Universidade Federal do Vale do São Francisco
Curso de Ciência da Computação

Análise de Padrões de Gênero em STEM com Técnicas de Aprendizado de Máquina e Explicabilidade de Modelos

Autora: Catarina Cysneiros Sampaio
Orientador: Débora Araújo

Salgueiro – PE
Junho de 2025

Sumário

Capítulo 1

Introdução

1.1 Motivação e Justificativa

De acordo com dados do Censo da Educação Superior de 2023 Brasil (2024), no Brasil, as mulheres representam a maioria tanto entre os ingressantes (59,4%) quanto entre os concluintes (59,6%) do Ensino Superior. No entanto, uma análise mais aprofundada revela um padrão significativo na distribuição de gênero entre as diferentes áreas do conhecimento. Enquanto cursos das áreas de Saúde e Educação apresentam predominância feminina, os cursos de Ciências Exatas e Tecnologias são majoritariamente ocupados por homens.

Apesar de serem maioria no Ensino Superior, a presença feminina permanece concentrada em determinadas áreas. Esse cenário leva à reflexão sobre os motivos pelos quais, mesmo com o avanço no acesso à educação superior pelas mulheres, a subrepresentação em campos como Ciência, Tecnologia, Engenharia e Matemática (STEM, na sigla em inglês) ainda persiste. A desigualdade de gênero nessas áreas não parece estar relacionada à ausência de mulheres no ambiente universitário, mas sim a fatores mais profundos que influenciam suas escolhas acadêmicas e profissionais.

Diversos estudos apontam que essas escolhas são moldadas desde cedo por uma combinação de fatores psicológicos Master, Meltzoff e Cheryan (2021), influências familiares Hsieh e Simpkins (2022) e condições socioeconômicas Morales, Grineski e Collins (2021). Desde a infância, meninas e meninos são expostos a estereótipos que moldam suas percepções sobre si mesmos e determinam expectativas sociais relacionadas ao comportamento e às áreas de interesse Master, Meltzoff e Cheryan (2021). Tais papéis de gênero, incentivados desde a infância, reforçam a ideia de funções distintas para homens e mulheres na sociedade, contribuindo diretamente para a disparidade nas áreas STEM (Science, Technology, Engineering and Mathematics — Ciência, Tecnologia, Engenharia e Matemática) McGuire et al. (2022). Esses estereótipos são disseminados em diversos contextos — familiar, educacional e midiático — sendo reconhecidos como

um dos principais agravantes da desigualdade de gênero nessas áreas Silva et al. (2022).

Além dos estereótipos, o suporte familiar, o ambiente escolar e o contexto socioeconômico exercem grande influência sobre as escolhas e interesses das meninas, podendo desmotivá-las a seguir carreiras nas áreas de Ciência e Tecnologia. Fatores como a necessidade de conciliar trabalho e estudo, a carga horária dos cursos, a valorização de formações mais voltadas ao mercado de trabalho e as diferenças entre as oportunidades disponíveis em cidades do interior e nas capitais também impactam essas decisões. Em muitos casos, as capitais oferecem uma maior variedade de cursos e instituições Brasil (2024), mas também apresentam um custo de vida mais elevado, o que pode representar uma barreira para estudantes de baixa renda.

Essas questões conduzem ao seguinte problema de pesquisa: **quais fatores influenciam a escolha do curso superior e como esses fatores se relacionam com a sub-representação feminina nas áreas de STEM?**

Buscar respostas para essa pergunta exige, antes de tudo, um olhar atento aos desafios enfrentados pelas meninas desde cedo — desafios que podem ser sociais, econômicos, culturais ou educacionais — e que acabam influenciando suas decisões acadêmicas e profissionais. Ao mesmo tempo, é fundamental compreender o que leva algumas meninas, mesmo diante de tantos obstáculos, a persistirem e ocuparem espaços em áreas onde ainda são minoria.

Tendo em vista esse contexto, este trabalho propõe mapear os perfis de meninas que estão cursando o Ensino Superior, com o objetivo de identificar os principais fatores que influenciaram suas escolhas pelos atuais cursos de graduação. Considerando as particularidades envolvidas na escolha de um curso superior, é necessário levar em conta os diferentes contextos socioeconômicos dos estudantes. Fatores como a necessidade de conciliar trabalho e estudo — o que pode influenciar a escolha por cursos com menor carga horária diária — e a busca por formações mais voltadas ao mercado de trabalho são aspectos relevantes nesse processo, além das oportunidades disponíveis para cada estudante.

A proposta é identificar os principais fatores que despertam o interesse ou contribuem para o afastamento das áreas de STEM, bem como os motivos que levam à escolha de outras áreas. Para isso, será utilizado um conjunto inicial de dados obtidos por meio da aplicação de questionários a estudantes do Ensino Superior em instituições públicas e privadas da cidade de Salgueiro — Pernambuco. A análise desses dados poderá servir de base para o desenvolvimento de modelos preditivos capazes de classificar e compreender os diferentes perfis de alunos, contribuindo para ações mais direcionadas de incentivo e suporte.

1.2 Objetivo Geral

Analisar os fatores que influenciam as percepções de pertencimento e exclusão de gênero em áreas STEM, a partir da aplicação de técnicas de aprendizado de máquina explicável e de agrupamento de dados, considerando o contexto de estudantes universitários do Sertão Central de Pernambuco.

1.3 Objetivos Específicos

- Elaborar e aplicar um questionário em instituições públicas e privadas de ensino superior da região do Sertão Central de Pernambuco, voltado para a identificação de fatores motivacionais e contextuais relacionados à escolha de cursos em STEM;
- Realizar uma análise exploratória dos dados coletados, identificando padrões e tendências nas respostas das participantes;
- Treinar um modelo de aprendizado de máquina supervisionado e explicável a fim de identificar padrões de resposta relacionados às percepções sobre áreas STEM.
- Aplicar técnicas de agrupamento para identificar perfis distintos de estudantes com base nas percepções de pertencimento e exclusão de gênero em áreas de STEM.
- Analisar e interpretar os grupos formados, discutindo as características predominantes em cada cluster e suas implicações para a compreensão dos fatores de exclusão e pertencimento de gênero em contextos universitários.

1.4 Contribuições do Trabalho

Este trabalho apresenta duas contribuições principais que visam tanto a divulgação de mulheres na ciência quanto o aprofundamento da análise sobre os fatores que influenciam suas escolhas acadêmicas.

A primeira contribuição consiste no desenvolvimento de uma plataforma digital de acesso aberto, voltada à divulgação de pesquisadoras brasileiras. Nessa plataforma, qualquer pessoa pode indicar mulheres que atuam na pesquisa científica, sem a necessidade de login, promovendo a visibilidade e o reconhecimento dessas profissionais. A ferramenta permite a filtragem por áreas de pesquisa, estado e país de atuação, etnia, entre outros critérios, facilitando a busca por pesquisadoras com diferentes perfis e contribuindo para a quebra de estereótipos de gênero na ciência.

A segunda contribuição é a realização de uma análise empírica com foco no município de Salgueiro–PE, envolvendo a aplicação de questionários e a posterior análise exploratória dos dados coletados. A partir dessas informações, serão aplicadas técnicas de agrupamento para identificar perfis distintos entre as alunas, considerando fatores como motivações, contexto socioeconômico e influências familiares. Esse mapeamento contribui para uma compreensão mais aprofundada das barreiras e incentivos que influenciam a presença feminina nas áreas de STEM no contexto específico do interior do Nordeste, oferecendo subsídios para políticas públicas e ações de incentivo mais eficazes e contextualizadas.

Capítulo 2

Referencial Teórico

2.1 Determinantes Psicossociais na Escolha de Carreira

2.1.1 Estereótipos de Gênero e sua Influência nas Escolhas Acadêmicas

Em primeiro lugar, é preciso estarmos cientes sobre o real cenário de participação feminina na graduação brasileira. Segundo dados do censo de educação superior, mulheres representam a maioria dos estudantes brasileiros de graduação e pós-graduação Brasil (2024). Porém, ao filtrarmos por áreas, percebemos que as mulheres se concentram em áreas específicas do conhecimento, como, por exemplo, relacionadas à saúde e educação (cursos de licenciatura). Ou seja, o cenário real do Brasil é que existem muitas mulheres na graduação, porém estão concentradas em determinadas áreas.

Se o problema não é a quantidade de mulheres na graduação, precisamos analisar quais são os fatores que influenciam a escolha de cursos específicos pelas mulheres.

Segundo um estudo realizado por Master, Meltzoff e Cheryan (2021), é evidente a influência dos estereótipos de gênero sobre o interesse e as habilidades atribuídas a meninas e meninos desde a infância. Observou-se que, já na educação infantil, as crianças tendem a reproduzir papéis de gênero nas diferentes áreas do conhecimento. Em especial, foi identificado que elas associam as áreas de STEM (Ciência, Tecnologia, Engenharia e Matemática) mais frequentemente aos homens, tanto em termos de interesse quanto de competência.

O estudo também diferencia os estereótipos de gênero em relação ao interesse e à habilidade: enquanto o interesse se refere à percepção de que determinadas áreas são mais atrativas ou adequadas para um gênero específico, a habilidade diz respeito à crença de que meninos ou meninas são naturalmente mais capazes ou competentes em

certas disciplinas.

Um segundo estudo sobre a presença de estereótipos entre estudantes McGuire et al. (2022) revela que os estereótipos de gênero relacionados às áreas de STEM estão presentes tanto em crianças quanto em adolescentes, do ensino fundamental ao médio. A manifestação desses estereótipos varia conforme o gênero dos estudantes: meninos tendem a reproduzi-los de forma mais acentuada, enquanto meninas os reproduzem em menor grau — embora ainda de maneira significativa.

