UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
FACULDADE DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

LUCAS MESQUITA RODRIGUES FERREIRA

ANÁLISE EXPLORATÓRIA DOS MICRODADOS DO POSCOMP: UM ESTUDO DE CASO COM OS CANDIDATOS DO PARÁ

Belém
2023

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
FACULDADE DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

LUCAS MESQUITA RODRIGUES FERREIRA

ANÁLISE EXPLORATÓRIA DOS MICRODADOS DO POSCOMP: UM ESTUDO DE CASO COM OS CANDIDATOS DO PARÁ

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Reginaldo Cordeiro dos Santos Filho

Belém
2023

Ferreira, Lucas Mesquita Rodrigues
Análise Exploratória dos Microdados do POSCOMP: Um Estudo de Caso com os Candidatos do Pará/ LUCAS MESQUITA RODRIGUES FERREIRA. – Belém, 2023.
61 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Reginaldo Cordeiro dos Santos Filho

Monografia – UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO, 2023.

1.Análise de Desempenho, 2.Análise exploratória, 3.Mineração de dados, 4.POSCOMP, 5.Pós Graduação. I. Título.

LUCAS MESQUITA RODRIGUES FERREIRA

ANÁLISE EXPLORATÓRIA DOS MICRODADOS DO POSCOMP: UM ESTUDO DE CASO COM OS CANDIDATOS DO PARÁ

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Ciência da Computação.

Data da Defesa: 29 de Maio de 2023
Conceito:

Banca Examinadora

Prof. Dr. Reginaldo Cordeiro dos Santos Filho
Faculdade de Computação - UFPA
Orientador

Profª. Dra. Marcelle Pereira Mota
Faculdade de Computação - UFPA
Membro da Banca

Prof. Dr. Filipe de Oliveira Saraiva
Faculdade de Computação - UFPA
Membro da Banca

Belém
2023

Aos meus pais, que sempre me apoiaram nessa jornada.

AGRADECIMENTOS

Agradeço à minha mãe Elis Regina, por sempre me apoiar no percurso acadêmico e pessoal, e pelo amor, conselhos e incentivo nos momentos mais difíceis desta jornada, a senhora sempre foi ativa na minha educação, procurando as melhores escolas para que eu pudesse me preparar, me dando suporte para que minha preocupação fosse somente o estudo, sem a senhora nada seria possível.

Agradeço ao meu pai Carlos José, por todo o esforço investido na minha educação, e pelo amor e apoio que foram concedidos, sempre me orientando sobre o caminho do estudo, seu esforço para que eu frequentasse boas escolas e para que eu focasse somente no estudo durante a graduação foi essencial para que eu chegasse a esse ponto.

Também agradeço ao restante de minha família pelo apoio, incentivo, e contribuição para a minha formação, através de informações e conselhos sobre oportunidades e o caminho do estudo.

Sou grato ao orientador do meu trabalho, professor Reginaldo Cordeiro, pela proposta da realização deste trabalho e pela confiança depositada na minha execução, obrigado por me manter motivado e pelas orientações durante todo o processo.

Também sou grato ao colega da pós-graduação Jean Carlos por me auxiliar na pesquisa e na construção de um artigo publicado. Aos meus amigos e eventuais colegas que me auxiliaram e apoiaram antes e durante a graduação, no âmbito pessoal, prestando conselhos e estando presentes em diversos momentos, e acadêmico, onde possibilitaram a conclusão deste e de outros trabalhos.

Ao CNPQ, por ter me dado a oportunidade de iniciar na produção científica através da iniciação científica. Sou grato também à Universidade Federal do Pará e o seu corpo docente que demonstrou estar comprometido com a qualidade e excelência do ensino, o que me permitiu o crescimento acadêmico.

Por fim, agradeço e dedico este trabalho a todos que me apoiaram, incentivaram e contribuíram de algum modo para minha entrada na universidade e formação acadêmica, essas ações permitiram meu crescimento pessoal e profissional.

“Sonhos... Todo homem tem sonhos...
Todo homem deseja perseguir seu sonho.
Isso o aflige, mas o sonho dá sentido à sua vida.”
(Griffith)

RESUMO

O POSCOMP é o exame nacional para ingresso na pós-graduação em computação, e atua como avaliador dos egressos dos cursos da área de computação e afins, onde a nota é usada como um dos requisitos para a entrada em programas de pós-graduação. Neste contexto avaliativo, é importante que os programas de pós-graduação conheçam um pouco do perfil dos candidatos ao processo de seleção, seja a partir de resultados obtidos em um tema, seja pelos interesses em determinadas especialidades manifestados por parte dos candidatos no ato da inscrição da prova. Dessa forma, este estudo tem como objetivo analisar e explorar os microdados do POSCOMP nas edições de 2016 a 2019 através de técnicas de mineração de dados, seguindo as etapas do processo KDD, a fim de extrair informações e conhecimento relevante, no intuito de disponibilizar os resultados para os gestores acadêmicos, candidatos, e demais interessados, elucidando os desempenhos e perfis dos candidatos ao programa de pós-graduação provenientes do estado do Pará no exame. Os resultados mostram que o desempenho dos candidatos residentes do Pará é próximo ao nacional em grande parte dos temas do exame, alcançando 80% do desempenho obtido nacionalmente em todos os temas, com exceção de lógica matemática, os resultados também mostram que engenharia de software e inteligência artificial são as especialidades mais pretendidas por parte dos candidatos residentes no Pará, e o número de mulheres inscritas no exame no estado apresentou alta desde 2018, seguindo tendências diferentes das nacionais.

Palavras-chave: Análise de Desempenho, Análise exploratória, Mineração de dados, POSCOMP, Pós-graduação.

ABSTRACT

POSCOMP is the national exam for entry into post-graduation programs in computing, and acts as an evaluator of graduates of computing courses and related areas, where the score is used as one of the requirements for entry into post-graduation programs. In this evaluative context, it is important for post-graduation programs to know a little about the profile of candidates for the selection process, either from the results obtained in a theme, or by the interests in certain specialties manifested by the candidates at the time of registration for the exam. Thus, this study aims to analyze and explore POSCOMP microdata from 2016 to 2019 editions with data mining techniques, following the steps of KDD process, in order to extract information and relevant knowledge, with the aim of making the results available to academic managers, candidates, and other interested parties, elucidating the performance and profile of the applicants to the post-graduation program from the state of Pará in the exam. The results show that the performance of candidates residing in Pará is close to the national performance in most of the subjects of the exam, reaching 80% of the national performance in all subjects, except mathematical logic, the results also show that software engineering and artificial intelligence are the most desired specialties by candidates, and the number of women enrolled in the exam on the state presented a rise since 2018, being different trends from the national ones.

Keywords: Performance analysis, Exploratory analysis, Data Mining, POSCOMP, Post-graduation.

LISTA DE ILUSTRAÇÕES

Figura 1 – Médias de acerto e Desvio Padrão do POSCOMP por edição. 16
Figura 2 – Etapas da exploração. 26
Figura 3 – Etapas de pré-processamento. 27
Figura 4 – Média de notas por edição. 37
Figura 5 – Inscritos brasileiros e residentes do Pará por edição. 38
Figura 6 – Média de acertos de residentes do Pará por tema normalizada considerando todos os anos (2016-2019). 40
Figura 7 – Média de acertos de brasileiros por tema normalizada considerando todos os anos (2016-2019). 41
Figura 8 – Média de acertos de residentes do Pará por tema normalizada em relação à média nacional considerando todos os anos (2016-2019). 42
Figura 9 – Média de acertos total dos residentes do Pará e Brasil por área normalizada considerando todos os anos (2016-2019). 43
Figura 10 – Porcentagem de acerto por tema (Brasil x Pará). 44
Figura 11 – 7 especialidades mais pretendidas no Pará considerando todos os anos (2016-2019). 46
Figura 12 – Alunos por especialidade de pós-graduação na UFPA. 46
Figura 13 – Nuvem de palavras com base nas especialidades mais pretendidas no Pará. . 47
Figura 14 – Número de inscritos por sexo (Pará). 48
Figura 15 – Número de inscritos por sexo (Brasil). 49
Figura 16 – Ausentes por edição (Brasil x Pará). 50
Figura 17 – Média de notas por edição em barra (Brasil x Pará) 58
Figura 18 – Média de acertos por tema por ano (Pará) 59
Figura 19 – Média de acertos por tema por ano (Brasil) 60

LISTA DE TABELAS

Tabela 1 – Áreas do POSCOMP e respectivos temas. 15
Tabela 2 – Tipos possíveis por atributo na base respostas. 28
Tabela 3 – Tipos possíveis por atributo na base gabarito. 29
Tabela 4 – Tipos possíveis por atributo na base contatos. 31
Tabela 5 – Tipos possíveis por atributo na base notas. 33
Tabela 6 – Especialidades pretendidas por residentes do Pará. 61

LISTA DE ABREVIATURAS E SIGLAS

ACTCOMP	Ambiente Colaborativo para Treinamento POSCOMP
API	*Application Programming Interface*
COPS	Coordenadoria de Processos Seletivos
COVID-19	Coronavírus/SARS-CoV-2
EDM	*Educational Data Mining*
ENEM	Exame Nacional do Ensino Médio
FUNDATEC	Fundação Universidade Empresa de Tecnologia e Ciências
KDD	*Knowledge-Discovery in Databases*
POSCOMP	Exame Nacional para Ingresso na Pós-Graduação em Computação
PPGCC	Programa de Pós-Graduação em Ciência da Computação
SBC	Sociedade Brasileira de Computação
SDK	*Software Development Kit*
SUS	*System Usability Scale*
SVG	*Scalable Vector Graphics*
UFMS	Universidade Federal de Mato Grosso do Sul
UFPA	Universidade Federal do Pará

SUMÁRIO

1 INTRODUÇÃO 13
1.1 Contextualização 14
1.2 Motivação 17
1.3 Justificativa 17
1.4 Objetivos 18
1.4.1 Objetivo Geral 18
1.4.2 Objetivos Específicos 18
1.5 Trabalhos Relacionados 18
1.6 Estrutura do Trabalho 20
2 FUNDAMENTAÇÃO TEÓRICA 22
2.1 KDD 22
2.2 Educational Data Mining - EDM 23
3 PROCESSO METODOLÓGICO 25
3.1 Metodologia 25
3.2 Base de dados 27
3.2.1 Repostas 28
3.2.2 Gabarito 29
3.2.3 Contatos 30
3.2.4 Notas 32
3.3 Pré-processamento 34
4 RESULTADOS E DISCUSSÃO 36
4.1 Desempenho por notas totais 36
4.2 Desempenho por tema 39
4.3 Dados sobre especialidades pretendidas 45
4.4 Dados sobre participantes por sexo 48
4.5 Dados sobre ausentes 50
5 CONSIDERAÇÕES FINAIS 52
REFERÊNCIAS 54

APÊNDICES **56**
APÊNDICE A – TRABALHOS PUBLICADOS PELO AUTOR **57**
APÊNDICE B – GRÁFICOS **58**
APÊNDICE C – TABELAS **61**

1 INTRODUÇÃO

No âmbito da computação, o Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP) figura como o exame responsável pela avaliação do conhecimento dos egressos da área de computação no Brasil, realizado desde 2002 pela Sociedade Brasileira de Computação (SBC), onde foi adotado a partir de 2006 no Peru, e posteriormente na Colômbia, onde ocorre desde 2012 (SBC, 2021b).

A nota do POSCOMP é adotada por diversas universidades do país como um requisito parcial para a entrada em cursos de pós-graduação em computação e áreas relacionadas, o que torna o exame relevante para os egressos da área que desejam adentrar na pós-graduação.

O Exame Nacional do Ensino Médio (ENEM) é aplicado desde 1998 e é responsável por avaliar o desempenho dos concluintes ou daqueles que continuam cursando o ensino médio no Brasil, sendo utilizado como único critério de entrada em diversas universidades do país, substituindo os vestibulares individuais, além de ser usado como requisito para a obtenção de bolsas e financiamento estudantil.

A realização de exames de cunho nacional para a avaliação de estudantes como o ENEM e o POSCOMP, em conjunto com a crescente utilização de tecnologias na educação, possibilitam a geração de grandes bases de dados, onde se torna viável a exploração para obtenção de dados não triviais como estatísticas sobre o desempenho dos alunos.

A exploração de tais bases de dados ainda é um desafio para as instituições de ensino, onde a abundância de dados existentes configura obstáculo para as instituições na obtenção de informações úteis (MARTINS; MIGUÉIS; FONSECA, 2018). A necessidade da adoção de técnicas computacionais para a exploração das bases de dados estudantis é reforçada no trecho: "A grande quantidade de dados presentes nas bases estudantis ultrapassa a habilidade humana de analisar e extrair as informações mais úteis sem a ajuda de técnicas de análise automatizada." (ALGARNI, 2016, p. 456).

A mineração de dados é um passo do processo *Knowledge Discovery on Databases* (KDD), onde se realiza a análise dos dados e a aplicação de algoritmos com o objetivo de encontrar padrões (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). A mineração de dados pode ser aplicada em dados provenientes de diversas áreas, incluindo a área educacional, sendo um processo que pode ser usado por parte das instituições para a obtenção de informações úteis.

A mineração de dados estudantis pode auxiliar alunos no processo de aprendizagem, e os professores podem prover ensino personalizado a partir de informações geradas pela mineração (HUI et al., 2020). O monitoramento constante acerca de índices acadêmicos é relevante para que alunos, professores, gestores acadêmicos e governo tomem um conjunto de ações com base nos resultados, como, por exemplo, a destinação de verbas para melhorias de infraestruturas, reformulação e revisão de projetos pedagógicos dos cursos, alteração de metodologias de ensino

e o entendimento real das necessidades dos candidatos (ROMERO; VENTURA, 2013).

No que diz respeito a edição de 2022 do POSCOMP, a FUNDATEC (2022) indica que houve 536 inscritos, com a presença de 401 candidatos, contando com uma abstenção de 25,19%. Nota-se que o número de inscritos foi bastante inferior aos apresentados nas demais edições analisadas, que serão discutidos com maior detalhe na subseção 4.1, configurando uma baixa adesão após a pandemia.

Em um levantamento bibliográfico preliminar, percebeu-se que não há trabalhos com a finalidade de analisar o desempenho dos participantes do POSCOMP, haja vista que a SBC não disponibiliza os microdados do exame de forma pública, que são dados detalhados dos participantes como respostas e áreas de interesse, inviabilizando a análise da base de dados e a realização de tais estudos. Caso os dados estivessem disponíveis ao público, diversas pesquisas poderiam ter sido realizadas com o objetivo de entender melhor o perfil dos candidatos aos programas de pós-graduação em computação do Brasil, a exemplo do que ocorre com outros exames. Assim, os trabalhos relacionados ao POSCOMP que foram encontrados na literatura somente buscam auxiliar os egressos na preparação para a prova através da criação de ferramentas computacionais, ou ainda, analisar o tema das provas em cada edição.

No entanto, para este trabalho, foi concedido o acesso aos microdados de maneira institucional, os quais foram solicitados através de um ofício para os colaboradores da SBC. Desse modo, o objetivo deste trabalho é analisar os microdados do POSCOMP obtidos via ofício, com enfoque nos egressos da área de computação residentes no estado do Pará, possuindo como recorte temporal as edições de 2016 a 2019, no intuito de gerar informações acerca do desempenho e perfil dos candidatos.

1.1 Contextualização

O POSCOMP é realizado desde 2002 pela Sociedade Brasileira de Computação (SBC), sendo responsável pela avaliação do conhecimento dos egressos da área da computação no Brasil, sendo adotado a partir de 2006 no Peru, e posteriormente na Colômbia, onde ocorre desde 2012 (SBC, 2021b).

A SBC define que o POSCOMP não está vinculado a nenhum programa de pós-graduação, logo, cada universidade decide como irá utilizar a nota em seu processo seletivo. Assim, a nota do exame é comumente utilizada como um dos requisitos para a entrada em diversos cursos de pós-graduação em computação e áreas correlatas no país, além de possuir também um papel de avaliador do conhecimento dos egressos da área de computação, onde os participantes podem indicar que estão fazendo o exame somente para autoavaliação.

O exame é realizado anualmente pela SBC, com o cancelamento das edições de 2020 e 2021 tendo ocorrido devido à pandemia da COVID-19 (SBC, 2021a), e é aplicado em todo

território nacional, permitindo a inscrição dos candidatos em instituições de outras regiões, sem a necessidade do deslocamento até a cidade onde a universidade pretendida está localizada. Após a realização da prova, são liberadas as transparências como o gabarito e a prova através do site da SBC, e os candidatos são informados acerca de seu desempenho, com a indicação de seus acertos e erros, sua média e desvio padrão, através do site da organizadora da prova.

A prova é composta por 70 questões objetivas, as quais se dividem em 3 áreas do conhecimento, sendo elas: Fundamentos da computação, Tecnologia de Computação e Matemática, onde a primeira possui 30 questões, e as demais possuem 20 questões cada. A SBC define ao todo 25 temas que são distribuídos de forma variada entre as 3 áreas do conhecimento citadas, distribuição observável na Tabela 1.

