UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO
CAMPUS SALGUEIRO - PE
CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Catarina Cysneiros Sampaio

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

Salgueiro - PE
2025

Catarina Cysneiros Sampaio

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

Trabalho de Conclusão de Curso do curso de Curso de graduação em Ciência da Computação apresentado ao Colegiado de Ciência da Computação como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.
Orientadora: Prof.ª Me. Débora da Conceição Araújo

Salgueiro - PE
2025

UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO - UNIVASF
Gabinete da Reitoria
Sistema Integrado de Bibliotecas (SIBI)
Av. José de Sá Maniçoba, s/n, Campus Universitário – Centro CEP 56304-917
Caixa Postal 252, Petrolina-PE, Fone: (87) 2101- 6760, biblioteca@univasf.edu.br

Sobrenome do autor, Prenome do autor
* Cutter Título do trabalho / Nome por extenso do autor. - local, ano.
xx (total de folhas antes da introdução em nº romano), 50 f.(total de folhas do trabalho): il. ; (caso tenha ilustrações) 29 cm.(tamanho do papel A4)

Trabalho de Conclusão de Curso (Graduação em nome do curso) - Universidade Federal do Vale do São Francisco, Campus, local, ano

Orientador (a): Prof.(a) titulação e nome do prof(a).

Notas (opcional)

1. Assunto. 2. Assunto. 3. Assunto. I. Título. II. Orientador (Sobrenome, Prenome). III. Universidade Federal do Vale do São Francisco.

* CDD

Ficha catalográfica elaborada pelo Sistema Integrado de Biblioteca SIBI/UNIVASF
Bibliotecário: Nome* e CRB*

* ***Dados inseridos pela biblioteca***

Exemplo:

Souza, José Augusto de
S729c Crianças com dificuldades de aprendizado: estudo nas escolas públicas da cidade de Juazeiro-BA / José Augusto de Souza. – Petrolina - PE, 2009.
xv, 140 f. : il. ; 29 cm.

Trabalho de Conclusão de Curso (Graduação em Psicologia) Universidade Federal do Vale do São Francisco, Campus Petrolina-PE, 2009.

Orientadora: Profª. Drª. Maria de Azevedo.

Inclui referências.

1. Crianças - Ensino. 2. Distúrbios da aprendizagem. 3. Escolas públicas – Juazeiro (BA). I. Título. II. Azevedo, Maria de. III. Universidade Federal do Vale do São Francisco.

370.15

Ficha catalográfica elaborada pelo Sistema Integrado de Biblioteca SIBI/UNIVASF
Bibliotecário: Nome e CRB.

Catarina Cysneiros Sampaio

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

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Bacharel em Ciência da Computação e aprovado em sua forma final pela banca examinadora.

Salgueiro - PE, 18 de dezembro de 2023.

Prof. Me. Ricardo Azevedo Moreira da Silva
Coordenador do Curso

Banca Examinadora:

Prof.ª Me. Débora da Conceição Araújo
Presidente da Banca

Prof. X Y Z, Me.
Avaliador
Universidade Federal do Vale do São Francisco

Prof. X Y Z, Dr.
Avaliador
Universidade Federal do Vale do São Francisco

AGRADECIMENTOS

Agradeço ao meu pai, à minha mãe e às minhas irmãs pelo apoio incondicional ao longo de toda a minha jornada acadêmica. Sem o incentivo e a compreensão de vocês, este trabalho não teria sido possível. Agradeço também à minha orientadora, Professora Débora, que não apenas foi fundamental para a realização deste trabalho, como também se tornou uma grande inspiração para mim.

RESUMO

Este trabalho investiga os fatores que influenciam o interesse e a autopercepção de estudantes universitários em relação às áreas de STEM no município de Salgueiro–PE. A pesquisa analisa como elementos psicossociais, educacionais e experiências de pertencimento moldam percepções distintas entre homens e mulheres. Para isso, são utilizadas técnicas de aprendizado de máquina explicável (Random Forest + SHAP) e métodos de clusterização para identificar perfis recorrentes entre os participantes. Os resultados apontam diferenças importantes na forma como cada gênero internaliza estereótipos, enfrenta experiências de exclusão e desenvolve interesse pela área, evidenciando desafios locais que impactam o acesso e a permanência em STEM no Sertão Central de Pernambuco.