O primeiro estudo também constatou que as meninas que reproduzem estereótipos em relação a STEM tendem a demonstrar menor interesse por atividades descritas de forma estereotipada. Ou seja, quanto mais as estudantes internalizam esses estereótipos, menos se identificam ou se engajam com determinadas tarefas que são socialmente associadas ao universo masculino.

A preferência por não se envolver em uma atividade descrita de forma estereotipada como masculina pode ser explicada pelo sentimento de pertencimento. Um estudo realizado com estudantes de 11 a 14 anos, em uma escola nos Estados Unidos Opps e Yadav (2022), revelou uma relação significativa entre o sentimento de pertencimento à área de ciência da computação e o nível de estereótipos que as alunas reproduzem. Observou-se que meninas que representam cientistas de forma estereotipada — especialmente com traços masculinos ou pejorativos — tendem a sentir-se menos pertencentes a essa área. Além disso, o estudo mostrou que, no que se refere aos estereótipos de aparência, as meninas demonstraram uma frequência significativamente maior do que os meninos ao retratar cientistas com características estereotipadas, como jaleco, óculos ou cabelos desalinhados.

2.1.2 Fatores Socioeconômicos e Geográficos na Escolha Acadêmica

No trabalho realizado por Park et al. (2024), os autores analisam o impacto do status socioeconômico (SES) sobre a relação entre aspirações ocupacionais e desempenho acadêmico. Verificou-se que, em grupos de SES alto e médio, aspirações elevadas exercem um efeito compensatório, contribuindo para a melhora do desempenho escolar. Contudo, entre estudantes de SES baixo, aspirações semelhantes não se traduzem em melhor desempenho, sendo o rendimento acadêmico prévio o principal indicador das aspirações futuras. O estudo aponta que, apesar desses estudantes de baixa renda almejarem ocupações de alta renda, a ausência de recursos econômicos, culturais e sociais limita a conversão dessas aspirações em resultados concretos. Além disso, ressalta-se que, em contextos de desvantagem, aspirar não basta: as aspirações frequentemente não são acompanhadas por estratégias, habilidades ou suporte necessários para sua re-

alização, configurando o que os autores chamam de "otimismo cruel" — uma esperança que, por ser difícil de concretizar, pode prejudicar o bem-estar dos estudantes.

A escolha de uma profissão com base na perspectiva de um alto retorno financeiro, por si só, não deve ser considerada um problema, mas sim uma escolha individual legítima. A preocupação surge, contudo, quando essa motivação não é fruto de uma preferência pessoal ou vocacional genuína, mas sim de uma necessidade imposta pelas condições socioeconômicas do estudante. Nesse caso, a decisão profissional deixa de ser uma escolha autônoma e passa a refletir a busca por estabilidade e sobrevivência, muitas vezes em detrimento de interesses e paixões pessoais.

Além dos fatores socioeconômicos, é fundamental compreender as oportunidades educacionais disponíveis na região onde esses estudantes vivem. Isso porque suas escolhas de curso estão fortemente condicionadas à viabilidade de frequentar presencialmente uma universidade, especialmente considerando que esta análise tem como foco os cursos presenciais.

De acordo com o Censo da Educação Superior de 2023, o município de Salgueiro oferece 17 cursos superiores presenciais. Desses, cerca de 7 estão diretamente relacionados às áreas de STEM (Ciência, Tecnologia, Engenharia e Matemática). Além disso, 8 cursos têm como foco a formação de professores, sendo ofertados na modalidade de licenciatura. Vale destacar que apenas duas opções de cursos STEM com grau de bacharelado são ofertadas por instituições públicas, o que pode representar uma limitação no acesso a essas áreas estratégicas para o desenvolvimento científico e tecnológico da região. Conforme o próprio plano pedagógico do curso de Ciência da Computação (2021), um dos objetivos de criação do curso é justamente atender às demandas técnicas e tecnológicas atuais e futuras do Sertão Central, contribuindo para o desenvolvimento regional.

2.1.3 Fatores Familiares na Escolha Acadêmica

Na revisão bibliográfica realizada por Gencel-Augusto et al. (2025), a influência parental foi identificada como um dos principais fatores que contribuem para a baixa representatividade de mulheres hispânicas nas áreas de STEM (Ciência, Tecnologia, Engenharia e Matemática). Observou-se que estudantes que recebem pouco incentivo ou reconhecimento por parte dos pais tendem a desenvolver uma percepção negativa sobre suas próprias habilidades em disciplinas como ciências e matemática, o que pode comprometer seu interesse e desempenho nessas áreas.

Esses dados evidenciam a relevância do papel da família no processo de escolha da carreira acadêmica. Quanto menor o incentivo recebido em determinadas áreas do conhecimento, menor tende a ser a autoconfiança dos estudantes para seguir uma trajetória profissional nessas áreas, uma vez que há uma maior propensão a acreditarem

que suas capacidades são inferiores às exigidas.

Além da falta de apoio em determinadas áreas, há também a influência significativa das expectativas familiares sobre o futuro dos estudantes. Em uma revisão bibliográfica realizada por Arshad e Arshad (2024), foi evidenciado que, em muitas culturas e contextos sociais, as decisões da família tendem a ter um peso maior do que os próprios desejos do estudante. Em outras palavras, a vontade familiar, em diversos casos, sobrepõe-se às aspirações individuais dos jovens, influenciando diretamente suas escolhas acadêmicas e profissionais.

2.1.4 Fatores Educacionais na Escolha Acadêmica

Segundo um panorama levantado por Pugliese sobre o estado da educação em STEM no Brasil, foi notado que ainda se trata de um movimento incipiente no país, com presença tímida na literatura acadêmica nacional e maior disseminação em escolas privadas e iniciativas de organizações não governamentais. Além disso, observa-se que, quando presente, o modelo é muitas vezes adotado como tendência estrangeira, com forte apelo de mercado, e não como uma proposta pedagógica adaptada às realidades e necessidades locais.

Esses fatores acabam afastando o interesse dos alunos, especialmente aqueles da rede pública, uma vez que o conteúdo frequentemente não dialoga com seu contexto social e cultural, nem considera as desigualdades estruturais que impactam seu acesso ao conhecimento. Sem estratégias inclusivas e contextualizadas, o ensino de STEM corre o risco de se tornar excludente, reforçando barreiras já existentes ao invés de superá-las.

Além disso, considerando os achados de Morales, Grineski e Collins (2021) sobre o impacto da discordância mentor-mentorado na intenção de estudantes Latinx de buscar a pós-graduação e na sua produtividade em pesquisa, fica evidente a crucial importância do mentor na trajetória acadêmica e profissional dos mentees. O estudo revela que, enquanto a discordância de gênero pode, surpreendentemente, estar associada a um aumento de (17%) na intenção de pós-graduação para os estudantes Latinx em geral, há uma nuance crítica: quando pareadas com mentores de gênero discordante, estudantes Latinas foram (70%) menos propensas a apresentar seus projetos de pesquisa em conferências profissionais. Isso sublinha que, embora a mentoria com diversidade de gênero possa, em certos aspectos, impulsionar as aspirações de longo prazo, ela pode, ao mesmo tempo, criar barreiras significativas para a participação ativa em marcos de produtividade de pesquisa de curto prazo. Complementarmente, a discordância de raça/etnia e, mais acentuadamente, a de status de primeira geração, está ligada a uma redução significativa da intenção de buscar a pós-graduação. Assim, a relação de mentoria não é neutra; a similaridade de experiências e backgrounds, especialmente em

dimensões sociais, raciais e de status familiar no ensino superior, pode ser um fator facilitador ou, na sua ausência (discordância), um obstáculo.

2.2 Análise de Dados na Educação e Estudos sobre Perfil de Estudantes

A análise de dados na educação emergiu como uma ferramenta transformadora, permitindo uma compreensão profunda e granular do processo de aprendizagem e do comportamento dos estudantes. Com a crescente digitalização do ambiente educacional, a vasta quantidade de dados gerados em plataformas de ensino e sistemas de gerenciamento da aprendizagem oferece um campo fértil para a extração de insights valiosos. Através da aplicação de técnicas avançadas de análise, é possível identificar padrões, prever tendências e, crucialmente, personalizar o ensino, adaptando-o às necessidades individuais de cada aluno. A capacidade de traçar perfis de estudantes, por exemplo, não apenas revela características demográficas, mas também expõe padrões de desempenho, estratégias de aprendizagem e trajetórias acadêmicas, capacitando as instituições a tomarem decisões baseadas em evidências e a otimizar a qualidade da educação.

O trabalho de Junior et al. (2023) destaca a importância da análise de dados no contexto educacional como uma ferramenta poderosa para obter insights sobre os estudantes. Segundo os autores, com o avanço da tecnologia e a crescente utilização de plataformas digitais e sistemas de gerenciamento de aprendizagem, tornou-se possível coletar, armazenar e processar grandes volumes de dados educacionais. A análise desses dados permite identificar padrões e tendências, personalizar o ensino de acordo com as necessidades individuais dos estudantes e tomar decisões baseadas em evidências com o objetivo de melhorar a qualidade da educação. Além disso, os autores ressaltam que técnicas como a análise preditiva e o uso de algoritmos de aprendizado de máquina ampliam ainda mais o potencial dessa abordagem.

Oliveira (2021) utiliza a análise de dados educacionais para compreender as transformações no perfil dos estudantes de graduação no Brasil entre os anos de 2001 e 2015. A autora realiza uma revisão da literatura e examina diversas bases de dados abertas, como a PNAD, o Censo da Educação Superior, a pesquisa da Andifes/Fonaprace e os resultados do ENADE. Por meio dessas fontes, são evidenciadas mudanças nos perfis dos estudantes quanto à raça/cor, renda e região de origem, revelando um processo de democratização do acesso à educação superior. A análise mostra, por exemplo, o aumento expressivo da participação de estudantes negros e de baixa renda, especialmente nas instituições públicas, como resultado das políticas de inclusão social implementadas

no período. Nas considerações finais, a autora destaca a riqueza dos dados disponíveis, especialmente da PNAD, para aprofundar estudos sobre o perfil dos estudantes de forma integrada entre setores público e privado.

