Submission date: 14/08/2023
Resubmission date: 15/10/2024
Camera ready submission date: 15/12/2024

1st round notification: 15/08/2024
2nd round notification: 01/12/2024
Available online: 26/12/2024
Publication date: 26/12/2024

Section: Regular article

Intelligent Vocational Guidance Based on Machine Learning Applied to POSCOMP Microdata

Orientação Vocacional Inteligente Baseada em Aprendizado de Máquina Aplicado aos Microdados do POSCOMP

Jean Carlos Carvalho Costa[1], **Lucas Mesquita Rodrigues Ferreira**[1], **Reginaldo Cordeiro dos Santos Filho**[1]

[1]Laboratório de Inteligência de Dados (LID) – Centro Paraense de Computação Distribuída de Alto Desempenho (CCAD) – Universidade Federal do Pará (UFPA)
Caixa Postal 66.075–110 – Belém – PA – Brasil

{jeancc.costa, mesquita.py}@gmail.com, regicsf@ufpa.br

Abstract. *The National Examination for Admission to Graduate Studies in Computing (POSCOMP) is applied by the Brazilian Society of Computing (SBC) to assess the knowledge of candidates to graduate programs in Computing in Brazil. As one of the main assessment instruments, the results of the POSCOMP exam, i.e., the database, have the potential to reveal relevant patterns about candidates. In this sense, this article aims to propose a clustering-based intelligent vocational guidance system from the exploratory analysis of the POSCOMP results between the years 2016 to 2019. This vocation system guides the candidate to follow a research area in post-graduation based on the performance obtained in the subjects of the test. The proposed approach aims to guide and support students in their academic decisions.*

Keywords. *Vocational Guidance; Educational Data Analysis; Graduate Studies in Computer Science; Clustering Algorithms; Data Mining.*

Resumo. *O Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP) é aplicado pela Sociedade Brasileira de Computação (SBC) para avaliar o conhecimento dos candidatos aos programas de pós-graduação em Computação no Brasil. Sendo um dos principais instrumentos avaliatórios, os resultados da prova do POSCOMP, i.e., a base de dados, tem o potencial de relevar padrões relevantes sobre os candidatos. Neste sentido, este artigo tem o objetivo de propor uma orientação vocacional inteligente baseado em clusterização a partir da análise exploratória dos resultados do POSCOMP entre os anos de 2016 a 2019. Este sistema de vocação orienta o candidato a seguir uma área de pesquisa na pós-graduação a partir do desempenho obtido nos temas da prova. A abordagem proposta busca orientar e apoiar os estudantes em suas decisões acadêmicas.*

Cite as: Costa, J. C. C.; Ferreira, L. M. R.; & Filho, R. C. S. (2023). Intelligent Vocational Guidance Based on Machine Learning Applied to POSCOMP Microdata (Orientação Vocacional Inteligente Baseado no Aprendizado de Máquina dos Microdados do POSCOMP). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), Vol 17(1), 19:1-19:26. DOI: 10.5753/isys.2024.3612

Palavras-Chave. *Orientação Vocacional; Análise de Dados Educacionais; Pós-Graduação em Computação; Algoritmos de Clusterização; Mineração de Dados.*

1. Introdução

O Exame Nacional para Ingresso na Pós-Graduação em Computação (POSCOMP) é uma prova realizada anualmente desde 2002, com o objetivo de proporcionar aos candidatos a oportunidade de ingressar em programas de pós-graduação na área de computação em todo o país [Sociedade Brasileira de Computação 2022]. O POSCOMP é uma avaliação objetiva e independente que visa avaliar o conhecimento dos candidatos. As instituições que oferecem programas de pós-graduação em computação adotam geralmente o POSCOMP como um dos critérios de seleção para preenchimento de vagas de mestrado e doutorado, ou ainda, para ranqueamento e distribuição de bolsas de pesquisa [Junior and Brancher 2014].

A prova do POSCOMP abrange três áreas de conhecimentos: Matemática, Fundamentos da Computação e Tecnologia da Computação, como ilustra a Tabela 1. Anualmente, o exame consiste em 70 questões de múltipla escolha que abordam diversos conteúdos. Os candidatos devem obter uma nota suficiente no exame para se qualificarem para ingressar em cursos de pós-graduação em computação [Sociedade Brasileira de Computação 2022].

Tabela 1. Conteúdos programáticos e número de questões do POSCOMP.

Áreas	Nº de Questões
Matemática	20 questões
Fundamentos da computação	30 questões
Tecnologia da computação	20 questões

Um fato importante a ser destacado é que o exame permite que os candidatos aos programas de pós-graduação realizem a prova com o mínimo de deslocamento necessário [Moura et al. 2012]. Isso desempenha um papel significativo na inclusão social dos candidatos, pois evita o aumento de custos relacionados à locomoção ao realizar a prova. Essa acessibilidade ampliada contribui consideravelmente para os candidatos terem acesso aos melhores programas de pós-graduação disponíveis no país.

A orientação vocacional refere-se aos processos e práticas que orientam os indivíduos na tomada de decisões informadas sobre suas carreiras e escolhas educacionais. Ela engloba uma série de atividades que visam alinhar interesses e habilidades pessoais com possíveis oportunidades de emprego ou acadêmicas, facilitando transições mais suaves para a força de trabalho e para a pesquisa. Os principais aspectos da orientação vocacional são: contexto histórico e influência política — evoluiu significativamente, influenciada por políticas públicas e estruturas educacionais [Melo-Silva et al. 2004]; integração educacional — a relação entre educação profissional e educação tradicional está cada vez mais difusa. Mudanças recentes enfatizam a necessidade de as escolas prepararem os discentes para um ambiente de pesquisa dinâmico, promovendo habilidades que sejam adaptáveis a diversas carreiras acadêmicas [Connell et al. 1994].

A análise dos dados dos participantes do POSCOMP tem importância significativa para as instituições de ensino do país, ao permitir a identificação das tendências, padrões e lacunas no conhecimento dos candidatos. Além disso, fornece uma visão valiosa sobre as áreas de interesse e competências dos proponentes, além de contribuir para elaboração de estratégias de redução da taxa de reprovações em disciplinas, desistências e outros desafios comumente enfrentados no âmbito educacional [Hui et al. 2020]. Isto não apenas ajuda as instituições a aperfeiçoar seus respectivos projetos pedagógicos e se posicionar de acordo com os interesses dos candidatos, mas também auxilia o próprio candidato a entender pormenorizadamente suas habilidades, contribuindo inclusive no apoio e tomada de decisão sobre a área que o mesmo deverá se aprofundar cientificamente [Carvalho et al. 2021].

Dessa forma, a criação de um sistema automatizado e inteligente torna importante no auxílio dos participantes na escolha de especialidades para o desenvolvimento de pesquisa na PPG em Computação [Imdad et al. 2017]. Como definem os autores [Russell and Norvig 2022], os agentes baseados em decisões buscam uma sequência de ações que alcance os objetivos. Portanto, a utilização de ferramentas inteligentes na orientação de escolha da especialidade, torna-se significativo tanto para as instituições quanto para os discentes. Pois, segundo [Ribeiro and Uvaldo 2007], considera que o desempenho de uma ocupação em harmonia com as aptidões, habilidades e interesses, tornaria o trabalho mais agradável, com uma maior produtividade e eficiência. Isso contribui para a progresso das pesquisas durante a jornada acadêmica dos discentes. Logo, buscam correlacionar as áreas de estudo com os discentes e com orientadores do programa [Droescher and Silva 2014].

Diante desse contexto, o presente trabalho tem como objetivo propor um sistema de orientação vocacional baseada em inteligência computacional, com técnicas de clusterização. Para atingir esse objetivo, será realizada uma análise exploratória dos microdados do POSCOMP, seguida da aplicação do algoritmo *K-Means* [Lloyd 1982, MacQueen 1967] para detectar padrões de interesse dos candidatos da prova. O intuito é descobrir quais são as áreas mais procuradas pelos proponentes e identificar os temas abordados na prova em que os participantes tem melhor e pior desempenho, em alguns casos considerando o sexo dos participantes. Em linhas gerais, a orientação vocacional proposto consegue prover recomendações de áreas de linhas de pesquisa com base nas notas dos candidatos que refletem as áreas de conhecimento abordadas nas provas do POSCOMP.

2. Trabalhos Relacionados

A literatura envolvendo a análise a posteriori dos microdados do POSCOMP é bastante escassa devido ao fato de que os dados não são publicizados após a divulgação das notas. Esta situação impossibilita que a comunidade científica tenha acesso aos dados e efetivamente elabore pesquisas científicas sobre o desempenho dos candidatos[1]. Entretanto, existem estudos que comparam o conteúdo exigidos no POSCOMP com a grade curricular de cursos de computação, além de pesquisas que oferecem auxílio para o aprendizado dos

[1]As bases de dados trabalhadas nesta pesquisa foram obtidas via ofício para as autoridades mantenedoras dos resultados da prova, salvaguardando o total sigilo aos dados sensíveis.

temas abordados na prova, seja oferecendo direcionamentos ou produzindo simulados via aplicativos de computador. Diante desta constatação, esta seção apresentará, por afinidade com o tema, alguns trabalhos que preservam algum tipo de relação com o desenvolvido por este artigo.

Os autores [Junior and Brancher 2014] conduziram uma pesquisa de opinião com 1239 participantes, incluindo professores, pesquisadores e profissionais da área de computação, para avaliar a relevância dos conteúdos programáticos abrangidos pelo POSCOMP. Os resultados quantitativos indicaram que certos conteúdos do exame apresentam uma relevância diferenciada, fornecendo direcionamentos valiosos para aprimorar os programas de ensino e as estratégias de preparação para o exame, de acordo com as áreas de conhecimento mais valorizadas pelos profissionais da área. Essa identificação de conteúdos relevantes é imprescindível para que os candidatos aumentem as chances de obter melhores desempenhos na prova, visto que a cobrança não é uniforme e, portanto, o estudo pode ser melhor direcionado.

Uma plataforma web, cunhada de *POSCOMP Coach*, foi desenvolvida pelos autores em [Marcelo Mendes et al. 2018]. A aplicação tem o objetivo de auxiliar candidatos na preparação para o exame e oferece uma base com 1120 questões distintas das provas do POSCOMP realizadas de 2002 a 2017, além de permitir a realização de simulados com controle de tempo e correção automática. O estudo também inclui os resultados da avaliação da plataforma, que envolve uma análise comparativa com soluções semelhantes e a avaliação de usabilidade. Os resultados destacam aspectos positivos da solução proposta, incluindo o número de questões disponíveis e a boa pontuação no teste de usabilidade realizado com discentes de todo o país.