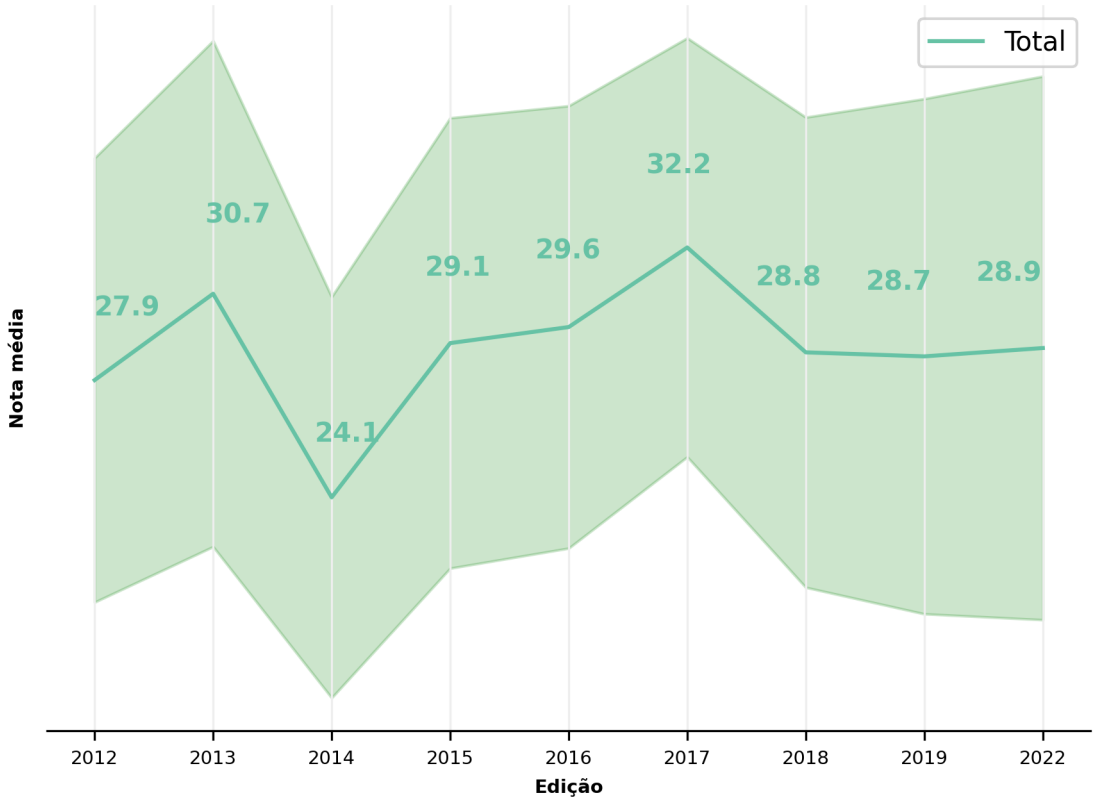
Tabela 1 – Áreas do POSCOMP e respectivos temas.

Matemática	**Fundamentos da computação**	**Tecnologia de computação**
Álgebra Linear	Análise de Algoritmos	Banco de Dados
Análise Combinatória	Algoritmos e Estruturas de Dados	Compiladores
Cálculo Diferencial Integral	Arquitetura e Organização de Computadores	Computação Gráfica
Geometria Analítica	Circuitos Digitais	Engenharia de Software
Lógica Matemática	Linguagens de Programação	Inteligência Artificial
Matemática Discreta	Linguagens Formais	Processamento de Imagens
Probabilidade e Estatística	Autômatos e Computabilidade	Rede de Computadores
	Organização de Arquivos e Dados	Sistemas Distribuídos
	Sistemas Operacionais	
	Técnicas de Programação	
	Teoria dos Grafos	

Fonte: De Sordi Jr., 2015.

No que diz respeito ao desempenho dos participantes na prova, a SBC disponibiliza uma Tabela que compreende as edições de 2012 a 2019, com a média de acertos total por ano e desvio padrão, incluindo também os acertos por área, a qual foi base para a geração da Figura 1. “A média de acertos nas provas é considerada baixa, refletindo o grau de dificuldade e exigência do exame.” (DE SORDI Jr., 2015, p.19), o que pode ser confirmado na Figura 1, visto que a média de acertos nunca superou 32,2 questões em relação ao total no período, expondo a dificuldade em relação aos temas do exame, o que pode estar ligado à grande quantidade de temas abordados

Figura 1 – Médias de acerto e Desvio Padrão do POSCOMP por edição.



Fonte: compilação do autor.

na prova, bem como à ausência de meios para a preparação para a mesma.

É importante destacar o pior desempenho no período, com a média de acerto chegando a 24,1 questões na edição de 2014, sendo um desempenho abaixo da média, mesmo considerando o baixo desempenho observado nas demais edições. As variações no desempenho entre edições podem se dar a partir da dificuldade das questões selecionadas naquele ano, ou ainda devido à preparação dos discentes no período.

Em conformidade, alguns temas não são abordados diretamente pelas estruturas curriculares dos cursos de graduação. Como exemplo é possível citar o tema compiladores, o qual não está presente nas estruturas curriculares dos cursos de ciência da computação e sistemas de informação da Universidade Federal do Pará (UFPA), no entanto, esse tema é cobrado no POSCOMP, o que demanda mais empenho, por exemplo, por parte dos candidatos do Pará que realizarão a prova para se candidatar aos programas de pós-graduação.

Segundo DE SORDI Jr. (2015) o exame foi executado pela Coordenadoria de Processos Seletivos (COPS) a partir de 2010, onde passou a informar o desvio padrão, além da média já informada nas edições anteriores. A partir de 2016 a responsável pela realização da prova passou a ser a Fundação Universidade Empresa de Tecnologia e Ciências (FUNDATEC).

1.2 Motivação

Tendo em vista a baixa quantidade de acertos do POSCOMP, assim como a ausência de estatísticas minuciosas em relação ao desempenho dos participantes, como seus desempenhos por tema, é pertinente a exploração de sua base de dados para a obtenção de informações relevantes, uma vez que diversos interessados podem ser beneficiados por tais informações.

As informações obtidas no trabalho podem ser utilizadas por coordenações e professores para a elaboração de estratégias que se adéquem às necessidades dos candidatos, observando em quais temas possuem maior dificuldade e entendendo seus desempenhos, o que pode acarretar mudanças na abordagem de tais temas e ações visando, por exemplo, a readequação e modernização de projetos pedagógicos de cursos de graduação. Os candidatos podem consultar as informações e atentar para determinados temas tendo em vista o histórico de seu estado, o que pode os auxiliar a reforçar os estudos referentes a tais temas, bem como podem obter acompanhamento para a melhora do desempenho nos temas em que são menos proficientes.

Os gestores acadêmicos da graduação terão suporte na tomada de decisão para a realização de ações visando o ingresso dos alunos na pós-graduação, bem como políticas para a redução da abstenção e aumento da participação feminina, tais ações podem envolver a destinação de recursos, alterações em diretrizes do ensino e realização de eventos para expor a importância da prova e da pós-graduação, incentivando os candidatos a realizarem o exame.

Com base nas informações obtidas, os gestores de programas de pós-graduação podem observar os interesses dos candidatos por determinada especialidade, o que permite analisar se a oferta é compatível com a demanda e planejar uma adequação aos interesses dos candidatos, além disso, outras melhorias nos programas de pós-graduação podem ser planejadas a partir das informações geradas. As melhorias propostas, quando implementadas, podem contribuir para o aperfeiçoamento do programa de pós-graduação, o que pode elevar sua nota em avaliações como a avaliação da Capes, a qual compreende notas de 1 a 7, sendo que uma nota menor que 3 configura a invalidação do curso a nível nacional, em contraponto, as notas 6 e 7 são possíveis somente para os programas que possuem doutorado e nível internacional de ensino.

1.3 Justificativa

O POSCOMP é o exame responsável pela avaliação do conhecimento dos egressos da área de computação no Brasil, onde sua nota é um requisito parcial comum para a entrada em programas de pós-graduação em computação em diversas universidades do país, o que o torna relevante para os egressos da área que desejam adentrar na pós-graduação. Apesar da grande quantidade de dados gerados e da importância do POSCOMP para a entrada na pós-graduação e como avaliador de desempenho, há poucos trabalhos relacionados ao exame, os quais se detém a criar ferramentas para auxiliar na preparação para a prova.

Tendo em vista a relevância do exame como avaliador e instrumento de entrada na pós-graduação, realizou-se uma pesquisa bibliográfica sobre o POSCOMP, onde se atestou que não há trabalhos com a finalidade de minerar os dados do exame, sendo as médias de acertos e os desvios padrões as únicas estatísticas disponíveis sobre o desempenho dos participantes, as quais são providas pela SBC em seu site.

Assim, este trabalho pretende analisar a base de dados do POSCOMP se utilizando de técnicas de mineração e visualização de dados, com a finalidade de gerar informações não triviais que poderão ser consultadas e utilizadas pela comunidade da computação e a sociedade em geral.

1.4 Objetivos

1.4.1 Objetivo Geral

Este trabalho tem por objetivo geral realizar uma análise exploratória dos microdados do POSCOMP a fim de detectar o perfil dos potenciais candidatos residentes do estado do Pará, no intuito de suportar as decisões e prover informações sobre interesse, desempenho e estatísticas descritivas a candidatos e gestores acadêmicos de pós-graduação.

1.4.2 Objetivos Específicos

Foram definidos os seguintes objetivos específicos a fim de guiar a elaboração do trabalho:

- Apresentar o número de ausentes por edição.
- Apresentar o número de candidatos por sexo.
- Relacionar e verificar o total de acertos dos residentes do estado do Pará aos acertos dos demais brasileiros.
- Relacionar e verificar os acertos por tema dos residentes do estado do Pará aos acertos dos demais brasileiros.
- Verificar quais áreas são pretendidas por parte dos candidatos.
- Verificar se o PPGCC da UFPA atende à demanda quanto às especialidades pretendidas pelos candidatos do estado.

1.5 Trabalhos Relacionados

Os trabalhos relacionados estão voltados, em sua maioria, ao desenvolvimento de ferramentas que visam auxiliar os candidatos na preparação para o exame, seja através de um sistema

web, aplicativo móvel ou ambos. Tal fato ocorre, pois a SBC disponibiliza somente as provas e gabaritos das edições realizadas, as quais os referidos trabalhos utilizam para elaboração de ferramentas de suporte educacional.

Em Batista et al. (2014) foi desenvolvido um aplicativo móvel para Android denominado POSCOMP, o qual reúne questões do exame, as quais são acrescidas com atualizações, e permite que o aluno escolha os temas para a realização de um simulado. Foi desenvolvido através da utilização de um framework local para aplicações móveis, o qual utilizou o Android SDK, que disponibiliza ferramentas para o desenvolvimento móvel. O aplicativo proposto visa auxiliar os estudantes para ingressar nos cursos de pós-graduação em computação. Foi testado em um grupo de estudos para o POSCOMP na Universidade Federal de Mato Grosso do Sul (UFMS).

Já no trabalho de Ribeiro e Junior (2020) foram desenvolvidos tanto um sistema web quanto um aplicativo móvel, onde estão disponíveis 1500 questões de exames anteriores do ENADE e POSCOMP, com objetivo de auxiliar na preparação para os exames. Ao gerar um simulado, o aluno pode personalizá-lo com questões dos dois exames, escolher o número de questões e definir o tempo. Além disso, são expostos gráficos e estatísticas de seu desempenho na página inicial. Para o desenvolvimento do web service, os autores utilizaram o Framework Spring Boot e a linguagem Java. Como API foram utilizados o Framework Hibernate e a Java Persistence API. No desenvolvimento do front-end foram utilizados Angular para o sistema web, e Kotlin para o mobile. Para avaliação dos softwares, os autores elaboraram um questionário com 10 perguntas para os professores e alunos. Com isso, o sistema obteve uma boa avaliação.

Na mesma linha do trabalho anterior, em Callegari e Oliveira (2020) também foi desenvolvido um aplicativo móvel com a finalidade de auxiliar os aspirantes ao POSCOMP através de simulados, porém o aplicativo tem como foco a área de matemática do exame, a qual possui baixa taxa de acertos. Assim, o diferencial do aplicativo está no foco em uma área de dificuldade dos alunos, bem como na apresentação de explicações detalhadas para cada questão, ao invés de somente apresentar a opção correta. Para o desenvolvimento do aplicativo, foi utilizado o React Native no front-end, com a utilização do firebase no back end, adicionalmente, foi utilizado o editor MathMagic, a fim de formatar as equações matemáticas com maior precisão.

Em DE SORDI Jr. (2015) foi desenvolvido um sistema web colaborativo denominado ACTCOMP, o qual permite que os usuários cadastrem novas questões, além das que já existem na base de dados da aplicação compreendem as questões do POSCOMP de 2012 a 2014, classificando-as em áreas e temas predefinidos. Também é possível que um usuário avalie questões propostas por outros usuários, caso se encontrem em temas com os quais têm proximidade, sendo possível reportar erros para o criador da questão, onde as questões só passam a ser utilizadas nos simulados a partir de 3 avaliações por diferentes usuários. Por fim, é possível realizar simulados padrões ou personalizados por área e subárea. O autor utilizou a linguagem PHP para o desenvolvimento do sistema, em conjunto com o Javascript, adicionalmente, foi utilizado o LaTeX para formatação de equações matemáticas. Como resultado, além da finalização da

aplicação, houve uma avaliação da mesma por 11 profissionais da área, a qual resultou em uma pontuação SUS de 83,40, onde se conclui que o sistema proporciona uma boa experiência para os usuários.

Em Mendes, Mendonça e Guedes (2018) foi desenvolvida uma plataforma web para auxiliar na preparação para o POSCOMP, onde as questões da base de dados de aplicação são provenientes das edições de 2002 a 2015, as questões foram transcritas manualmente para LATEX e posteriormente transformadas para SVG, a fim de manter ilustrações e caracteres da área de matemática. Além disso, a aplicação permite a realização de simulados padrões e personalizados por área, exibindo os desempenhos geral e por área. Para o desenvolvimento da plataforma foram utilizados o framework Django, MySQL, Apache, Bootstrap e Chartjs.

Em DE SORDI Jr. e Brancher (2014) foi aplicado um questionário a fim de analisar a relevância do POSCOMP para a comunidade, formada por professores, pesquisadores e profissionais da área de computação: onde os autores identificaram que a comunidade considera importantes temas como banco de dados e análise de algoritmos, bem como temas importantes para o mercado, no caso de engenharia de software. Outra conclusão apontada é de que as temas que envolvem matemática não são consideradas relevantes no geral, com exceção de lógica matemática, a qual é a segunda melhor avaliada.

Augusto et al. (2021) realiza uma comparação do currículo do curso de ciência da computação definido na SBC com o tema do POSCOMP, e o faz através da incidência de eixos de formação nas edições do exame, como também por meio da incidência de temas que formam os eixos no exame. Como resultado, os autores observaram que o eixo 1 (Resolução de Problemas) é o mais presente no exame, sendo aquele que teve mais temas contemplados no exame. Da mesma forma, os autores identificaram que os temas mais influentes são Arquitetura e Organização de Computadores e Sistemas Operacionais, onde a influência considera a incidência e consistência dos temas nas edições do exame.

1.6 Estrutura do Trabalho

O capítulo 2 trata da fundamentação teórica utilizada para a compreensão do processo de mineração de dados, em especial a mineração de dados estudantis, que permitiu a definição de etapas e guiou a utilização de métodos para a exploração das bases de dados do POSCOMP.

O capítulo 3 apresenta o processo metodológico definido para a realização de revisão bibliográfica e exploração das bases de dados, com a definição de ferramentas, ambientes, etapas que guiaram a exploração, e descrição das operações realizadas nas bases a fim de alcançar os objetivos pretendidos.

O capítulo 4 apresenta os resultados obtidos, a partir dos quais se faz uma discussão, onde são realizadas comparações entre os dados obtidos no estado do Pará e no Brasil, elencando

estatísticas e gráficos a partir dos objetivos definidos.

O capítulo 5 apresenta as considerações finais do trabalho e cita possíveis trabalhos futuros a serem desenvolvidos.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os principais conceitos e autores utilizados como embasamento teórico do trabalho, os quais fornecem uma base sólida de conhecimento que possibilita a compreensão dos temas pesquisados, bem como apresentam processos e métodos e consolidados que apoiam a mineração de dados e a exploração de bases de dados estudantis.

2.1 KDD

A utilização de um processo bem definido é necessária na exploração de bases de dados, a fim de gerar informações úteis a partir de dados de diversos campos como saúde, educação, e afins. O *Knowledge Discovery in Databases* (KDD) é uma área que desenvolve técnicas e métodos que podem transformar os dados em informações úteis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A definição mais famosa acerca do KDD é proveniente de um artigo de Fayyad, Shapiro e Smyth, a qual indica: “o Knowledge Discovery in Databases é um processo não trivial que identifica padrões válidos, originais, potencialmente úteis, e em última análise compreensíveis em dados” (FAYYAD et al., 1996, 83).

Maimon e Rokach (2010) definem KDD como a exploração e análise automática de grandes bases de dados, sendo um processo organizado que permite a identificação de padrões úteis e válidos em bases de dados complexas.

O KDD é um processo composto por 9 etapas, onde são realizadas ações como a definição de objetivos para a mineração, seleção e refinamento dos dados, a mineração e posterior interpretação e utilização dos dados. O processo é interativo, sendo possível retornar para a etapa anterior a qualquer momento, a fim de realizar a limpeza em relação a determinados atributos, transformar o formato dos mesmos, entre outros, onde se pode realizar diversas iterações a fim de alcançar um resultado satisfatório. A partir disso, as etapas do processo KDD são:

1. Entender o domínio do problema e os conhecimentos já existentes sobre ele, definindo um objetivo para o processo.
2. Selecionar uma base de dados ou amostra como alvo, que será utilizada na descoberta de conhecimento.
3. Realizar o pré-processamento dos dados, onde é realizada a limpeza da base, com o tratamento de dados faltosos e afins.
4. Reduzir a base através da redução de dimensionalidade e seleção dos atributos relevantes.

