação.

4.4 Regressão Logística após atribuir novos rótulos às classes

Com as novas classes, houve uma melhora significativa na F1-Score do modelo, chegando a 71%. A Figura 4.12 mostra a Matriz de Confusão após a atribuição dos novos rótulos. Agora é possível notar a visível mudança na precisão do modelo.

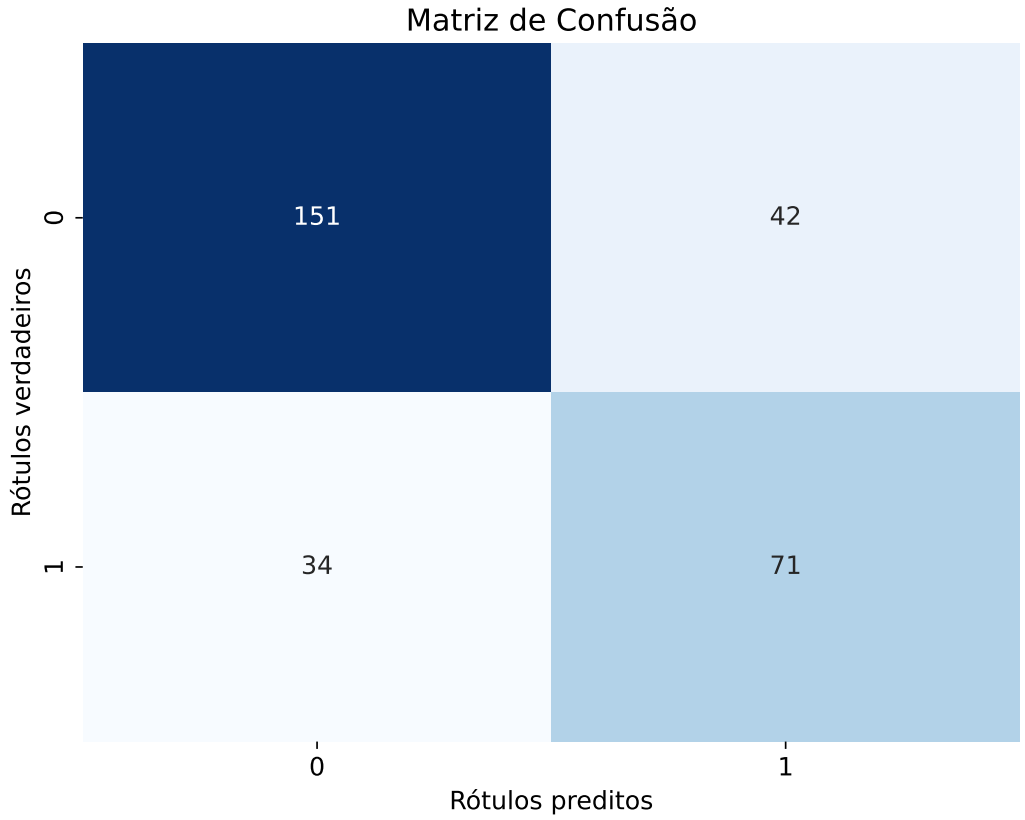

Figura 4.12 – Matriz de confusão da classificação após a atribuição dos novos rótulos

A Tabela 4.11 mostra os coeficientes da Regressão Logística. Os coeficientes representam a influência de cada característica sobre a probabilidade de um determinado resultado.

Para este caso, se o coeficiente é positivo, significa que um aumento no valor da característica levará a uma maior probabilidade de o resultado tender para a classe 1, e se caso tivermos um coeficiente negativo, um aumento no valor da característica levará a uma maior probabilidade de o resultado tender para a classe 0.

Tendo isso em mente, pode-se sugerir que a característica I-10 Index tem uma forte influência na probabilidade de se predizer a classe 1, ao contrário da característica de número de publicações em cinco anos, que tem uma forte influência na probabilidade de se predizer a classe 0.

Tabela 4.11 – Coeficientes da Regressão Logística

Característica	**Coeficiente**
Média de citações em 2 anos	-0.118073
Número de citações em 5 anos	-0.339638
Número de publicações em 5 anos	-0.037546
Número de citações	0.147540
H Index	0.302345
I10 Index	0.462498
Ano de doutorado	-0.219935
Bolsa de produtividade	0.148556
Número de publicações	-0.214852

5 Considerações Finais

De forma quadrienal a CAPES faz uma coleta de dados e aplica uma avaliação Programas de Pós Graduação. O objetivo principal da presente monografia é coletar e usar dados bibliométricos, visando entender se é possível ou não prever as notas que a CAPES atribui aos programas.

Neste trabalho, usamos do Mann Whitney U para realizar testes estatísticos, com o intuito de inferir em quais características bibliométricas de Programas de Pós-Graduação na área de Computação de notas diferentes se diferem. Também utilizamos de um modelo de Regressão Logística para tentar predizer as notas dos cursos em que autores dos PPGs estão inseridos.

Para cada característica, o Mann Whitney U foi capaz de distinguir corretamente quais grupos podem ser considerados advindos da mesma distribuição. Por isso, pode-se encontrar algumas características que tornam as notas dos Programas de Pós Graduação voltados à área de Ciência da Computação o que são. Também é possível perceber a discrepância entre os programas mais bem ranqueados em relação à programas com notas mais baixas, visto que seus números são bem maiores.

