Descobrindo Perfis Interseccionais de Mulheres na Computação: Uma Análise Baseada em Clusterização K-Means

Anna Beatriz Marques
beatriz.marques@ufc.br
Universidade Federal do Ceará (UFC)
Russas, Ceará, Brasil

Valéria Maria da S. Pinheiro
valeria.pinheiro@ufc.br
Universidade Federal do Ceará (UFC)
Russas, Ceará, Brasil

Maria Elanne M. Rodrigues
elannemendes@alu.ufc.br
Universidade Federal do Ceará (UFC)
Fortaleza, Ceará, Brasil

Abstract

Women's participation in Computing remains marked by structural inequalities that intensify when race, territory, and other social dimensions are considered together. Although the literature discusses intersectionality in this field, most research remains theoretical or relies solely on descriptive statistics, with a scarcity of empirical studies that employ quantitative methods to identify intersectional profiles that account for the overlap of intersectional dimensions. This study applies clustering techniques to institutional data on students enrolled in Computing courses at a Brazilian public university, adopting a two-step approach: (i) clustering of the general population and (ii) clustering restricted to women students. The results show that aggregate analysis is insufficient to reveal minority female profiles due to the severe population imbalance, which reproduces dominant patterns. In contrast, clustering restricted to women identifies three distinct intersectional clusters, marked especially by differences in race/ethnicity and academic pathways. These findings highlight the potential of quantitative data mining methods to reveal intersectional dynamics that are often invisible in traditional analyses, providing support for equity policies that are more sensitive to the specificities of women students in Computing.

CCS Concepts

• **Social and professional topics → User characteristics**; • **Theory of computation → Unsupervised learning and clustering**.

Keywords

Interseccionalidade, Mulheres, Clusterização, Computação

ACM Reference Format:

Anna Beatriz Marques, Valéria Maria da S. Pinheiro, and Maria Elanne M. Rodrigues. 2026. Descobrindo Perfis Interseccionais de Mulheres na Computação: Uma Análise Baseada em Clusterização K-Means. In *Proceedings of 17a Edição do Computer on the Beach (XVII Computer on the Beach).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introdução

A participação de mulheres nos cursos de Computação tem crescido no Brasil, mas ainda permanece significativamente inferior à de homens, conforme indicam os dados do Censo da Educação Superior [10]. Esse cenário reflete desigualdades estruturais que se manifestam de forma combinada quando se consideram marcadores sociais como raça/etnia, território e trajetória educacional. A perspectiva interseccional proposta por Crenshaw [5] sustenta que essas desigualdades não atuam de forma isolada, mas se sobrepõem e produzem experiências distintas, mesmo entre mulheres inseridas em um mesmo ambiente formativo.

A redução das desigualdades estruturais na educação superior, especialmente em áreas historicamente masculinizadas como a Computação, é central para o avanço do Objetivo de Desenvolvimento Sustentável 5 (ODS 5) da Agenda 2030 da ONU, que visa "alcançar a igualdade de gênero e empoderar todas as mulheres e meninas" [8]. Esse esforço se articula também ao Objetivo de Desenvolvimento Sustentável 11 (ODS 11), que busca "tornar as cidades e comunidades mais inclusivas, seguras, resilientes e sustentáveis" [7]. Instituições de ensino superior desempenham papel estratégico na promoção do desenvolvimento local e na formação de profissionais para atuar em setores essenciais à transformação digital de cidades e comunidades. Assim, compreender desigualdades interseccionais na formação em Computação contribui tanto para a equidade educacional quanto para a construção de comunidades acadêmicas e profissionais mais inclusivas, fortalecendo os pilares sociais dos ODS 5 e 11.

Embora estudos nacionais recentes contribuam para caracterizar a diversidade nos cursos de Computação [13, 16], muitos deles baseiam-se predominantemente em estatísticas descritivas, o que limita a compreensão de padrões estruturais mais profundos. Outras pesquisas abordam interseccionalidade a partir de perspectivas qualitativas [12, 14], evidenciando desafios enfrentados por mulheres na área, mas sem explorar técnicas quantitativas capazes de revelar agrupamentos latentes ou perfis interseccionais emergentes. Assim, persiste uma lacuna metodológica relacionada ao uso de métodos de análise de dados para identificar, de forma sistemática, como diferentes marcadores sociais interagem no contexto educacional da Computação.

Métodos de clusterização não supervisionada oferecem uma oportunidade promissora para avançar nessa discussão, já que permitem identificar padrões ocultos e grupos estruturalmente semelhantes em bases institucionais. No entanto, técnicas como K-Means são sensíveis a desbalanceamentos populacionais [9, 11], podendo reproduzir padrões dominantes e invisibilizar grupos minoritários, fenômeno alinhado às discussões de Criado-Perez sobre vieses e apagamentos algorítmicos [6]. No contexto da Computação brasileira, a predominância masculina pode dificultar a identificação de subgrupos entre as mulheres quando a análise é realizada de forma agregada.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

XVII Computer on the Beach, Florianópolis, SC, Brasil

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

Neste estudo, aplicamos técnicas de clusterização a dados institucionais de estudantes matriculados em cursos da área de Computação da Universidade Federal do Ceará - Campus de Russas[1]. A análise foi conduzida em duas etapas: (i) clusterização da população geral e (ii) clusterização restrita às estudantes mulheres. Essa abordagem permite comparar como padrões majoritários influenciam a formação dos agrupamentos e como a concentração masculina pode mascarar diferenças internas ao grupo feminino.