5. Escolher um método de mineração com base nos objetivos definidos, o qual pode ser um método de clusterização, regressão ou outro.

6. Selecionar o algoritmo e o método que serão utilizados na mineração, tendo em vista a natureza dos dados.

7. Minerar os dados e buscar padrões com base no algoritmo e método definidos.

8. Analisar os padrões obtidos, onde é possível utilizar a visualização de dados a fim de auxiliar a análise.

9. Utilizar o conhecimento obtido de forma concreta, o que pode ocorrer através da disseminação da informação para os interessados.

Data Mining ou mineração de dados é uma etapa do processo KDD onde os dados são analisados e é possível produzir padrões a partir dos mesmos, em condições adequadas de eficiência computacional (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Contudo, a mineração de dados também é tratada como um sinônimo para o processo KDD, e pode ser aplicada a diversas áreas e objetivos. A partir dessa concepção de mineração de dados, áreas mais específicas surgiram, como a *Educational Data Mining* (EDM), que define padrões e métodos específicos para a mineração de bases de dados estudantis, e será discutida na próxima seção.

2.2 Educational Data Mining - EDM

Em virtude da necessidade de ferramentas e informações que suportem o monitoramento estudantil, a área de *Educational Data Mining* (EDM) foi identificada como pertinente e escolhida como guia para a exploração, definindo aplicações, interessados e métodos a serem utilizados.

Baker et al. (2010) define EDM como uma área de desenvolvimento de métodos que visam explorar dados provenientes de bases estudantis, os quais são utilizados para gerar melhor entendimento sobre os estudantes e suas condições.

Romero e Ventura (2013) definem a EDM como uma área que procura desenvolver, pesquisar e aplicar métodos para análise de padrões em grandes conjuntos de dados, sendo a aplicação de Data Mining em bases educacionais.

Assim, a EDM está preocupada com a mineração de bases de dados estudantis, a fim de gerar conhecimento e identificar comportamentos. A geração de conhecimento afeta diversos interessados, os quais podem ser, segundo Romero e Ventura (2013): alunos, através da geração de recomendações de matérias, do acompanhamento para melhora do desempenho, e do entendimento das necessidades; educadores, pelo entendimento dos desempenhos dos alunos e suas causas, o que pode provocar alterações em métodos e comportamentos; pesquisadores, através do desenvolvimento de métodos e estudos das técnicas de mineração de dados, que podem

indicar a eficiência de uma técnica em cada caso; gestores, pelo entendimento das necessidades dos alunos, bem como sobre técnicas eficazes, auxiliando na oferta de disciplinas, gestão de recursos e tomada de decisão.

Segundo Anoopkumar e Rahman (2016), alguns obstáculos podem impedir ou dificultar a utilização da EDM, dentre eles, cita-se: tipos diferentes de dados, os quais possuem relações entre si, bem como a necessidade de adequação de técnicas que normalmente não são aplicadas a dados categóricos como os estudantis.

Peña-Ayala (2014) define diversas aplicações de EDM, dentre as quais, pode-se destacar a modelagem de estudantes, a qual tem como finalidade a criação ou otimização de um modelo estudantil a partir de características agregadas sobre os mesmos, como também os fatores que podem afetar negativamente a vida estudantil, os quais podem ser a causa de erros; predição de desempenho dos estudantes, que utiliza os dados para prever as notas finais do aluno, como também a possibilidade de aprovação, retenção, e a capacidade de aprendizado; geração de recomendações, a qual realiza sugestões de tarefas, matérias e demais atividades com base no perfil do aluno e progresso atual; comunicar aos interessados, que está voltada para os diretores, coordenadores e professores, os quais podem se utilizar da informação acerca dos perfis dos estudantes e seu progresso para tomar decisões no processo de ensino, seja na realização das atividades de uma matéria, ou ainda na mudança do plano do curso. Dentre as aplicações citadas, este trabalho se enquadra na comunicação aos interessados, visto que gera informações que podem ser usadas para suportar decisões dos gestores acadêmicos em diversos níveis.

Tendo em vista as diferentes aplicações de EDM, diversos métodos de mineração de dados podem ser aplicados, onde alguns são mais adequados a determinadas aplicações. Baker et al. (2010) define que os principais grupos de métodos incluem a predição, clusterização, entre outros. Este trabalho se utiliza do método de destilação de dados, onde se busca organizar os dados de forma que humanos consigam efetuar julgamentos, como identificar padrões e características através da estatística descritiva e visualização. Entende-se que a utilização de representações gráficas e dashboards auxilia no processo de tomada de decisão.

3 PROCESSO METODOLÓGICO

Neste capítulo, serão apresentadas a metodologia adotada no desenvolvimento deste trabalho, a qual compreende as tecnologias utilizadas para pesquisa e desenvolvimento, a descrição dos atributos da base de dados analisada, as etapas do processo KDD aplicadas para a geração de conhecimento para os interessados, bem como uma descrição granular acerca do pré-processamento realizado na base de dados, a fim de torna-lá apropriada para a exploração.

3.1 Metodologia

A metodologia utilizada foi composta de revisão bibliográfica, a fim de obter entendimento acerca dos trabalhos existentes que abordam o POSCOMP, das técnicas padrão e objetivos definidos pela EDM, e também pela aplicação do processo de descoberta de conhecimento de Banco de Dados (KDD).

As bases de dados utilizadas no presente estudo foram cedidas, em formato tabular, pela Sociedade Brasileira de Computação (SBC) por meio de um ofício, compreendendo as edições de 2016 a 2019 do POSCOMP. Os detalhes sobre os dados obtidos estão descritos na seção 3.2.

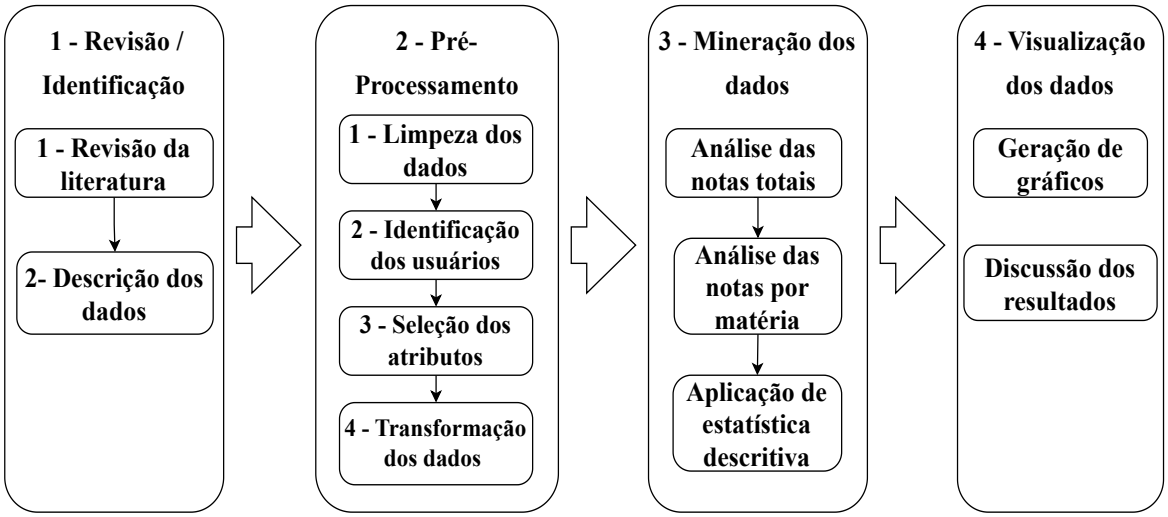
O presente estudo é de natureza quantitativa, tendo em vista que visa a mineração de bases de dados para a geração de estatísticas e gráficos. A pesquisa bibliográfica foi conduzida nos motores de busca Google Scholar, IEEE Xplore, ACM Library e no repositório da SBC, tendo em vista trabalhos com o intuito de visualizar e minerar séries de dados estudantis, especialmente aqueles ligados ao POSCOMP, à graduação e pós-graduação.

A mineração dos dados foi realizada utilizando a linguagem Python 3, e o ambiente de desenvolvimento Jupyter Notebook, com o auxílio de bibliotecas auxiliares como Pandas e Numpy para a manipulação e transformação dos dados, e as bibliotecas Matplotlib e Seaborn para a visualização dos dados, entre outras. A análise foi realizada a partir de estatística descritiva e visualização dos dados obtidos, as quais deram suporte para a discussão e geraram informações.

O processo KDD possui etapas definidas por Fayyad et al. (1996), que foram adaptadas por Romero e Ventura (2013) visando a exploração de bases estudantis. Tais etapas foram adotadas neste trabalho, tendo em vista a exploração das bases de dados do POSCOMP e a geração de gráficos e informação para os interessados.

Além das etapas definidas pelos autores, foi realizada a revisão e identificação da literatura pertinente, bem como dos dados, a fim de entender o estado da arte nas áreas de interesse e os atributos existentes na base. A partir disso, foram definidas 4 etapas, as quais compreendem a metodologia deste trabalho, e estão descritas na Figura 2:

Figura 2 – Etapas da exploração.



Fonte: compilação do autor.

1. Revisão / Identificação:

    1.1 Revisão da Literatura: a revisão da literatura foi realizada tendo em vista estudos que explorassem a base de dados do POSCOMP, através da geração de estatística descritiva e visualização de dados, bem como estudos relacionados ao KDD e à EDM, a fim de compreender os processos e métodos utilizados.

    1.2 Descrição dos dados: um dicionário foi criado para descrever os dados provenientes da base do POSCOMP, onde houve a divisão por base de dados e descrição de certas características dos dados. A partir desta etapa, foram definidos os seguintes objetivos para a exploração:

    - Relacionar e verificar o total de acertos dos residentes do estado do Pará aos acertos dos demais brasileiros.
    - Relacionar e verificar os acertos por área dos residentes do estado do Pará aos acertos dos demais brasileiros
    - Apresentar o número de estudantes por sexo no estado do Pará e no Brasil.
    - Apresentar o número de ausentes no estado do Pará e no Brasil.
    - Verificar quais áreas são pretendidas por parte dos egressos no Pará.

2. Pré-processamento: as etapas de pré-processamento dos dados foram baseadas no processo apresentado em Peña-Ayala (2014), como expõe a Figura 3:

Figura 3 – Etapas de pré-processamento.

Fonte: Peña-Ayala, 2014.

2.1 Limpeza dos dados: a preparação das bases de dados foi realizada tendo em vista dados faltosos, uma vez que os alunos podem ter sido ausentes, não respondido a questões, bem como podem não ter sido inseridas informações de cadastro. Também foi analisado se existiam inconsistências na base, como dados conflitantes.

2.2 Identificação dos usuários: houve a devida precaução quanto à exposição de dados sensíveis dos participantes contidos nas bases.

2.3 Seleção dos atributos: as bases possuem muitos atributos, onde foram selecionados somente aqueles pertinentes à pesquisa.

2.4 Transformação dos dados: pode ser preciso que o formato da base seja alterado, ou ainda que algumas colunas devam ser transformadas, a fim de facilitar a análise e a tornar mais homogênea.

3. Mineração de dados: verificação de estatísticas descritivas das bases, como média, mediana e desvio padrão para itens como a quantidade de alunos, quantidade de acertos por tema, notas totais, ausência no exame, questões deixadas em branco e áreas pretendidas.

   Nesta etapa se utilizou das transformações e pré-processamento dos dados para gerar estatísticas em consonância com os objetivos definidos na etapa de revisão/identificação, moldando os dados para que fossem utilizados também na etapa de visualização e interpretação dos resultados, onde foram gerados os gráficos e informações sobre os egressos que realizaram o exame.

4. Visualização e Interpretação dos resultados: geração de gráficos comparativos e informação a partir das estatísticas originadas na etapa 3 e discussão dos resultados, baseada na etapa de avaliação definida no KDD.

3.2 Base de dados

Esta seção busca descrever as bases de dados obtidas, para melhor compreensão acerca dos temas das mesmas e sua utilização na exploração. Os dados obtidos do POSCOMP compreendem as edições de 2016 a 2019, onde para cada edição há 4 bases de dados:

1. Respostas: compreende as respostas dos alunos aos itens.

2. Gabarito: contém o gabarito do POSCOMP da referida edição.
3. Contatos: contém os dados dos candidatos preenchidos na inscrição.
4. Notas: reúne as notas obtidas pelos participantes, apresentando notas por áreas do conhecimento.

As características de cada base serão descritas nas subseções a seguir, onde os atributos serão comentados individualmente.

3.2.1 Repostas

A base de dados de respostas possui 75 colunas, onde as 5 primeiras se referem a dados dos participantes, como:

1. NUM: o ID do participante.
2. NOME DO PARTICIPANTE: nome completo do participante.
3. SALA: a sala onde foi realizada a prova.
4. NÚMERO DE INSCRIÇÃO: número de inscrição do participante.
5. SITUAÇÃO: se o participante esteve presente ou ausente.

As 70 colunas restantes são numeradas e se referem às respostas dos participantes na prova, onde o campo pode conter valores de “A” a “E”, o valor “**-**”, que aparece quando o participante foi ausente, e também o valor “*”, que representa que uma questão foi anulada. A Tabela 2 apresenta os tipos de cada atributo, onde as colunas de 1 a 70 que possuem atributo nulo são referentes a questões não respondidas por alunos presentes.

Tabela 2 – Tipos possíveis por atributo na base respostas.

RESPOSTAS	
Atributo	**Tipos possíveis**
NUM	Inteiro
NOME CANDIDATO	String
SALA	Inteiro
NÚMERO DE INSCRIÇÃO	Inteiro
SITUAÇÃO	String
1-70	String, Null

Fonte: compilação do autor.

3.2.2 Gabarito

A base de dados de gabarito possui 12 colunas, a Tabela 3 indica os tipos possíveis para cada atributo da base, os quais estão descritos a seguir:

1. COD CARGO: indica o cargo, o valor é 1 para Mestrado/Doutorado e 2 para Autoavaliação.
2. CARGO: indica o cargo, o qual pode ser Mestrado/Doutorado ou Autoavaliação.
3. QUESTAO: indica o número da questão, seus valores vão de 1 a 70.
4. MATERIA: indica o tema o qual a questão pertence.
5. RESPOSTA: indica a resposta para a questão, seus valores vão de A a E, ou ainda asterisco, caso a questão tenha sido anulada.
6. ASSINATURA ELETRONICA: indica a presença de assinatura eletrônica, a coluna está em branco.
7. TIPO GABARITO: indica o tipo do gabarito, somente o valor D foi encontrado.

Tabela 3 – Tipos possíveis por atributo na base gabarito.

GABARITO	
Atributo	**Tipos possíveis**
COD CARGO	Inteiro
CARGO	String
QUESTAO	Inteiro
MATERIA	String
RESPOSTA	String
ASSINATURA ELETRONICA	Null
TIPO GABARITO	String
CONCURSO	Inteiro
COD MATERIA	Inteiro
PESO QUESTAO	Inteiro
STATUS QUESTAO	Inteiro
TIPO PROVA APLICADA	Inteiro

Fonte: compilação do autor.

8. CONCURSO: indica o número do concurso, associado a edição.
9. COD MATERIA: Indica o código do tema o qual pertence à questão, os valores podem ser de 1 a 25.

10. PESO QUESTAO: indica o peso da questão, se houver.
11. STATUS QUESTAO: indica se a questão é válida (1) ou foi anulada (2).
12. TIPO PROVA APLICADA: indica o tipo da prova aplicada, na edição de 2017 houve dois tipos de prova, onde essa coluna poderia assumir os valores 1 e 2.

3.2.3 Contatos

A base de dados de contatos possui 31 colunas, contendo dados dos participantes preenchidos na inscrição, os tipos de seus atributos podem ser observados na Tabela 4 e seus atributos são descritos a seguir:

1. NOME: nome completo do participante.
2. CPF: cpf do participante, apresenta pontos e traço.
3. DATA NASC.: data de nascimento do participante, está no formato dd/mm/yyyy.
4. SEXO: o sexo do participante, assume os valores M ou F.
5. TIPO DE DOCUMENTO: o tipo de documento cadastrado pelo participante, assume valores de 99 e RG.
6. DOCUMENTO: o Id do documento cadastrado.
7. NOME DO PAI: nome do pai do participante, se houver.
8. NOME DA MÃE: nome da mãe do participante.
9. ENDERECO: o endereço do participante.
10. NUMERO: o número da casa do participante.
11. COMPL. : o complemente do endereço do participante.
12. BAIRRO: bairro no qual o participante reside.
13. CIDADE: cidade na qual o participante reside.
14. ESTADO: estado no qual o participante reside, representado pela sigla.
15. CEP: cep referente ao endereço do participante, apresenta traço.
16. EMAIL: email do participante.
17. DDD: referente ao número do participante, que indica estado ou região do número.
18. FONE: número de telefone fixo do participante, se houver.

19. DDD CEL: referente ao número de celular do participante, que indica estado ou região do número.

20. CELULAR: número de celular do participante, se houver.

21. SENHA: a senha utilizada pelo participante.

Tabela 4 – Tipos possíveis por atributo na base contatos.

CONTATOS	
Atributo	**Tipos possíveis**
NOME, CPF, SEXO, NOME DA MÃE, EMAIL	String
DATA NASC.	Data
TIPO DE DOCUMENTO, DOCUMENTO	Inteiro, String
NOME DO PAI, DESCRICAO DA NECESSIDADE	String, Null
ENDERECO, CEP, INSCRICAO, CÓDIGO, DS CARGO	Inteiro, String
NUMERO, SENHA	Inteiro, String, Data
COMPL.	String, Inteiro, Null
BAIRRO, CIDADE, ESTADO	String
DDD, FONE	Inteiro, Float, Null
DDD CEL, CELULAR, ESTRANGEIRO	Null
DATA INSCRIÇÃO, CARGO	Data, String
DEFICIENTE	Float, Null
LOCAL DE PROVA	String, Float
NECESSIDADES ESPECIAIS	String

Fonte: compilação do autor.