Palavras-chave: STEM. Gênero. XAI. SHAP. Random Forest. Clusterização. Pertencimento acadêmico.

ABSTRACT

This study investigates the factors that influence university students' interest and self-perception regarding STEM fields in the municipality of Salgueiro, Brazil. The research examines how psychosocial, educational, and belonging-related experiences shape gender-specific perceptions. Explainable machine learning techniques (Random Forest + SHAP) and clustering methods are applied to identify recurring student profiles. The results reveal meaningful differences in how men and women internalize stereotypes, experience exclusion, and develop interest in STEM, highlighting local challenges that affect access and persistence in these fields within the Sertão Central region of Pernambuco.

Keywords: STEM. Gender. XAI. SHAP. Random Forest. Clustering. Academic belonging.

LISTA DE FIGURAS

Figura 1 – Número de respondentes por gênero 34
Figura 2 – Número de respondentes por curso 34
Figura 3 – Número de respondentes por área geral do curso 35
Figura 4 – Gráfico de SHAP para o modelo de classificação de gênero 36
Figura 5 – Gráfico de SHAP para o modelo de classificação sobre visão de futuro em STEM - Feminino 40
Figura 6 – Gráfico de SHAP para o modelo de classificação sobre visão de futuro em STEM - Masculino 41
Figura 7 – Dendrograma da Análise de Clusters Hierárquica 42
Figura 8 – Análise do Cotovelo para Determinação do Número de Clusters 43
Figura 9 – Gráfico de radar comparativo entre os clusters 46

LISTA DE TABELAS

Tabela 1 – Fichamento de trabalhos relacionados 28
Tabela 2 – Fichamento das questões das features identificadas pelas análises com Random Forest e SHAP values 55
Tabela 3 – Tabela de tradução da variável feelingExcludedTech 57
Tabela 4 – Tabela de tradução da variável activityGenderRestriction 57
Tabela 5 – Tabela de tradução da variável professionsByGenderOpinion 57
Tabela 6 – Tabela de tradução da variável feltSTEMNotForMeEver 57
Tabela 7 – Tabela de tradução da variável stemCareerInterest 58
Tabela 8 – Tabela de tradução da variável admiredExactTeachersGender 58
Tabela 9 – Tabela de tradução da variável schoolExactInterestByGender 58
Tabela 10 – Tabela de tradução da variável familySchoolPerformanceValue 58
Tabela 11 – Tabela de tradução da variável feelingExcludedExact 58
Tabela 12 – Tabela de tradução da variável stemCareerEverInterest 59
Tabela 13 – Tabela de tradução da variável childhoodHobbiesTime 59
Tabela 14 – Tabela de tradução da variável familyStudyIncentive 59
Tabela 15 – Tabela de tradução da variável preCollegeExactInterestLevel 59
Tabela 16 – Tabela de tradução da variável teacherPerformanceValue 59
Tabela 17 – Tabela de tradução da variável gender 60
Tabela 18 – Tabela de tradução da variável feelingExcludedTech 60
Tabela 19 – Tabela de tradução da variável activityGenderRestriction 60
Tabela 20 – Tabela de tradução da variável professionsByGenderOpinion 60
Tabela 21 – Tabela de tradução da variável feltSTEMNotForMeEver 60
Tabela 22 – Tabela de tradução da variável stemCareerInterest 61
Tabela 23 – Tabela de tradução da variável admiredExactTeachersGender 61
Tabela 24 – Tabela de tradução da variável schoolExactInterestByGender 61
Tabela 25 – Tabela de tradução da variável familySchoolPerformanceValue 61

LISTA DE ABREVIATURAS E SIGLAS

IA	Inteligência Artificial
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LGPD	Lei Geral de Proteção de Dados
OOB	*Out-of-Bag*, técnica utilizada em florestas aleatórias para estimar o erro do modelo com base nas amostras não selecionadas em cada bootstrap
RF	Random Forest, algoritmo de aprendizado de máquina baseado em árvores de decisão
SHAP	SHapley Additive exPlanations
STEM	Ciência, Tecnologia, Engenharia e Matemática (do inglês, *Science, Technology, Engineering and Mathematics*)
XAI	*Explainable Artificial Intelligence* (Inteligência Artificial Explicável)