As principais contribuições deste artigo são: (i) aplicação sistemática de K-Means para explorar características demográficas em cursos de Computação; (ii) identificação de agrupamentos interseccionais entre mulheres, destacando diferenças relacionadas à raça/etnia e ao ano/período de ingresso; (iii) proposição de clusters interpretáveis que podem subsidiar análises futuras sobre permanência, participação e políticas de equidade.

Ao integrar métodos quantitativos de análise de dados com fundamentos da interseccionalidade, este trabalho busca contribuir para um entendimento mais refinado das desigualdades presentes na formação em Computação e apoiar o desenvolvimento de ações institucionais mais sensíveis às especificidades das estudantes mulheres. Assim, este trabalho contribui com estudos quantitativos sobre interseccionalidade de gênero na Computação, necessidade apontada por diversos trabalhos sobre interseccionalidade na área [2, 14, 17]. Ao revelar perfis distintos entre mulheres, considerando raça/etnia, território e trajetória educacional, o presente estudo oferece evidências empíricas que podem subsidiar políticas de equidade alinhadas tanto ao ODS 5 quanto ao ODS 11, contribuindo para comunidades acadêmicas mais inclusivas, resilientes e socialmente sustentáveis.

2 Trabalhos Relacionados

Esta seção apresenta pesquisas que exploram diferentes aplicações de análise de dados e técnicas de clusterização em contextos de diversidade, inclusão e compreensão de perfis de usuários.

Novais et al. [12] investigaram as experiências de mulheres na Computação a partir de entrevistas do podcast "Emílias Podcast - Mulheres na Computação", visando identificar os principais desafios enfrentados e as motivações que impulsionam sua entrada e permanência na área. Utilizando uma abordagem qualitativa baseada em análise de conteúdo, as entrevistas foram transcritas, codificadas e analisadas, incluindo o uso de técnicas de clusterização para identificar padrões semânticos nos depoimentos. Os resultados revelam obstáculos como estereótipos de gênero, falta de pertencimento, ausência de representatividade, discriminação e barreiras institucionais. Em contraste, emergem motivações importantes, como interesse por tecnologia, desejo de gerar impacto social, apoio de redes e comunidades, e influência de mentoras. O estudo conclui que essas motivações ajudam a contrabalançar as dificuldades enfrentadas, reforçando a importância de iniciativas que ampliem a visibilidade e promovam ambientes mais inclusivos para mulheres na Computação.

Barakyo [3] aplicou técnicas de ciência de dados para apoiar o programa de diversidade e inclusão de uma empresa de celulose, utilizando análises de correlação entre variáveis demográficas e indicadores organizacionais, redução de dimensionalidade por meio da Análise de Componentes Principais (PCA) e clusterização com K-means para identificar grupos de colaboradores com perfis semelhantes. A combinação dessas técnicas permitiu revelar padrões de representatividade e desigualdades relacionadas a gênero, raça, deficiência, promoções, treinamentos e outros fatores, evidenciando áreas de sub-representação e acesso desigual a oportunidades. O estudo também empregou visualizações e data *storytelling* para facilitar a interpretação dos resultados, mostrando que fatores de diversidade influenciam trajetórias distintas dentro da empresa e indicando pontos que podem orientar ações internas de inclusão e equidade.

No trabalho de Branco et al. [4], os autores investigaram uma abordagem para a geração automática de personas femininas a partir de dados coletados por questionários e analisados por meio do método de Clusterização, utilizando a ferramenta WEKA (*Waikato Environment for Knowledge Analysis*), com o objetivo de reduzir o esforço manual típico do processo tradicional de criação de personas. O estudo coletou respostas de 130 alunas de cursos de Computação da Universidade Federal do Ceará - Campus de Russas e identificou quatro personas, cada persona representa um perfil distinto de alunas, considerando renda, semestre, motivações, dificuldades e percepções sobre preconceito e desistência do curso. O estudo conclui que a técnica de clusterização contribuiu significativamente para identificar padrões relevantes e gerar personas de boa qualidade.

Considerando os trabalhos apresentados, nota-se que todos exploram, em diferentes contextos, a análise de dados ou métodos qualitativos para entender desigualdades, perfis de usuários ou padrões de comportamento. Isso se aproxima do objetivo desta pesquisa, que também busca identificar estruturas ocultas em grupos sociais específicos. No entanto, esses estudos variam em relação ao foco, ao grupo investigado e ao propósito da análise.

O presente trabalho avança ao combinar técnicas quantitativas de clusterização com o conceito de interseccionalidade, aplicando essa abordagem a dados institucionais de cursos de Computação para identificar desigualdades estruturais entre estudantes, especialmente entre as mulheres. Assim, embora se relacione com os trabalhos anteriores por usar métodos de agrupamento, ela se diferencia por aplicar essas técnicas de forma sistemática para revelar perfis interseccionais no contexto educacional. Dessa forma, contribui tanto metodológica quanto teoricamente para preencher lacunas ainda existentes nas discussões sobre diversidade e gênero na Computação.

3 Metodologia

A metodologia deste estudo foi estruturada para investigar padrões demográficos e interseccionais entre estudantes dos cursos de Computação da Universidade Federal do Ceará (UFC) - campus de Russas por meio de técnicas de clusterização. Uma metodologia em duas etapas foi definida: (i) clusterização da população geral de estudantes e (ii) clusterização restrita às estudantes mulheres. Essa estratégia permitiu avaliar tanto os padrões dominantes presentes no conjunto completo da amostra quanto a forma como esses padrões ocultam ou distorcem características específicas de mulheres, por serem grupos minoritários da população estudantil da área.