22. INSCRICAO: código de inscrição do participante, pode conter traço.

23. DATA INSCRIÇÃO: contém a data e hora da inscrição do participante, onde a data está no formato dd/mm/yyyy e a hora hh:mm.

24. CARGO: indica o número e o cargo pretendido.

25. CÓDIGO: código referente ao cargo, o valor é 1 para Mestrado/Doutorado e 2 para Autoavaliação.

26. DS CARGO: indica a descrição do cargo.

27. DEFICIENTE: se o participante é portador de deficiência.

28. LOCAL DE PROVA: indica a cidade e estado de realização da prova, no formato Cidade – Sigla Estado.

29. NECESSIDADES ESPECIAIS: se o participante possui necessidades especiais, assume os valores de N ou S.
30. DESCRICAO DA NECESSIDADE: descrição do que o participante precisa, caso possua necessidade especial.
31. ESTRANGEIRO: se o participante é estrangeiro.

3.2.4 Notas

Na edição de 2016, esta base de dados possuía 75 colunas, onde haviam 2 colunas destinadas à nota da prova discursiva, e a observações sobre a mesma, sendo as duas nulas. A partir da edição de 2017, a base passou a ter 73 colunas, excluindo aquelas referentes à prova discursiva. Os tipos de cada atributo podem ser observados na Tabela 5 e a descrição das colunas se segue:

1. CONCURSO: se refere ao código da edição do POSCOMP.
2. NOME: nome completo do participante.
3. DATA NASC: data de nascimento do participante, está no formato dd/mm/yyyy.
4. CPF: cpf do participante, apresenta pontos e traço.
5. DOCUMENTO: o ID do documento cadastrado.
6. NEC ESP: se o participante possui necessidades especiais, pode assumir os valores N e S.
7. HIPO: se o participante apresenta carência financeira, com necessidade de isenção das taxas, está em branco em todos os documentos.
8. DEF: se o participante possui alguma deficiência, está em branco em todos os documentos.
9. AFRO: se o participante é afrodescendente, está em branco em todos os documentos.
10. INDIO: se o participante é indígena, pode assumir os valores N e S.
11. COD CARGO: código referente ao cargo, o valor é 1 para Mestrado/Doutorado e 2 para Autoavaliação.
12. CARGO: indica o cargo pretendido, assume os valores Mestrado/Doutorado ou Autoavaliação.
13. CÓD. ESPECIALIDADE: código referente à especialidade pretendida pelo participante, por exemplo, inteligência artificial.
14. ESPECIALIDADE: nome da especialidade pretendida.

15. INSCRICAO: número de inscrição do participante.
16. IDENTIFICADOR: número de identificação do participante.
17. PRESENTE DIA1 : se o participante esteve presente no primeiro dia, pode assumir os valores S e N.
18. PRESENTE DIA2 : se o participante esteve presente no segundo dia, pode assumir os valores S e N, está em branco.
19. PRESENTE DIA3 : se o participante esteve presente no terceiro dia, pode assumir os valores S e N, está em branco.
20. PARTICIPAÇÃO EM JÚRI: se o participante participou de júri, está em branco.

Da coluna 21 a 70, são discriminados em pares os acertos e a nota do participante em cada tema, onde a coluna com os acertos é exibida primeiro, seguida da coluna com a nota em determinado tema. Por fim, as últimas três colunas da base de notas são:

71. ACERTOS TOTAL: total de acertos do participante.
72. NOTA TOTAL TO: nota total do participante.
73. ESTRANGEIRO: se o participante é estrangeiro, está em branco.

Tabela 5 – Tipos possíveis por atributo na base notas.

NOTAS	
Atributo	**Tipos possíveis**
CPF	String
NEC ESP	String
INDIO	String
COD CARGO	Inteiro
CARGO	String
CÓD. ESPECIALIDADE	Float, Null
ESPECIALIDADE	String, Float
PRESENTE DIA1	String
ACERTOS Tema	Float, Null
NOTA Tema	Float, Null
ACERTOS TO	Float, Null
NOTA TOTAL TO	Float, Null

Fonte: compilação do autor.

3.3 Pré-processamento

Esta seção descreve com detalhes as operações realizadas na exploração das bases de dados, a partir das etapas de pré-processamento e mineração de dados definidas na seção anterior.

A etapa de pré-processamento dos dados se iniciou na base de notas, com a identificação das colunas que possuíam todos os valores nulos, identificou-se que as colunas [HIPO, DEF, AFRO, PRESENTE DIA2, PRESENTE DIA3, PARTICIPAÇÃO EM JÚRI, NOTA TOTAL Prova Discursiva, OBSERVAÇÃO Prova Discursiva, ESTRANGEIRO] possuíam todos os valores nulos, sendo retiradas das bases referentes as notas. Ademais, a coluna INDIO só possuía valores "N" em todas as edições, o que indica que nenhum indígena realizou o exame no período, sendo assim, a coluna também foi retirada. A base de notas iniciou a etapa de pré-processamento com 73 atributos, restando 60 atributos ao fim da etapa.

A existência de ausentes na base de notas provocava a existência de muitos valores nulos, onde foi realizada uma seleção para separar os ausentes dos presentes, criando uma base de dados somente com os presentes e outra somente com os ausentes, utilizadas posteriormente para analisar o número de ausentes no exame.

Verificou-se que na base de notas há várias linhas com dados do mesmo participante, com alteração somente no atributo de especialidade, indicando que o participante informou diversas especialidades que deseja adentrar na pós-graduação. Para a realização da análise dos ausentes foi preciso retirar os dados duplicados tanto da base de ausentes quanto de presentes, a fim de obter a porcentagem de ausentes em relação ao total.

Mesmo após a retirada de ausentes da base de notas, ainda restaram dois atributos que possuíam algum valor nulo, os quais eram [CÓD. ESPECIALIDADE, ESPECIALIDADE], e que se referem a qual área de pós-graduação o participante deseja adentrar. A partir disso, verificou-se que os valores nulos eram referentes aos participantes que se inscreveram para autoavaliação, o que é lógico, visto que não possuem interesse em aplicar para um programa de pós-graduação. Os valores nulos foram tratados posteriormente, quando foi realizada a análise das especialidades pretendidas no Pará, sendo filtrados somente os participantes com o valor de cargo "Mestrado/Doutorado", o que gerou uma nova base sem valores nulos.

Na base de contatos, de 31 colunas foram mantidas somente as 3 colunas necessárias para as análises, sendo as colunas [INSCRICAO, SEXO, ESTADO], com a exclusão das demais, as quais se referem a dados sensíveis dos participantes.

Na base de notas, colunas com dados sensíveis dos participantes também foram retiradas, sendo mantidas somente aquelas pertinentes à análise, como as colunas com a especialidade pretendida e as notas. O estudo tem como foco os estudantes do Pará e também do Brasil, logo foram identificados e retirados da base de notas estudantes estrangeiros como peruanos e colombianos.

Após a seleção dos presentes na base de dados de notas, foram selecionadas somente as notas dos participantes residentes no Pará, o que pôde ser realizado através do atributo "ESTADO" presente na base de contatos, sendo selecionados somente os números de inscrição pertencentes ao estado, os quais foram utilizados para a realização da seleção na base de notas.

Na base de dados de notas, verificou-se que existiam pares de colunas que possuíam o mesmo valor, sendo referentes ao número de acertos por tema e a nota obtida no respectivo tema. Logo, de um total de 50 atributos, foram mantidas somente 25 colunas referentes a cada um dos temas.

As bases de dados de notas possuíam disparidades em relação à nomenclatura dos atributos que representam os temas em diferentes anos, as quais variam em acentuação, letras maiúsculas e minúsculas, presença de conectivos como "e", presença de "ç" e afins, como exemplo se tem o tema "Cálculo Diferencial e Integral", o qual apareceu em outro ano como "Cálculo Diferencial Integral", sem o "e". Dessa forma, foi preciso padronizar a nomenclatura dos temas para que fosse possível agrupar e analisar os dados, o que foi feito a partir da retirada de acentos e afins, com a padronização quanto à nomenclatura nas bases em todos os anos.

Ademais, a normalização das médias de acerto dos residentes do Pará em relação ao Brasil foi realizada através da transformação dos dicionários que contém as médias em séries, permitindo a divisão das médias do Pará por aquelas obtidas nacionalmente.

Semelhantemente aos nomes dos temas, as especialidades presentes nas bases de notas também apresentavam disparidades entre si, com diferença na acentuação, havendo também especialidades semelhantes escritas de forma diferente, ou ainda que englobam uma à outra, como exemplo se tem "inteligência computacional" e "inteligência artificial". Também foram encontradas linhas em que o candidato escreveu mais de uma especialidade, que eram separadas por espaços, vírgulas, traços, o conectivo "e" e afins.

Por fim, diversas outras transformações dos dados foram realizadas, com a alteração do formato dos dados para dicionários, listas e afins, bem como por funções visando a normalização e padronização dos dados, possibilitando a exploração dos dados.

4 RESULTADOS E DISCUSSÃO

Neste capítulo serão apresentados e discutidos os resultados obtidos após a mineração da base de dados do POSCOMP, com base na interpretação dos gráficos e demais informações geradas a partir da visualização dos dados do exame. Os resultados apresentados são referentes aos candidatos residentes no estado do Pará, onde são realizadas comparações com os desempenhos e características nacionais, relevantes para dimensionar os resultados do estado, a fim de observar como o mesmo se apresenta frente ao desempenho do país, exceto pela seção de especialidades pretendidas. Cada uma das subseções trata sobre uma medida de desempenho ou característica dos candidatos.

A seção 4.1 se refere ao desempenho por notas totais, onde são apresentadas informações acerca das notas totais brutas, isto é, o total de questões certas por participante, onde a principal informação encontrada é a média total de acertos.

A seção 4.2 trata das notas por tema, visto que o POSCOMP é dividido em 25 temas, que podem ser consultados na Tabela 1. A análise do desempenho por tema é justificada, pois permite identificar com maior precisão a proficiência e deficiência dos egressos da área. Nessa seção serão encontradas estatísticas como as porcentagens de acertos em cada tema dos residentes do Pará, do Brasil, bem como comparações entre ambos.

A seção 4.3 verifica as especialidades de mestrado e doutorado mais pretendidas pelos residentes do Pará no período analisado, onde a demanda proveniente do resultado é relacionada com a oferta de uma das universidades federais do estado.

A seção 4.4 apresenta informações sobre participantes do exame por sexo, tanto no que se refere a residentes do Pará, quanto aos demais Brasileiros, analisando também as tendências de crescimento e decrescimento dos participantes em relação às nacionais.

A seção 4.5 apresenta informações sobre participantes ausentes, verificando as tendências de aumento e diminuição no número de ausentes dos residentes do Pará em relação às nacionais.

4.1 Desempenho por notas totais

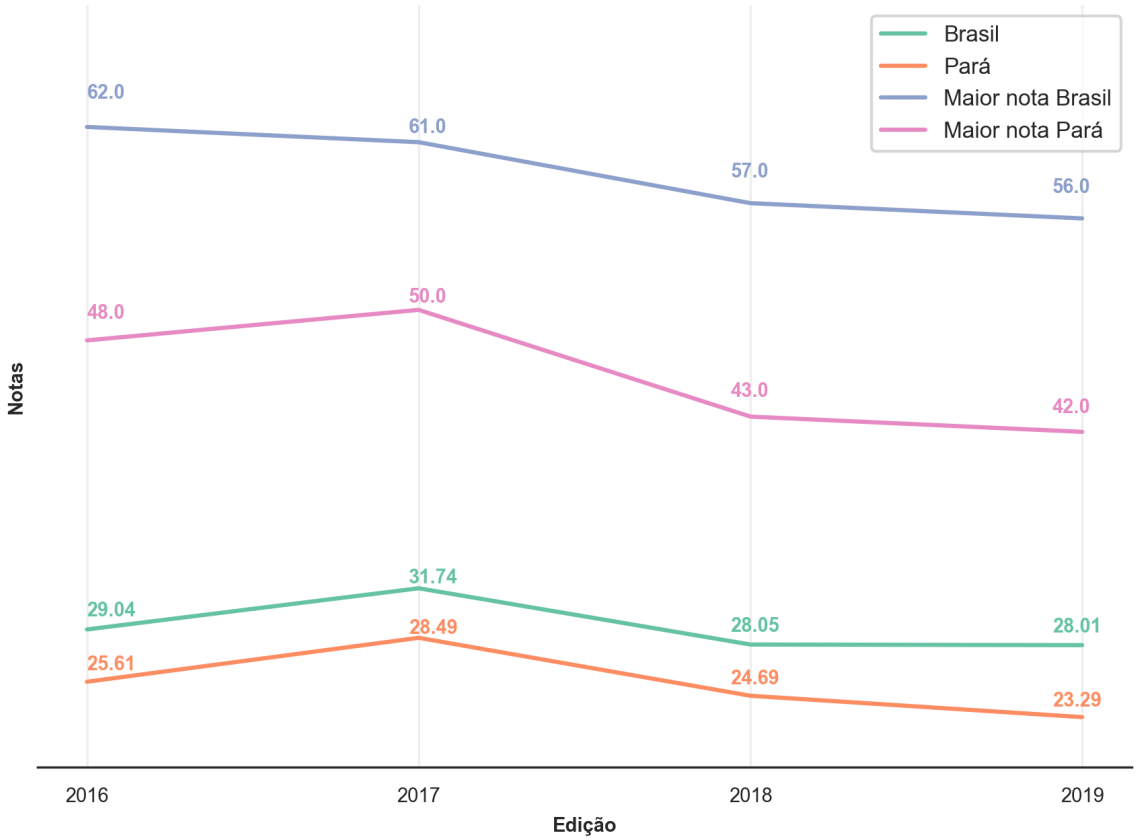
A nota total consiste no total de questões corretas obtidas pelo participante, a qual pode assumir valores de 0 a 70. O desempenho em relação às notas totais é medido através do cálculo da média de notas totais de todos os residentes do Pará e do Brasil.

Na Figura 4 é perceptível que as médias de notas totais dos residentes do Pará se mantêm abaixo das nacionais em todo o período analisado, onde mesmo as médias nacionais são consideradas baixas, uma vez que alcançam menos de 50% de questões corretas. Em adição, percebe-se que a média dos residentes do Pará acompanha as tendências nacionais, apresentando crescimento do ano de 2016, onde era 25,61, até o ano de 2017, onde houve aumento para 28,49.

Porém, a partir do ano de 2018 a média de notas totais de residentes do Pará apresentou queda, sendo de 24,69 em 2018, havendo redução para 23,29 em 2019, o menor desempenho no período, redução que pode estar ligada ao decréscimo no número de inscritos no exame nos períodos citados, comportamento que pode ser observado na Figura 5. É importante destacar, ainda, que a média de notas totais do Brasil presente na Figura 1 é diferente daquela presente na Figura 4, pois a primeira considera também o desempenho de estrangeiros como Peruanos.

Ademais, é visível que as maiores notas obtidas por candidatos residentes do Pará em cada edição também são menores que aquelas auferidas nacionalmente, e que as maiores notas alcançadas no estado seguem a tendência nacional, exceto pelo ano de 2017, onde houve crescimento na nota em relação ao ano anterior. Assim, os candidatos que obtiveram a maior nota em cada edição analisada foram provenientes dos estados de Santa Catarina em 2016, Rio Grande do Norte em 2017, Ceará e Mato Grosso em 2018, e Sergipe e Pernambuco em 2019.

Figura 4 – Média de notas por edição.



Fonte: compilação do autor.

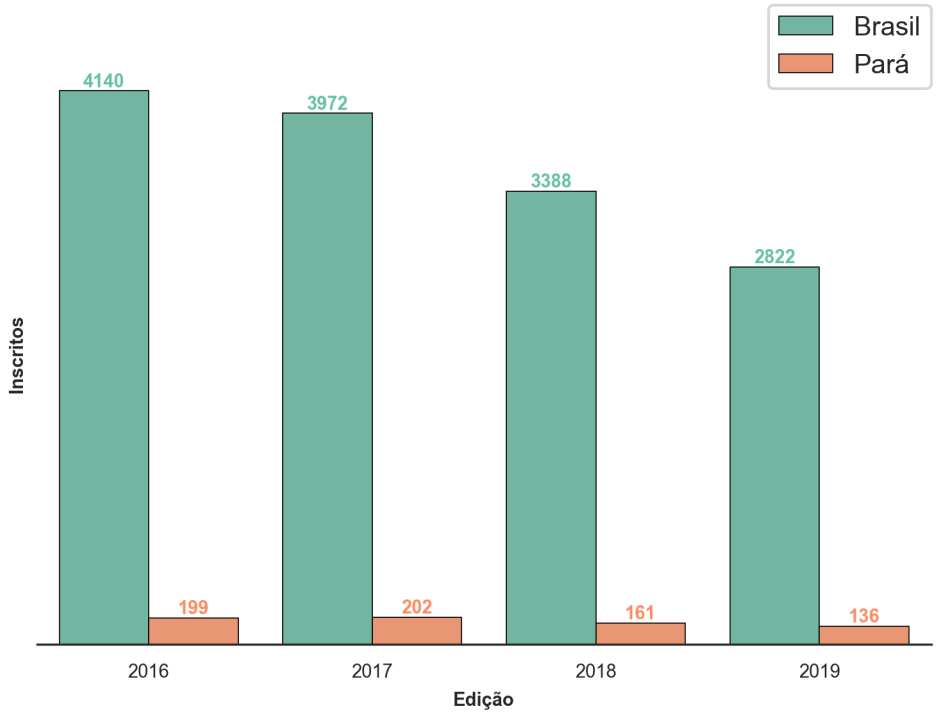
O trabalho de Cunha, Sales e Santos (2021) apresentou os desempenhos médios de acertos em cada tema da prova do Exame Nacional de Desempenho dos Estudantes (ENADE), considerando todos os anos em que a prova foi aplicada para o curso bacharelado em ciência da computação da UFPA e os comparou com os desempenhos médios, nos mesmos temas, a nível

nacional. Sabe-se a partir desse estudo que, por exemplo, os discentes da UFPA possuem 36%[1] de dificuldade em temas como: compiladores, linguagens formais autômatos e computabilidade, lógica matemática e cálculo diferencial integral, entre outros.