O trabalho de Schel e Drechsel (2025) investiga as competências de autorregulação da aprendizagem em estudantes de formação em docência. O estudo tem como objetivo identificar diferentes perfis de aprendizagem autorregulada entre esses estudantes. Para alcançar esse objetivo, os autores empregam a análise de perfis latentes (Latent Profile Analysis - LPA), uma técnica estatística multivariada. A LPA é utilizada para agrupar indivíduos em subgrupos (perfis) com base em suas respostas a questionários e testes que avaliam diversas competências de autorregulação da aprendizagem. Ao invés de analisar cada competência isoladamente, a LPA permite identificar combinações de competências que caracterizam grupos distintos de alunos, revelando padrões de pontos fortes e fracos em suas estratégias de aprendizagem. Os resultados da análise de perfis latentes revelaram a existência de quatro perfis distintos de estudantes de formação em docência em relação às suas competências de autorregulação da aprendizagem. Esses perfis são:

1. Estudantes com altas competências de autorregulação em todos os domínios: Este grupo demonstra um elevado nível de proficiência em todas as facetas da autorregulação da aprendizagem.

2. Estudantes com deficiências em estratégias cognitivas: Este perfil se caracteriza por ter dificuldades específicas no uso de estratégias cognitivas eficazes para a aprendizagem.

3. Estudantes com deficiências em estratégias metacognitivas: Este grupo apresenta lacunas nas suas habilidades de monitoramento e regulação do próprio processo de aprendizagem (metacognição).

4. Estudantes com deficiências em estratégias de regulação de recursos: Este perfil é marcado por dificuldades em gerenciar recursos de aprendizagem, como tempo e ambiente de estudo.

2.3 Técnicas de Agrupamento como Estratégia de Mapeamento de Perfis

O agrupamento, ou análise de *clusters*, é uma técnica fundamental da mineração de dados voltada para a descoberta de padrões e estruturas ocultas em conjuntos de dados. Trata-se do processo de organizar objetos em grupos de tal forma que os elementos pertencentes ao mesmo grupo sejam altamente semelhantes entre si, enquanto

apresentem grande dissimilaridade em relação aos elementos de outros grupos. Segundo Han, Kamber e Pei (2011), a avaliação dessas similaridades ou dissimilaridades ocorre com base nos atributos dos objetos, sendo comum o uso de medidas de distância como critério quantitativo para definir a proximidade entre os dados.

O agrupamento se diferencia de outras técnicas de aprendizado de máquina por ser uma abordagem de aprendizado não supervisionado. Isso significa que, ao contrário de métodos supervisionados como classificação, em que os dados de entrada estão associados a rótulos de classe previamente definidos, o agrupamento opera sem qualquer informação prévia sobre categorias. Como afirmam os autores Han, Kamber e Pei (2011), o agrupamento é, portanto, uma forma de "aprendizado por observação", sendo especialmente útil em contextos nos quais não se conhece previamente a estrutura dos dados ou os agrupamentos naturais existentes.

A aplicação da análise de agrupamento é particularmente relevante no contexto da educação para a criação de perfis de estudantes. Um exemplo claro dessa aplicação é o trabalho de Oliveira et al. (2022). Neste estudo, os autores realizam uma revisão da literatura e uma análise de diferentes algoritmos de clusterização com o objetivo de identificar e compreender padrões de engajamento de estudantes em ambientes de aprendizagem online. Ao agrupar estudantes com comportamentos de engajamento semelhantes, é possível traçar perfis específicos que revelam, por exemplo, alunos altamente ativos, moderadamente engajados ou aqueles com baixo nível de participação.

Expandindo essa aplicação para um contexto de saúde e bem-estar, Martins et al. (2021) demonstraram o potencial do agrupamento na identificação de perfis de risco. O objetivo do trabalho foi agrupar adolescentes escolares com características semelhantes em relação à predisposição ao uso de substâncias psicoativas. Ao aplicar técnicas de clusterização sobre dados de variáveis sociodemográficas, comportamentais e relacionadas à saúde, os pesquisadores conseguem segmentar a população estudada em perfis distintos, como aqueles com maior vulnerabilidade devido a fatores familiares, sociais ou psicológicos, e aqueles com menor risco.

EXPLICAR AQUI O MODELO DE AGRUPAMENTO UTILIZADO NO SEU TRABALHO. EXPLICAR A TÉCNICA UTILIZADA PARA A ESCOLHA DA QUANTIDADE DE GRUPOS. - METODO CLUSTERIZAÇÃO HIERARQUICA - METODO WARD NA CLUSTERIZAÇÃO HIERARQUICA - SHAP VALUES? -

(CITAR LIVRO AQUI) A clusterização hierárquica é uma técnica de agrupamento que organiza os objetos de um conjunto de dados em uma estrutura de níveis, formando uma hierarquia de clusters. Diferentemente dos métodos particionais, que particionam os dados diretamente em um número pré-determinado de grupos, a clusterização hierárquica constrói uma representação em forma de árvore que evidencia como os objetos se agrupam ou se separam ao longo das etapas do processo. Essa representação gráfica

é denominada dendrograma, o qual fornece uma visualização clara dos relacionamentos entre os dados em diferentes níveis de granularidade, permitindo compreender tanto a formação de grupos mais amplos quanto suas subdivisões internas.

A construção dessa hierarquia pode ocorrer de duas maneiras, que definem os principais tipos de métodos hierárquicos: aglomerativo e divisivo.

O método aglomerativo, também conhecido como abordagem bottom-up, inicia considerando cada objeto como um cluster isolado. Em seguida, os clusters mais semelhantes são iterativamente fundidos, resultando em grupos progressivamente maiores. Esse processo continua até que todos os objetos estejam reunidos em um único cluster ou até que um critério de parada seja alcançado. O dendrograma desse tipo de método registra cada fusão e sua ordem, permitindo observar como os grupos emergem e se relacionam ao longo do processo. Um aspecto característico dessa abordagem é sua irreversibilidade: uma vez realizada a fusão entre dois clusters, essa decisão não pode ser revertida.

O método divisivo, ou abordagem top-down, segue a lógica oposta. O algoritmo inicia com todos os objetos reunidos em um único cluster e procede realizando divisões sucessivas, segmentando o conjunto em clusters menores. As divisões continuam de maneira recursiva até que os clusters atendam a um nível desejado de homogeneidade ou até que reste apenas um objeto por grupo. Assim como no método aglomerativo, o processo registrado no dendrograma mostra a ordem e o nível em que cada divisão ocorre, embora essa abordagem tipicamente demande maior custo computacional.

Em ambos os casos, a utilização do dendrograma como representação permite visualizar a estrutura hierárquica dos dados, identificar níveis de similaridade e determinar, de forma interpretável, quantos clusters são mais adequados à análise. (CITAR LIVRO AQUI)

Medidas de Ligação em Clusterização Hierárquica

De acordo com Han, Kamber e Pei (2012), os métodos hierárquicos aglomerativos dependem de uma função de distância para decidir quais clusters devem ser fundidos a cada etapa. As quatro medidas clássicas de ligação (*linkage*) entre dois clusters C_i e C_j são apresentadas a seguir. Nesta descrição, $p \in C_i$ e $p' \in C_j$ representam objetos pertencentes aos clusters, m_i e m_j são os centróides dos clusters, e n_i e n_j são os respectivos tamanhos.

Single Linkage (mínima distância)

A distância entre dois clusters é definida como a menor distância entre quaisquer dois objetos pertencentes aos clusters distintos:

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \|p - p'\|. \tag{2.1}$$

Esse método é também conhecido como *nearest-neighbor* e pode formar cadeias

alongadas devido à sua sensibilidade a ruídos.

Complete Linkage (máxima distância)

A medida de ligação é definida como a maior distância entre pares de objetos pertencentes aos dois clusters:

$$d_{\max}(C_i, C_j) = \max_{p \in C_i,\, p' \in C_j} \|p - p'\|. \tag{2.2}$$

Este método tende a produzir clusters mais compactos, pois considera a pior distância entre os grupos.

Centroid Linkage (distância entre centróides)

A distância entre dois clusters é medida pela distância entre seus centróides:

$$d_{\text{centroid}}(C_i, C_j) = \|m_i - m_j\|. \tag{2.3}$$

Apesar de simples, este método pode ocasionar inversões no dendrograma (*reversals*).

Average Linkage (distância média)

A distância média entre todos os pares de objetos dos dois clusters é utilizada como critério de fusão:

$$d_{\text{avg}}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} \|p - p'\|. \tag{2.4}$$

Esse método é menos sensível a ruídos do que o single ou complete linkage, sendo um compromisso entre ambos.

Método de Ward (mínima variância)

Além das medidas clássicas, o método de Ward constitui um tipo distinto de ligação, baseado na minimização da variância interna dos clusters. Segundo Randriamihamison et al. (2020), o método Ward (ou Ward2) seleciona, a cada etapa, a fusão de clusters que acarreta o menor aumento na soma das distâncias quadráticas intra-cluster.

Supondo que C_i e C_j sejam candidatos à fusão, o critério de Ward é expresso como:

$$\Delta(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \|m_i - m_j\|^2, \tag{2.5}$$

onde $\Delta(C_i, C_j)$ representa o aumento na soma das variâncias internas (*within-cluster sum of squares*) decorrente da fusão. O método seleciona o par (C_i, C_j) que minimiza esse valor.

O método Ward difere dos demais por não basear a ligação em distâncias diretas entre pontos ou médias aritméticas, mas sim na otimização da compactação dos clusters, sendo especialmente útil quando se deseja obter grupos homogêneos e com pequena

variabilidade interna.