SUMÁRIO

1 INTRODUÇÃO 12
1.1 QUESTÕES DE PESQUISA 13
1.2 OBJETIVOS 13
1.2.1 Objetivo Geral 13
1.2.2 Objetivos Específicos 13
1.3 JUSTIFICATIVA 14
1.4 ORGANIZAÇÃO DO TRABALHO 14
2 FUNDAMENTAÇÃO TEÓRICA 16
2.1 DETERMINANTES PSICOSSOCIAIS NA ESCOLHA DE CARREIRA . . 16
2.1.1 Estereótipos de Gênero e sua Influência nas Escolhas Acadêmicas 16
2.1.2 Fatores Socioeconômicos e Geográficos na Escolha Acadêmica . . 17
2.1.3 Fatores Familiares na Escolha Acadêmica 17
2.1.4 Fatores Educacionais na Escolha Acadêmica 18
2.2 TÉCNICAS DE AGRUPAMENTO COMO ESTRATÉGIA DE MAPEAMENTO DE PERFIS 19
2.3 EXPLICABILIDADE E INTERPRETABILIDADE EM MODELOS DE AM 22
2.4 TRABALHOS CORRELATOS 26
2.4.1 Análise de Dados na Educação e Estudos sobre Perfil de Estudantes 26
2.4.2 Fatores determinantes no interesse e autopercepção em STEM . . 28
3 DELINEAMENTO METODOLÓGICO 31
4 RESULTADOS 33
4.1 CARACTERIZAÇÃO DA AMOSTRA E ANÁLISE EXPLORATÓRIA DOS DADOS 33
4.2 MODELO DE CLASSIFICAÇÃO E ANÁLISE DAS VARIÁVEIS MAIS IMPORTANTES 35
4.2.1 Análise com ambos os gêneros 35
4.2.1.1 Interpretação do Padrão de Cores no Gráfico de SHAP 36
4.2.1.2 Sentimento de Exclusão em Tecnologia (*feelingExcludedTech*) 36
4.2.1.3 Restrição de Atividades por Motivo de Gênero (*activityGenderRestriction*) 37
4.2.1.4 Opinião sobre Profissões por Gênero (*professionsByGenderOpinion*) 37
4.2.1.5 Percepção de que STEM Nunca Foi Para Mim (*feltSTEMNotForMeEver*) 37
4.2.1.6 Interesse em Carreira STEM (*stemCareerInterest*) 37
4.2.1.7 Admiração por Professores (*admiredExactTeachersGender*) 37
4.2.1.8 Interesse Escolar por Gênero (*schoolExactInterestByGender*) 37
4.2.1.9 Percepção do Desempenho Escolar Familiar (*familySchoolPerformanceValue*) 38
4.2.1.10 Resumo 38

4.2.2 Análise Comparativa por Gênero: Modelos para Mulheres e Homens 38
4.2.2.1 Diversidade dos Gráficos e Diferença no Volume de Respostas 38
4.2.2.2 Similaridades Entre os Modelos: Interesse Prévio em STEM 38
4.2.2.3 Sentimento de "STEM Não é Para Mim" 39
4.2.2.4 Estereótipos de Profissão e Interesse Masculino em STEM 39
4.2.2.5 Sentimento de Exclusão: Presente no Modelo Masculino, Ausente no Feminino 39
4.2.2.6 Admiração por Professores de Exatas 40
4.3 FORMAÇÃO E INTERPRETAÇÃO DOS GRUPOS 41
4.3.1 Construção da Estrutura Hierárquica 41
4.3.2 Determinação do Número Ideal de Clusters 42
4.3.3 Estrutura e Tamanho dos Clusters 43
4.3.4 Interpretação dos Perfis Identificados 43
4.3.5 Síntese Interpretativa dos Grupos 45
4.3.6 Visualização Comparativa dos Perfis por Cluster 45
5 CONCLUSÕES 47
5.1 TRABALHOS FUTUROS 48
REFERÊNCIAS 49
APÊNDICE A – ESTRUTURA DAS PERGUNTAS 55
APÊNDICE B – TABELAS DE RESULTADOS DOS CLUSTERS GERAL 57
APÊNDICE C – TABELAS DE TRADUÇÃO DE FEATURES . . 60

1 INTRODUÇÃO

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 sub-representação 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; CHERYAN, 2021), influências familiares (HSIEH; SIMPKINS, 2022) e condições socioeconômicas (MORALES; GRINESKI; 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; 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; HOFFMAN *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.