[1] http://www.campusrussas.ufc.br/

A metodologia contemplou limpeza e preparação dos dados, codificação de variáveis categóricas, normalização, seleção de hiperparâmetros e validação dos agrupamentos. As decisões adotadas foram fundamentadas na literatura especializada sobre técnicas de clusterização [1][11][9]. Desta forma, buscou-se garantir a reprodutibilidade, a transparência e o rigor metodológico.

A análise foi conduzida no ambiente do Google Colab, utilizando Python. O pré-processamento, a transformação das variáveis e a aplicação dos algoritmos de clusterização foram implementados com **pandas** e **numpy** para manipulação dos dados, **scikit-learn** para normalização, *one-hot encoding*, K-Means e t-SNE, e **matplotlib** e **seaborn** para geração das visualizações.

3.1 Base de dados

Este estudo utilizou os dados disponibilizados pelo Plano de Dados Abertos (PDA) da UFC[2]. Especificamente, utilizou-se o conjunto de dados de Estudantes matriculados (Graduação), que contém informações sobre estudantes matriculados em cursos de graduação. Tendo em vista o foco em interseccionalidade de gênero, as variáveis selecionadas para o estudo foram: sexo, raça/etnia, estado de naturalidade, tipo de nacionalidade, ano de ingresso e período de ingresso. Além disso, essas variáveis compõem dados estruturais amplamente disponíveis em bases de dados institucionais brasileiras [10].

3.2 Pré-processamento

O pré-processamento dos dados foi realizado para garantir que todas as variáveis estivessem adequadas ao uso de algoritmos de clusterização baseados em distância. Inicialmente, foram padronizados os valores ausentes ou inconsistentes nas variáveis categóricas (como sexo, raça/etnia, naturalidade e nacionalidade), convertendo diferentes representações de ausência (como "nan", "NULL", strings vazias ou espaços) em valores faltantes reconhecidos pelo **pandas**. A padronização e o tratamento explícito de valores faltantes são práticas recomendadas para evitar vieses decorrentes de inconsistências ou de registros parcialmente preenchidos [1].

As variáveis categóricas foram então transformadas por meio de *one-hot encoding*, criando colunas binárias para cada categoria. Essa técnica é amplamente utilizada para permitir que algoritmos como o K-Means — que operam em espaços numéricos — processem informações originalmente textuais [9]. As variáveis numéricas relacionadas ao ingresso (ano e período) foram convertidas para formato numérico e, quando necessário, os valores faltantes foram imputados pela mediana, uma prática recomendada para variáveis ordinais e temporais [18].

Por fim, todas as características foram normalizadas com o *StandardScaler*, que centraliza e normaliza os dados. A normalização evita que variáveis com amplitude maior exerçam influência desproporcional no processo de agrupamento [11]. Esse processo de pré-processamento prepara a base para a aplicação consistente e comparável dos métodos de clusterização utilizados no estudo.

3.3 Clusterização da População Geral

A primeira etapa de análise consistiu em aplicar técnicas de clusterização não supervisionada à população geral de estudantes dos cursos analisados: Ciência da Computação e Engenharia de Software do campus de Russas da UFC. O objetivo dessa etapa foi identificar padrões gerais na distribuição dos perfis demográficos, sem ainda introduzir recortes por sexo. Esse procedimento é recomendado em estudos exploratórios de agrupamento para observar como os dados se organizam naturalmente antes da aplicação de filtros específicos [9].

Para essa análise utilizou-se o algoritmo K-Means, amplamente empregado em aplicações educacionais e em estudos de segmentação quando os dados já foram convertidos para um espaço numérico consistente [11]. Como a clusterização baseada em distância é sensível à escala e ao formato das variáveis, aplicou-se previamente o processo de normalização descrito na seção anterior.

A escolha do número de clusters foi conduzida por meio dos métodos clássicos de apoio à decisão, como o coeficiente de silhouette e o método do cotovelo. Esses critérios, amplamente utilizados na literatura, ajudam a identificar valores de k que produzam agrupamentos coerentes e estáveis sem impor complexidade excessiva ao modelo [9][11].

Essa clusterização geral teve caráter essencialmente exploratório e serviu como base comparativa para a segunda etapa da metodologia, dedicada exclusivamente às estudantes mulheres.

3.4 Clusterização Restrita às Estudantes Mulheres

Após a clusterização geral, optou-se por realizar uma segunda etapa concentrada exclusivamente nas estudantes mulheres. Essa decisão metodológica se baseia em duas premissas fundamentais descritas na literatura:

Desbalanceamento de grupos provoca distorções nos agrupamentos: Técnicas de clusterização são sensíveis à distribuição das classes e tendem a privilegiar padrões mais frequentes. Quando um grupo é minorizado numericamente, como observado no caso das mulheres nos cursos de Computação, seus padrões específicos tendem a ser absorvidos pelos clusters dominantes, tornando-se estatisticamente invisíveis [9][11].

A importância de preservar minorias em análises demográficas: Pesquisas sobre boas práticas de preparação de dados destacam a necessidade de cuidados específicos para evitar que padrões minoritários sejam apagados ou superagregados durante o processamento [1]. Decisões metodológicas devem evitar que grupos sub-representados desapareçam nas médias ou nos agrupamentos gerais.

A clusterização referente apenas às estudantes mulheres seguiu o mesmo processo da etapa anterior, garantindo consistência metodológica. Foram reutilizados os mesmos procedimentos de normalização, codificação e preparação dos dados.