A escolha do número adequado de clusters é um passo essencial em algoritmos de agrupamento. Um dos critérios mais utilizados para essa finalidade é o *Elbow Method*, o qual, conforme descrito no artigo *"A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm"*, baseia-se na avaliação da soma dos erros quadráticos (*Sum of Squared Errors* – SSE) em função do número de clusters k. Para cada partição gerada, calcula-se a variância intra-cluster como a soma das distâncias quadráticas entre cada observação e o centróide de seu agrupamento. Formalmente, para um cluster C_j com n_j elementos e centróide μ_j, tem-se

$$\mathrm{SSE}(C_j) = \sum_{i=1}^{n_j} \|x_i - \mu_j\|^2,$$

e a soma total da variância intra-cluster para k clusters é dada por

$$\mathrm{SSE}(k) = \sum_{j=1}^{k} \mathrm{SSE}(C_j).$$

O método consiste em plotar $\mathrm{SSE}(k)$ contra o número de clusters e identificar o ponto em que a redução de SSE deixa de ser substancial, formando uma inflexão visual conhecida como "cotovelo". Como ressaltado em (AUTHOR; AUTHOR, 2020), esse ponto representa um equilíbrio entre a compactação dos clusters e a complexidade do modelo, sendo interpretado como o número apropriado de grupos para a estrutura dos dados.

Tabela 2.1: Fichamento de trabalhos relacionados

Fonte	**Descrição**
(MILLER et al., 2024) Master, Meltzoff e Cheryan (2021) McGuire et al. (2022) Master et al. (2023)	Estereótipos de gênero em torno das habilidades e interesses de meninos e meninas em STEM se formam desde a infância e podem moldar escolhas acadêmicas e de carreira ao longo do tempo.
Emran et al. (2020) Liu, Liu e He (2024)	Pais com formação científica podem expor seus filhos a uma visão mais realista e crítica da ciência, enfatizando suas incertezas, criatividade e natureza interpretativa.
Bohrnstedt et al. (2024)	Apesar de meninas demonstrarem menor identidade e autoeficácia matemática em comparação aos meninos, ambos apresentam desempenho matemático semelhante.

Fonte	Descrição
Menezes e Santos (2021)	A presença feminina na Computação ainda é minoritária devido à combinação de estereótipos de gênero, falta de representatividade, desinformação sobre a área e ausência de incentivo familiar e escolar — sendo que iniciativas como oficinas, projetos e parcerias com universidades têm mostrado potencial para reverter esse cenário desde o Ensino Médio.
Zúñiga-Mejías e Huincahue (2024)	Estereótipos de gênero em STEM afetam meninas desde o ensino fundamental, influenciam suas aspirações profissionais e são perpetuados por pais, professores e colegas — sendo necessárias intervenções precoces.
Tellhed, Björklund e Strand (2023)	Estudantes do ensino fundamental na Suécia apresentam estereótipos implícitos e explícitos que associam tecnologia aos homens e cuidados às mulheres, o que reduz o interesse de meninas por educação tecnológica — especialmente entre aquelas que internalizam mais fortemente esses estereótipos.
Su, Putka e Rounds (2023)	Estereótipos sobre perfis de interesse em ciência da computação não refletem a realidade da área e propõem estratégias para tornar STEM mais inclusiva e alinhada à diversidade de interesses, especialmente entre mulheres.
Opps e Yadav (2022)	O estudo revela que meninas do ensino fundamental reproduzem mais estereótipos visuais sobre cientistas da computação do que meninos, e que isso pode impactar negativamente seu senso de pertencimento na área.
Silva et al. (2022)	A evasão de mulheres na computação está ligada a estereótipos, discriminação e baixo sentimento de pertencimento, e as soluções atuais ainda são insuficientes.
Elvira-Zorzo, Gandarillas e Martí-González (2025)	Mulheres universitárias relatam maiores dificuldades psicossociais, menor autonomia e mais problemas de saúde mental no processo de aprendizagem do que homens.

Fonte	Descrição
Feige et al. (2025)	O autoconceito matemático das crianças foi influenciado principalmente pelas expectativas e encorajamento dos pais, especialmente dos pais homens, com efeitos mais fortes sobre os meninos e impacto duradouro sobre as meninas.
Dulce-Salcedo, Maldonado e Sánchez (2022)	A maior exposição de alunas a professoras de STEM no ensino médio está associada a um aumento na matrícula dessas jovens em cursos universitários da área, sugerindo um efeito positivo de modelos femininos na escolha de carreira.
McGuire et al. (2021)	Interações com educadoras mulheres em espaços informais de ciência aumentam o interesse de meninas por matemática e reduzem estereótipos de que meninos são melhores na área, evidenciando o papel positivo de modelos femininos.

2.4 Explicabilidade e Interpretabilidade em Modelos de AM

1. Conceitos de Explainable Artificial Intelligence (XAI) Explainable Artificial Intelligence (XAI) refere-se à capacidade de sistemas de inteligência artificial de fornecer explicações claras, compreensíveis e significativas sobre seus processos de decisão. Segundo o relatório *EDPS TechDispatch on Explainable Artificial Intelligence* (European Data Protection Supervisor, 2021), muitos modelos modernos, especialmente aqueles baseados em aprendizagem profunda, operam como verdadeiras "caixas-pretas", dificultando a compreensão de sua lógica interna tanto por usuários quanto pelos próprios engenheiros responsáveis. Essa opacidade pode ocultar vieses, erros ou correlações espúrias, gerando riscos significativos para indivíduos afetados por decisões automatizadas. Nesse contexto, o XAI busca tornar o comportamento dos modelos mais acessível ao ser humano, promovendo transparência, interpretabilidade e accountability. O documento destaca que a explicabilidade deve permitir compreender competências do sistema, justificar decisões específicas e revelar informações relevantes sobre o processo decisório.

Os princípios de transparência, interpretabilidade e explicabilidade constituem elementos centrais no desenvolvimento de sistemas de Inteligência Artificial responsáveis.

Conforme destaca o relatório *EDPS TechDispatch on Explainable Artificial Intelligence* (European Data Protection Supervisor, 2021), a transparência refere-se à capacidade de compreender o funcionamento geral do sistema, sua finalidade, seus limites e as condições sob quais suas decisões são produzidas, permitindo que usuários e autoridades saibam "o que o sistema faz" e "como o faz". A interpretabilidade, por sua vez, diz respeito ao grau em que seres humanos conseguem entender as relações entre entradas, processamento interno e saídas do modelo, reduzindo o efeito de "caixa-preta" associado a abordagens opacas de aprendizado de máquina. Já a explicabilidade envolve fornecer justificativas claras, significativas e contextualizadas para decisões específicas tomadas pelo sistema, revelando as razões e fatores que contribuíram para determinado resultado.

2. SHAP como métrica de contribuição e interpretabilidade

Os valores SHAP (SHapley Additive exPlanations) constituem um método de explicabilidade baseado na teoria dos valores de Shapley, permitindo quantificar a contribuição individual de cada variável para a predição de um modelo. Conforme apresentado por *Nanal2024*, a previsão do modelo para a instância i pode ser decomposta como a soma aditiva das contribuições de cada variável, expressa pela Equação (2):

$$\hat{y}_i = shap_0 + shap(X_{1i}) + shap(X_{2i}) + \cdots + shap(X_{ji}),$$

na qual $\hat{y}_i$ representa a predição do modelo para o catchment i, enquanto $shap(X_{ji})$ corresponde ao valor SHAP associado à j-ésima variável dessa instância. O termo $shap_0$ é definido na Equação (3),

$$shap_0 = E(\hat{y}_i),$$

sendo a média global das predições em todos os catchments. Assim, como descreve explicitamente o artigo, os valores SHAP permitem interpretar a saída do modelo ao decompor sua predição em efeitos individuais atribuídos a cada variável, fornecendo transparência e suporte à interpretabilidade no contexto de Explainable Artificial Intelligence (XAI).

3. Uso de XAI para entender vieses de gênero, justiça algorítmica e tomadas de decisão baseadas em dados.

4. Explicar Random Forest e seu uso para detectar as variáveis importantes com SHAP

Capítulo 3

Metodologia

Este estudo adota uma abordagem quantitativa para investigar o fenômeno da escolha de cursos no ensino superior. O objetivo é mensurar características e comportamentos de estudantes, com foco especial em compreender por que alguns indivíduos optam por áreas STEM (Ciência, Tecnologia, Engenharia e Matemática) enquanto outros escolhem cursos de outras áreas do conhecimento. A pesquisa focará na identificação de padrões e na construção de perfis de estudantes com base em suas escolhas. Trata-se de um estudo transversal, o que significa que a coleta de dados ocorrerá em um único período de tempo, oferecendo um panorama das variáveis em questão no momento da pesquisa.

A população-alvo desta pesquisa é composta por estudantes do ensino superior, matriculados em diferentes áreas do conhecimento (incluindo STEM e não-STEM), nas instituições públicas e privadas localizadas na cidade de Salgueiro-PE. A amostra será definida por conveniência, buscando a participação voluntária dos alunos. Todos os participantes serão informados sobre os objetivos do estudo e terão sua anonimidade e confidencialidade, em conformidade com a LGPD.

Os dados serão coletados por meio de um questionário estruturado e autoadministrado, aplicado de forma online via plataforma autoral disponibilizada na web. O instrumento será composto majoritariamente por questões objetivas e fechadas, organizadas em quatro eixos temáticos principais: influências familiares, influências educacionais, fatores psicológicos e características socioeconômicas e demográficas. Essas seções são projetadas para coletar informações detalhadas que abordam os fatores que influenciam as escolhas de curso (sejam eles na área STEM ou não-STEM), além de outras variáveis relevantes para a formação dos perfis dos estudantes. Haverá questões abertas apenas para os casos em que for necessário compreender o conceito ou a forma de pensamento dos participantes sobre determinado assunto.