Na tentativa de relacionar as duas provas, ENADE e POSCOMP, considerando também especificamente os candidatos do Pará e que fizeram graduação na UFPA em Ciência da Computação, em um cenário onde estes candidados fizeram a prova do ENADE e POSCOMP no mesmo ano, e por fim, admitindo que os temas cobrados na prova do POSCOMP estão distribuídos de forma uniforme nas 70 questões, conclui-se que esses candidatos terão dificuldade em 36% da prova. Dito de outra forma, esperava-se que a média de notas dos candidatos residentes do Pará fosse de aproximadamente 45 pontos[2]. Pode-se perceber a partir da Figura 4 que a média dos residentes do Pará em 2019 foi de 23,29, indicando que foi obtida uma nota média abaixo da esperada, considerando as premissas acima.

Ademais, o decréscimo no número de inscritos por edição ilustrado pela Figura 5 pode estar ligado à escolha dos egressos da computação pela entrada no mercado de trabalho, os quais deixam de realizar a prova do POSCOMP visando uma vaga no mercado. Observa-se que após a pandemia o número de inscritos foi ainda menor, com apenas 536 inscritos na edição de 2022, onde a ausência do exame nos anos de 2020 e 2021 pode ter impactado na inscrição dos egressos.

Figura 5 – Inscritos brasileiros e residentes do Pará por edição.



Fonte: compilação do autor.

[1] Em 9 temas dos 25, a média local da UFPA está abaixo da média nacional.
[2] Média de pontos = (1 - 0,36) * 70

4.2 Desempenho por tema

O desempenho por tema é encontrado na base de dados com as notas individuais dos egressos, onde há duas colunas para cada um dos 25 temas que compõem o POSCOMP, que expõem os acertos e notas de cada aluno no referido tema.

Para o cálculo da média de acertos por tema, primeiramente foram calculadas as médias de acerto por tema em cada edição, onde se somou a nota de todos os candidatos da região analisada, realizando a divisão pela quantidade de alunos que realizaram o exame na determinada edição. Após a obtenção das médias de acertos das 4 edições analisadas, foi possível calcular a média de acerto em todo o período, somando as médias de acertos por tema de cada edição obtidas e dividindo pelo número de edições analisadas.

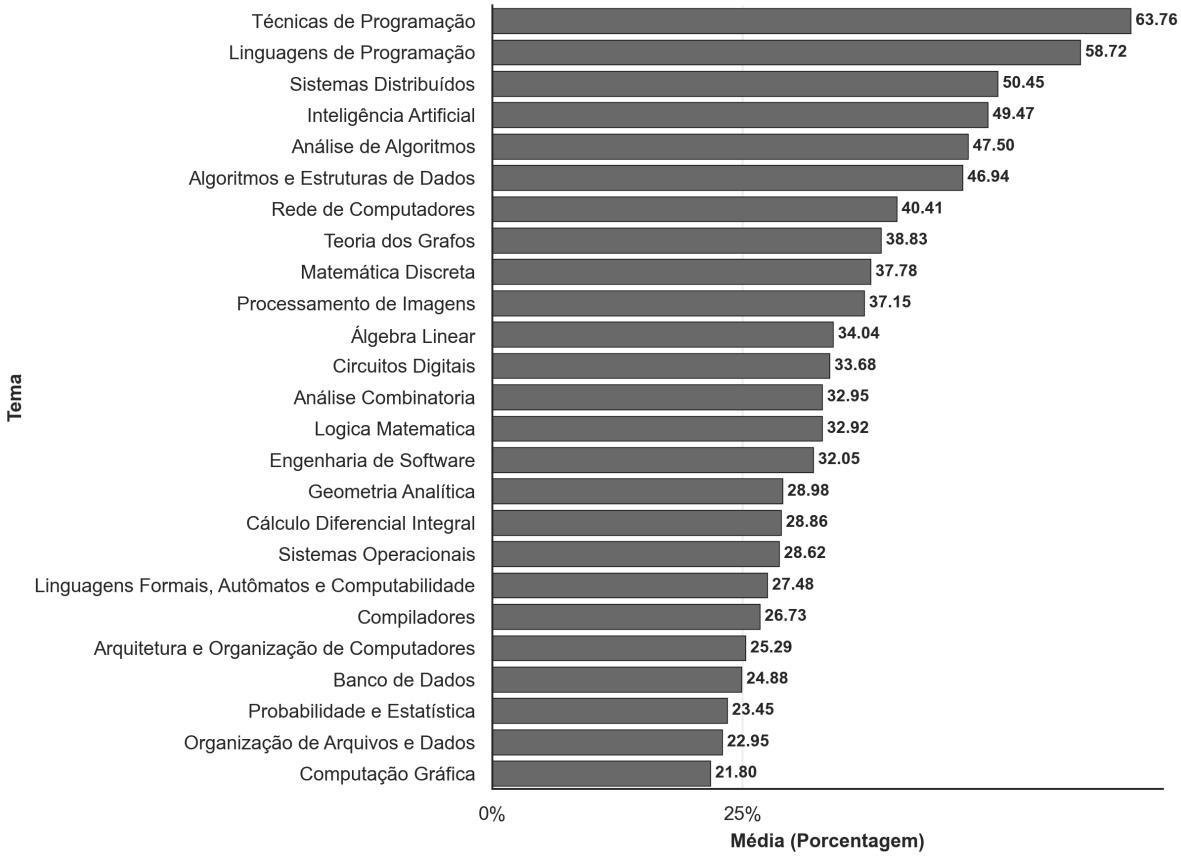
Contudo, observa-se que o número de questões não é uniforme para todos os temas, variando entre 2 a 3 questões por tema, onde aqueles que possuem apenas 2 questões em todas as edições analisadas são: Inteligência Artificial, Probabilidade e Estatística, Processamento de Imagens, Computação Gráfica e Compiladores. Assim, para o cálculo da média total de acertos em cada tema foram adotados dois passos:

- Na primeira, a média de acertos de cada tema em todo o período foi dividida pelo número de questões referentes ao mesmo, onde o resultado foi multiplicado por 100 para a geração da porcentagem em relação ao total.
- Na segunda, as médias totais de acertos em todo o período foram divididas pelas médias totais do Brasil, realizando uma normalização dos valores em comparação aos nacionais.

A Figura 6 foi gerada no primeiro passo, e representa a média total de acertos dos residentes do Pará por tema, a média de cada ano pode ser consultada na Figura 18 presente nos apêndices. Na figura é possível notar que os dois temas com maior média de acertos estão relacionados à programação, prática comumente realizada durante todo o período dos cursos de computação, sendo os temas técnicas de programação e linguagens de programação, com porcentagens de acerto de 63,76% e 58,72%, respectivamente. É importante salientar que um tema comumente visto ao fim do curso é sistemas distribuídos, pois requisita conhecimentos de disciplinas anteriores como redes, arquitetura de computadores e outras, o tema apresenta desempenho de 50,45%, ficando em terceiro lugar.

Os temas que apresentaram os menores desempenhos na Figura 6 são análogos aos que o trabalho de Cunha, Sales e Santos (2021) apresentou, onde os autores analisaram o desempenho dos alunos da UFPA em 5 edições do ENADE, e os temas “banco de dados”, “compiladores” e “probabilidade e estatística” ficaram abaixo do desempenho nacional, apresentando médias maiores que 0.8, porém menores que a média nacional, a qual foi representada por 1. A relação entre os dois trabalhos reforça a deficiência dos alunos residentes no estado em tais temas, visto que o ENADE só é realizado obrigatoriamente por alunos concluintes da graduação.

Figura 6 – Média de acertos de residentes do Pará por tema normalizada considerando todos os anos (2016-2019).



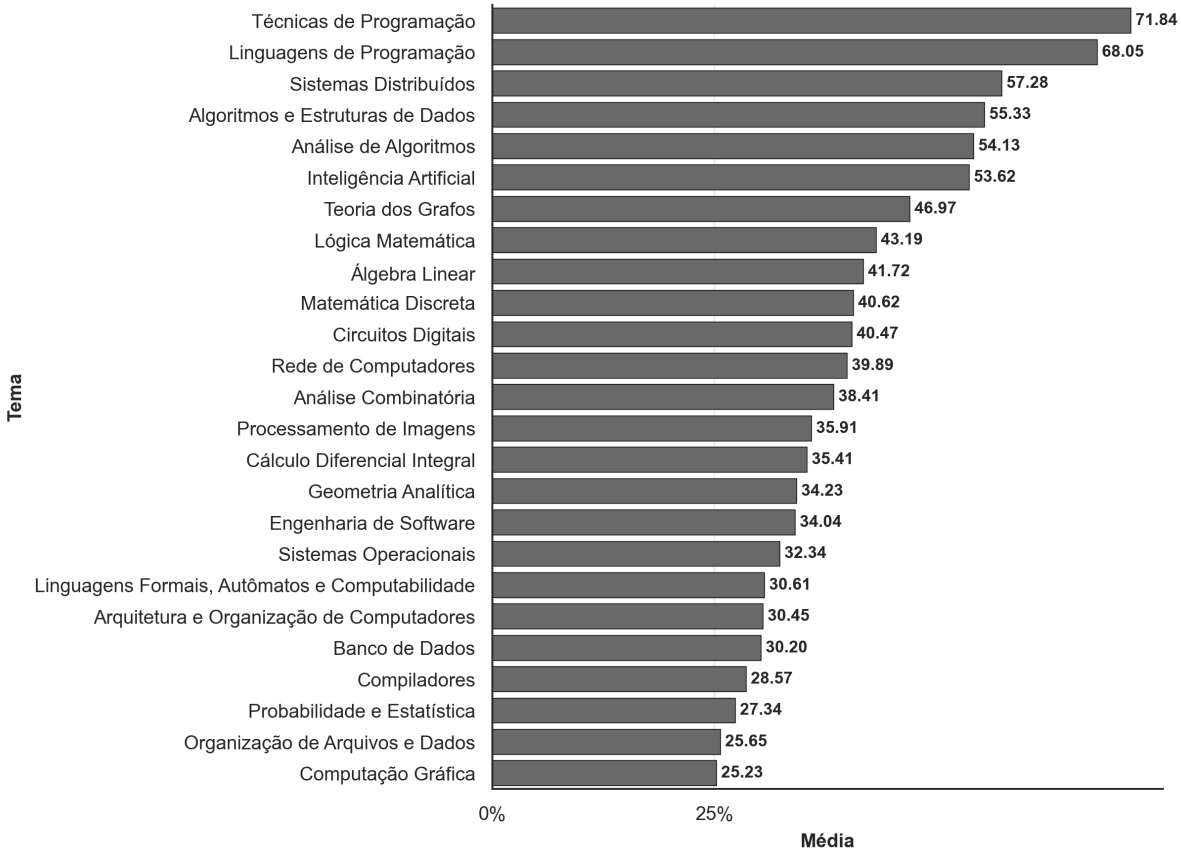
Fonte: compilação do autor.

Ainda na figura, em contraponto, temas iniciais como cálculo diferencial integral e arquitetura e organização de computadores aparecem com menos de 30% de média de acertos, o que pode indicar o esquecimento de tais temas por parte dos alunos ao longo do tempo. Dos temas que pertencem à área de matemática, aquele com maior número de acertos é matemática discreta, com 37,78% de acertos, outros temas da área aparecem com resultados próximos a 30%, com probabilidade e estatística sendo aquela com a menor média de acertos com 23,45%. Em relação aos temas com as piores porcentagens de acerto, percebe-se que os dois que apresentam menor desempenho são computação gráfica e organização de arquivos e dados, os quais podem ser temas que não recebem a atenção necessária dos egressos na preparação para o exame.

A Figura 7, proveniente do primeiro passo, representa a média total de acertos dos brasileiros por tema do exame, a média de cada ano pode ser consultada na Figura 19 presente nos apêndices. Com base na figura, pode-se verificar que os dois temas com maior porcentagem de acertos nacionalmente são os mesmos da figura anterior, reforçando a ideia de que a prática da programação leva ao acerto em tais temas, e o desempenho nacional em técnicas de programação supera em 8% aquele obtido por parte dos residentes do Pará, enquanto a porcentagem de acertos em linguagens de programação supera em quase 10%.

A tendência em relação a temas iniciais dos cursos de computação se mantém a percebida no gráfico anterior, apesar de temas como cálculo diferencial integral e arquitetura e organização de computadores superarem o desempenho dos residentes do Pará em 6% e 5% respectivamente, suas colocações em relação aos outros temas permanecem as mesmas. Já os temas da área de matemática apresentam melhora em relação ao apresentado pelos residentes do Pará, onde os temas lógica matemática, álgebra linear e matemática discreta possuem desempenho superior a 40%. Os temas com as piores taxas de acertos são os mesmos da figura anterior, onde o desempenho em computação gráfica supera o dos residentes do Pará em apenas 3%.

Figura 7 – Média de acertos de brasileiros por tema normalizada considerando todos os anos (2016-2019).

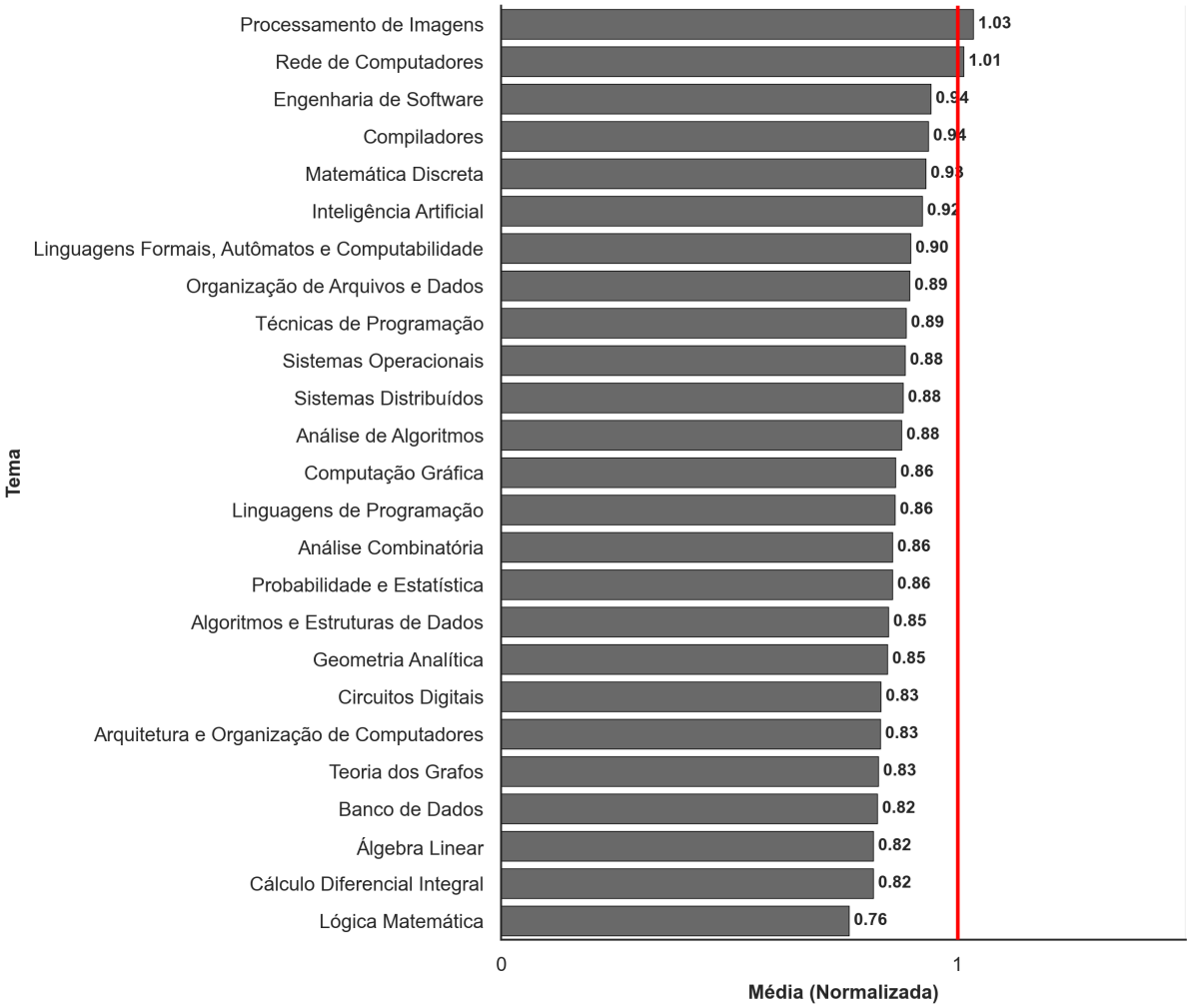


Fonte: compilação do autor.