Assim como na etapa geral, o algoritmo K-Means foi aplicado após testar diferentes valores de k e avaliar a qualidade dos agrupamentos. A identificação do número de clusters mais adequado utilizou os mesmos critérios descritos anteriormente.

Durante esta fase, adotou-se também uma etapa adicional para garantir a qualidade e a interpretabilidade dos agrupamentos: a identificação e a exclusão de clusters muito pequenos. A literatura sobre clusterização descreve que grupos com tamanho extremamente reduzido podem refletir mais ruído do que padrões reais

[2]https://dados.ufc.br/organization/prograd

e tendem a ser instáveis a pequenas perturbações nos dados [9]. Como recomendação prática, clusters com quantidade mínima insuficiente, geralmente menos de 5% do conjunto analisado, devem ser tratados como outliers estruturais, não como agrupamentos interpretáveis.

Por fim, os clusters válidos resultantes dessa etapa foram avaliados por meio de visualizações, como t-SNE, técnicas amplamente utilizadas para quantificar a coerência e a separabilidade dos agrupamentos.

4 Resultados

4.1 Visão Geral da População Estudada

Antes da aplicação das técnicas de clusterização, realizou-se uma análise exploratória para caracterizar o perfil sociodemográfico e acadêmico da população estudada. O conjunto original de dados era composto por 810 estudantes matriculados nos cursos de Computação analisados. Após a remoção de registros com dados faltantes nas variáveis selecionadas, 772 estudantes foram considerados para as análises subsequentes.

A Figura 1 mostra a distribuição por sexo dos estudantes. A amostra é fortemente desigual: 80,1% (n = 649) são do sexo masculino e 19,9% (n = 161) do sexo feminino. Essa discrepância ressalta a sub-representação feminina nos cursos de Computação.

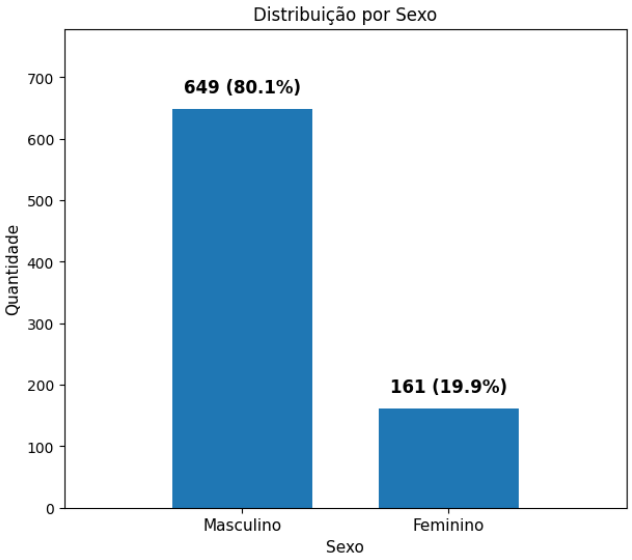

Figure 1: Distribuição dos estudantes por sexo.

A distribuição racial dos estudantes é apresentada na Figura 2. Observa-se que a maioria dos estudantes se autodeclara parda (48,3%; n = 391) ou branca (40,7%; n = 330). Grupos racializados historicamente sub-representados, como pretos (5,8%) e indígenas (0,1%) aparecem com baixa representação. Registros não informados correspondem a 4,3% da amostra.

A Figura 3 apresenta a distribuição de estudantes por ano de ingresso (2019–2024). Como os dados analisados correspondem aos estudantes matriculados, há uma tendência esperada de que a

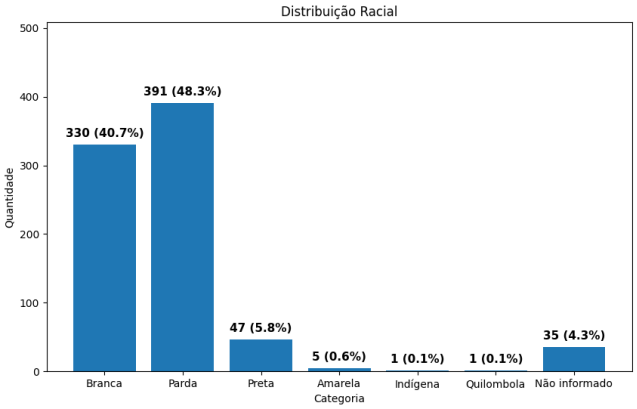

Figure 2: Distribuição racial dos estudantes.

maioria tenha ingressado nos anos mais recentes. De fato, 25,8% dos estudantes ingressaram em 2024, seguido de 21,9% em 2023.

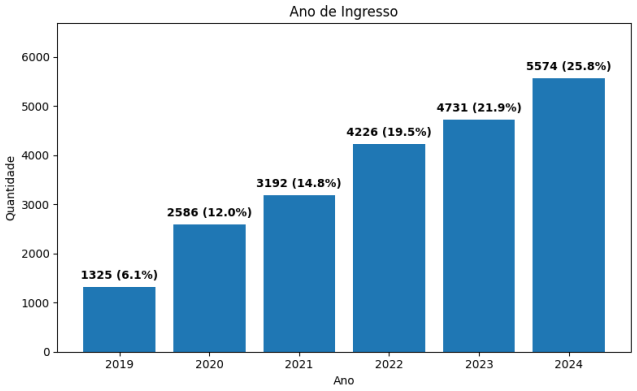

Figure 3: Distribuição dos estudantes por ano de ingresso.