1.1 QUESTÕES DE PESQUISA

O desenvolvimento desse trabalho foi elaborado com objetivo de responder as seguintes questões de pesquisa:

QP02 De que forma os fatores psicossociais, familiares, educacionais e socioeconômicos identificados se associam ao interesse declarado em seguir carreira nas áreas de STEM, considerando possíveis moderações por gênero e contexto regional?

QP03 Quais variáveis — e em que direção — emergem como mais influentes para distinguir gêneros e o interesse por STEM no ensino superior do Sertão Central Pernambucano?

QP04 Como métodos de explicabilidade de IA (por exemplo, Random Forest combinada com SHAP) podem esclarecer as contribuições individuais e globais dessas variáveis, validar a interpretação dos perfis obtidos por clusterização e subsidiar recomendações de intervenção educativa locais?

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.

1.2 OBJETIVOS

Os objetivos deste trabalho são subdivididos em objetivos gerais e objetivos específicos. Estes são:

1.2.1 Objetivo Geral

nvestigar os fatores que contribuem para as diferenças de gênero e para as percepções de pertencimento em áreas STEM, utilizando técnicas de aprendizado de máquina explicável e métodos de clusterização, no contexto de estudantes universitários do Sertão Central de Pernambuco.

1.2.2 Objetivos Específicos

Os objetivos específicos são:

- Elaborar e aplicar um questionário em instituições públicas e privadas de ensino superior da região do Sertão Central de Pernambuco, concebido para identificar fato-

res motivacionais, contextuais e psicossociais (incluindo medidas de pertencimento, exclusão e interesse por STEM) e variáveis demográficas relevantes;

- Realizar uma análise exploratória dos dados coletados, identificando padrões, tendências e associações entre variáveis, bem como realizando tratamento de dados, codificação e verificação de consistência para suportar etapas posteriores de modelagem;
- Treinar e validar um modelo de aprendizado de máquina supervisionado e explicável (por exemplo Random Forest), com separação treino/teste e validação cruzada, visando identificar padrões associados a gênero e interesse em STEM; estimar importâncias globais (permuta / OOB) e produzir explicações locais e globais usando SHAP (TreeExplainer) para interpretar direção e magnitude das contribuições das variáveis;
- Aplicar técnicas de agrupamento (por exemplo, clusterização hierárquica com método de Ward) para identificar perfis distintos de estudantes com base nas variáveis mais relevantes; selecionar o número de clusters por métodos como elbow e silhouette e integrar os resultados de clusterização com as explicações de IA para validação interpretativa dos perfis;
- Analisar e interpretar os grupos formados, descrevendo características predominantes em cada cluster, relacionando-as às explicações do modelo.

1.3 JUSTIFICATIVA

Este trabalho se justifica por buscar compreender as razões que sustentam a baixa representatividade feminina nas áreas STEM, mesmo em um contexto nacional no qual as mulheres já são maioria no Ensino Superior. Embora esse fenômeno seja amplamente discutido em grandes centros urbanos, pouco se sabe sobre como ele se manifesta em regiões interioranas, como o Sertão Central de Pernambuco — e, de forma ainda mais específica, na cidade de Salgueiro. Assim, este estudo pretende contribuir para a compreensão das barreiras e motivações que moldam o interesse por STEM no interior nordestino, unindo análise de dados, técnicas explicáveis de aprendizado de máquina e identificação de perfis de estudantes a partir de métodos de clusterização.

1.4 ORGANIZAÇÃO DO TRABALHO

O trabalho está organizado em cinco capítulos. A seguir apresenta-se um resumo do conteúdo de cada um deles:

- **Capítulo 1 – Introdução:** apresenta a motivação, o problema de pesquisa, os objetivos geral e específicos, as contribuições do trabalho e a justificativa metodológica e territorial.
- **Capítulo 2 – Fundamentação Teórica:** reúne a revisão da literatura sobre estereótipos de gênero, fatores socioeconômicos e familiares, aspectos educacionais, trabalhos sobre análise de dados educacionais, técnicas de agrupamento e princípios de explicabilidade em modelos de aprendizado de máquina.
- **Capítulo 3 – Delineamento Metodológico:** descreve o desenho da pesquisa, o instrumento de coleta (questionário), os procedimentos de pré-processamento e codificação dos dados, bem como os métodos analíticos empregados (clusterização hierárquica, critérios para seleção do número de clusters, treinamento de Random Forest e cálculo de valores SHAP).
- **Capítulo 4 – Resultados:** apresenta a caracterização da amostra, a análise exploratória, os resultados da clusterização e a interpretação dos perfis identificados, além das análises do modelo de classificação e das explicações por SHAP, discutindo implicações e limitações.
- **Capítulo 5 – Conclusões:** sintetiza os principais achados, discute limitações do estudo, destaca contribuições teóricas e práticas e propõe direções para pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