Antes da análise, os dados coletados passarão por um processo de pré-processamento, essencial para garantir a qualidade e a consistência dos resultados. Inicialmente, será

realizada uma inspeção para identificar e tratar valores ausentes, inconsistentes ou duplicados. As variáveis numéricas serão normalizadas utilizando a técnica de normalização min-max, a fim de padronizar a escala dos dados entre 0 e 1, prevenindo que variáveis com escalas maiores dominem os algoritmos de agrupamento.

Grande parte das respostas às questões fechadas segue uma escala ordinal de intensidade ou frequência (por exemplo, de "não gostava nem me identificava"até "sempre tive muito interesse"), o que permite sua transformação em valores numéricos com base na ordem natural das alternativas. Essa conversão preservará a semântica da resposta, permitindo que tais variáveis sejam utilizadas diretamente nos métodos quantitativos, como o agrupamento.

As variáveis nominais, como sexo, modalidade de ensino ou instituição, serão codificadas por meio de one-hot encoding, possibilitando a inclusão dessas informações nos algoritmos que exigem entrada numérica. As respostas abertas — como as relativas às percepções sobre ciência e matemática — serão analisadas qualitativamente ou, caso haja volume suficiente e padrão de recorrência, poderão ser categorizadas para análises complementares. Dependendo da consistência e representatividade dessas respostas, elas poderão ser incluídas ou não na etapa de clusterização.

Após o pré-processamento, será conduzida uma análise estatística descritiva para caracterizar o perfil geral da amostra. Serão calculadas frequências absolutas e relativas, medidas de tendência central (como média e mediana) e de dispersão (como desvio padrão), conforme o tipo de variável. Os dados serão apresentados por meio de tabelas e gráficos descritivos, incluindo gráficos de barras, histogramas e diagramas de setores, facilitando a visualização e interpretação das principais características dos participantes.

Essa etapa visa identificar tendências gerais da amostra, como proporção de estudantes em áreas STEM e não-STEM, níveis de interesse por disciplinas de exatas, influência familiar percebida e variáveis socioeconômicas. Também serão observadas correlações iniciais entre as variáveis, com o intuito de levantar hipóteses para as etapas posteriores de análise.

Para a construção dos perfis de estudantes, serão aplicadas técnicas de agrupamento (clustering), com o objetivo de identificar grupos com características semelhantes em relação às variáveis investigadas. Inicialmente, serão realizados testes exploratórios com diferentes métodos de agrupamento, como Análise Hierárquica e Mapas Auto-organizáveis (Self-Organizing Maps — SOM). A escolha do método definitivo será guiada pela qualidade dos agrupamentos gerados, interpretabilidade dos resultados e compatibilidade com os dados.

A definição do número ideal de grupos será feita com base em métricas de avaliação como o método do cotovelo (elbow method), que observa a redução da inércia intra-

grupo em função do número de clusters, e o índice de silhueta (silhouette score), que avalia a coesão interna e separação entre os grupos. As variáveis utilizadas no processo de clusterização incluirão fatores como perfil acadêmico (interesse por disciplinas, valorização escolar), influências familiares, trajetória escolar, indicadores socioeconômicos e psicológicos.

Após a definição dos clusters, será realizada uma análise estatística inferencial com o objetivo de verificar se há diferenças estatisticamente significativas entre os grupos formados. Para isso, serão aplicados testes como:

- **Teste do qui-quadrado de independência**, para verificar associações entre variáveis categóricas (ex: área do curso e grupo de perfil);
- **Análise de variância (ANOVA)** ou o **teste de Kruskal-Wallis** (caso os dados não atendam aos pressupostos de normalidade), para comparar médias de variáveis numéricas entre os clusters;
- **Correlação de Spearman**, quando pertinente, para avaliar relações entre variáveis ordinais.

Essas análises permitirão não apenas validar os agrupamentos encontrados, mas também oferecer uma compreensão mais profunda das distinções entre os perfis de estudantes, destacando quais fatores contribuem mais significativamente para as escolhas de carreira no ensino superior.

Capítulo 4

RESULTADOS E DISCUSSÕES

4.1 Caracterização da Amostra e Análise Exploratória dos Dados

A pesquisa foi desenvolvida por meio da aplicação de um questionário estruturado, direcionado a estudantes do Ensino Superior público da cidade de Salgueiro, localizada no Sertão Central de Pernambuco. Ao todo, foram obtidas 88 respostas válidas, constituindo a base de dados utilizada nas análises subsequentes. O instrumento contemplou questões voltadas à trajetória de interesse dos respondentes em áreas STEM, tanto ao longo do tempo quanto no momento atual, bem como perguntas que investigavam fatores associados ao afastamento (ou permanência) nessas áreas. Além disso, o questionário abrangeu aspectos relacionados às motivações que influenciaram a escolha do curso superior escolhido e incluiu variáveis demográficas, como sexo, escolaridade dos pais, entre outras. – explicar tratamento dos dados

Figura 4.1: Legenda

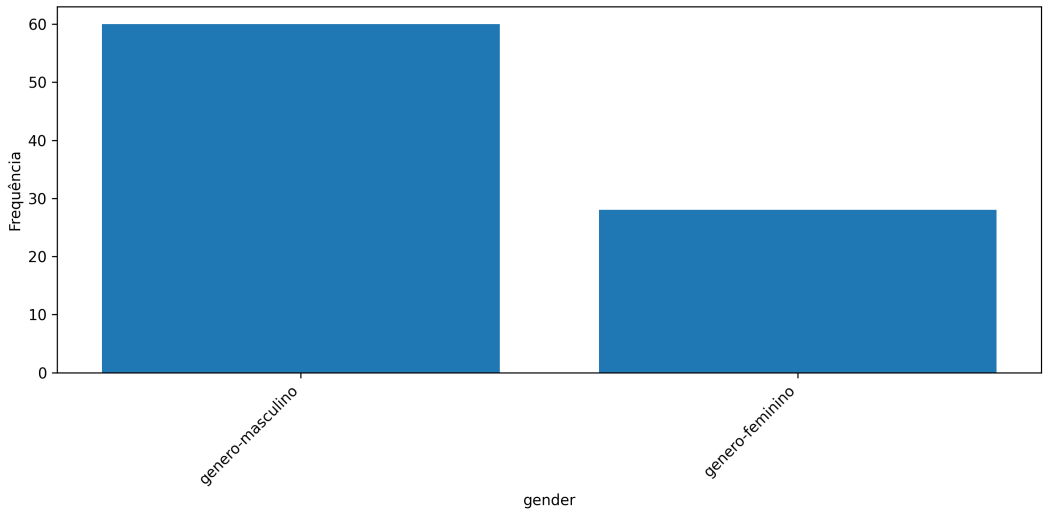


Figura 4.2: Legenda

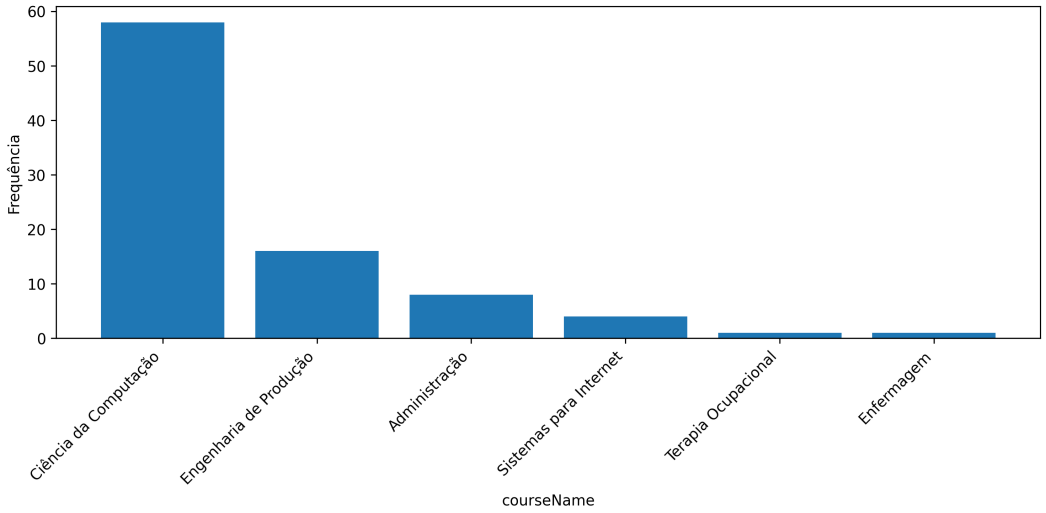
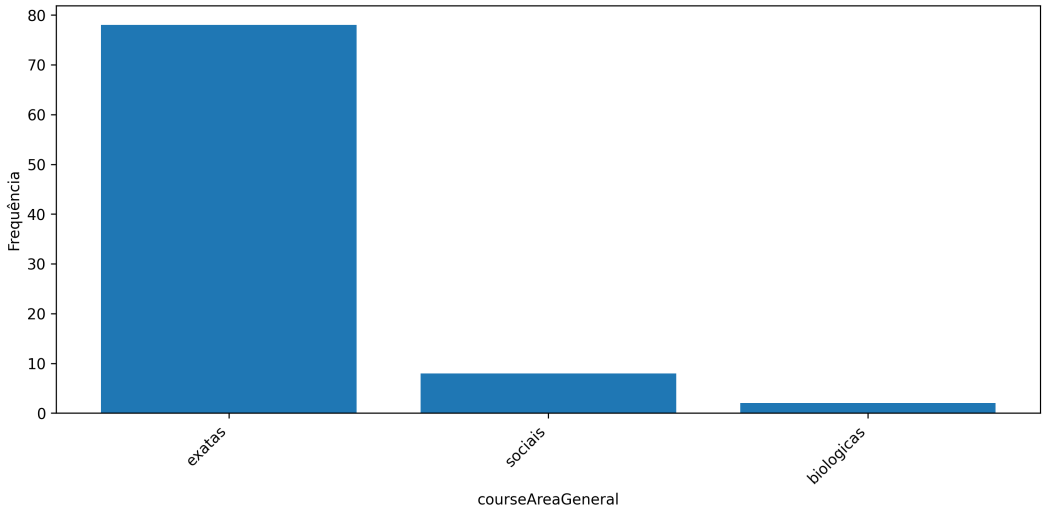