A partir da segundo passo foi gerada a Figura 8, que representa o desempenho por tema dos residentes do Pará em comparação com o nacional. No que diz respeito à comparação com o desempenho nacional, nota-se que o desempenho dos residentes do Pará é superior apenas em dois temas, sendo eles, processamento de imagens e rede de computadores, superando os resultados nacionais em 0,03 e 0,01, respectivamente. Os demais temas possuem um desempenho relativamente próximo em relação aos obtidos nacionalmente, com grande parte dos temas ficando acima de 0,80, com exceção de lógica matemática. Nota-se que os 3 temas com pior desempenho em relação ao desempenho nacional são pertencentes à área da matemática,

reforçando a deficiência na área no estado do Pará. De modo geral, mesmo que o desempenho dos residentes do Pará sejam inferiores na maioria dos temas, as médias se mantém próximas às nacionais, independente da área a qual estejam atrelados.

Figura 8 – Média de acertos de residentes do Pará por tema normalizada em relação à média nacional considerando todos os anos (2016-2019).



Fonte: compilação do autor.

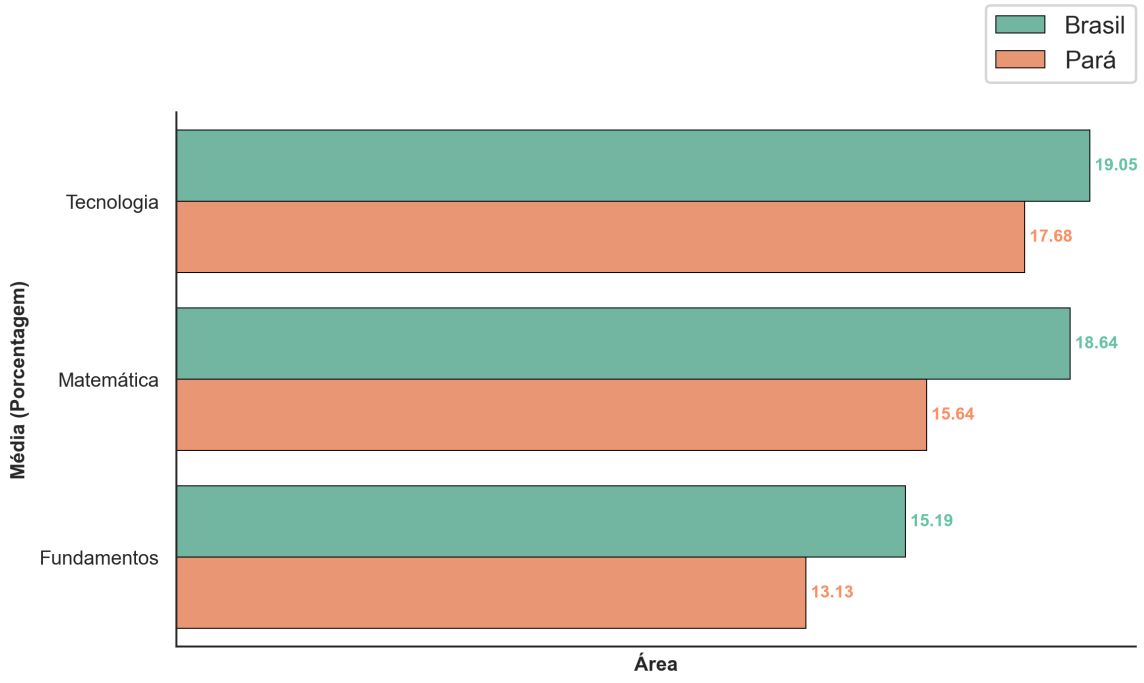
Para além dos 25 temas, a SBC também define 3 grandes áreas que os comportam, as quais podem ser observadas na Tabela 1. A análise por área traz uma visão mais objetiva sobre o desempenho dos egressos, visto que uma área agrega temas de cunho semelhante, o que pode facilitar a análise visual, bem como a adoção de medidas baseadas nos dados por parte de coordenadores e diretores, as quais influenciariam um conjunto de temas.

A partir da divisão dos temas definidas pela SBC, foi gerada a Figura 9, a qual representa a média total de acertos de residentes do Pará em cada área em todo o período analisado. A figura elucida que a área de fundamentos, quando comparada às demais, é aquela que aparece com menor desempenho, já que possui maior número de questões e áreas sob seu domínio em relação às demais, uma explicação para tal desempenho poderia ser o grande leque de conhecimentos

cobrados nesta área. Já a área "Matemática" aparece em segundo, onde sua média está a cerca de 2% do tema "Tecnologia", resultado esperado visto que a área "Matemática" possui médias menores por tema como expresso na Figura 6. A área "Tecnologia" possui a melhor média de acertos por área se considerados os números de questões que cada uma possui, tal desempenho pode ter correlação com a pretensão dos egressos expressa na Figura 11, já que a área de tecnologia é composta pelas 3 áreas mais pretendidas por residentes do Pará na pós-graduação.

No que diz respeito aos acertos por área no país, a Figura 9 expõe um comportamento semelhante ao que ocorre com os residentes do Pará, onde a área de fundamentos aparece com o pior desempenho, possuindo 15,20% de acerto médio nacionalmente, sendo 2% maior do que o obtido pelos residentes do Pará, já a área de matemática aparece em segundo, com média de acertos de 18,65%, enquanto a área de tecnologia foi aquela com o melhor desempenho, sendo de 19,04%, sendo superiores às médias do Pará em 3% e 1% respectivamente. Percebe-se também que a distância entre a porcentagem de acertos nas duas primeiras áreas é menor do que a observada no gráfico 9, expondo um equilíbrio maior nas áreas de competência em âmbito nacional.

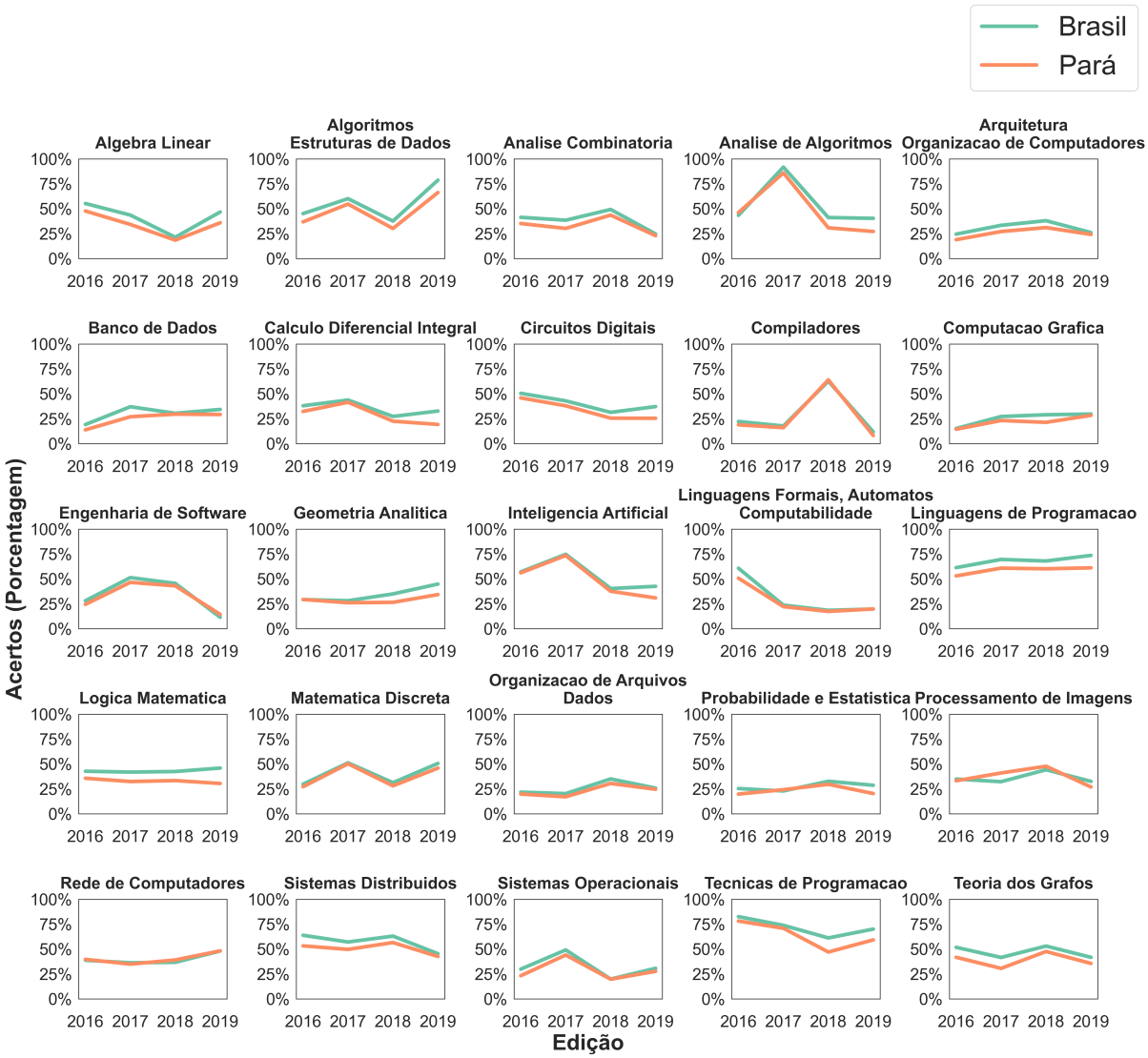
Figura 9 – Média de acertos total dos residentes do Pará e Brasil por área normalizada considerando todos os anos (2016-2019).



Fonte: compilação do autor.

A Figura 10 foi gerada a partir da média de acertos por tema em cada edição, e apresenta a comparação individual entre a porcentagem média de acertos em cada tema de residentes do Pará e do Brasil, onde o Pará apresenta desempenho igual ou inferior ao nacional em todo o período, apresentando desempenho maior que o nacional somente nos temas "processamento de imagens" na edição de 2017, e "rede de computadores" na edição de 2018.

Figura 10 – Porcentagem de acerto por tema (Brasil x Pará).



Fonte: compilação do autor.

Embora o desempenho dos residentes do Pará esteja abaixo das médias nacionais em grandes partes dos temas, o perfil dos candidatos se assemelha ao perfil nacional em relação ao desempenho na prova, onde é visível que as curvas de desempenho do Pará seguem, em geral, a tendência nacional, como exemplo se tem o tema “Algoritmos e Estruturas de Dados” onde tanto o desempenho nacional quanto no estado apresentaram aumento de 2016 a 2017, queda de 2017 a 2018 e novo aumento de 2018 a 2019.

4.3 Dados sobre especialidades pretendidas

Ao se inscrever no POSCOMP, os egressos têm a oportunidade de informar em quais áreas de mestrado ou doutorado gostariam de adentrar, onde podem ser informadas diversas áreas pretendidas, a partir da digitação livre em um campo de texto do formulário de inscrição.

A partir da padronização dos interesses descrita na subseção 3.3, verificou-se que 70 especialidades foram informadas por residentes do Pará. Ademais, como os inscritos escrevem a especialidade em um campo de texto livre, é possível que a escrita ocorra de forma diferente para cada uma, dificultando o agrupamento. Dessa forma, foram realizados processos de normalização das especialidades pretendidas, onde foram removidas divisões realizadas por barras e outros caracteres, bem como foi feita a uniformização de acentuação e letras maiúsculas e minúsculas.

Após a normalização das especialidades pretendidas, foi realizado um agrupamento por intermédio da correlação das especialidades, onde especialidades semelhantes como "Inteligência artificial" e "Inteligência computacional" foram agrupadas em uma área comum, o que reduziu as 70 especialidades para 32. Após a redução, foi calculado o número de egressos que pretendem adentrar em cada área, viabilizando a análise em relação às especialidades pretendidas no estado do Pará. Dentre as 32 especialidades, foram escolhidas apenas 7 para a composição do gráfico, tendo em vista que seus números superam 10 candidatos, as demais especialidades e seus respectivos participantes podem ser consultados na Tabela 6, presente nos apêndices.

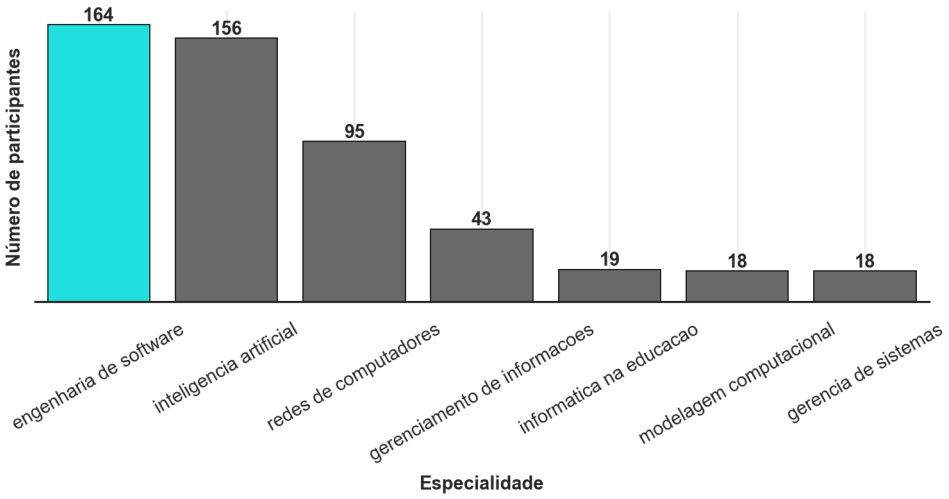
Na Figura 11, nota-se que engenharia de software e inteligência artificial são as especialidades com maior pretensão por parte dos egressos no estado do Pará, com 164 e 156 candidatos com pretensão de entrada em cursos da área, respectivamente. Redes de computadores e gerenciamento de informações também possuem grande pretensão, com mais de 40 candidatos com pretensão de entrada, por fim, informática na educação, gerência de sistemas e modelagem computacional aparecem com números semelhantes.

Além disso, foi realizada uma análise em relação aos alunos ativos no Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Pará, levando em consideração o período de 2017 a 2022, no intuito de verificar e comparar se as especialidades pretendidas por pate dos candidatos do POSCOMP são aderentes às especialidades cursadas pelo corpo discente atual do PPGCC.

A Figura 12 foi gerada a partir da lista de alunos ativos no PPGCC da UFPA[3], os alunos ativos são aqueles que estão cursando o programa atualmente, onde foram contabilizados os alunos por especialidade cursada com base no currículo lattes do seu orientador principal, sendo realizado um agrupamento de especialidades semelhante ao já feito na base de dados do estado, como exemplo a especialidade "inteligência computacional - bioinformática" foi incluída em "inteligência computacional" na contagem.

[3] Link para o website: <https://sigaa.ufpa.br/sigaa/public/programa/alunos.jsf?lc=pt_BR&id=338>

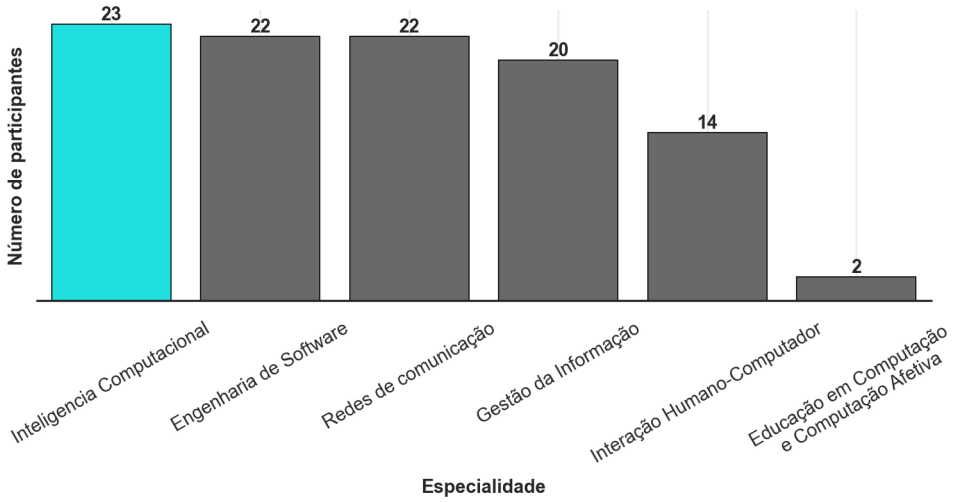
Figura 11 – 7 especialidades mais pretendidas no Pará considerando todos os anos (2016-2019).



Fonte: compilação do autor.

Com base no exposto na Figura 12, o Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Pará atende às principais especialidades pretendidas pelos egressos. Constatou-se que a UFPA atende a 5 das 7 especialidades, onde não foram encontrados projetos de pesquisa ou grupos de pesquisa com foco em gerência de sistemas e modelagem computacional, assim, conclui-se que a UFPA atende de forma aceitável às principais demandas dos egressos da região, supondo que os mesmos desejassem aplicar para o programa de pós-graduação da instituição.

Figura 12 – Alunos por especialidade de pós-graduação na UFPA.