A Figura 4 mostra a origem geográfica dos estudantes por estado. A maioria expressiva (87,1%; n = 617) é oriunda do estado onde a universidade está localizada (Ceará), enquanto os demais estados têm participação muito menor. Somente o estado de São Paulo não está localizado na mesma região do Brasil que os demais estados. Esse padrão sugere uma forte concentração regional na origem dos estudantes.

4.2 Clusterização Geral

A análise inicial foi realizada sobre a população completa de estudantes, utilizando o algoritmo K-Means para explorar padrões gerais na base demográfica. Após a etapa de remoção das linhas com dados faltantes, uma população de 772 estudantes foi considerada. Em seguida, o número ótimo de clusters foi estimado por meio dos métodos do cotovelo e do coeficiente de silhouette, conforme ilustrado na Figura 5.

Após a formação dos agrupamentos, aplicou-se o critério de exclusão de clusters muito pequenos (inferiores a 5% da população), seguindo as recomendações de Hennig et al. [9] e Aguinis et al. [1].

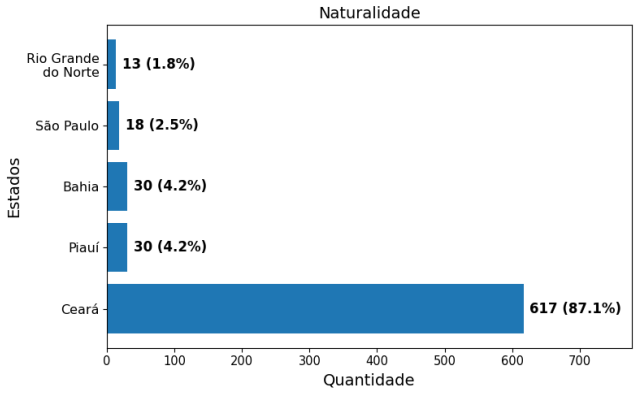

Figure 4: Naturalidade dos estudantes por estado.

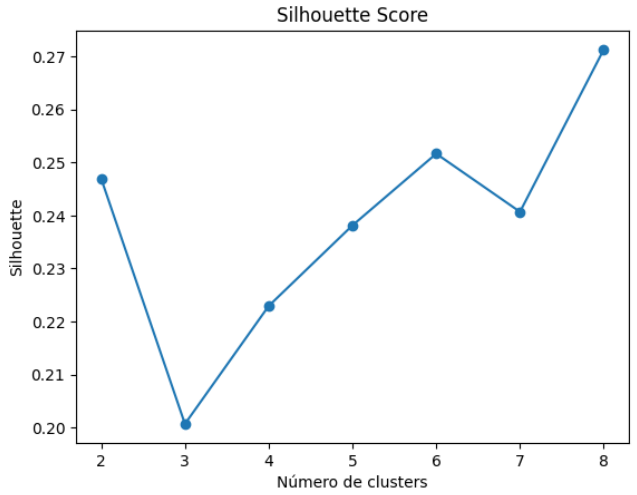

Figure 5: Estimativa do número ótimo de clusters na população completa utilizando o coeficiente silhouette.

Com isso, dois clusters instáveis foram removidos da interpretação. A separação espacial dos clusters foi visualizada por meio de uma projeção bidimensional via *t*-SNE, apresentada na Figura 6. O *t*-SNE é útil para inspeção qualitativa da estrutura dos agrupamentos e confirma que a distribuição dos clusters gerais reflete padrões homogêneos, com pouca separação entre grupos no que se refere às características de sexo e raça/etnia.

Os agrupamentos válidos resultantes capturaram majoritariamente padrões demográficos dominantes na população, reproduzindo a predominância de estudantes homens, autodeclarados pardos ou brancos e naturais do estado do Ceará. A Tabela 1 apresenta um resumo dos clusters considerados estáveis.

Notadamente, observou-se apenas um agrupamento específico relacionado às estudantes mulheres, o que não representa sua diversidade étnico-racial, indicando que o desbalanceamento da população geral impacta diretamente a capacidade do algoritmo de identificar padrões minoritários. Esse achado fundamenta a etapa subsequente de clusterização restrita às mulheres.

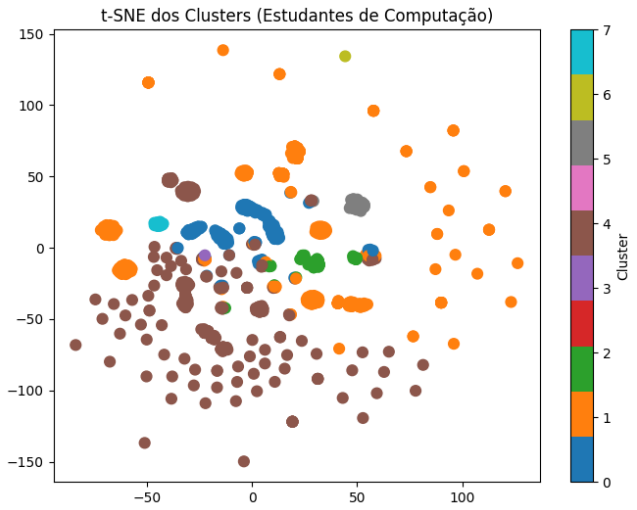

Figure 6: Projeção bidimensional dos clusters gerais por meio de *t*-SNE.