A pesquisa disponível em [Silveira et al. 2021] apresenta um processo de avaliação de cursos de graduação, com o objetivo de analisar as questões do ENADE dos anos de 2008 a 2017 e do POSCOMP dos anos de 2014 a 2018, que propõe determinar quais matérias são exigidas nas questões. Os resultados apresentados no trabalho demonstraram a necessidade de uma carga horária de formação abrangente no eixo de fundamentos da computação, a fim de proporcionar uma base sólida aos estudantes e, consequentemente, aumentar as chances de sucesso dos candidatos à pós-graduação.

Uma aplicação móvel destinada ao sistema operacional Android também está disponível na literatura [Batista et al. 2014]. O software tem o objetivo de auxiliar candidatos na preparação para o POSCOMP a partir de uma coleção de questões da prova, organizadas por tópicos, onde cada questão é acompanhada de sua resolução. Isso permite aos usuários estudarem de forma autônoma, utilizando o aplicativo como uma ferramenta de aprendizado. A utilização da computação móvel para apoiar a educação tem se mostrado atrativa, permitindo que os usuários acessem informações de estudo e conhecimento em qualquer lugar e a qualquer momento.

O trabalho de [Arcanjo Augusto et al. 2021] realizou uma análise comparativa entre as edições de 2014 a 2019 do POSCOMP, para avaliar os egressos da área de Computação, e o Currículo de Referência (CR) da SBC, homologado em 2016. A partir dessa comparação foram identificados que aproximadamente 60% dos conteúdos do CR não foram abordados nos exames do POSCOMP, que também apenas 14 conteúdos apre-

sentaram incidências significativas e consistentes nas edições do exame, e ainda, os eixos de formação estabelecidos pelo CR apresentaram diferenças significativas no número de questões correspondentes exploradas ao longo das edições do exame. Essas descobertas são significativas para os futuros candidatos, visto que não somente apontam os conteúdos mais frequentes nas provas, como também indicam aqueles de menor incidência, indicando novamente que os estudos conduzidos pelos candidatos deve ser não-uniforme, portanto, reforçando aquelas temáticas mais frequentes na prova.

Para além dos trabalhos relacionados diretamente com o POSCOMP, também foram encontrados estudos que se relacionam com o tema de mineração em dados educacionais, os quais abordam a análise desempenho dos estudantes em instituições de ensino, utilizando dados públicos disponíveis em sites governamentais e em bancos de dados das próprias instituições de ensino.

Os autores [Fernando Raguro et al. 2022] em seu trabalho buscaram avaliar o desempenho dos alunos para que levasse a melhora dos cursos, com isso utilizaram técnicas de mineração de dados educacionais, na qual abordam possuir métodos para extrair informações úteis dos desempenhos dos alunos e prever resultados futuros utilizando técnicas de aprendizado de máquina. Já em [Amazona and Hernandez 2019], os autores usaram abordagem de mineração de dados educacional para modelar os desempenhos dos alunos ao aplicar modelos de classificação, tais como: *Naïve Bayes, Decision Tree e Deep Learning in Neural Network*. O trabalho foi desenvolvido utilizando dados do curso de bacharelado em tecnologia da informação realizado em um período de seis semestres.

Os pesquisadores [Ahmed et al. 2020] utilizaram as técnicas de mineração de dados educacionais para analisar e prever o desempenho acadêmico dos alunos de graduação, para propor uma intervenção na melhoria do desempenho. O objetivo da pesquisa dos autores é calcular o desempenho acadêmico dos discentes de graduação usando técnicas de mineração de dados, precisamente, algoritmos de classificação, para o registro de 800 alunos do curso de Ciência da Computação. Para avaliar o desempenho, foram utilizados quatro métodos de seleção de características: algoritmos genéticos, razão de ganho, relevo e ganho de informação, e para algoritmos de classificação: *K-Nearest Neighbor, Naïve Bayes, Bagging, Random Forest e J48 Decision Tree*. Diante disso, os resultados experimentais dos trabalhos mostraram que o método de algoritmos genéticos fornece a melhor acurácia de 91,37% com o classificador KNN.

Outros trabalhos, como os disponíveis em [Carrillo and Parraga-Alava 2018, Islam et al. 2019, Nabil et al. 2021], também preveem a realização de um estudo sobre o desempenho dos alunos a partir da aplicação de técnicas de aprendizado de máquina e mineração de dados em base de dados voltados à área da Computação, com vistas para entender as competências e fraquezas dos alunos e, por fim, permitir que os resultados encontrados contribuam para a elaboração de estratégias educacionais. Neste sentido, estes trabalhos estão em coerência com o desenvolvido neste artigo, pois estão no mesmo guarda-chuva de prover melhorias para as instituições de ensino e todos os entes que fazem parte do processo educacional.

Diante do exposto, esta pesquisa busca analisar as notas finais dos candidatos que prestaram o POSCOMP nos anos de 2016 a 2019, utilizando uma metodologia seme-

lhante ao processo de descoberta do conhecimento (em inglês, *Knowledge-Discovery in Databases – KDD*). Isso permitirá realizar uma análise exploratória dos dados e verificar o desempenho nos conteúdos da área da Computação. A partir desta análise exploratória dos dados, será possível propor uma orientação vocacional que recomendará áreas de pesquisa de acordo com as notas obtidas no POSCOMP.

3. Metodologia de pesquisa

A metodologia deste artigo segue uma versão adaptada do KDD comumente utilizado em cenários que se precisa extrair algum tipo conhecimento pertinente e presente em uma base de dados. De acordo com o artigo [Fayyad et al. 1996], o KDD refere-se a um processo geral de descoberta de conhecimento, compreendido em etapas que vão desde a coleta de dados até a interpretação de conhecimento propriamente dita. Conforme destacado pelos autores em [Silva Guerra et al. 2018], os dados brutos podem ser definidos como fatos, valores documentados ou resultados de medições. Quando esses dados adquirem um sentido ou significado, eles se transformam em informações. E quando essas informações são assimiladas por um agente, tornando-o consciente e capacitado para tomar decisões com base nelas, surge o conhecimento.

Tomando como base o resultado do POSCOMP, contendo os acertos e erros dos candidatos em cada questão aplicada em cada ano (de 2016 a 2019), define-se a metodologia deste trabalho de acordo com a Figura 1.

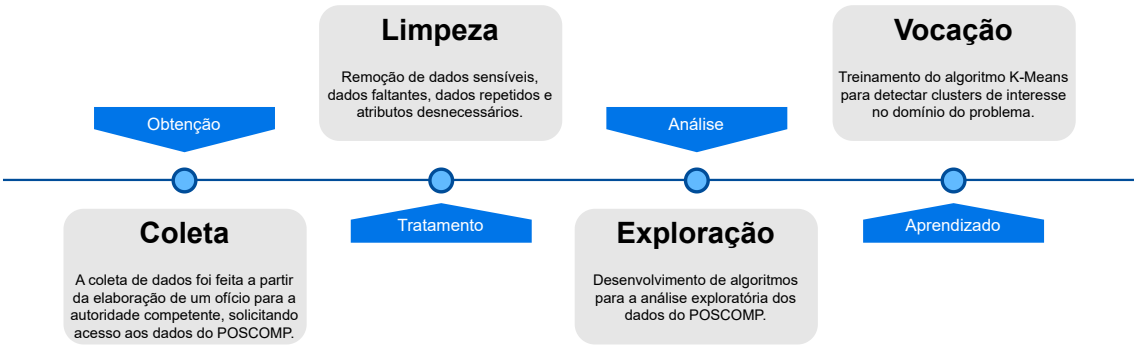

Figura 1. Metodologia da pesquisa aplicada neste artigo.

A Figura 1 divide em etapas o manejamento da base de dados, em princípio, tendo como objetivo realizar não somente a análise exploratória dos dados, mas também desenvolver uma orientação vocacional com base no desempenho que o candidato teve na prova. Sendo assim, destacam-se os seguintes procedimentos:

1. **Obtenção e Coleta**: a base de dados foi enviada pela autoridade competente a partir de um pedido feito em ofício redigido pelos autores do artigo. Os dados recebidos vieram separados por ano, de 2016 a 2019, e protegido por uma credencial de acesso.
2. **Limpeza e Tratamento**: realizado um tratamento de limpeza, remoção dos valores ausentes ou inconsistentes. Da mesma forma, a aplicação de técnicas de

padronização, normalização e transformação foram aplicadas na base de dados. Além disso, dados sensíveis foram removidos por serem irrelevantes para esta pesquisa. Esse processo é fundamental para obter resultados confiáveis e significativos na análise posterior.

3. **Análise e Exploração**: os dados são explorados utilizando técnicas estatísticas e de visualização da informação para identificar padrões, tendências, correlações e características relevantes. Essa análise preliminar auxilia os pesquisadores a entender a base e vislumbrar possíveis técnicas de aprendizado de máquina para serem aplicados à base. Nesta etapa, detectou-se, por exemplo, ser possível aplicar o algoritmo de clusterização para separar os participantes da prova em grupos, i.e., *clusters*, que representam as áreas de interesse.
4. **Aprendizado e Vocação**: feita a seleção de atributos da base de dados que serão utilizadas nesta etapa, o algoritmo de clusterização *K-Means* é aplicado para encontrar os *clusters* propriamente ditos e, com isso, conduzir a orientação vocacional. Foram utilizadas as notas obtidas em cada conteúdo da prova (matemática, fundamentos e tecnologia) de cada candidato para servir de base de treinamento do algoritmo *K-Means*.

Por fim, foi utilizado um conjunto de ferramentas computacionais que desempenharam um papel fundamental para o processo de análise exploratória dos dados e descoberta de conhecimento. A linguagem de programação *Python* foi utilizada para desenvolver os códigos, enquanto as bibliotecas *Pandas* e *NumPy* foram empregadas para realizar a análise dos dados. Além disso, as bibliotecas *Matplotlib* e *Seaborn* foram utilizadas para realizar a visualização da informação. É importante destacar que essas ferramentas são amplamente adotadas em pesquisas nas áreas de mineração de dados e aprendizado de máquina. Em seguida, descrevem-se com mais detalhes sobre a base de dados, o pré-processamento e preparação da base para a aplicação do algoritmo de clusterização.