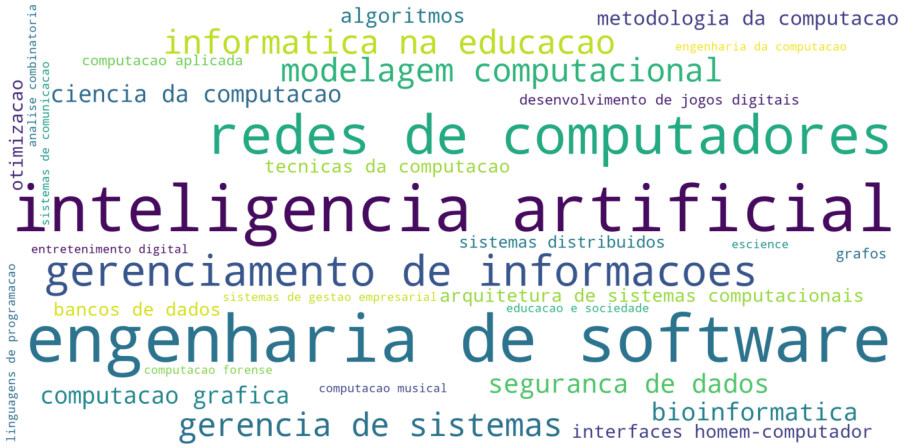
Fonte: compilação do autor.

Por fim, decidiu-se pela geração de uma nuvem de palavras como representação alternativa das especialidades mais pretendidas por residentes do Pará, tendo em vista a natureza dos

dados acerca de especialidades pretendidas, compostos por especialidades que apresentam certa frequência, o que é característico das nuvens de palavras.

A nuvem de palavras presente na Figura 13 foi gerada a partir da Tabela 6, que contém todas as especialidades pretendidas por residentes do Pará. Quanto mais ocorrências a especialidade possui, isto é, quanto mais participantes declararam interesse na área, maior a fonte representada na nuvem, o que torna mais visual o resultado obtido no gráfico anterior, com as especialidades engenharia de software e inteligência artificial possuindo fontes maiores, enquanto especialidades como análise combinatória e educação e sociedade possuem fontes menores.

Figura 13 – Nuvem de palavras com base nas especialidades mais pretendidas no Pará.



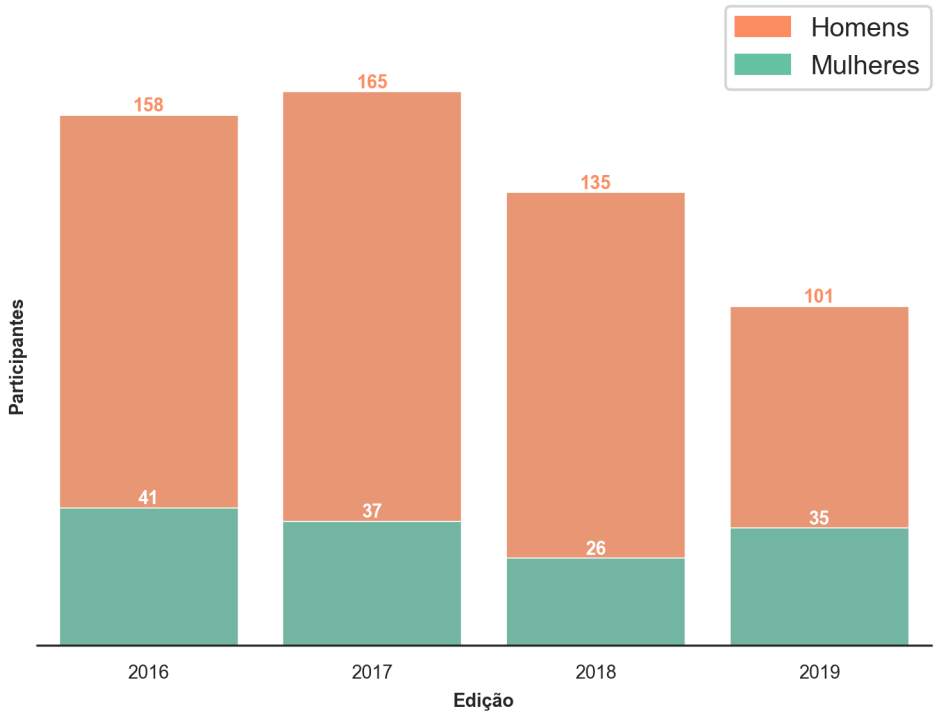
Fonte: compilação do autor.

4.4 Dados sobre participantes por sexo

Verificar a distribuição de participantes por sexo é relevante, pois, a presença de mulheres na computação tem se mostrado pequena em relação ao número de homens, e segundo Freitas, Cosme e Nascimento (2019) o entendimento de aspectos que aumentam tal disparidade é necessário para que se criem ações voltadas à inserção de mulheres na área, assim, a visualização dos resultados permite que coordenações de cursos e outros interessados analisem os impactos de implementações de políticas de incentivo à adesão de candidatas.

As figuras 14 e 15 foram geradas a partir do campo "Sexo" presente nos dados de inscrição, onde no caso da Figura 14 foram filtrados somente os residentes do Pará. Na Figura 14, que representa o número de inscritos por sexo no Estado do Pará, nota-se que o número de homens inscritos no exame foi superior ao de mulheres em todo o período, onde o número de inscritas nunca chegou a 30% do total. O número de inscritas no exame apresentou queda de 41 na edição de 2016 para 26 em 2018, representando uma diminuição de 4,46%, havendo aumento para 35 no ano de 2019, sendo o ano com maior porcentagem de mulheres em relação ao total de participantes, contando com 25,73% de inscritas.

Figura 14 – Número de inscritos por sexo (Pará).



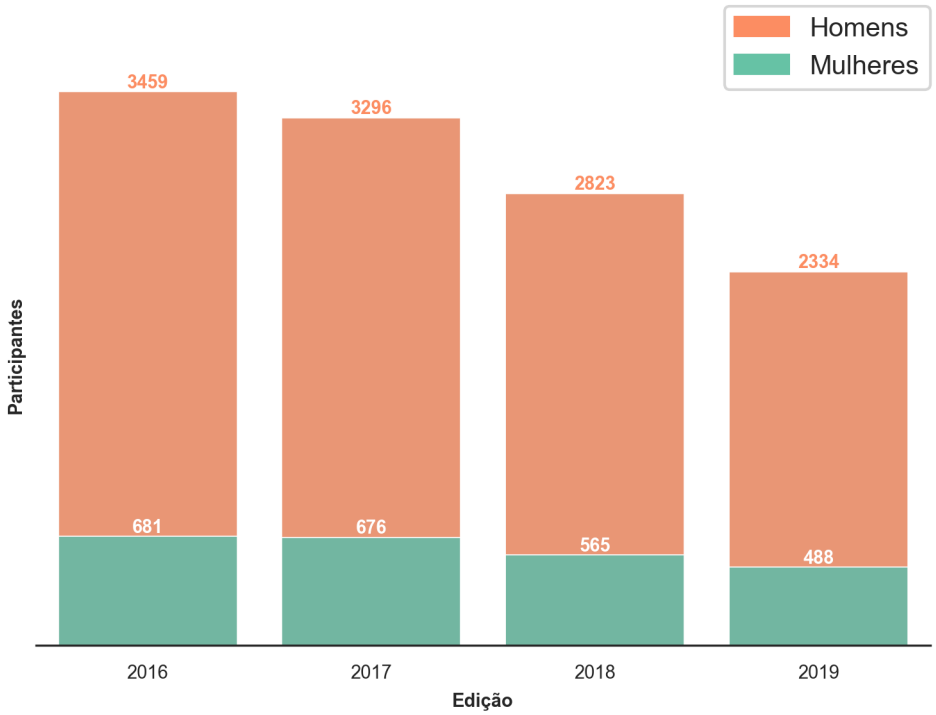
Fonte: compilação do autor.

O aumento de inscritas residentes no Pará em 2019, em contraponto ao número de homens, pode estar ligado a iniciativas existentes para a inclusão de mulheres na computação como o Programa Meninas Digitais proveniente da SBC, o qual tem como objetivo levar a área de computação através de dinâmicas, palestras e cursos para despertar o interesse de

meninas estudantes do ensino médio na área. Além disso, o programa possui projetos parceiros pertencentes a diferentes instituições, os quais podem ter contribuído para esse aumento.

Na Figura 15, que representa o número de inscritos por sexo no país, é visível que o número total de inscritos somente diminuiu em todo o período, o que se aplica tanto para homens quanto para mulheres. Houve aumento no número total de inscritos no Pará de 2016 a 2017, bem como houve aumento no número de inscritas no período de 2018 a 2019, contrariando as tendências nacionais. Contudo, a partir de 2017, o Pará acompanhou a tendência nacional de queda no número total de inscritos. A edição de 2019 foi aquela com a maior presença de mulheres em relação ao total, com 17,29% de inscritas no exame, contudo, isso se deve mais ao fato de que o número de homens inscritos apresentou queda de 32,52% em relação ao primeiro ano analisado, enquanto o número de mulheres caiu 28,34% em relação ao mesmo ano.

Figura 15 – Número de inscritos por sexo (Brasil).



Fonte: compilação do autor.

O número total de inscritos brasileiros no exame era de 4140 em 2016, caindo para 2822 em 2019, representando uma queda de 31,45% no total de inscritos, uma possibilidade é que os egressos de computação estejam optando pela inserção no mercado ou outros cursos de capacitação, ao invés de realizarem a prova para auxiliar na entrada na pós-graduação. Já no Pará, o total de inscritos era de 199 em 2016, número reduzido a 136 em 2019, o que configurou uma queda de 31,65% nos inscritos do estado.

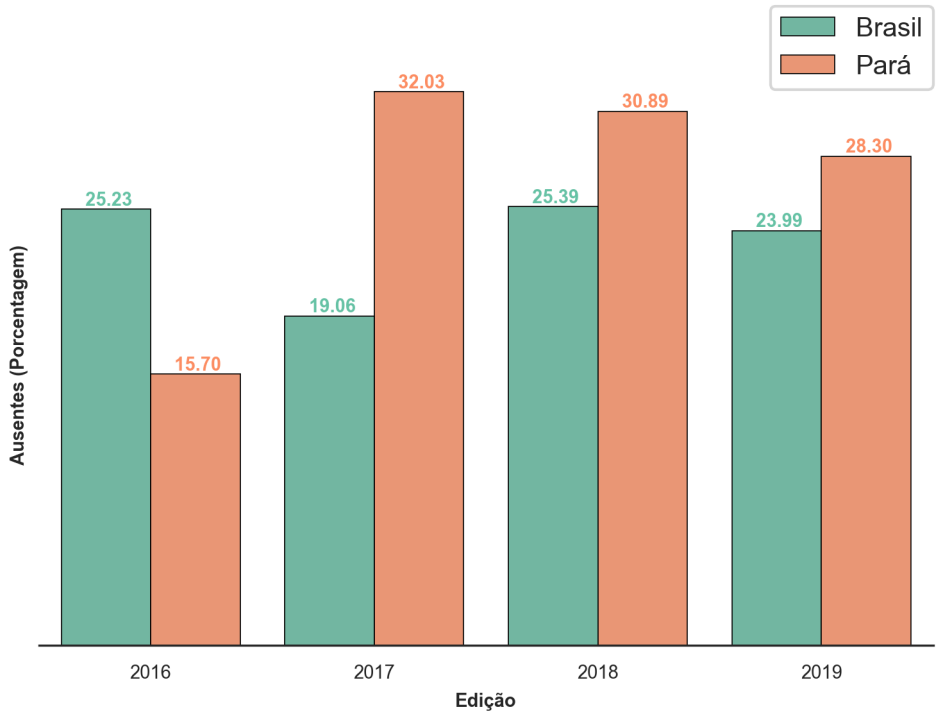
4.5 Dados sobre ausentes

Mensurar os ausentes é importante para verificar a adesão à prova, sendo necessário ressaltar o custo da inscrição, o qual é acima de cem reais, onde outro ponto relevante para a análise é que não são necessários deslocamentos interestaduais para a realização da prova, mesmo que o candidato deseje utilizar a nota em universidades de outros locais, contudo, ainda é possível que deslocamentos intra estaduais sejam necessários, visto que apenas algumas cidades aplicam o exame.

Para obter o número de ausentes, foi selecionado o campo de presença do dia 1 de prova, tendo em vista que os demais campos se encontram vazios. Para a geração do gráfico de ausentes foi criada uma base de dados contendo os atributos [ano, ausentes, local], a qual contém as porcentagens médias de ausentes no Pará e no Brasil por edição do exame. Foram obtidos os números de candidatos presentes e os ausentes, a partir dos quais foram adquiridas as razões de egressos ausentes em relação ao total de participantes, as quais foram multiplicadas por 100, o que resultou na porcentagem de ausentes em cada edição, onde o mesmo processo foi aplicado para o Brasil e também para os residentes do Pará.

No que diz respeito aos inscritos ausentes, é possível observar na Figura 16 que, com exceção de 2016, a porcentagem de ausentes do estado do Pará supera a nacional nas demais edições, possuindo comportamento dissidente do nacional.

Figura 16 – Ausentes por edição (Brasil x Pará).



Fonte: compilação do autor.

A porcentagem de ausentes dobrou no período de 2016 a 2017 no Pará, contrariando o

comportamento nacional, que apresentou queda no mesmo período. A porcentagem de ausentes no Brasil foi de 19,03% no ano de 2017 para 25,39% em 2018, mostrando aumento do número de ausentes, enquanto a porcentagem relacionada ao Pará foi de 32,03% em 2017 para 30,89% em 2018, expondo a redução dos ausentes nesse período, redução que se manteve na edição de 2019, que teve 28,30% dos inscritos ausentes. A FUNDATEC (2022) expõe que 25,19% dos candidatos totais estiveram ausentes na edição de 2022, configurando um crescimento superior a 1% em relação à edição de 2019.

Percebe-se que a porcentagem de ausentes permanece alta mesmo após a pandemia, com 1/4 dos inscritos ausentes na edição de 2022 do exame. As possíveis causas para a ausência podem ser o desinteresse pela prova, onde oportunidades ligadas ao mercado de trabalho podem ter surgido para os egressos no período anterior a realização do exame, ou ainda a possibilidade de entrada na pós-graduação por meios como processos seletivos específicos.

5 CONSIDERAÇÕES FINAIS

Este trabalho realizou a exploração das bases de dados do POSCOMP, a fim de gerar informações acerca dos desempenhos e perfis dos candidatos da prova que residem no estado do Pará, com o recorte temporal das edições de 2016 a 2019. As informações geradas pelo trabalho podem beneficiar diversos interessados, como gestores acadêmicos, candidatos, e demais interessados, principalmente no que tange aos programas de pós-graduação em computação.

A pesquisa se iniciou com a realização de uma revisão bibliográfica para entendimento das principais técnicas e conceitos relacionados ao KDD e à EDM, bem como sobre os trabalhos já existentes acerca do exame, onde se constatou que só haviam trabalhos visando suportar os alunos na realização da prova, não sendo realizada a análise ou exploração da base de dados.

Os conceitos pesquisados suportaram a exploração da base de dados do POSCOMP, tendo em vista a geração de informações e gráficos que podem levar a um melhor entendimento acerca do desempenho e características dos candidatos no estado, como também podem suportar a tomada de decisão por parte dos interessados. A exploração foi realizada a partir de objetivos que visam gerar informações sobre as notas totais, notas por matéria, especialidades pretendidas, participantes por sexo, e ausentes, resultados que foram descritos com maior detalhe na seção 4, com a apresentação de gráficos e discussão das informações obtidas.

Como principais resultados, verificou-se que a média das notas totais dos residentes do Pará foram menores que as nacionais em todo o período analisado, acompanhando a tendência nacional de queda a partir de 2017, onde o comportamento se repetiu quando se compararam as maiores notas obtidas no estado e nacionalmente em cada edição. Já no desempenho por tema, observou-se que os dois temas com maior porcentagem de acertos de residentes do Pará estão relacionados à programação, e que o desempenho dos residentes do Pará foi maior que o nacional em 2 temas em todo o período, onde os demais temas apresentam resultados próximos aos nacionais, reforçando o resultado obtido nas notas totais.

Ainda em relação aos resultados, na análise de especialidades pretendidas por residentes do Pará, as três especialidades mais pretendidas foram engenharia de software, inteligência artificial e redes de computadores, sendo constatado também que os resultados são aderentes aos alunos ativos da UFPA, a qual atende a demanda em 5 das 7 especialidades mais pretendidas por residentes do estado, onde as duas especialidades não atendidas são gerência de sistemas e modelagem computacional. Para atender também essas duas especialidades, sugere-se, por exemplo, que seja considerada a criação de mestrados profissionais nessas áreas, com a alocação de professores em futuras chamadas para participação no PPGCC, de forma que esses novos docentes lecionem nesses campos, ou ainda, que os professores que atuem no programa assistam também para essas especialidades, abrangendo suas respectivas áreas de atuação nos próximos editais do programa ou na página do próprio programa de pós-graduação.

Quanto aos inscritos por sexo, nacionalmente, o número de mulheres apresentou queda em todo o período, em contraponto, o número de inscritas residentes no Pará apresentou alta no período de 2018 a 2019. Por fim, no que tange aos ausentes, tem-se que o número de ausentes entre residentes do Pará mais que dobrou de 2016 a 2017, contudo, o número de ausentes tem apresentado queda a partir de então, com tendência diferente da nacional

Em relação a possíveis trabalhos futuros, o escopo da análise pode ser expandido, com a realização de uma análise por região ou mesmo nacional, onde também é possível que os dados sejam minerados com o objetivo de identificar padrões das notas dos candidatos conforme as áreas de conhecimentos, com a utilização de clusterização, ou mesmo com a finalidade de prever as médias dos candidatos nas próximas edições do exame, o que pode ser feito através de algoritmos de regressão, ou ainda, pode-se analisar as possíveis causas para o decréscimo de candidatos no exame nas últimas edições.