Figura 4.3: Legenda



4.1.1 Estrutura das perguntas

Informações Pessoais

Tabela 4.1: Fichamento das questões – Informações Pessoais

Questionamento	**Tipo**	**Opções**	**Feature**
Me sinto excluído(a) em conversas sobre tecnologia.	Seleção única	Escala likert	feelingExcludedTech

Questionamento	**Tipo**	**Opções**	**Feature**
Já deixei de participar de alguma atividade por achar que não era comum para o gênero com o qual me identifico.	Seleção única	Escala likert	activityGenderRestriction
Na sua opinião, existem profissões que são mais adequadas para homens e outras para mulheres?	Seleção única	Escala likert	professionsByGenderOpinion
De modo geral, ao longo da sua trajetória escolar, você sentia que seus professores valorizavam seu desempenho?	Seleção única	Sim, a maioria valorizava; Alguns valorizavam; Neutro; Poucos valorizavam; Desempenho ignorado	teacherPerformanceValue
Antes de entrar no ensino superior, você se interessava por disciplinas da área de exatas?	Seleção única	Muito interesse; Algum interesse; Pouco interesse; Nenhum interesse	preCollegeExactInterestLevel
Minha família me incentiva mais a continuar meus estudos do que a começar a trabalhar logo.	Seleção única	Escala likert	familyStudyIncentive
Tenho interesse em me tornar um(a) profissional nas áreas de ciência, engenharia ou tecnologia.	Seleção única	Escala likert	stemCareerInterest
Durante sua infância e adolescência, você sentia que tinha tempo e condições para praticar seus hobbies e atividades de interesse pessoal?	Seleção única	Sim; Às vezes; Raramente; Não	childhoodHobbiesTime

4.2 Modelo de Classificação e Análise das Variáveis mais Importantes

4.2.1 analise com ambos os generos

Para esta etapa da análise, foi realizada uma seleção prévia de variáveis, mantendo-se apenas aquelas diretamente relacionadas a STEM e aos fatores identificados na literatura como influenciadores do interesse — ou da falta de interesse — por carreiras nessa área. Essa filtragem inicial teve o objetivo de garantir que o modelo analisasse exclusivamente os elementos associados ao interesse em STEM, permitindo que a interpretação com SHAP refletisse apenas esses fatores específicos, sem interferência de variáveis externas ao fenômeno investigado. Além disso, considerando o número reduzido de respondentes (n = 88), optou-se por trabalhar com apenas oito variáveis, assegurando maior estabilidade nas estimativas e evitando sobreajuste.

Após a definição desse subconjunto de variáveis, o conjunto de dados é carregado e a variável-alvo gender é separada das demais variáveis explicativas selecionadas. Em seguida, os dados são divididos em treino e teste, e o modelo de Random Forest é ajustado com base nesse conjunto filtrado. Uma vez treinado, o modelo é interpretado por meio dos valores SHAP, aplicados sobre as amostras de teste, permitindo identificar de forma transparente como cada um desses fatores relacionados ao interesse por STEM contribui para as predições da variável gender.

Figura 4.4: Legenda

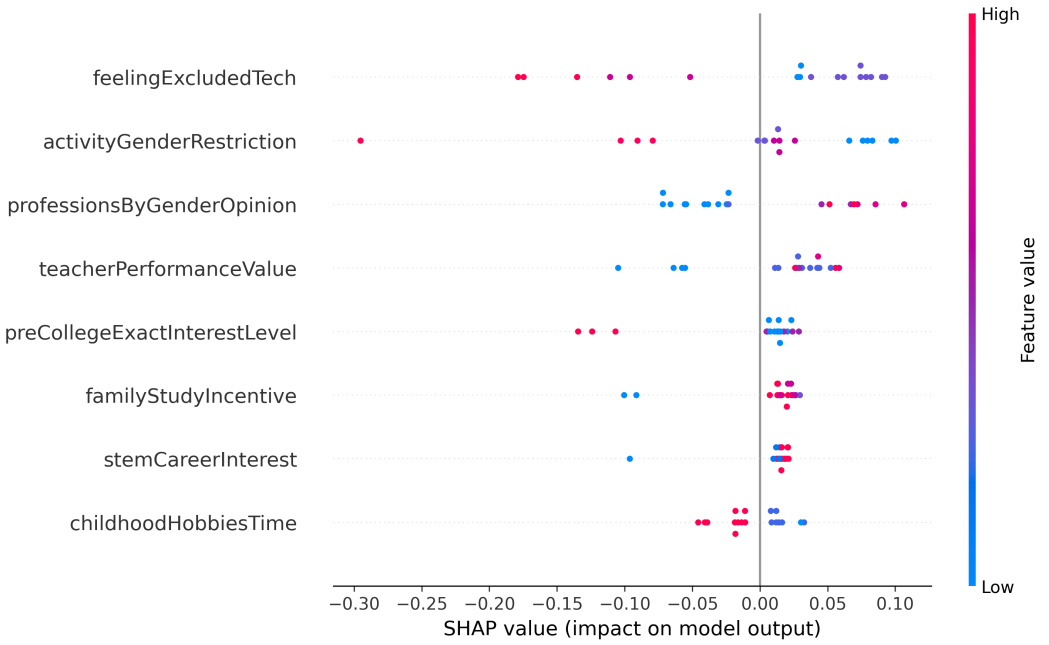


Tabela 4.2: Tabela de tradução da variável gender

Valor Original	**Codificado**	**Normalizado**
genero-feminino	0	0.00
genero-masculino	1	1.00

Tabela 4.3: Tabela de tradução da variável feelingExcludedTech

Valor Original	**Codificado**	**Normalizado**
1	0	0.00
2	1	0.33
3	2	0.67
4	3	1.00

Tabela 4.4: Tabela de tradução da variável activityGenderRestriction

Valor Original	**Codificado**	**Normalizado**
1	0	0.00
2	1	0.25
3	2	0.50
4	3	0.75
5	4	1.00

Tabela 4.5: Tabela de tradução da variável professionsByGenderOpinion

Valor Original	**Codificado**	**Normalizado**
1	0	0.00
2	1	0.25
3	2	0.50
4	3	0.75
5	4	1.00

Tabela 4.6: Tabela de tradução da variável teacherPerformanceValue

Valor Original	**Codificado**	**Normalizado**
alguns-valorizavam	0	0.25
desempenho-ignorado	1	1.00
maioria-valorizava	2	0.00
poucos-valorizavam	3	0.75
tratavam-neutro	4	0.50

Tabela 4.7: Tabela de tradução da variável preCollegeExactInterestLevel

Valor Original	**Codificado**	**Normalizado**
muito-interesse	1	0.00
algum-interesse	0	0.33
pouco-interesse	3	0.67
nenhum-interesse	2	1.00

Tabela 4.8: Tabela de tradução da variável familyStudyIncentive

Valor Original	**Codificado**	**Normalizado**
1	0	0.00
2	1	0.25
3	2	0.50
4	3	0.75
5	4	1.00

Tabela 4.9: Tabela de tradução da variável stemCareerInterest

Valor Original	**Codificado**	**Normalizado**
1	0	0.00
2	1	0.25
3	2	0.50
4	3	0.75
5	4	1.00

Tabela 4.10: Tabela de tradução da variável childhoodHobbiesTime

Valor Original	Codificado	Normalizado
as-vezes	0	0.67
nao-raremente	1	0.00
raramente-responsabilidades	2	0.33
sim-tempo-apoio	3	1.00

A análise inicial foi conduzida utilizando a base completa de respondentes, sem separação por gênero, tomando como variável alvo o próprio gênero (0 = feminino, 1 = masculino). O gráfico de SHAP resultante oferece uma visão global dos principais fatores que distinguem os respondentes do gênero feminino e masculino, permitindo observar tanto a relevância relativa de cada variável quanto a direção de seus efeitos.

Interpretação do Padrão de Cores no Gráfico de SHAP

Nos gráficos de interpretação de SHAP, os pontos são coloridos conforme o valor original da feature: valores altos são representados na cor vermelha, enquanto valores baixos são representados em azul. No caso da variável alvo "gênero", valores baixos correspondem ao gênero feminino e valores altos ao masculino. Dessa forma, pontos azuis no gráfico indicam perfis associados ao público feminino, enquanto pontos vermelhos representam perfis associados ao público masculino. Essa distinção visual permite compreender a direção das relações entre as variáveis explicativas e a probabilidade de o modelo predizer um determinado gênero.

Sentimento de Exclusão em Tecnologia como Fator Mais Relevante

A variável mais importante identificada pelo modelo para diferenciar pessoas de gêneros distintos foi o *sentimento de exclusão em áreas de tecnologia*. Os valores de SHAP indicam que valores mais altos dessa variável (representando maior sensação de exclusão) contribuem para aumentar a probabilidade de o respondente ser classificado como feminino, enquanto valores baixos estão principalmente associados ao público masculino. Esse resultado é compatível com a literatura especializada, que aponta que meninas e mulheres tendem a vivenciar com maior frequência percepções de exclusão, insegurança ou desestímulo relacionados a ambientes de tecnologia e STEM.

Restrição de Atividades por Motivo de Gênero

A segunda variável mais relevante corresponde à ocorrência de restrições de atividades devido ao gênero. Valores altos dessa variável — indicando que a pessoa já deixou

de realizar alguma atividade por questões de gênero — foram mais frequentes entre respondentes do gênero feminino. Entretanto, observou-se que parte dos respondentes masculinos também sinalizou valores altos, o que pode indicar uma interpretação distinta da pergunta ou uma compreensão mais abrangente sobre o conceito de restrição. Esse comportamento diferencia-se de achados já consolidados na literatura e aponta para possíveis ambivalências na interpretação da questão.