3.1. Obtenção e tratamento dos dados

Os dados utilizados neste estudo foram coletados das edições do POSCOMP realizadas no período de 2016 a 2019. Devido à natureza limitada e à ausência de acesso público aos dados, foi necessário entrar em contato com a SBC para obter acesso aos dados por meio de ofício e realizar a análise pretendida. A escolha do período dos dados levou em consideração as edições ao longo desses anos, com a Fundação Universidade Empresa de Tecnologia e Ciências (FUNDATEC) assumindo a responsabilidade a partir de 2016.

Para realizar a análise exploratória dos dados desejado e efetivamente construir um modelo computacional que sugere a vocação do candidato, iniciou-se selecionando os atributos considerados relevantes para os objetivos deste artigo. Esses atributos incluem: o estado de origem de cada candidato, as notas obtidas em cada tema aplicado nas provas, o sexo dos candidatos e a sua especialidade, i.e., um campo que o candidato informou no ato da inscrição[2] e a presença dos candidatos na prova. A título de simplificação, foram criadas mais cinco colunas derivadas de outros atributos: “ano da prova”, “região do país”, “matemática”, “fundamentos da computação” e “tecnologia da computação”,

[2]Um mesmo candidato pode indicar mais de uma especialidade

sendo estas três últimas valores numéricos correspondentes as notas nas áreas abordadas pelo POSCOMP. Estes atributos foram extraídos a partir da soma das notas obtidas pelos candidatos em cada tema específico relacionado a área em questão. A Tabela 2 apresenta os temas da prova associados às respectivas áreas. Esta seleção e extração de atributos permite uma análise abrangente acerca do desempenho dos candidatos e seus interesses acadêmicos de pesquisa.

Tabela 2. Áreas do POSCOMP e os respectivos temas.

Áreas	**Matemática**	**Fundamentos da Computação**	**Tecnologia da Computação**
Temas	Álgebra Linear Análise Combinatória Cálculo Dif. e Int. Geometria Analítica Lógica Matemática Matemática Discreta Prob. e Estatística	Análise de Algoritmos Alg. e Est. de Dados Arqt. e Org. de Comp. Circuitos Digitais Ling. de Programação Ling. Formais, Aut. e Comp. Org. de Arq. e Dados Sistemas Operacionais Técnicas de Programação Teoria dos Grafos	Banco de Dados Compiladores Computação Gráfica Engenharia de Software Inteligência Artificial Processamento de Imagens Rede de Computadores Sistemas Distribuído

Seguindo com o tratamento dos dados, realizou-se também a remoção de inscrições duplicadas, o tratamento de valores preenchidos incorretamente, a transformação de dados categóricos em numéricos para posterior extração de medidas estatísticas e, por fim, a remoção de inscrições com muitos dados faltantes, visando preservar a qualidade dos dados para análise.

A especialidade informada pelo candidato no ato da inscrição é um campo onde deve ser indicado um conjunto de áreas computacionais que tem interesse em pesquisar, caso seja aprovado em um programa de pós-graduação. Em particular, este é um atributo que requer uma padronização, visto que, por exemplo, especialidades como “inteligência artificial” e “inteligência computacional” estavam presentes nos dados, sendo necessário unificá-las e renomeá-las para “inteligência artificial”. Além disso, entende-se que a primeira especialidade indicada pelo candidato aponta para a sua maior preferência, portanto apenas ela é considerada nas análises futuras deste artigo.

3.2. Estatísticas descritivas em dados tratados

Após a etapa de tratamento de dados foram aplicadas um conjunto de estatísticas descritivas, separando os candidatos por sexo, para informar o número de inscritos, presentes e ausentes, considerando inclusive filtros por estado. Também são extraídas informações acerca da porcentagem de acertos em cada tema da prova, considerando todos os anos. Por fim, foram reveladas as especialidades mais frequentes no ato da inscrição dos candidatos, seja considerando todo o país, seja por estados, ou ainda, seja por sexo. Essas visualizações são apresentadas na seção de resultados deste trabalho.

3.3. Classificação da vocação dos candidatos via algoritmo *K-Means*

A orientação vocacional inteligente proposta neste artigo consiste em fornecer uma recomendação da linha de pesquisa mais adequada para o candidato, levando em consideração suas notas no exame. A Figura 2 ilustra o passo-a-passo da proposta, desde o recebimento do cartão resposta do candidato contendo as marcações de cada questão até a orientação vocacional. A partir da figura, percebe-se que a nota do candidato perpassa por uma transformação de espaço, saindo de 70 questões binárias de acerto e erro para apenas 3 notas reais, representando as áreas de matemática, fundamentos da computação e tecnologia da computação, conforme Tabela 2.

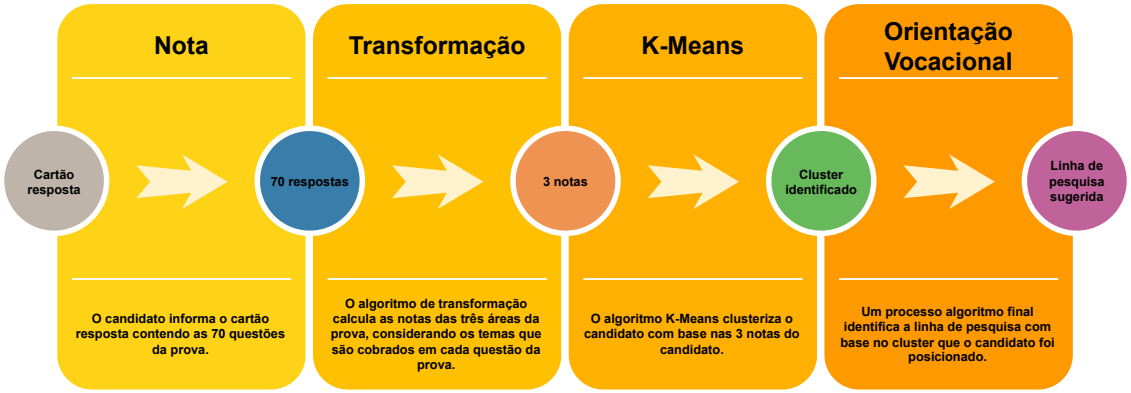

Figura 2. Passo-a-passo da orientação vocacional inteligente a partir da nota do candidato no POSCOMP.

Ainda na Figura 2, nota-se também a etapa de aplicação do algoritmo de clusterização *K-Means* [Faceli et al. 2023]. Para que o passo-a-passo funcione, é necessário construir o modelo computacional de clusterização e isso requer uma base de dados contendo os registros de notas de participantes que prestaram o exame em anos anteriores. Aqui neste artigo, foi utilizada uma base de dados contendo os cartões de resposta de candidatos que fizeram a prova nos anos de 2016 a 2019. O algoritmo de clusterização *K-Means* foi utilizado para encontrar relações de similaridade e dissimilaridade entre esses candidatos. É importante destacar que o algoritmo foi aplicado no espaço tri-dimensional, i.e., considerando a transformação das notas de cada candidato. E ainda, ao aplicar um algoritmo de clusterização em uma base de dados, é necessário determinar o número de *clusters* que se deseja encontrar, i.e., o parâmetro K do algoritmo, e esse é um aspecto crítico na aplicação do algoritmo *K-Means*.

Os métodos AIC e BIC são abordagens comumente utilizadas para selecionar o número de *clusters* em análise de clusterização. Essa abordagem visa encontrar um equilíbrio entre a complexidade do modelo e sua adequação aos dados [Fraley and Raftery 2002]. Neste trabalho, o valor de K foi determinado por meio desses dois métodos que são critérios de seleção de número de *clusters* amplamente utilizados na literatura. Sendo assim, foi possível determinar quantos *clusters* existem na base de dados trabalhada.

Após a etapa de clusterização da base de dados, tornou-se viável determinar a linha de pesquisa associada a cada cluster, i.e., para cada cluster foi necessário identificar a maioria absoluta das especialidades (atributo da base de dados) indicadas pelos candidatos e assumir o cluster com aquele significado, doravante chamado de linha de pesquisa. Essa análise proporcionou uma compreensão mais aprofundada dos interesses e habilidades dos candidatos agrupados.

Posteriormente, já com a identificação de cada cluster, e o passo-a-passo da Figura 2 completo, será possível sugerir uma linha de pesquisa para quaisquer candidatos que realizarem a prova em qualquer ano, bastando apenas indicar para o *pipeline* o cartão resposta contendo as 70 questões respondidas. Essa recomendação de linha de pesquisa tem como objetivo auxiliar o candidato na seleção de programas de pós-graduação em computação que estejam alinhados com suas aptidões e objetivos profissionais.

4. Resultados e discussões

Nesta seção serão apresentados os resultados da análise exploratória dos dados do POSCOMP e a orientação vocacional inteligente com base no algoritmo de clusterização *K-Means*. Inicialmente, será apresentada uma análise exploratória dos anos de 2016 a 2019, extraindo algumas frequências absolutas do número de inscritos, considerando o sexo dos candidatos ao longo dos anos. Também serão discutidas as notas dos candidatos com base nos temas da prova, i.e., outro espaço de características que revela o desempenho dos candidatos na prova. Por fim, uma análise de estados do Brasil é considerada, inclusive identificando as especialidades indicadas pelos participantes.

4.1. Análise exploratória dos dados

Esta análise exploratória possibilita uma compreensão generalizada das características dos candidatos que realizaram o POSCOMP entre os anos de 2016 a 2019. A Figura 3 mostra o número total de candidatos inscritos, por ano, que realizaram a prova, juntamente com aqueles que não se fizeram presentes. De maneira geral, percebe-se uma diminuição do número de inscritos ao longo dos anos, revelando também uma diminuição no número

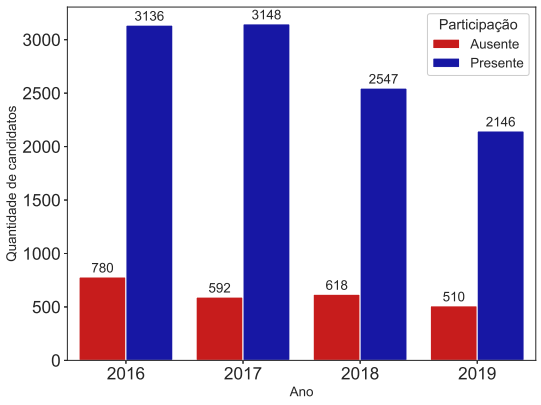

Figura 3. Candidatos que realizaram a prova e os faltantes.