Em relação à Regressão Logística, a predição não foi bem sucedida usando as notas originais pois estamos tentando fazer a predição de uma nota de um curso a partir dos dados de um único autor. Um PPG pode ter uma nota alta e, ao mesmo tempo, ter professores com características abaixo da média, fazendo com que a predição em si não tenha uma boa acurácia. A principal solução encontrada foi criar novas classes (grupos), o que fez com que o modelo conseguisse melhorar a predição de forma satisfatória. Ao terminar a predição, os coeficientes da regressão nos ajudaram a determinar as características que melhor contribuem para predizer as notas dos programas. No caso, as características que mais contribuem para prever professores que estão contidos no grupo da Classe 0 são o número de citações em cinco anos e também o ano de doutorado do professor. Já as características que mais ajudam a prever professores que estão contidos no grupo da Classe 1 são o H-Index e o I10-Index.

Para trabalhos futuros, podemos incluir junto das características, o ano de criação dos PPGs, pois se um programa recente colocar professores mais antigos, poderá causar ruído durante a classificação. Podemos também criar melhores classes para cada autor, ajudando na predição. Uma outra possível solução é utilizar um novo classificador, que consiga lidar com os problemas já mencionados.

Referências

CAPES. **Documento da Área - Área 02: Ciência da Computação**. 2019. Acesso em 19 de novembro de 2023. Disponível em: <https://www.gov.br/capes/pt-br/centrais-de-conteudo/ccomp-pdf#:~:text=Enquanto%20a%20graduaÃğÃčo%20jÃą%20estÃą,dados%20de%20abril%20de%202019>.

CAPES. **Avaliação Quadrienal segue até 23 de dezembro de 2022**. 2021. Acesso em: 17 de novembro de 2023. Disponível em: <https://www.gov.br/capes/pt-br/assuntos/noticias/avaliacao-quadrienal-segue-ate-23-de-dezembro-de-2022>.

CAPES. **Resultado da Avaliação Quadrienal 2017-2020**. 2022. Acesso em: 24 de janeiro de 2024. Disponível em: <https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/avaliacao/avaliacao-quadrienal/resultado-da-avaliacao-quadrienal-2017-2020>.

Conselho Federal de Educação. **Parecer CFE n°977/65**. 1965. Acesso em 19 de novembro de 2023. Disponível em: <https://www.scielo.br/j/rbedu/a/NsLTtFBTJtpH3QBFhxFgm7L/?format=pdf&lang=pt>.

DAV/CAPES. **Sobre a Quadrienal**. 2021. Acesso em: 20 de novembro de 2023. Disponível em: <https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/avaliacao/avaliacao-quadrienal/sobre-a-quadrienal>.

FORTUNATO, S. *et al.* Science of science. **Science**, v. 359, n. 6379, p. eaao0185, 2018. Disponível em: <https://www.science.org/doi/abs/10.1126/science.aao0185>.

GERON, A. **Mãos à obra: aprendizado de maquina com Scikit-Learn, Keras & Tensorflow**. [S.l.]: Rio de Janeiro: Alta Books, 2021.

GOENNER, C. F.; SNAITH, S. M. Predicting graduation rates: An analysis of student and institutional factors at doctoral universities. **Journal of College Student Retention: Research, Theory & Practice**, v. 5, n. 4, p. 409–420, 2004. Disponível em: <https://doi.org/10.2190/LKJX-CL3H-1AJ5-WVPE>.

JAMES, G. *et al.* **An introduction to statistical learning: With applications in python**. [S.l.]: Springer Nature, 2023.

JONES, E.; OLIPHANT, T.; PETERSON, P. Scipy: Open source scientific tools for python. 01 2001.

KRAUSE, G.; MONGEON, P. Measuring data re-use through dataset citations in openalex. In: . [s.n.], 2023. Disponível em: <https://dapp.orvium.io/deposits/6442d8d30f5efe988a0e1d67/view>.

MCKNIGHT, P. E.; NAJAB, J. Mann-whitney u test. In: ____. **The Corsini Encyclopedia of Psychology**. John Wiley Sons, Ltd, 2010. p. 1–1. ISBN 9780470479216. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470479216.corpsy0524>.

MOREIRA, E. F. **Regressão Logística em R**. 2019. Acesso em: 26 de janeiro de 2024. Disponível em: <https://rpubs.com/dudubiologico/545528>.

MURPHY, K. P. **Machine learning: a probabilistic perspective**. [S.l.]: MIT press, 2012.

PRIEM, J.; PIWOWAR, H.; ORR, R. **OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts**. 2022. Acesso em 10 de outubro de 2023.

SHULMAN, L. S. The carnegie classification of institutions of higher education. **Menlo Park: Carnegie Publication**, Citeseer, 2001.

SOARES, P. B. *et al.* Análise bibliométrica da produção científica brasileira sobre tecnologia de construção e edificações na base de dados web of science. **Ambiente Construído**, Associação Nacional de Tecnologia do Ambiente Construído - ANTAC, v. 16, n. 1, p. 175–185, Jan 2016. ISSN 1678-8621. Disponível em: <https://doi.org/10.1590/s1678-86212016000100067>.

STEINER, J. E. Qualidade e diversidade institucional na pós-graduação brasileira. **Estudos Avançados**, Instituto de Estudos Avançados da Universidade de São Paulo, v. 19, n. 54, p. 341–365, May 2005. ISSN 0103-4014. Disponível em: <https://doi.org/10.1590/S0103-40142005000200019>.