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; 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; HOFFMAN *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; 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

De acordo com (PARK *et al.*, 2024), estudantes de menor nível socioeconômico tendem a enfrentar maiores dificuldades para transformar aspirações elevadas em resultados concretos, pois carecem de recursos e suporte que possibilitem alcançar essas metas. Mesmo quando almejam ocupações de maior prestígio ou renda, suas aspirações nem sempre se convertem em trajetórias acadêmicas ou profissionais viáveis, refletindo limitações impostas pelo contexto econômico e social.

A compreensão do interesse em STEM entre estudantes do Sertão Central de Pernambuco exige considerar as condições socioeconômicas e, sobretudo, as oportunidades educacionais disponíveis na região.

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 (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; 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, 2020) 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; COLLINS, 2021) sobre o impacto da discordância mentor-mentorado na intenção de estudantes latinos de buscar a pós-graduação e na sua produtividade em pesquisa, fica evidente a crucial importância do mentor na trajetória acadêmica e profissional dos mentorados. 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 latinos 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 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; 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; 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.

De acordo com (HAN; KAMBER; PEI, 2011), 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. (HAN; KAMBER; PEI, 2011)

Ainda de acordo com (HAN; KAMBER; PEI, 2011), os métodos hierárquicos aglomerativos utilizam medidas de distância ou similaridade para determinar quais grupos devem ser fundidos em cada etapa. A seguir apresentam-se, de forma descritiva e sem fórmulas, as quatro medidas clássicas de ligação (*linkage*) entre clusters, o que é suficiente para contextualizar a escolha do método adotado neste trabalho.

Single linkage (mínima distância)

Nesse procedimento, a ligação entre dois grupos é determinada pela menor distância observada entre quaisquer dois pontos pertencentes a esses grupos. Por considerar apenas o par de pontos mais próximos, o método favorece a formação de cadeias e estruturas

alongadas, sendo sensível a ruídos e pontos isolados que podem ligar clusters distintos.

Complete linkage (máxima distância)

Aqui a fusão é conduzida pela maior distância entre pares de observações dos dois grupos. Esse critério tende a gerar clusters mais compactos e homogêneos, já que exige que todos os pontos do grupo resultante fiquem relativamente próximos; por outro lado, pode ser influenciado por outliers ao forçar partições mais restritas.

Centroid linkage (distância entre centróides)

O critério baseia-se na distância entre os centróides (médias) dos clusters. É uma abordagem simples e intuitiva, porém pode provocar inversões (*reversals*) no dendrograma em determinadas configurações, ou seja, a ordem de fusões pode não refletir monotonicamente a proximidade original entre elementos.

Average linkage (distância média)

Neste método, a ligação é calculada a partir da média das distâncias entre todos os pares de pontos pertencentes aos dois grupos. Funciona como um compromisso entre os extremos representados pelo single e pelo complete linkage, apresentando maior robustez a ruídos e formando clusters com grau intermediário de compactação.

Método de Ward (mínima variância)

Em contraste com esses critérios baseados em distâncias diretas, o método de Ward utiliza outro princípio para decidir a fusão de clusters. Segundo (RANDRIAMIHAMISON; VIALANEIX; NEUVIAL, 2021), o método adota uma perspectiva de minimização da variância interna, avaliando, a cada etapa, qual fusão provoca o menor aumento na soma das dispersões intracluster (within-cluster sum of squares). Assim, em vez de observar apenas pares específicos de pontos ou a média entre eles, o algoritmo considera o efeito global que uma fusão teria na homogeneidade e na compactação dos grupos.