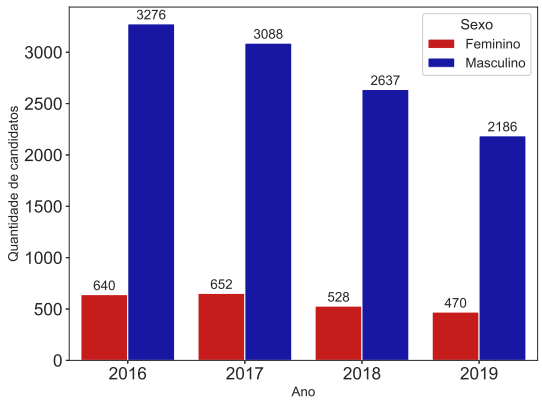

Figura 4. Candidatos inscritos por sexo.

de candidatos presentes na prova, assim como os ausentes. Já na Figura 4 é possível observar de maneira mais clara a distribuição dos candidatos por sexo, em cada ano. Verifica-se que, embora haja uma participação relevante do sexo masculino, o número de inscritos vem diminuindo de 2016 a 2019. Por outro lado, a participação feminina ainda é baixa, entretanto mantém-se relativamente constante ao longo dos anos, evidenciando uma tendência de estabilidade.

As Figuras 5 e 6 mostram o número de participantes presentes e ausentes por sexo, respectivamente, masculino e feminino, durante os anos de 2016 a 2019. Os gráficos mostram que, de maneira geral, o público ausente mantém-se abaixo de 20%, independentemente do sexo do candidato. Além disso, é possível identificar que historicamente o público feminino se faz mais presente na prova do que o público masculino, apesar da baixa representatividade, conforme mostrado anteriormente na Figura 4. Embora não haja estatísticas conhecidas sobre o porquê alguns candidatos faltem ao exame, imagina-se que em algumas situações há a dificuldade de transporte para o local do exame, o custo relacionado ao deslocamento para outra cidade, entre outros fatores que influenciam na desistência em realizar a prova.

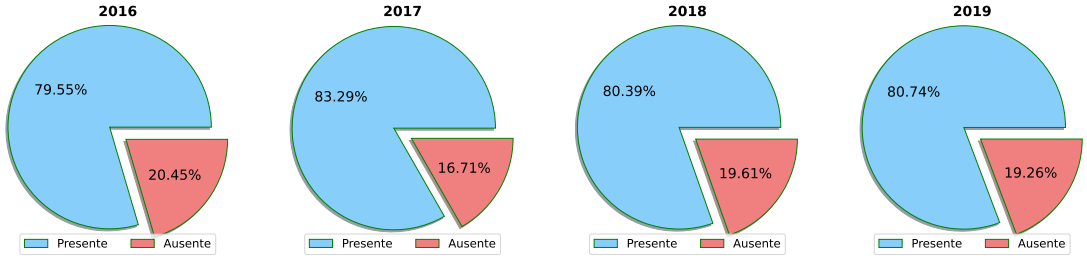

Figura 5. Porcentagens de presentes e ausentes masculinos ao longo dos anos.

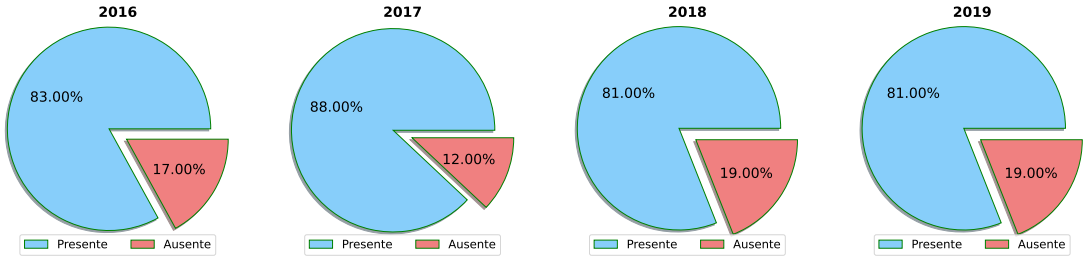

Figura 6. Porcentagens de presentes e ausentes femininos ao longo dos anos.

Observa-se também, ainda nas Figuras 5 e 6, que o ano de 2017 apresentou a maior porcentagem de presentes tanto masculinos quanto femininos. De acordo com a Figura 3, esse ano foi o que obteve também o maior número de candidatos inscritos e, consequentemente, isso pode ter colaborado para se ter um aumento de presentes na prova. Em relação à porcentagem de ausentes, tem-se o ano de 2016 como aquele com maior número de ausentes para o sexo masculino e distribuídos de forma igualitária, os anos de 2018 e 2019 como aqueles com o maior número de ausentes do sexo feminino. Não foram observados na literatura e no noticiário da época os possíveis fatores que possam indicar

a causa dessas porcentagens. Entretanto, as porcentagens reveladas neste artigo podem suscitar novos trabalhos investigativos com vistas a mitigar esse alto índice de ausentes em uma prova que poderá decidir se um candidato entra ou não em um programa de pós-graduação em computação. Ainda que o PPG não use a nota do POSCOMP como fator eliminatório, sabe-se que muitos desses programas ranqueiam os candidatos a receber bolsas de pós-graduação a partir das notas da prova, ou seja, ao não realizar a prova, o candidato nem será ranqueado e, consequentemente, não receberá a bolsa. Portanto, ainda que o inscrito seja aprovado no programa de pós-graduação, o mesmo poderá desistir visto que não será financiado por um órgão de fomento.

A Figura 7 apresenta o número de candidatos inscritos por unidade da federação. Como pode ser visto, os números revelam uma distribuição desigual de candidatos entre os estados brasileiros. Isso ocorre por inúmeros fatores, tais como a qualidade do programa de pós-graduação que é oferecida pela instituição do estado, pela concentração de pós-graduandos, pelas oportunidades inclusive de empregos em pesquisa oferecidos no estado, pela densidade demográfica, pelo maturidade da instituição de ensino, pelo surgimento do PPG, entre outros fatores. Observa-se que os estados de São Paulo, Mato Grosso e Rio Grande do Sul se destacam com maior número de participantes, enquanto os estados de Acre e Rondônia registram menor quantidade de inscritos. Importante mencionar que, de acordo com as informações disponíveis no site [PPGCC/UFAC 2023], da Universidade Federal do Acre (UFAC) teve início em 2018, o que pode impactar no número de participantes do estado. Em relação ao estado de Rondônia, não foi identificado PPGs em Computação, o que pode levar ao número baixo de participantes no POSCOMP.

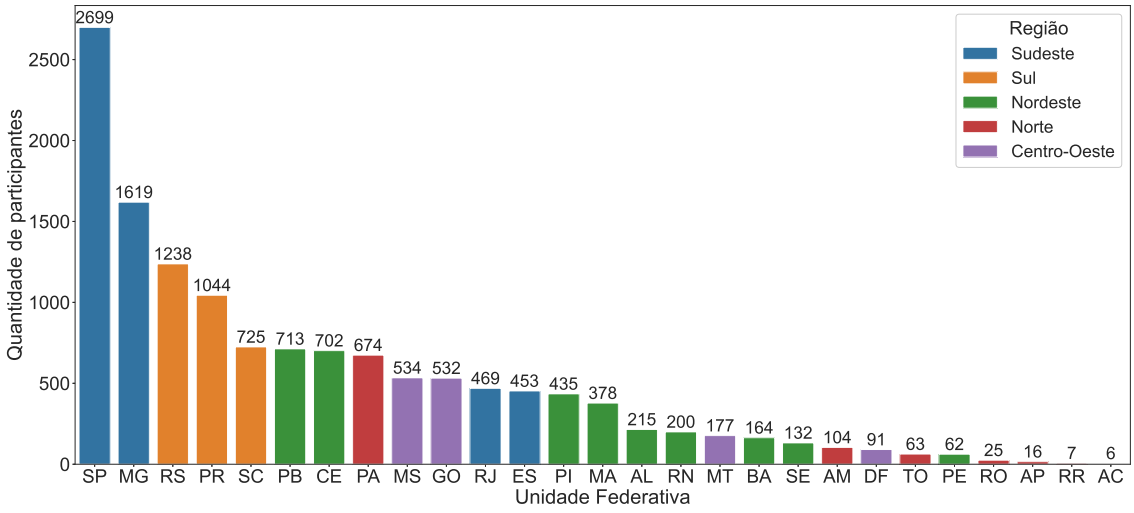

Figura 7. Total absoluto de inscritos por estados do Brasil de 2016 a 2019.

Com o objetivo de mostrar a distribuição de inscritos por região do Brasil com mais distinção, ilustra-se na Figura 8 a somatória de inscritos com destaques para as regiões sudeste, nordeste e centro-oeste, pois respectivamente encontram-se no primeiro, segundo e terceiro lugar. Esta distribuição de inscritos segue, inclusive, parcialmente a distribuição de população do Brasil, onde o sudeste e o nordeste são, respectivamente, as

duas regiões mais populosas do país [Brasil 2023]. Figurando no último lugar em número de inscritos, encontra-se a região sul que contém menos estados e não necessariamente usa o POSCOMP como requisito parcial para adentrar a uma pós-graduação em computação.

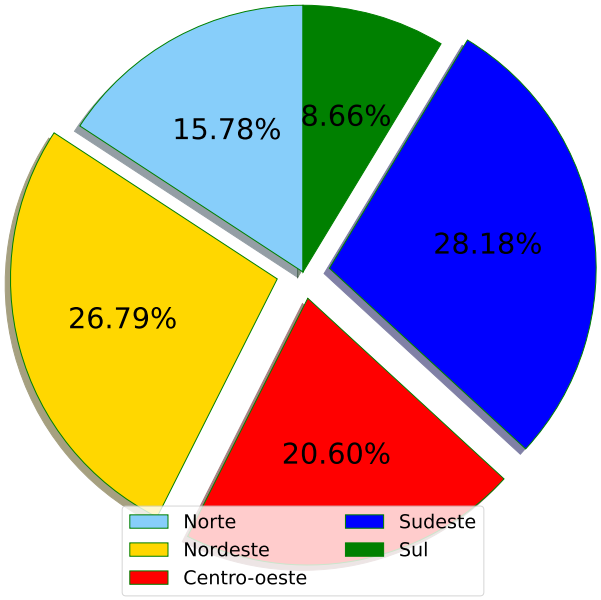

Figura 8. Porcentagens de inscritos por região do Brasil de 2016 a 2019, considerando um total absoluto de 13.477 candidatos.