Opinião sobre Profissões por Gênero

A terceira variável de maior importância foi a opinião acerca da existência de profissões consideradas mais adequadas a um gênero específico. Valores elevados — indicando maior concordância com estereótipos de gênero em profissões — foram mais associados ao público masculino, enquanto valores baixos — indicando menor concordância — foram mais associados ao público feminino. Esse resultado reforça achados prévios de que mulheres tendem a rejeitar mais intensamente concepções estereotipadas sobre papéis profissionais.

Incentivo Familiar: Estudo para Meninas e Trabalho para Meninos

Outra variável de destaque refere-se ao *incentivo familiar*. A análise dos valores de SHAP mostra que meninas tendem a relatar maior incentivo familiar para estudar, continuar os estudos e seguir trajetórias acadêmicas mais longas. Entre os meninos, entretanto, o incentivo familiar aparece mais frequentemente associado à expectativa de trabalhar cedo ou priorizar atividades laborais. Esse contraste reflete padrões socioculturais observados em diversas regiões do país, onde há uma tendência de atribuir às meninas maior responsabilidade escolar e, aos meninos, maior responsabilidade econômica. Assim, o modelo evidencia que o tipo de incentivo familiar recebido constitui um dos fatores relevantes na diferenciação entre os gêneros.

Interesse em Seguir Carreira em STEM

A variável *stemCareerInterest* surgiu como a sétima mais relevante para distinção entre os gêneros. Os valores de SHAP mostram que meninos tendem a responder com maior segurança e convicção quanto ao interesse em seguir carreira em STEM, frequentemente optando por respostas de concordância máxima. Já entre as meninas, observa-se maior hesitação: embora muitas indiquem interesse, poucas selecionam opções de certeza absoluta. Esse comportamento pode sugerir menor autoconfiança, menor percepção de pertencimento ou maior cautela ao declarar intenções futuras em áreas de STEM, o que está alinhado com evidências amplamente discutidas na literatura de gênero e educação em ciências.

4.2.2 Análise Comparativa por Gênero: Modelos para Mulheres e Homens

Para aprofundar a compreensão das diferenças entre percepções e fatores associados ao interesse em STEM, os dados foram separados por gênero. Para cada grupo foi treinado um modelo de Random Forest com o objetivo de prever o interesse em seguir carreira em STEM, e em seguida aplicaram-se gráficos de SHAP individuais. Essa abordagem permite observar como diferentes fatores influenciam de maneira distinta o interesse de mulheres e homens por carreiras STEM.

Similaridades Entre os Modelos: Interesse Prévio em STEM

Em ambos os modelos (feminino e masculino), a variável correspondente ao interesse prévio em STEM aparece como um dos fatores de maior impacto na previsão da intenção de carreira em STEM. Essa relação é intuitiva, uma vez que estudantes que demonstram afinidade e gosto pela área tendem naturalmente a considerar uma carreira no mesmo campo. Além disso, em oitavo lugar nos dois modelos, aparece a variável que avalia a importância do reconhecimento acadêmico por parte de professores. Essa variável indica que sentir-se reconhecido, incentivado ou valorizado pelos docentes está positivamente associado ao interesse por seguir uma trajetória em STEM, tanto para meninas quanto para meninos, sugerindo que práticas pedagógicas de apoio podem desempenhar papel relevante na motivação dos estudantes.

Sentimento de "STEM Não é Para Mim"

Um dos achados mais expressivos está na variável *feltSTEMNotForMe*. Esse item aparece como a terceira variável mais importante entre as meninas, e como a quinta entre os meninos. Esse resultado indica que o sentimento de que STEM "não é para elas" exerce influência significativamente mais forte sobre a decisão das meninas. Esse achado está em consonância com a análise global, onde sentimentos de exclusão se mostraram fortemente associados ao gênero feminino. Embora o valor absoluto dessa variável não apresente diferenças tão acentuadas entre meninas que gostam ou não de STEM, o modelo consegue identificar nuances suficientes que tornam essa percepção mais determinante para elas.

Diversidade dos Gráficos e Diferença no Volume de Respostas

O gráfico de SHAP masculino apresenta maior diversidade e espalhamento de pontos, decorrente da maior quantidade de respostas de pessoas do gênero masculino. Esse desequilíbrio amostral influencia a variabilidade visual representada no gráfico, tornando-o mais denso e heterogêneo.

Estereótipos de Profissão e Interesse Masculino em STEM

Entre os meninos, a variável relacionada à opinião sobre profissões por gênero aparece como a segunda mais importante. O SHAP indica que meninos interessados em STEM tendem a reproduzir menos estereótipos de gênero, enquanto aqueles com baixo interesse na área tendem a concordar mais com tais estereótipos. Esse achado sugere que visões mais igualitárias sobre papéis profissionais estão associadas a maior inclinação masculina para carreiras STEM.

Sentimento de Exclusão: Presente no Modelo Masculino, Ausente no Feminino

Outro ponto relevante é que a variável *feelingExcludedTech* aparece entre as principais no modelo masculino, mas não figura entre as mais relevantes para o modelo feminino. Isso ocorre porque a grande maioria das meninas — inclusive aquelas que afirmam gostar de STEM — indica algum nível de sentimento de exclusão em áreas tecnológicas. Assim, dentro do grupo feminino, essa variável apresenta baixa variabilidade, tornando-se menos útil para distinguir quem tem interesse em STEM de quem não tem. Já entre os meninos, essa percepção apresenta diferenças mais marcantes, o que faz com que a variável contribua de forma mais significativa para o modelo masculino.

Admiração por Professores de Exatas

Na análise feminina, destaca-se a presença da variável *admiredExactTeacherGender*, ausente entre as mais relevantes no modelo masculino. Esse resultado indica que meninas que se enxergam em STEM tendem a admirar tanto professores homens quanto professoras mulheres de áreas de exatas. Esse achado reforça a importância de modelos de referência — especialmente figuras femininas — para que meninas se sintam representadas e visualizem sua possibilidade de pertencimento em carreiras STEM.

4.3 Formação e Interpretação dos Grupos

Figura 4.5: Legenda

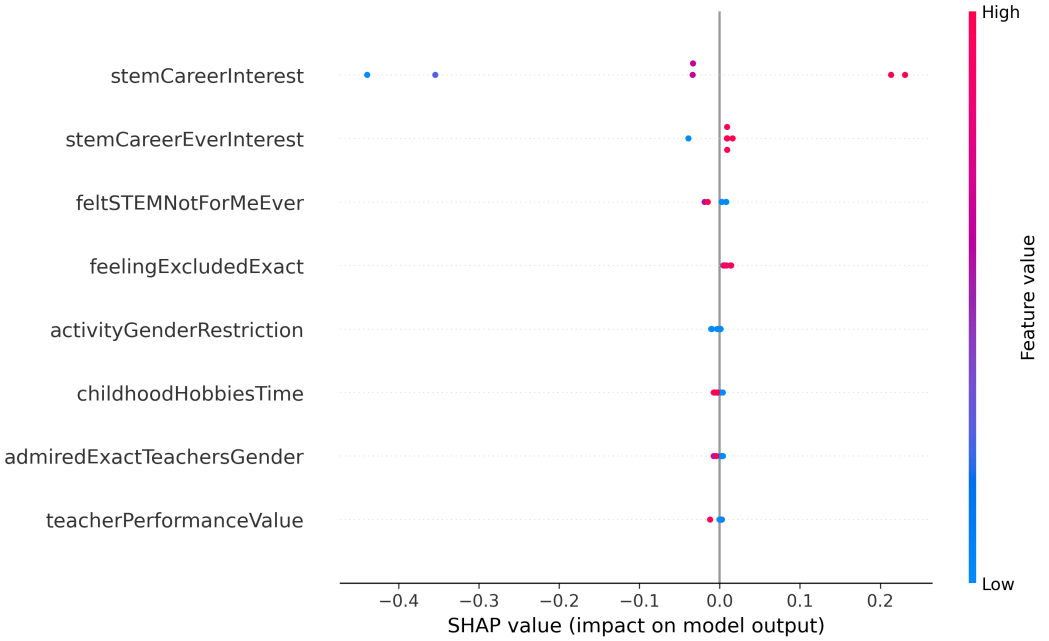
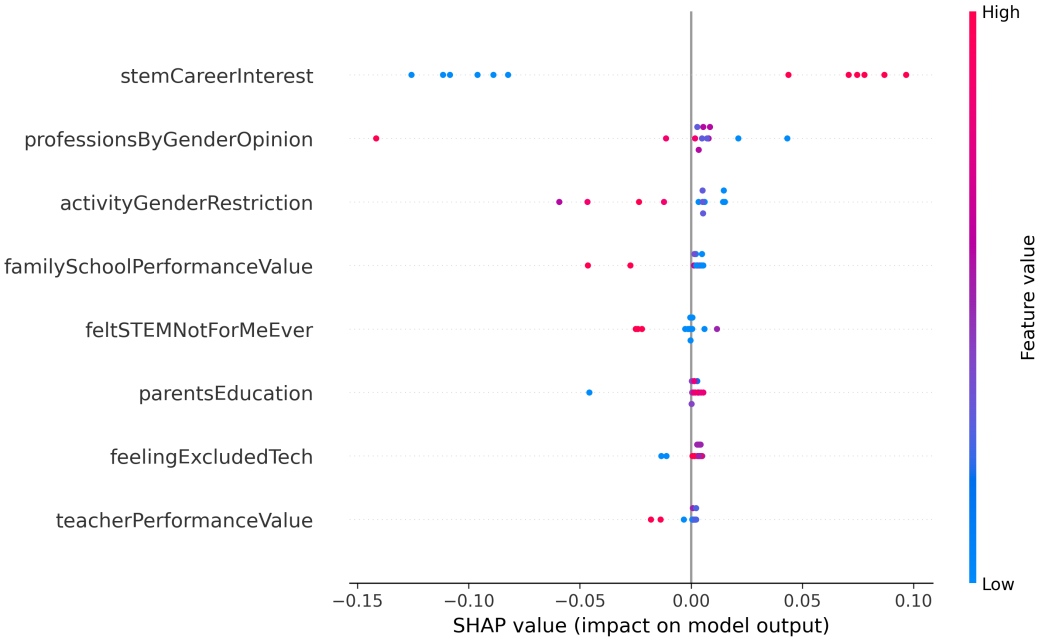


Figura 4.6: Legenda



Capítulo 5

CONSIDERAÇÕES FINAIS

1. Retomar brevemente a motivação do estudo
2. Retomar os objetivos alinhando-os ao que foi atendido
3. Reforçar os métodos utilizados e os principais achados dos resultados e discussões
4. Indicar as principais limitações da pesquisa
5. Sumarizar as principais contribuições da pesquisa
6. Indicar perspectivas e trabalhos futuros.