4.3 Clusterização Restrita às Estudantes Mulheres

Após a análise geral, procedeu-se à etapa de clusterização considerando apenas o subconjunto de estudantes do sexo feminino. Essa etapa adicional foi motivada pela baixa representatividade de mulheres na população em geral, o que impediu a formação de agrupamentos significativos na análise inicial. A clusterização considerou a população de 161 estudantes mulheres matriculadas nos cursos de Computação analisados. Após a remoção das linhas com dados faltantes, uma população de 159 registros foi considerada. O melhor valor encontrado foi $k = 7$, conforme apresentado na Figura 7.

Após a formação dos sete agrupamentos, aplicou-se o critério de exclusão de clusters muito pequenos, definido como menos de 5% da população, conforme diretrizes discutidas anteriormente. Quatro clusters foram removidos desta etapa por apresentarem tamanho insuficiente (Clusters 1, 2, 4 e 6). A separação espacial dos clusters foi visualizada por meio da projeção bidimensional via *t*-SNE, exibida na Figura 8. Observa-se que os três clusters válidos formam regiões densas e bem definidas, com baixa sobreposição entre si. O agrupamento correspondente às estudantes pardas aparece como o maior e mais coeso, enquanto os clusters de estudantes brancas e pretas formam grupos igualmente distinguíveis, ainda que de tamanhos diferentes. Essa separação gráfica reforça a estabilidade dos três agrupamentos identificados.

Os três clusters remanescentes apresentaram perfis demográficos distintos e massas críticas adequadas para interpretação. A Tabela 2 apresenta um resumo dos agrupamentos válidos.

O padrão observado indica que cada cluster concentra predominantemente estudantes de um único grupo racial, reforçando a natureza interseccional dos agrupamentos identificados. De forma geral, a análise restrita às mulheres permitiu revelar padrões não identificáveis na clusterização geral devido à sub-representação

Table 1: Clusters válidos após exclusão de grupos pequenos (população completa).

Cluster	Sexo Predominante	Raça/Etnia	Naturalidade	Ano de Ingresso (Mediana)	Tamanho
0	Feminino	Parda	Ceará	2023	139
1	Masculino	Parda	Ceará	2023	287
2	Masculino	Preta	Ceará	2022	44
4	Masculino	Branca	Ceará	2023	257
5	Masculino	Parda	Bahia	2023	30
7	Masculino	Branca	Rio Grande do Norte	2023	13

Table 2: Clusters válidos após exclusão de grupos pequenos (clusterização da população de estudantes mulheres).

Cluster	Raça/Etnia Predominante	Naturalidade	Ano de Ingresso (Mediana)	Tamanho
0	Parda	Ceará	2023	77
3	Preta	Ceará	2021	10
5	Branca	Ceará	2023	61

Figure 7: Estimativa do número ótimo de clusters na população de estudantes mulheres utilizando o coeficiente silhouette.

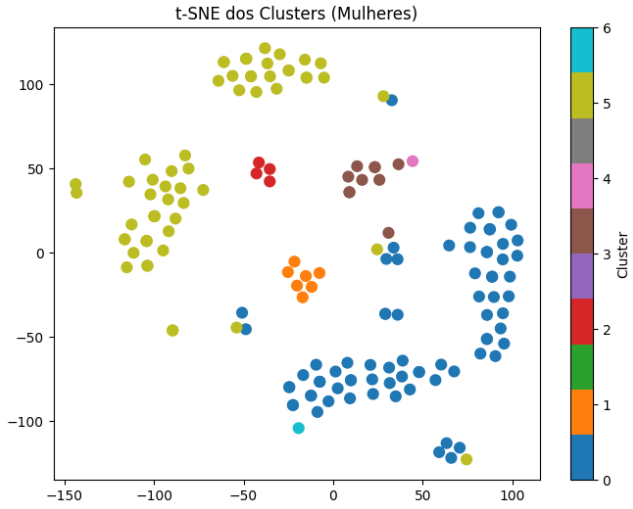

Figure 8: Projeção bidimensional dos clusters femininos por meio de *t*-SNE.

feminina. Os três agrupamentos válidos identificados apresentam diferentes composições raciais e trajetórias de ingresso.

5 Discussão

A literatura sobre interseccionalidade argumenta que as desigualdades não atuam de forma isolada, mas se combinam e se reforçam mutuamente nas experiências de grupos sociais específicos [5]. No campo da Computação, os estudos interseccionais conduzem revisões da literatura que discutem os potenciais impactos de raça, classe e gênero na formação e permanência de estudantes [2][15]. Ainda há uma lacuna significativa de estudos empíricos que utilizem métodos quantitativos e técnicas de análise de dados para identificar, de forma sistemática, como esses marcadores se manifestam nos cursos de Computação. O presente estudo contribui para preencher essa lacuna ao aplicar técnicas de clusterização para revelar padrões interseccionais entre estudantes mulheres de forma exploratória e baseada em dados reais.

Os resultados apresentados neste estudo evidenciam a importância de adotar uma perspectiva interseccional nas análises de gênero na Computação. Poucos estudos anteriores sobre a participação feminina em cursos da área adotam esta perspectiva [13, 16] ao explorar dimensões de diversidade além do gênero, como raça/etnia, idade e deficiência, ainda que na forma de estatística descritiva, sem explorar outros métodos de análise de dados, como a clusterização. Embora tais abordagens contribuam para descrever a sub-representação feminina, não capturam a complexidade das vivências que emergem quando diferentes dimensões de diversidade são consideradas conjuntamente.