A Figura 9 informa o número de inscritos discriminando o sexo informado no cadastro. Pode-se destacar que a participação feminina acompanha o panorama nacional anteriormente discutido e carece ainda de uma participação maior em números absolutos, apesar dos dados terem revelados que as mulheres faltam menos ao exame. Como pode ser visto no gráfico o estado de São Paulo é o que possui maior representatividade feminina de inscritos no país, enquanto que os estados do Acre, Paraná e Amapá possuem a menor representatividade em termos percentuais.

Esses dados ressaltam a importância de investigar os fatores que podem influenciar a participação de candidatos de diferentes sexos em exames do POSCOMP. Além disso, destaca-se a necessidade de promover a igualdade e incentivar a participação feminina na área da computação em todas as regiões do país. Sendo assim, percebe-se a importância dos eventos nacionais promovidos pela SBC para as mulheres na área de tecnologia da informação, especialmente o evento WIT (*Women in Information Technology*), que tem como objetivo reunir mulheres pesquisadoras de todo o Brasil, ressaltado no site Meninas Digitais da SBC [Digitais 2023].

Com o intuito de identificar mais profundamente o desempenho e o interesse dos inscritos presentes no POSCOMP ao longo dos anos, as figuras seguintes permitem capturar a qualidade das notas, as especialidades mais frequentes, considerando todo o país e por estado, além de classificar os interesses por sexo. Esses gráficos evidenciam, por

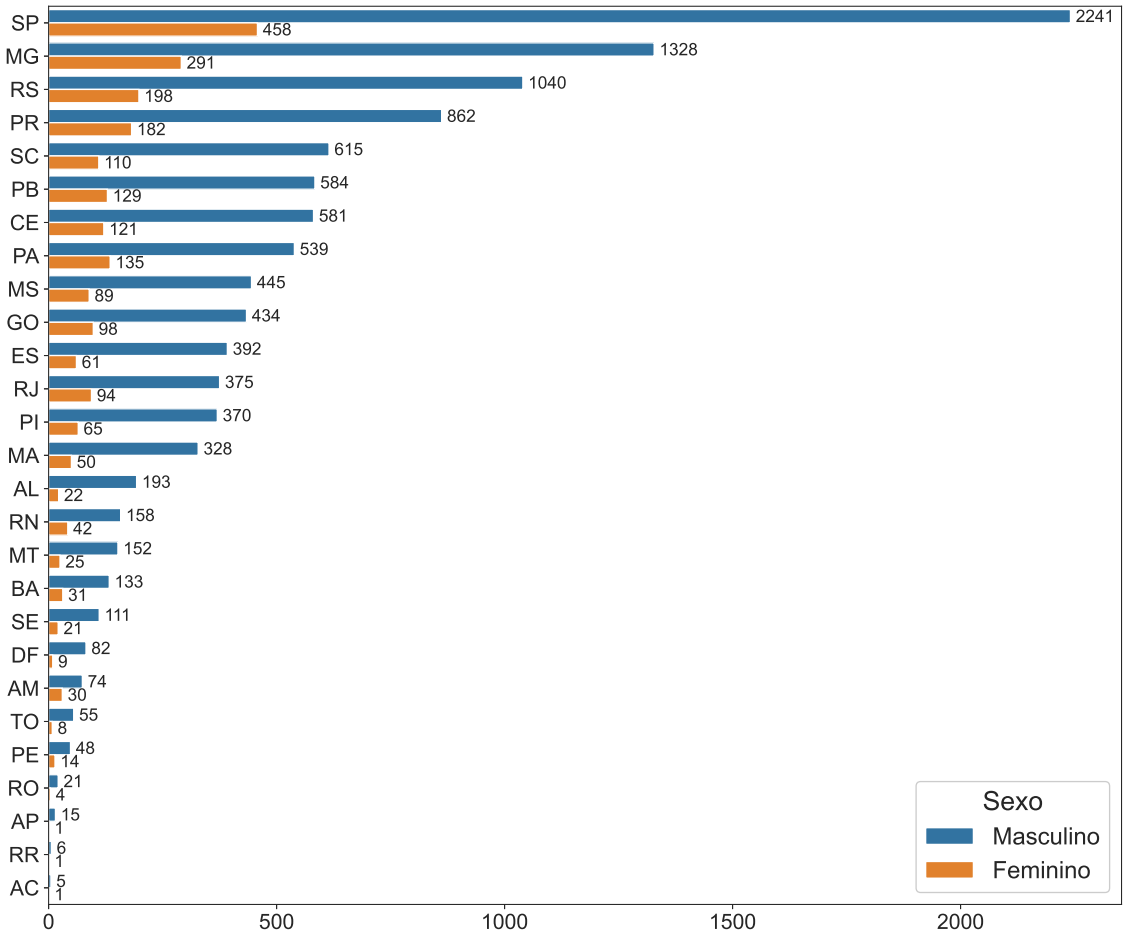

Figura 9. Total de inscritos por sexo em cada estado do Brasil.

exemplo, o conhecimento adquirido nas graduações em computação e podem apontar para melhorias do projeto pedagógico dos cursos, tanto da pós-graduação, mas especialmente da graduação.

O gráfico de barras horizontais da Figura 10 fornece a média de acertos, em porcentagem, dos candidatos em cada tema de estudo aplicado na prova. É importante ressaltar que cada questão da prova requer um conhecimento particular sobre um tema comumente associado a área da computação, portanto é possível, para cada questão, classificá-la em um tema particular. Sabendo do total de questões de um tema, pode-se derivar a porcentagem de acerto por tema. Desta forma, ao detectar que o tema "Técnicas de Programação" possui maior porcentagem, sabe-se que, em média, os candidatos acertaram 73.21% das questões com esta temática. Por outro lado, o tema com menor média de acertos é "Computação gráfica", indicando que a temática precisa de uma atenção maior por parte das instituições de ensino de computação. E de maneira geral, o gráfico revela também que poucos temas possuem média superior a 50%, i.e., são em apenas 6 temas que os candidatos acertam mais da metade das questões. Observando estes temas em particular, nota-se que eles preservam relação direta com programação e estrutura de dados, além

Figura 10. Média geral, em porcentagem, dos temas aplicados no POSCOMOP.

de ser corroborado por técnicas de estimar a complexidade computacional dos algoritmos. Estes temas são normalmente abordados desde o ensino básico, com a realização de projetos e eventos que incentivem os alunos a programar, assim como durante a graduação, por meio de disciplinas que aprimorem habilidades e aprofundem conhecimentos em algoritmos. Portanto, considera-se como um destaque positivo do perfil dos candidatos a pós-graduação em computação no Brasil.

Seguindo com a análise exploratória dos dados, sabe-se que no ato da inscrição, cada candidato insere, em um campo no formulário, as especialidades de seu interesse. Com base nessa informação, realizou-se um processamento de texto para construir uma nuvem de palavras contendo as frequências das especialidades mais mencionadas como interesse dos participantes, a qual pode ser observada na nuvem de palavras da Figura 11. As palavras simples ou compostas maiores representam aquelas com maior citação por parte dos candidatos. Observando o resultado que a figura ilustra, nota-se que as especialidade mais frequente indicadas pelos candidatos são, nesta ordem: inteligência artificial, engenharia de software, sistemas de computação e sistemas de informação. Um fator importante a se destacar é que a inteligência artificial é um dos temas em que os candidatos acertam, em média, acima de 50% da prova (Figura 10), enquanto que na

Figura 11. Nuvem de palavras contendo as especialidades mais frequentes informadas pelos candidatos no ato da inscrição.

engenharia de software, os candidatos apresentam maiores dificuldades, não acertando nem 36% da prova.

Ao observar o interesse por estado, a Figura 12 revela que há uma diversidade de especialidades nas buscas pelos candidatos em cada estado, o que ressalta amplitude e abrangência da área de estudo, evidenciando novamente a predominância da inteligência artificial nos estados de São Paulo, Minas Gerais e Paraná.

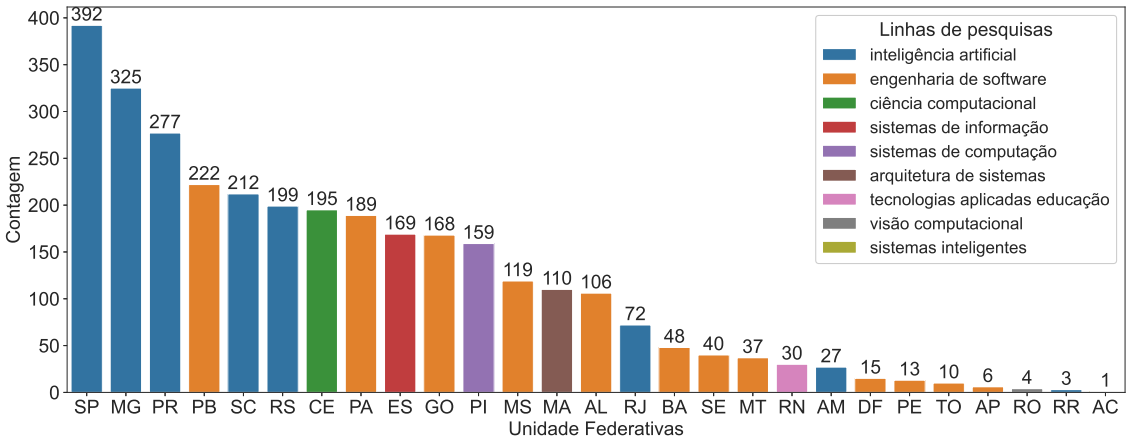

Figura 12. Especialidade mais frequentes por estado.

Os resultados evidenciam que algumas especialidades têm maior domínio em determinados estados. Por exemplo, em Alagoas, a área de engenharia de software foi a mais requisitada, enquanto no Espírito Santo, sistemas de informação teve maior representatividade. Já no Mato Grosso do Sul, engenharia de software foi a especialidade mais

frequente, assim como o estado do Pará. Essas descobertas fornecem informações importantes para a orientação vocacional e o planejamento das instituições de ensino, bem como auxiliar na tomada de decisões relacionadas à oferta de cursos e especializações em cada estado.