Referências Bibliográficas

ARSHAD, A.; ARSHAD, A. Navigating educational pathways: How collectivistic cultural norms shape educational choices in college students. *Journal of Professional & Applied Psychology*, v. 5, n. 4, p. 697–710, 2024. Accepted 28 September 2024. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/josi.12655>.

AUTHOR, A.; AUTHOR, B. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Journal of Clustering Methods*, 2020. Placeholder entry — please replace with full bibliographic details.

BOHRNSTEDT, G. W. et al. Mathematics motivation and mathematics performance: Does gender play a role? *AERA Open*, v. 10, n. 1, p. 1–17, 2024. Disponível em: <https://doi.org/10.1177/23328584241298272>.

BRASIL. *Censo da Educação Superior 2023: notas estatísticas*. Brasília, DF, 2024.

COMPUTAçãO, B. em Ciência da. *Projeto Pedagógico do Curso (PPC): Bacharelado em Ciência da Computação*. 2021. Disponível em: https://portais.univasf.edu.br/ccicomp/curso/projeto-pedagogico-do-curso-ppc. Acesso em: 24 jun. 2025. Aprovado em 1 de julho de 2021. Disponível em: <https://portais.univasf.edu.br/ccicomp/curso/projeto-pedagogico-do-curso-ppc>.

DULCE-SALCEDO, O. V.; MALDONADO, D.; SáNCHEZ, F. Is the proportion of female stem teachers in secondary education related to women's enrollment in tertiary education stem programs? *International Journal of Educational Development*, v. 91, p. 102591, 2022. Disponível em: <https://doi.org/10.1016/j.ijedudev.2022.102591>.

ELVIRA-ZORZO, M. N.; GANDARILLAS, M. Ángel; MARTí-GONZáLEZ, M. Psychosocial differences between female and male students in learning patterns and mental health-related indicators in stem vs. non-stem fields. *Social Sciences*, v. 14, n. 2, p. 71, 2025. Disponível em: <https://doi.org/10.3390/socsci14020071>.

EMRAN, A. et al. Understanding students' perceptions of the nature of science in the context of their gender and their parents' occupation. *Science & Education*, v. 29, p. 237–261, 2020. Disponível em: <https://doi.org/10.1007/s11191-020-00103-z>.

European Data Protection Supervisor. *EDPS TechDispatch on Explainable Artificial Intelligence*. 2021. <https://edps.europa.eu>. Placeholder entry — replace with the authoritative reference/URL.

FEIGE, P. et al. Impact of mothers' and fathers' math self-concept of ability, child-specific beliefs and behaviors on girls' and boys' math self-concept of

ability. *PLOS ONE*, v. 20, n. 2, p. e0317837, 2025. Disponível em: <https://doi.org/10.1371/journal.pone.0317837>.

GENCEL-AUGUSTO, J. et al. Underrepresentation of hispanic women in science, technology, engineering, mathematics, and medicine. *CA: A Cancer Journal for Clinicians*, v. 75, n. 1, p. 1–20, 2025. Accepted 15 October 2024. Disponível em: <https://doi.org/10.3322/caac.21875>.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. [S.l.]: Morgan Kaufmann, 2011.

HSIEH, T.-Y.; SIMPKINS, S. D. Longitudinal associations between parent degree/occupation, parent support, and adolescent motivational beliefs in stem. *Journal of Adolescence*, v. 94, n. 5, p. 728–747, 2022. Disponível em: <https://doi.org/10.1002/jad.12059>.

JUNIOR, J. C. G. et al. AnÁlise de dados educacionais: Como a tecnologia pode ser usada para obter insights sobre o desempenho dos alunos. *Revista Contemporânea*, v. 3, n. 8, p. 11056–11072, 2023. Accepted 08/08/2023.

LIU, R.; LIU, C.; HE, P. Chinese grades 1–9 students' views of the nature of science: Do they differ by grade level, gender, and parents' occupation? *Science & Education*, 2024. Disponível em: <https://doi.org/10.1007/s11191-024-00519-x>.

MARTINS, D. D. et al. Clusterização do perfil de adolescentes escolares com predisposição ao uso de substância psicoativas. *Research, Society and Development*, v. 10, n. 2, p. e37510212528, 2021. Publicado: 19/02/2021.

MASTER, A.; MELTZOFF, A. N.; CHERYAN, S. Gender stereotypes about interests start early and cause gender disparities in computer science and engineering. *Proceedings of the National Academy of Sciences*, v. 118, n. 48, p. e2100030118, 2021. Disponível em: <https://doi.org/10.1073/pnas.2100030118>.

MASTER, A. et al. Gender equity and motivational readiness for computational thinking in early childhood. *Early Childhood Research Quarterly*, v. 64, p. 242–254, 2023. Disponível em: <https://doi.org/10.1016/j.ecresq.2023.03.004>.

MCGUIRE, L. et al. Gender stereotypes and peer selection in stem domains among children and adolescents. *Sex Roles*, v. 87, p. 455–470, 2022. Disponível em: <https://doi.org/10.1007/s11199-022-01327-9>.

MCGUIRE, L. et al. Science and math interest and gender stereotypes: The role of educator gender in informal science learning sites. *Frontiers in Psychology*, v. 12, p. 503237, 2021. Disponível em: <https://doi.org/10.3389/fpsyg.2021.503237>.

MENEZES, S. K. de O.; SANTOS, M. D. F. dos. Gender in computer education in brazil and the entry of girls into the area - a systematic review of literature. *Revista Brasileira de Informática na Educação*, v. 29, p. 456–484, 2021. Disponível em: <https://doi.org/10.5753/RBIE.2021.29.0.456>.

MILLER, D. I. et al. The development of children's gender stereotypes about stem and verbal abilities: A preregistered meta-analytic review of 98 studies. *Psychological Bulletin*, v. 150, n. 12, p. 1363–1396, 2024. Disponível em: <https://doi.org/10.1037/bul0000456>.

MORALES, D. X.; GRINESKI, S. E.; COLLINS, T. W. Effects of mentor-mentee discordance on latinx undergraduates' intent to pursue graduate school and research productivity. *Annals of the New York Academy of Sciences*, v. 1499, n. 1, p. 54–69, 2021. Disponível em: <https://doi.org/10.1111/nyas.14602>.

OLIVEIRA, A. L. M. de. Perfil dos estudantes de graduação entre 2001 e 2015: uma revisão. *Avaliação (Campinas)*, v. 26, n. 1, p. 237–252, 2021. Aprovado em: 13 de novembro de 2020.

OLIVEIRA, P. L. S. de et al. Identificação de pesquisas e análise de algoritmos de clusterização para a descoberta de perfis de engajamento. *Revista Brasileira de Informática na Educação*, v. 30, p. 01–19, 2022. Published: 13/Feb/2022.

OPPS, Z.; YADAV, A. Who belongs in computer science? In: *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education (SIGCSE '22)*. ACM, 2022. p. 383. Disponível em: <https://doi.org/10.1145/3478431.3499301>.

PARK, J. et al. Occupational aspirations and academic achievement: Rethinking the direction of effects and the role of socioeconomic status in middle childhood and adolescence. *Journal of Social Issues*, v. 80, n. 4, p. 1408–1432, 2024. Accepted 20 November 2024. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/josi.12655>.

SCHEL, J.; DRECHSEL, B. A latent profile analysis for teacher education students' learning: an overview of competencies in self-regulated learning. *Frontiers in Psychology*, v. 16, n. 1527438, 2025. ACCEPTED 07 April 2025.

SILVA, U. F. et al. Problemas enfrentados por alunas de graduação em ciência da computação: uma revisão sistemática. *Educação em Revista (Educ. Pesqui.)*, v. 48, p. e236643, 2022. Disponível em: <https://doi.org/10.1590/S1678-4634202248236643>.

SU, R.; PUTKA, D. J.; ROUNDS, J. Computer science work and interest profiles: stereotype vs. realities. *Scientific Reports*, v. 13, p. 21910, 2023. Disponível em: <https://doi.org/10.1038/s41598-023-47963-3>.

TELLHED, U.; BJöRKLUND, F.; STRAND, K. K. Tech-savvy men and caring women: Middle school students' gender stereotypes predict interest in tech-education. *Sex Roles*, v. 88, p. 307–325, 2023. Disponível em: <https://doi.org/10.1007/s11199-023-01353-1>.

ZúñIGA-MEJíAS, V.; HUINCAHUE, J. Gender stereotypes in stem: a systemic review of studies conducted at primary and secondary school. *Educação e Pesquisa*, v. 50, p. e258677, 2024. Disponível em: <https://doi.org/10.1590/S1678-4634202450258677>.