O presente trabalho demonstra empiricamente que análises agregadas podem invisibilizar grupos minoritários no próprio grupo de mulheres, uma vez que a clusterização geral, dominada pela maioria masculina, não foi capaz de identificar padrões diferenciados entre as mulheres. Essa invisibilização tem sido amplamente discutida na literatura sobre vieses algorítmicos e análise de grupos minoritários [6]. Os achados reforçam que a aplicação direta de algoritmos de clusterização à população completa resulta na reprodução dos padrões dominantes, especialmente em bases fortemente desbalanceadas por sexo, como é característico dos cursos de Computação no Brasil [10].

Em contraste, a clusterização restrita às mulheres revelou perfis significativos e socialmente relevantes, diferenciados principalmente por raça/etnia e por características temporais de ingresso. A emergência de três grupos principais — mulheres pardas, mulheres brancas e mulheres pretas — demonstra a heterogeneidade do grupo feminino na universidade analisada e que essa heterogeneidade pode ser capturada de forma objetiva por meio de técnicas de mineração de dados com sensibilidade interseccional. Esse resultado é particularmente expressivo, dado que mulheres pretas e pardas costumam enfrentar desigualdades adicionais no acesso e na permanência no ensino superior, aspecto amplamente discutido em estudos qualitativos, mas raramente analisado com base em dados institucionais estruturados [14, 17].

Ao considerar a interseccionalidade como a sobreposição dinâmica de múltiplos marcadores sociais tais como gênero, raça, território e trajetórias educacionais [5], torna-se fundamental adotar métodos analíticos capazes de refletir essa multiplicidade de camadas. A clusterização aplicada neste estudo demonstra que técnicas de análise de dados podem operar como instrumentos para estruturar e representar, de forma objetiva, essas camadas interseccionais, agrupando estudantes que compartilham condições sociais semelhantes. Esses agrupamentos não apenas revelam diferenças internas no grupo de mulheres, mas também criam unidades analíticas que podem ser utilizadas em análises subsequentes de desempenho acadêmico, permanência, participação em atividades curriculares e trajetórias educacionais. Assim, os clusters interseccionais tornam-se um recurso metodológico valioso para orientar intervenções institucionais mais direcionadas, permitindo que políticas de apoio estudantil sejam sensíveis às desigualdades que emergem da combinação entre gênero e outros marcadores estruturais.

É importante reconhecer as limitações relacionadas às variáveis disponíveis na base institucional utilizada. A identificação das estudantes foi realizada com base na variável *sexo*, e não na *gênero*, pois os dados fornecidos pela universidade seguem o padrão de coleta do INEP, que utiliza uma categorização binária. Assim, embora este estudo discuta "mulheres", essa definição está restrita às opções oferecidas pelos dados institucionais, não contemplando identidades de gênero diversas. Além disso, observou-se a presença de dados faltantes em relação à raça/etnia, uma vez que a autodeclaração é opcional e socialmente sensível, especialmente no contexto das políticas de cotas no ensino superior. Esse fenômeno pode influenciar a decisão de declarar ou não a identidade racial, devendo os valores ausentes serem interpretados à luz dessas dinâmicas sociais, e não como ausência de marcador racial.

A discussão deste estudo indica que a análise de gênero em cursos de Computação se beneficia substancialmente de uma abordagem interseccional orientada por dados. Ao integrar métodos empíricos de clusterização com reflexões sobre desigualdades estruturais, este trabalho ao revelar dinâmicas frequentemente negligenciadas no debate sobre diversidade na Computação.

5.1 Ameaças à Validade

Este estudo apresenta algumas limitações que devem ser consideradas na interpretação dos resultados.

Os dados utilizados provêm de bases institucionais públicas da universidade analisada. Embora sejam dados oficiais, algumas variáveis apresentam valores faltantes, especialmente na autodeclaração de raça/etnia. Essa ausência pode refletir dinâmicas sociais relacionadas ao contexto das políticas de cotas e à natureza voluntária da declaração.

A variável utilizada para identificar mulheres corresponde ao atributo *sexo* presente na base institucional, e não à identidade de gênero. Essa limitação decorre do padrão de coleta adotado em bases educacionais brasileiras e pode restringir a representação de identidades de gênero diversas.

Os resultados referem-se especificamente aos cursos de Computação da Universidade Federal do Ceará – campus de Russas. Assim, embora os padrões identificados sejam consistentes com tendências descritas na literatura, a generalização para outras instituições ou regiões deve ser realizada com cautela.

A clusterização depende das variáveis disponíveis na base analisada. Outras dimensões relevantes para análises interseccionais — como renda familiar, escolaridade dos pais ou políticas de acesso — não estavam disponíveis nos dados utilizados.

6 Conclusões e Trabalhos Futuros

O estudo teve como objetivo identificar perfis interseccionais entre estudantes de Computação e compreender como diferentes marcadores sociais, especialmente raça/etnia e gênero, se combinam para moldar suas experiências. Os resultados mostraram que a clusterização é um método adequado para identificar a sobreposição de marcadores sociais que geralmente passam despercebidos quando analisamos dados de forma agregada. A predominância masculina nos cursos de Computação acabou ocultando diferenças importantes entre as próprias mulheres. Ao analisar apenas esse grupo, emergiram três perfis distintos, caracterizados por raça/etnia, origem e trajetória de ingresso. Esses achados reforçam a ideia de que a interseccionalidade é essencial para compreender desigualdades educacionais e de que métodos quantitativos ajudam a revelar padrões que refletem realidades sociais diversas.