Por fim, a partir dos resultados encontrados anteriormente, as Figuras 13 e 14 mostra as especialidades indicadas por cada sexo. Por exemplo, na linha de pesquisa de inteligência artificial, há uma maior representação masculina, enquanto que em engenharia de software a presença feminina é mais expressiva. Essas informações são relevantes para compreender a participação de cada sexo nas diferentes áreas de estudos dentro dos cursos de Computação.

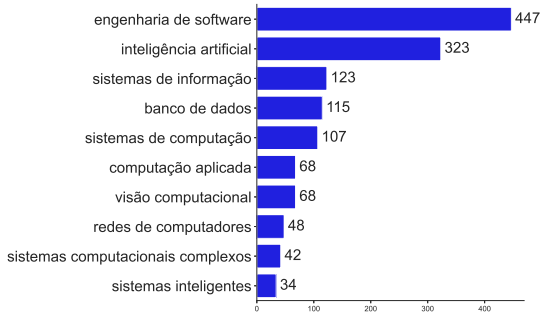

Figura 13. Especialidades mais comuns entre participantes femininos.

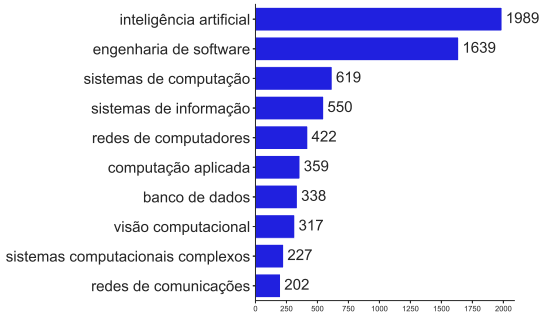

Figura 14. Especialidades mais comuns entre participantes masculinos.

Tendo como base as especialidades informadas pelos candidatos no ato da inscrição, as seções seguintes explicam como construiu-se a orientação vocacional para que, com base na nota obtida no POSCOMP, o candidato recebe uma orientação para seguir uma linha de pesquisa na pós-graduação e assim aumentar as suas chances de sucesso. A motivação fundamental para construir esse modelo computacional de aprendizado de máquina vem da observação de que é comum encontrar candidatos que indicam áreas descorrelacionadas ou distantes do ponto de vista científico, por exemplo, há candidatos que apontam as especialidades de engenharia de software e redes de computadores, indicando talvez que o candidato ainda não sabe ou não decidiu sua linha de pesquisa e, portanto, pode acabar escolhendo um seguimento que não seja sua vocação. Neste sentido, a orientação vocacional pode auxiliar o participante a tomar a decisão de maneira mais embasada, i.e., com base na nota obtida na prova. Por óbvio, todo aprendizado de máquina está sujeito a uma determinada acurácia e que, portanto, pode conter erros. Então sugere-se que os candidatos que fizerem uso do modelo construído neste artigo tenham somente a finalidade de obter um direcionamento preliminar e que faça sua escolha levando em considerações as questões regionais, estaduais e locais de sua realidade.

4.2. Identificação do número de *clusters*

O algoritmo de aprendizado de máquina *K-Means* é amplamente utilizado na literatura de inteligência artificial, pois permite identificar *clusters* em base de dados. Um cluster

acomoda todas as amostras que possuem similaridade entre si, ao passo que amostras que pertencem a *clusters* diferentes devem relevar disimilaridades. Sua aplicação é amplamente difundida, permitindo a identificação de padrões de desempenho e características dos grupos de candidatos [Marbouti et al. 2021, Awat and Ballera 2018]. Além disso, o *K-Means* oferece eficiência computacional, o que é essencial ao lidar com conjuntos de dados extensos, como é o caso do POSCOMP.

O algoritmo *K-Means* foi aplicado no espaço de características contendo três dimensões que representam as áreas do POSCOMP: matemática, fundamentos da computação e tecnologia da computacao. Esses atributos representam as somas das áreas dos eixos temáticos aplicadas no POSCOMP, como mencionado anteriormente.

Para determinar o número de *clusters* presentes na base de dados foram empregados os métodos *Akaike Information Criterion* (AIC) e *Bayes Information Criterion* (BIC), que são critérios de seleção e auxiliam na identificação do número ideal de *clusters* [Fraley and Raftery 2002]. Para determinar o número de *clusters* é necessário executar o algoritmo de clusterização para um intervalo de número de *clusters*. Aqui neste trabalho o intervalo foi configurado entre 1 e 10 *clusters*. A cada modelo construído é calculado os valores (i.e., *scores*) do AIC e BIC obtidos a partir da disposição dos centróides de cada cluster. O número ideal de *clusters* é indicado pelo menor *score* obtido em cada critério de seleção. Como pode ser visto na Figura 15, o número de *clusters* ideal indicado pelo método BIC é 2 e pelo método AIC é 6. Entretanto, é possível perceber que a partir de 2 *clusters* não se vê uma diminuição significativa no *score* do critério AIC, mostrando pouca punição ao aumentar o número de *clusters*. Neste sentido, consideração uma boa relação de compromisso entre complexidade do modelo computacional e o domínio do problema, optou-se por considerar o número 2 como sendo o ideal de *clusters* presentes na base de dados do POSCOMP.

Sabendo o número de *clusters* presentes na base de dados, o próximo passo consistiu em construir o modelo computacional de clusterização responsável por realizar indicar a vocação dos candidatos do POSCOMP. Com vistas a reprodutibilidade deste trabalho, destacam-se a seguir a configuração dos principais hiperparâmetros do algoritmo *K-Means* utilizados neste trabalho:

- **K = 2**: define o número de *clusters* desejados a partir dos critérios de seleção.
- **init = "random"**: os centróides dos *clusters* são inicializados aleatoriamente a partir de uma distribuição uniforme considerando todo o espaço de característica.
- **T = 100**: define o número máximo de iterações permitidas para a convergência do algoritmo.

4.3. Características dos *clusters* identificados

Os atributos de matemática, fundamentos da computação e tecnologia da computação foram extraídos dos resultados dos candidatos. No contexto da análise com o algoritmo *K-Means*, esses valores de notas foram utilizados como atributos, possibilitando a identificação e formação de *clusters* de candidatos com interesses similares em relação às linhas de pesquisa desejadas. A Figura 16 apresenta a divisão dos *clusters* obtidos após a aplicação do algoritmo *K-Means*. Os *clusters* foram identificados da seguinte forma:

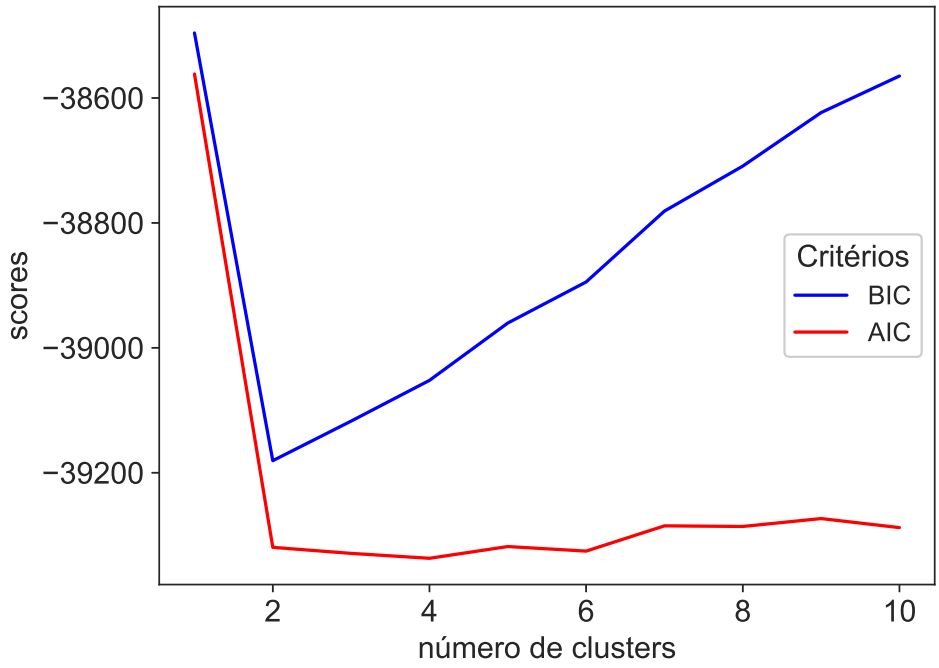

Figura 15. Análise de critérios de informação para determinar o número de *clusters* no *K-means*

o cluster 0 foi associado à cor azul, enquanto o cluster 1 foi associado à cor vermelha. Esses *clusters* refletem as notas obtidas pelos candidatos em cada área da computação e indicam a qual cluster o candidato pertence, com base em sua nota.

No primeiro momento, percebe-se que, em sua maioria, os *clusters* representam candidatos que obtiverem uma melhor nota em relação ao outro. Para visualizar esta projeção de maneira mais objetiva, a Figura 17 informa o desempenho dos candidatos em cada área dentro de cada cluster, exibindo os histogramas com o desempenho das notas dos candidatos em todo o Brasil. Embora não haja uma distinção absoluta entre os histogramas dos *clusters*, há ligeiro desvio do cluster 0 para a direita, indicando que esses candidatos obtiveram um melhor desempenho no POSCOMP, como já era esperado devido ao resultado da clusterização.