REFERÊNCIAS

ALGARNI, A. Data mining in education. **International Journal of Advanced Computer Science and Applications**, Science and Information (SAI) Organization Limited, v. 7, n. 6, 2016.

ANOOPKUMAR, M.; RAHMAN, A. M. Z. A review on data mining techniques and factors used in educational data mining to predict student amelioration. In: IEEE. **2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)**. [S.l.], 2016. p. 122–133.

AUGUSTO, A. L. A. et al. Comparação dos conteúdos do poscomp com o currículo de referência dos cursos de computação da sbc. **Revista ComInG - Communications and Innovations Gazette**, v. 5, n. 3, p. 14–23, nov. 2021. Disponível em: <https://periodicos.ufsm.br/coming/article/view/67894>.

BAKER, R. et al. Data mining for education. **International encyclopedia of education**, Elsevier Oxford, UK, v. 7, n. 3, p. 112–118, 2010.

BATISTA, E. J. S. et al. Desenvolvimento de um aplicativo para android com questões do poscomp como um objeto de aprendizagem para o auxílio no ingresso a programas de pós-graduação. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2014. v. 3, n. 1, p. 127.

CALLEGARI, B. V.; OLIVEIRA, K. A. d. Desenvolvimento de um aplicativo para a análise de questões do poscomp na área de matemática e suas recorrências. 2020.

CUNHA, R.; SALES, C.; SANTOS, R. Análise automática com os microdados do enade para melhoria do ensino dos cursos de ciência da computação. In: **Anais do XXIX Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 208–217. ISSN 2595-6175. Disponível em: <https://sol.sbc.org.br/index.php/wei/article/view/15912>.

DE SORDI Jr., F. **Desenvolvimento de um ambiente colaborativo de treinamento preparatório para o POSCOMP**. Dissertação (Mestrado) — Universidade Estadual de Londrina, 2015. Disponível em: <http://www.bibliotecadigital.uel.br/document/?code=vtls000202187>.

DE SORDI Jr., F.; BRANCHER, J. Uma pesquisa de opinião sobre a relevância dos conteúdos abrangidos pelo poscomp. In: SBC. **Anais do XXII Workshop sobre Educação em Computação**. [S.l.], 2014. p. 90–99.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996.

FAYYAD, U. M. et al. Knowledge discovery and data mining: Towards a unifying framework. In: **KDD**. [S.l.: s.n.], 1996. v. 96, p. 82–88.

FREITAS, B.; COSME, L.; NASCIMENTO, M. Exame nacional de desempenho de estudantes (enade): Análise do perfil das mulheres dos cursos da área de computação. In: SBC. **Anais do XIII Women in Information Technology**. [S.l.], 2019. p. 179–183.

FUNDATEC. **Gabaritos disponíveis: Exame POSCOMP 2022 foi aplicado em 24 cidades brasileiras no último domingo (18/09)**. 2022. FUNDATEC. Disponível em: <https://www2.fundatec.org.br/2022/09/19/gabaritos-disponiveis-exame-poscomp-2022-foi-aplicado-em-24-cidades-brasileiras-no-ultimo-domingo-18-09/>. Acesso em: 20 dez. 2022.

HUI, H. et al. Application of student achievement analysis based on apriori algorithm. In: IEEE. **2020 2nd International Conference on Information Technology and Computer Application (ITCA)**. [S.l.], 2020. p. 19–22.

MAIMON, O.; ROKACH, L. Introduction to knowledge discovery and data mining. In: ____. **Data Mining and Knowledge Discovery Handbook**. Boston, MA: Springer US, 2010. p. 1–15. ISBN 978-0-387-09823-4. Disponível em: <https://doi.org/10.1007/978-0-387-09823-4_1>.

MARTINS, M. P.; MIGUÉIS, V.; FONSECA, D. Data mining educacional: uma revisão da literatura. In: INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS. **13th Iberian Conference on Information Systems and Technologies (CISTI)**. [S.l.], 2018.

MENDES, F. M.; MENDONÇA, A. P.; GUEDES, E. B. Poscomp coach: Plataforma web para apoio ao ingresso na pós-graduaçao em computaçao. **RENOTE**, v. 16, n. 1, 2018.

PEÑA-AYALA, A. Educational data mining. **Studies in Computational Intelligence**, Springer, v. 524, 2014.

RIBEIRO, T. C.; JUNIOR, W. F. C. Plataforma para auxílio na preparação de estudantes para as avaliações do enade e poscomp. 2020.

ROMERO, C.; VENTURA, S. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 3, n. 1, p. 12–27, 2013.

SBC. **Comitê Gestor do POSCOMP decide pela não realização da prova em 2021**. 2021. POSCOMP. Disponível em: <https://www.sbc.org.br/noticias/2302-comissao-especial-de-informatica-na-educacao-lanca-portal-para-criar-diretorio-de-especialistas>. Acesso em: 22 out. 2022.

SBC. **Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP)**. 2021. POSCOMP. Disponível em: <https://www.sbc.org.br/index.php?option=com_content&view=article&layout=edit&id=458>. Acesso em: 06 out. 2022.

Apêndices

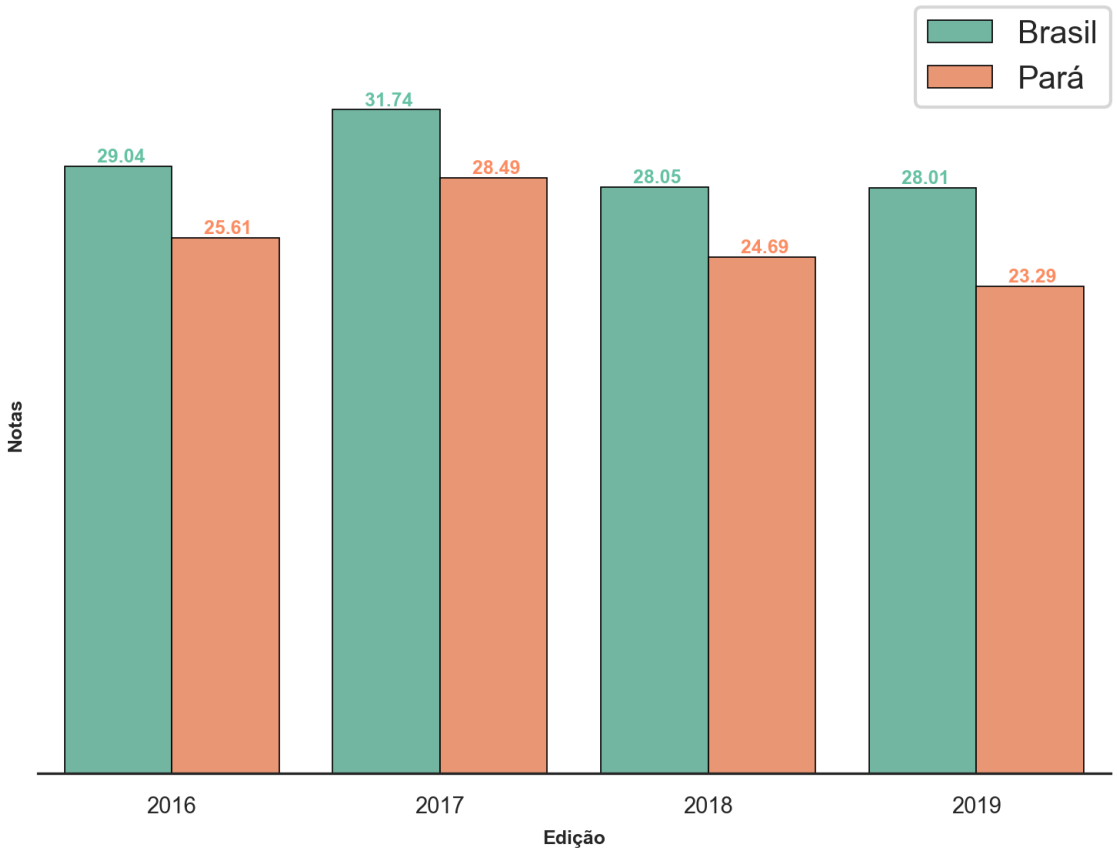
APÊNDICE A – TRABALHOS PUBLICADOS PELO AUTOR

1. FERREIRA, L. M. R.; FILHO, R. C. dos S.; COSTA., J. C. de C. **EXPLORATORY ANALYSIS OF THE POSCOMP RESULTS: A CASE STUDY OF CANDIDATES FROM PARÁ ANÁLISE EXPLORATÓRIA DOS RESULTADOS DO POSCOMP: UM ESTUDO DE CASO DOS CANDIDATOS DO PARÁ**. 2022. Disponível em: <http://contecsi.tecsi.org/index.php/contecsi/19CONTECSI/paper/view/7128>.

APÊNDICE B – GRÁFICOS

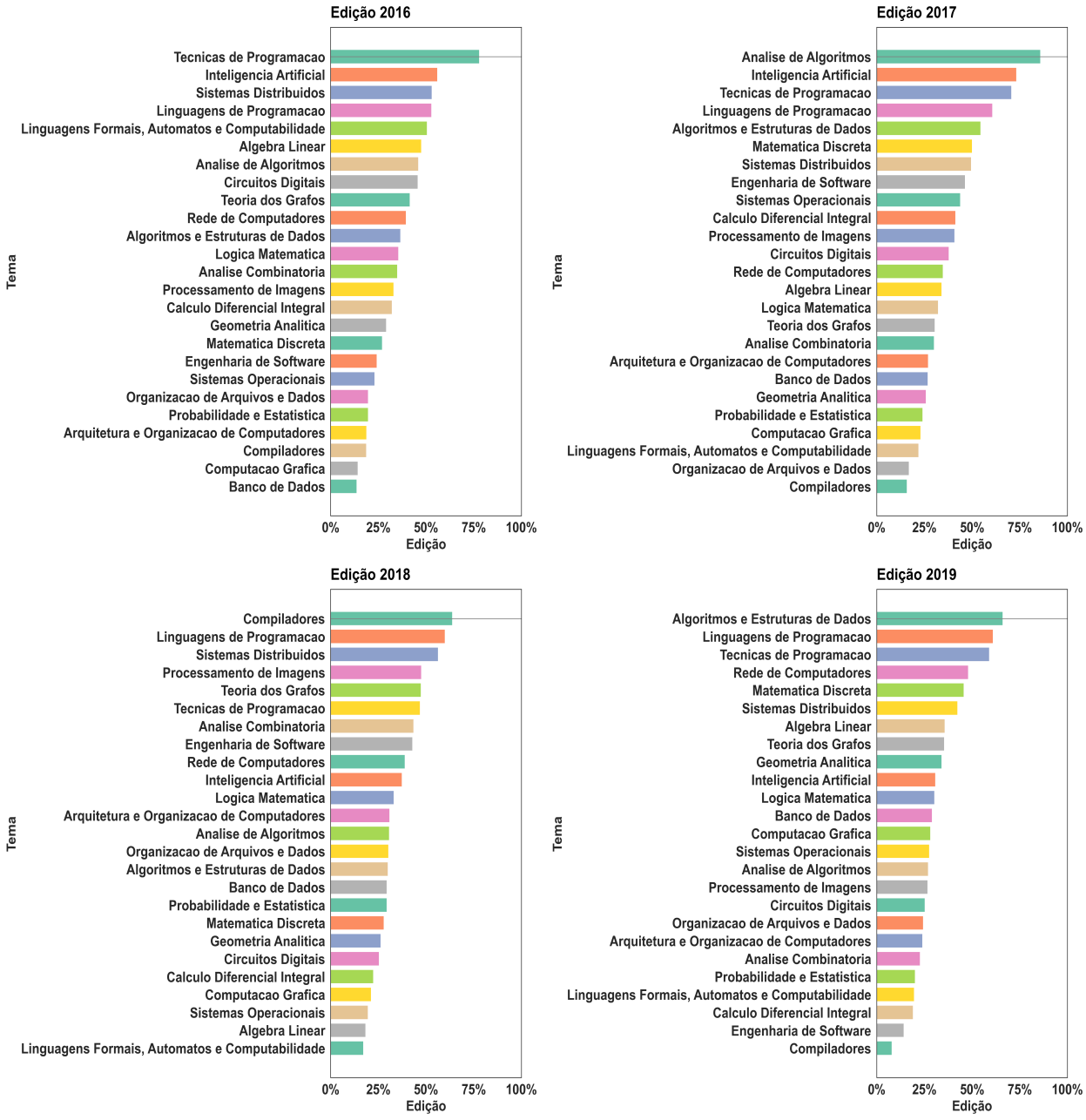
Gráficos complementares, mostram informações em detalhe ou com outra visão.

Figura 17 – Média de notas por edição em barra (Brasil x Pará)



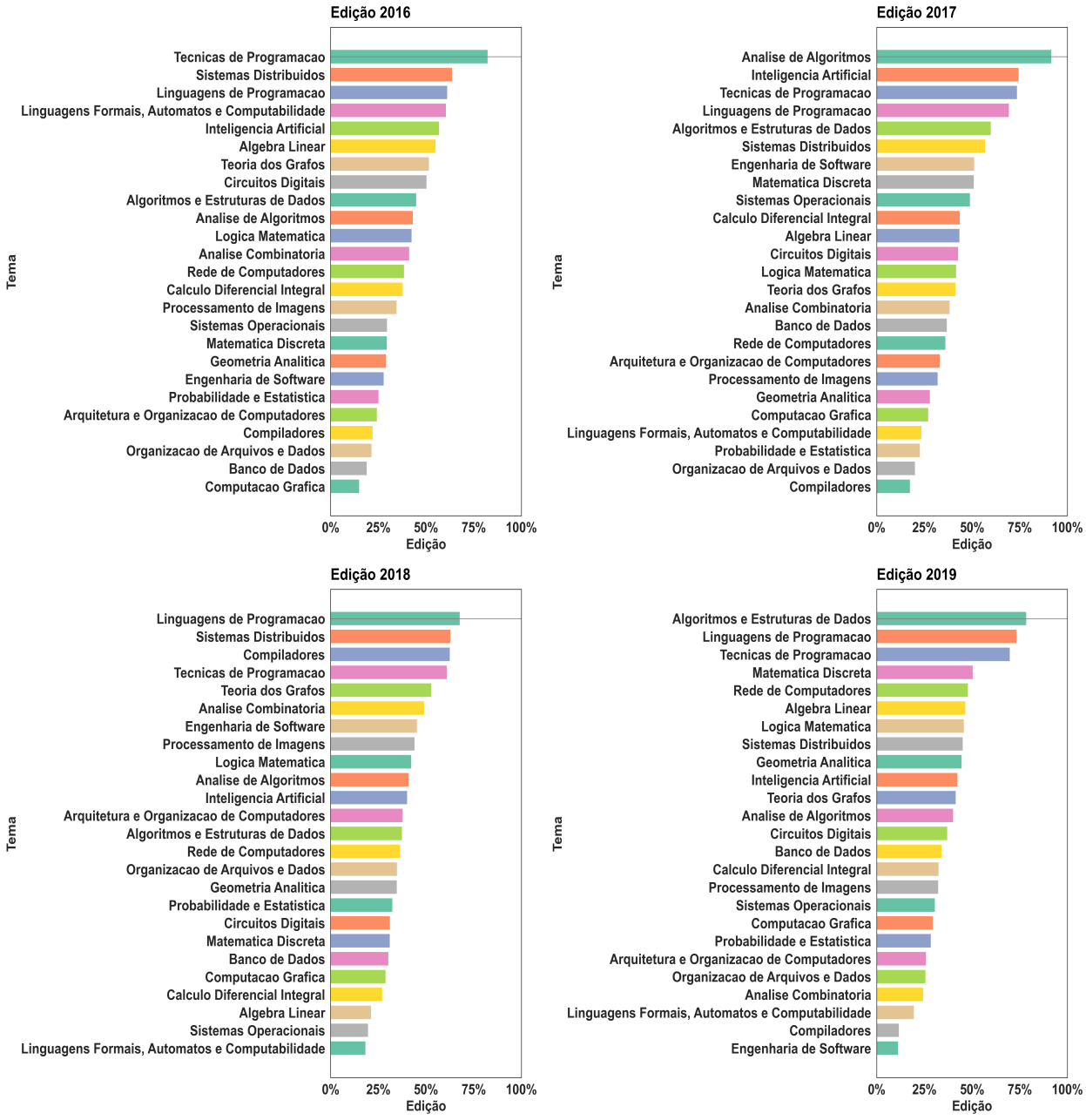
Fonte: compilação do autor.

Figura 18 – Média de acertos por tema por ano (Pará)



Fonte: compilação do autor.

Figura 19 – Média de acertos por tema por ano (Brasil)



Fonte: compilação do autor.

APÊNDICE C – TABELAS

Tabela 6 – Especialidades pretendidas por residentes do Pará.

ESPECIALIDADES PRETENDIDAS	
Especialidade	**Alunos**
engenharia de software	164
inteligência artificial	156
redes de computadores	95
gerenciamento de informações	43
informática na educação	19
modelagem computacional	18
gerência de sistemas	18
segurança de dados	9
bioinformática	8
computação gráfica	7
ciência da computação	7
metodologia da computação	5
algoritmos	5
interfaces humano-computador, otimização	4
arquitetura de sistemas computacionais	4
bancos de dados	4
técnicas da computação	4
sistemas distribuídos	3
computação aplicada, grafos	2
desenvolvimento de jogos digitais	2
analise combinatória	1
escience	1
computação musical	1
computação forense	1
entretenimento digital	1
linguagens de programação	1
sistemas de gestão empresarial	1
engenharia da computação	1
sistemas de comunicação	1
educação e sociedade	1

Fonte: compilação do autor.