A partir desses resultados, o estudo sugere que Instituições de Ensino Superior adotem análises interseccionais em seus diagnósticos, usem clusters para orientar políticas de permanência e ampliem a coleta de informações sobre seus estudantes. Também aponta caminhos para pesquisas futuras, como incluir variáveis acadêmicas e socioeconômicas, investigar evasão, acompanhar trajetórias ao longo do tempo e considerar políticas de cotas. Além disso, pesquisas futuras podem explorar diferentes algoritmos de clusterização, bem como análises de estabilidade e de validação cruzada dos agrupamentos, ampliando a robustez das investigações e aprofundando a compreensão sobre a experiência estudantil em Computação no Brasil.

Acknowledgments

As autoras agradecem o apoio financeiro da Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP) por meio do processo PRH-0212-00011.01.00/23.

References

[1] Herman Aguinis, N Sharon Hill, and James R Bailey. 2021. Best practices in data collection and preparation: Recommendations for reviewers, editors, and authors. *Organizational Research Methods* 24, 4 (2021), 678–693.

[2] Marília Amaral and Leander Oliveira. 2024. Como Abordamos a Interseccionalidade na Computação? Busca por Valores Interseccionais em uma Revisão Sistemática de Literatura na Base SOL. In *Anais do XVIII Women in Information Technology* (Brasília/DF). SBC, Porto Alegre, RS, Brasil, 183–194. doi:10.5753/wit.2024.2605

[3] Daniele Assis Lucas Baraky. 2024. *Análise de dados para impulsionar o programa de diversidade e inclusão na CENIBRA.* Monografia (Especialização em Ciência de Dados). Instituto de Ciências Exatas e Aplicadas, Universidade Federal de Ouro Preto, João Monlevade.

[4] Karina da S. C. Branco, Rhenara A. Oliveira, Francisco L. Q. da Silva, Jacilane de H. Rabelo, and Anna B. S. Marques. 2020. Does this persona represent me? investigating an approach for automatic generation of personas based on questionnaires and clustering. In *Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems* (Diamantina, Brazil) *(IHC '20)*. Association for Computing Machinery, New York, NY, USA, Article 44, 6 pages. doi:10.1145/3424953.3426648

[5] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *University of Chicago Legal Forum*, Vol. 140. 139–167.

[6] Caroline Criado-Perez. 2019. *Invisible Women: Data Bias in a World Designed for Men.* Penguin Books, London.

[7] Organização das Nações Unidas no Brasil. 2015. ODS 11: Cidades e Comunidades Sustentáveis. https://brasil.un.org/pt-br/sdgs/11. Acesso em: nov. 2025.

[8] Organização das Nações Unidas no Brasil. 2015. ODS 5: Igualdade de gênero. https://brasil.un.org/pt-br/sdgs/5. Acesso em: nov. 2025.

[9] Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci. 2015. *Handbook of cluster analysis.* CRC press.

[10] INEP. 2024. Censo da Educação Superior 2024: Painel de Estatísticas. https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior/resultados Acesso em: nov. 2025.

[11] Anil K Jain. 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters* 31, 8 (2010), 651–666.

[12] Tainara Silva Novaes, Larissa Behrens Soares, Adolfo Neto, Mariangela Setti, and Maria Claudia Figueiredo Pereira Emer. 2025. Desafios e Motivações de Mulheres na Computação-Análise de Entrevistas de um Podcast. In *Women in Information Technology (WIT).* SBC, 207–217.

[13] Claudia Pereira, José Figuerêdo, Thiago Alves, Naiara Santos, Nelma Galvão, and Teófilo Galvão Filho. 2024. (In)visibilidade da Diversidade nos Cursos Presenciais de Computação e Tecnologias da Informação e Comunicação: Um Panorama das Universidades Públicas da Bahia. In *Anais do IV Simpósio Brasileiro de Educação em Computação* (Evento Online). SBC, Porto Alegre, RS, Brasil, 90–101. doi:10.5753/educomp.2024.237512 http://doi.org/10.5753/educomp.2024.237512.

[14] Yolanda A Rankin and Jakita O Thomas. 2020. The intersectional experiences of Black women in computing. In *Proceedings of the 51st ACM technical symposium on computer science education.* 199–205.

[15] Karolyne Rodrigues, Rayane Duarte, Ádina Nascimento, Fernanda Ferreira do Nascimento, and Rogério César. 2025. Interseccionalidade e Tecnologia: Um Mapeamento Sistemático de Publicações em Português sobre Gênero, Raça e Classe na Participação Feminina. In *Anais do XIX Women in Information Technology* (Maceió/AL). SBC, Porto Alegre, RS, Brasil, 24–34. doi:10.5753/wit.2025.8067

[16] Maria Santos, Laís Vossen, Daniella Vasconcellos, Guilherme Borchardt, Roger Venson Junior, Eric Silveira, Marily Silva, and Isabela Gasparini. 2023. Panorama da diversidade nos cursos presenciais de Computação e Tecnologias da Informação e Comunicação das universidades públicas de Santa Catarina. In *Anais do III Simpósio Brasileiro de Educação em Computação* (Evento Online). SBC, Porto Alegre, RS, Brasil, 69–78. doi:10.5753/educomp.2023.228188 http://doi.org/10.5753/educomp.2023.228188.

[17] Anna Szlavi, Marit Fredrikke Hansen, Sandra Helen Husnes, and Tayana Uchôa Conte. 2023. Intersectionality in computer science: a systematic literature review. In *2023 IEEE/ACM 4th Workshop on Gender Equity, Diversity, and Inclusion in Software Engineering (GEICSE).* IEEE, 9–16.

[18] Stef Van Buuren. 2018. *Flexible imputation of missing data.* CRC press.