A título de aprofundamento nas análises dos candidatos presentes em cada *clusters*, a Figura 18 mostra o desempenho médio dos candidatos em relação aos temas abordados nas provas ao longo dos anos. Através dessa visualização, é possível observar que, em alguns temas de estudo, e.g., “Matemática Discreta”, “Geometria Analítica”, “Algoritmos e Estrutura de Dados”, “Linguagens de Programação”, “Redes de Computadores”, “Computação Gráfica” e “Banco de Dados”, os candidatos apresentaram um desempenho crescente ao longo dos anos. Aqui é importante destacar que “Computação Gráfico” é o tema com menor média de acertos por candidato (Figura 10, mas aqui evidencia-se positivamente que o tema está melhorando a cada ano. No entanto, na grande maioria dos temas, os candidatos tiveram um desempenho baixo independentemente do cluster em que eles foram acomodados, destacando portanto a necessidade de uma análise mais

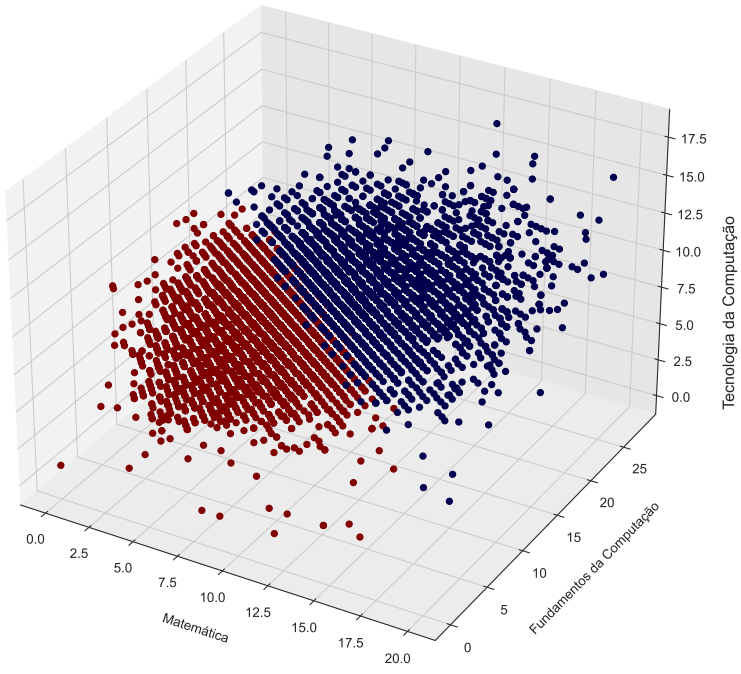

Figura 16. Resultado da aplicação da técnica *K-Means* com a identificação de 2 *clusters*.

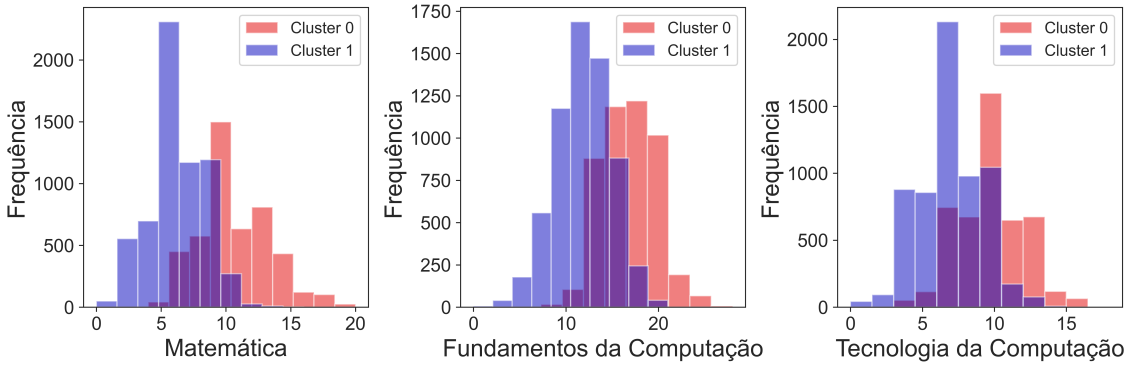

Figura 17. Distribuição de Desempenho por Cluster.

aprofundada sobre as possíveis causas desse cenário. Por fim, as curvas do cluster 0 estão acima das curvas do cluster 1, o que já era esperado dado o fato de que os candidatos situados no cluster 0 obtiveram uma nota melhor do que aqueles do cluster 1.

4.4. Representação significativa dos *clusters*

Observa-se a partir da Figura 19 a especialidade indicada por todos os candidatos que fazem parte do cluster 0, i.e., aquele cluster onde foram identificados os participantes com maior nota. Percebe-se que esses candidatos demonstram grande interesse em realizar

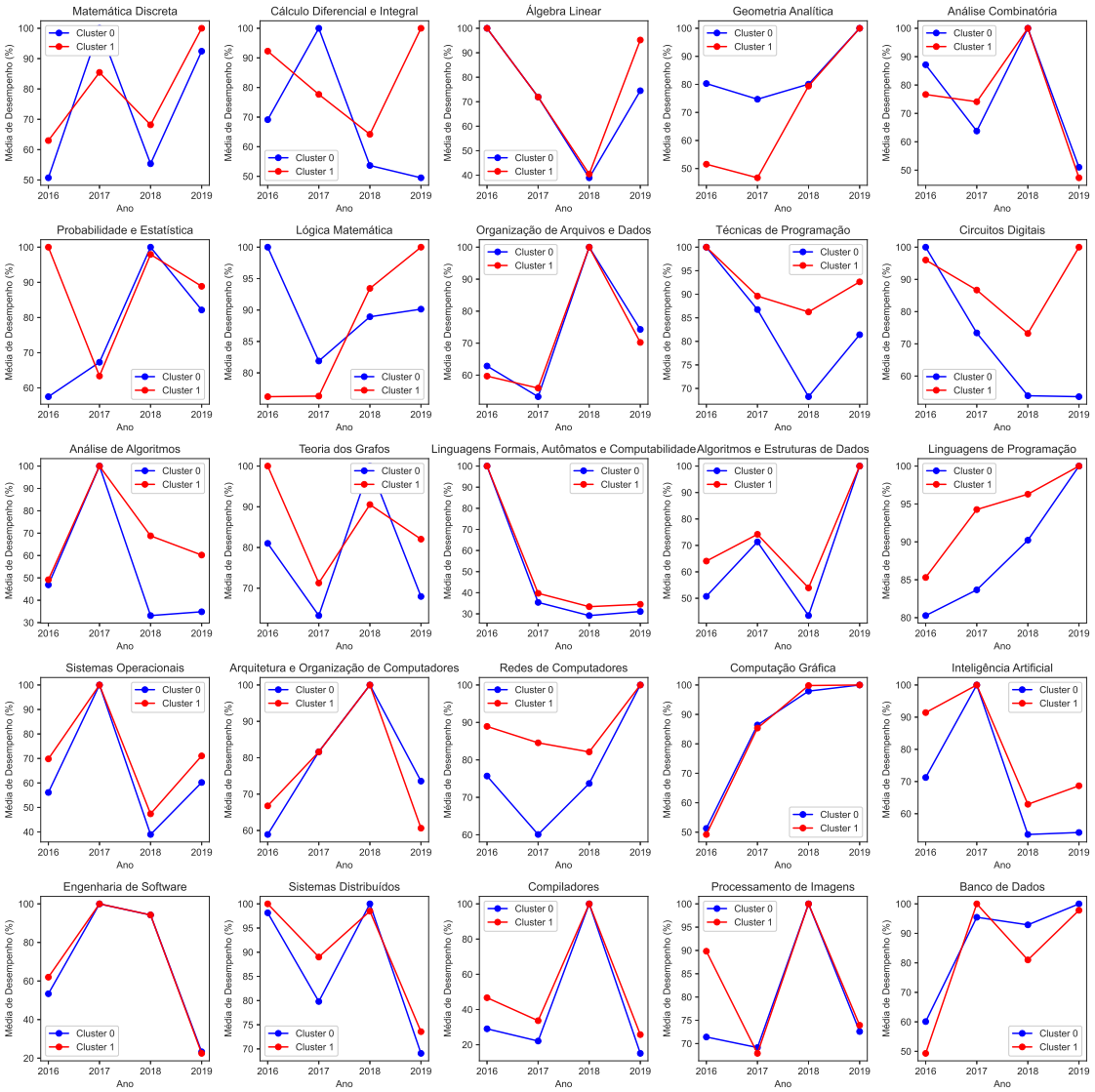

Figura 18. Desempenho dos candidatos por tema e por ano, considerando a divisão entre *clusters*.

pesquisas na linha de pesquisa de inteligência artificial, seguido de engenharia de software e sistemas de informação.

Já em relação ao cluster 1, i.e., aqueles que obtiveram menores notas em relação ao cluster 0, observa-se que os candidatos têm um forte interesse em estudos relacionados à engenharia de software, como ilustra na 20. Isso pode ser devido à crescente demanda por profissionais qualificados nessa área, bem como às oportunidades de carreira e inovação tecnológica que a engenharia de software oferece.

A partir da frequência absoluta das especialidades indicadas pelos próprios candidatos foi possível identificar o significado de cada cluster, definindo-se da seguintes forma: cluster 0 associado a linha de pesquisa de inteligência artificial e cluster 1 rela-

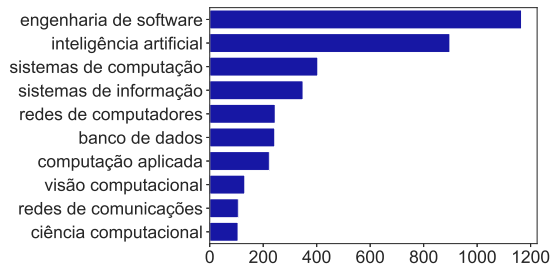

Figura 19. Áreas de interesse do cluster 0.

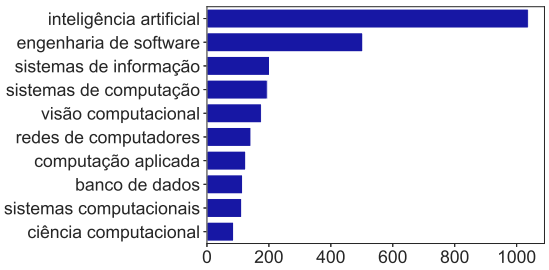

Figura 20. Áreas de interesse do cluster 1.

cionado à engenharia de software. Portanto, de posse de uma nova nota do POSCOMP, i.e., o cartão resposta do candidato, será sempre possível seguir o *pipeline* descrito na anteriormente na Figura 2 e, assim, sugerir uma linha de pesquisa para o candidato baseado na transformação do cartão resposta, onde contém as notas temáticas, para as três áreas cobradas na prova. A classificação de um novo candidato está baseada no fato de que também será sempre possível definir a qual cluster o candidato pertence, 0 ou 1, calculando a menor distância Euclidiana do novo vetor tridimensional, i.e., a nota do candidato, para os centródides dos dois *clusters*. Os vetores tridimensionais dos centróides encontrados são descritos na Figura 3 e poderão ser utilizados pela comunidade científica:

Tabela 3. Valores de Centroides dos Atributos por Cluster no Algoritmo K-means

Cluster	Matemática	Fundamentos da Computação	Tecnologia da Computação
0	0.3057	0.4226	0.3720
1	0.5264	0.5999	0.5246

5. Considerações Finais

O objetivo deste trabalho consistiu em realizar uma análise exploratória dos microdados do POSCOMP nos anos de 2016 a 2019, a fim de minerar padrões relevantes a partir do desempenho dos candidatos que realizaram as provas ao longo dos anos. Foi objeto de estudo também relevar os principais interesses informados pelos próprios participantes em relação às especialidades oferecidas pelos Programas de Pós-Graduação na área de Ciência da Computação do país. Após análise exploratória, aplicou-se o algoritmo de aprendizado de máquina chamado *K-Means* para identificar *clusters* que preservassem a similaridade intra-cluster e dissimilaridade inter-cluster, considerando novamente os desempenhos dos participantes do exame para que assim fosse possível sugeri-lo uma linha de pesquisa.

A análise exploratória dos dados revelou que em todas os anos que a prova foi aplicada, o público masculino apresentou uma participação maior do que o público feminino. Esse padrão indica uma disparidade de gênero, infelizmente já esperada para a área de computação no Brasil, sugerindo que ainda existem muitos desafios a serem enfrentados

pelas gestores acadêmicos, sociedade civil e políticas públicas para trazer mais mulheres para a área da computação. Essa questão é relevante para a compreensão das dinâmicas de gênero na educação e na formação acadêmica, assim como à promoção da igualdade de oportunidades e a busca por uma maior representatividade feminina em áreas tradicionalmente dominadas por homens. A necessidade de incentivar e apoiar a participação das mulheres em áreas de Ciência, Tecnologia, Engenharia e Matemática (STEM) torna-se evidente, buscando ampliar as oportunidades e quebrar preconceitos de gênero que ainda persistem na sociedade. É fundamental que instituições educacionais e pesquisadores continuem investigando e buscando soluções para diminuir essa disparidade de gênero, proporcionando um ambiente mais inclusivo e igualitário. A promoção de programas de orientação vocacional, como o apresentado nesta pesquisa, o combate aos estereótipos de gênero e a criação de espaços de diálogo e apoio são alguns caminhos que podem ser explorados.

A orientação vocacional ao candidato que realizou a prova foi realizada a partir dos resultados pelo algoritmo *K-Means*, onde é possível sugerir ao concluinte de graduação uma das linhas de pesquisa comumente seguidas nas pós-graduações em computação. Essa orientação pode ser útil para auxiliar a tomada de decisão por parte dos candidatos em relação à escolha de sua área de estudo durante o curso de pós-graduação. Novamente, os modelos computacionais estão sujeitos a erros. Neste sentido, sugere-se que os candidatos que fizerem uso do modelo construído neste trabalho tenham apenas a intenção de obter um direcionamento preliminar e que faça sua escolha final levando em considerações também outros fatores relevantes para a sua carreira acadêmica. Vale destacar que as áreas designadas neste trabalho, i.e., inteligência artificial e engenharia de software, são altamente valorizadas no mercado, e os candidatos que buscam atuar nessas áreas podem encontrar diversas oportunidades de emprego como desenvolvedores, cientistas de dados, analistas, entre outras profissões.

Essa pesquisa também contribui significativamente para o desenvolvimento dos estudantes de graduação, fornecendo uma ferramenta útil para identificar as áreas em que possuem maior vocação e considerá-la como linha de pesquisa na pós-graduação. Além disso, os resultados têm grande relevância para gestores e professores, pois permitem identificar os discentes com aptidão em áreas específicas do conhecimento.

Referências

[Ahmed et al. 2020] Ahmed, M. R., Tahid, S. T. I., Mitu, N. A., Kundu, P., and Yeasmin, S. (2020). A comprehensive analysis on undergraduate student academic performance using feature selection techniques on classification algorithms. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6.

[Amazona and Hernandez 2019] Amazona, M. V. and Hernandez, A. A. (2019). Modelling student performance using data mining techniques: Inputs for academic program development. In *Proceedings of the 2019 5th International Conference on Computing and Data Engineering*, ICCDE' 19, page 36–40, New York, NY, USA. Association for Computing Machinery.

[Arcanjo Augusto et al. 2021] Arcanjo Augusto, A. L., Viana, B. M. d. F., Oliveira, D. E. C., Bezerra, G. A., Neto, J. M. L., Maciel, K. W. d. S., Ferraz, P. C., Silva, R. S., and da Costa, U. S. (2021). Comparação dos conteúdos do poscomp com o currículo de referência dos cursos de computação da sbc. *Revista ComInG - Communications and Innovations Gazette*, 5(3):14–23.

[Awat and Ballera 2018] Awat, K. A. S. and Ballera, M. A. (2018). Applying k-means clustering on questionnaires item bank to improve students' academic performance. In *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM)*, pages 1–6.

[Batista et al. 2014] Batista, E. J. S., dos Santos Batista, J., de Souza, R. S., and Primo, W. M. (2014). 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 *Workshops do Congresso Brasileiro de Informática na Educação*. Sociedade Brasileira de Computação - SBC.

[Brasil 2023] Brasil, C. I. d. (2023). População do brasil passa de 203 milhões, mostra censo 2022.

[Carrillo and Parraga-Alava 2018] Carrillo, J. M. and Parraga-Alava, J. (2018). How predicting the academic success of students of the ESPAM MFL?: A preliminary decision trees based study. In *2018 IEEE Third Ecuador Technical Chapters Meeting (ETCM)*, pages 1–6.

[Carvalho et al. 2021] Carvalho, H. P. d., Soares, M. V., Carvalho, S. M. d. L., and Telles, T. C. K. (2021). O professor e o ensino remoto: tecnologias e metodologias ativas na sala de aula. *Revista Educação Pública*, 21(28).

[Connell et al. 1994] Connell, H., Lowe, N., Skilbeck, M., and Tait, K. (1994). *2. The Vocational Quest: New Directions in Education and Training*.

[Digitais 2023] Digitais, M. (2023). Women in information technology.

[Droescher and Silva 2014] Droescher, F. D. and Silva, E. L. d. (2014). O pesquisador e a produção científica. *Perspectivas em Ciência da Informação*, 19(1):170–189.

[Faceli et al. 2023] Faceli, K., Lorena, A. C., Gama, J., de Almeida, T. A., and de Leon Ferreira de Carvalho, A. C. P. (2023). *Inteligência artificial : uma abordagem de aprendizado de máquina*. LTC, Rio de Janeiro, second edition.

[Fayyad et al. 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37.

[Fernando Raguro et al. 2022] Fernando Raguro, M., Carpio Lagman, A., P. Abad, L., and S. Ong, P. L. (2022). Extraction of lms student engagement and behavioral patterns in online education using decision tree and k-means algorithm. In *2022 4th Asia Pacific Information Technology Conference*, APIT 2022, page 138–143, New York, NY, USA. Association for Computing Machinery.

[Fraley and Raftery 2002] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.

[Hui et al. 2020] Hui, H., Ming-jie, T., Qing-tao, Z., and Xiao-liang, Z. (2020). Application of student achievement analysis based on apriori algorithm. In *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, pages 19–22.

[Imdad et al. 2017] Imdad, U., Ahmad, W., Asif, M., and Ishtiaq, A. (2017). Classification of students results using knn and ann. In *2017 13th International Conference on Emerging Technologies (ICET)*, pages 1–6.

[Islam et al. 2019] Islam, R., Sazid, M. T., Mahmud, S. R., Ferdous, C. N., Reza, R., and Hossain, S. A. (2019). Parametric study of student learning in it using data mining to improve academic performance. In *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pages 286–290.

[Junior and Brancher 2014] Junior, F. D. S. and Brancher, J. (2014). Uma pesquisa de opinião sobre a relevância dos conteúdos abrangidos pelo poscomp. In *Anais do XXII Workshop sobre Educação em Computação*, pages 90–99, Porto Alegre, RS, Brasil. SBC.

[Lloyd 1982] Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Trans. Inf. Theory*, 28(2):129–136.

[MacQueen 1967] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J., editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

[Marbouti et al. 2021] Marbouti, F., Ulas, J., and Wang, C.-H. (2021). Academic and demographic cluster analysis of engineering student success. *IEEE Transactions on Education*, 64(3):261–266.

[Marcelo Mendes et al. 2018] Marcelo Mendes, F., P. Mendonça, A., and B. Guedes, E. (2018). Poscomp coach: Plataforma web para apoio ao ingresso na pós-graduação em computação. *Revista Novas Tecnologias na Educação*, 16(1).

[Melo-Silva et al. 2004] Melo-Silva, L. L., Lassance, M. C. P., and Soares, D. H. P. (2004). 1. a orientação profissional no contexto da educação e trabalho.

[Moura et al. 2012] Moura, N., Gordiano, R. S., Silva, R. K. J., and Santos, S. S. (2012). Poscomp e a importância da pós-graduação para aprimoramento profissional. *Centro Federal de Educação Tecnológica de Minas Gerais*.

[Nabil et al. 2021] Nabil, A., Seyam, M., and Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access*, 9:140731–140746.

[PPGCC/UFAC 2023] PPGCC/UFAC (2023). Programa de pós-graduação em ciência da computação.

[Ribeiro and Uvaldo 2007] Ribeiro, M. A. and Uvaldo, M. d. C. C. (2007). Frank parsons: Trajetória do pioneiro da orientação vocacional, profissional e de carreira. *Revista Brasileira de Orientação Profissional*, 8(1):19–31.

[Russell and Norvig 2022] Russell, S. and Norvig, P. (2022). *Inteligência Artificial - Uma Abordagem Moderna*. GEN LTC.

[Silva Guerra et al. 2018] Silva Guerra, M., Asseiss Neto, H., and Azevedo Oliveira, S. (2018). A case study of applying the classification task for students' performance prediction. *IEEE Latin America Transactions*, 16(1):172–177.

[Silveira et al. 2021] Silveira, M. E. R. d., Boll, D. C. C., Matozinho, F. L., Negrizoli, I. F., Vanzin, L., Camara, M. K., Ismael, M. N., Alcantara, R. A. d. S., and Oyamada, M. S. (2021). Classificação por matérias das questões do enade e poscomp. *Revista ComInG - Communications and Innovations Gazette*, 5(2):9–19.

[Sociedade Brasileira de Computação 2022] Sociedade Brasileira de Computação (2022). Exame nacional para ingresso na pós-graduação em computação (poscomp).