Enquanto os métodos clássicos podem favorecer estruturas alongadas (single linkage), muito compactas (complete linkage) ou instáveis (centroid linkage), o método de Ward produz clusters mais equilibrados, coesos e com baixa variabilidade interna, justamente por otimizar continuamente a qualidade da partição. Como destacam (RANDRIAMIHAMISON; VIALANEIX; NEUVIAL, 2021), essa característica confere ao método maior estabilidade e interpretabilidade, além de dendrogramas mais consistentes, especialmente em conjuntos de dados onde a homogeneidade dos grupos é um objetivo central.

Em síntese, o método de Ward difere fundamentalmente dos demais porque não se apoia apenas em medições de distância entre objetos ou centroides, mas em um critério estatístico de compacidade mínima, funcionando como um processo de otimização que privilegia a formação de clusters internamente homogêneos.

A determinação do número adequado de clusters é uma etapa fundamental em algoritmos de agrupamento, especialmente naqueles que exigem a definição prévia desse parâmetro. Entre os métodos mais utilizados para essa finalidade, destaca-se o Método do Cotovelo (Elbow Method), cuja popularidade se deve à sua simplicidade e caráter visual.

Como descrito por (SHI *et al.*, 2021), o Método do Cotovelo baseia-se em observar como a distorção interna dos clusters — isto é, o grau de compactação dos grupos — se comporta à medida que o número de clusters aumenta. Em geral, adicionar mais clusters tende a tornar cada grupo mais homogêneo, reduzindo o erro interno. No entanto, essa redução não é constante: após certo ponto, criar novos clusters deixa de gerar uma melhoria significativa.

O Método do Cotovelo explora exatamente esse comportamento. Ele consiste em executar o algoritmo de clustering para diferentes valores de k e, em seguida, construir um gráfico que relaciona o número de clusters com a medida de erro. A interpretação visual busca identificar um ponto de inflexão — o "cotovelo" — onde a curva deixa de apresentar quedas acentuadas e passa a diminuir mais lentamente. Esse ponto representa um equilíbrio entre ganho de qualidade e simplicidade do modelo, sendo interpretado como o número apropriado de clusters para a estrutura dos dados.

(SHI *et al.*, 2021) ressaltam, porém, que essa identificação depende da clareza do formato da curva. Quando o gráfico apresenta um cotovelo evidente, a escolha é relativamente simples; entretanto, quando a curva é muito suave, a identificação do ponto ideal torna-se subjetiva e pode variar entre analistas. Essa limitação motivou os autores a propor um método quantitativo alternativo para detectar automaticamente o cotovelo, evidenciando que, apesar de amplamente utilizado, o Método do Cotovelo nem sempre fornece uma indicação precisa em situações de baixa contrastividade na curva.

Em síntese, o Método do Cotovelo é uma ferramenta intuitiva e amplamente empregada para estimar o número de clusters, fundamentando-se na análise visual da redução do erro. No entanto, como discutem (SHI *et al.*, 2021), sua natureza subjetiva pode limitar sua eficácia em cenários em que o formato da curva não apresenta um ponto de inflexão claramente definido.

2.3 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 (responsabilização e capacidade de prestar contas pelos resultados e decisões produzidos pelo modelo). 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_shap*, a previsão do modelo para a instância i pode ser decomposta como a soma aditiva das contribuições de cada variável, expressa pela Equação (2):

$$\hat{y}_i = \text{shap}_0 + \text{shap}(X_{1i}) + \text{shap}(X_{2i}) + \cdots + \text{shap}(X_{ji}),$$

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

$$\text{shap}_0 = E(\hat{y}_i),$$

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).

O algoritmo *Random Forest*, proposto por (BREIMAN, 2001), é um método de aprendizado supervisionado baseado na combinação de múltiplas árvores de decisão. Em vez de treinar uma única árvore, o modelo constrói centenas ou milhares delas, cada uma gerada a partir de uma amostra aleatória do conjunto de dados original, obtida pelo método de *bootstrap*. Esse procedimento consiste em selecionar exemplos aleatoriamente com reposição, de modo que cada árvore é treinada com um subconjunto ligeiramente diferente dos dados. Além disso, em cada divisão interna da árvore, apenas um subconjunto aleatório das variáveis é disponibilizado para escolha da melhor divisão. Essa dupla

aleatoriedade — nos exemplos e nos atributos — reduz a correlação entre as árvores, aumenta a diversidade da floresta e torna o modelo mais robusto, estável e resistente ao *overfitting* (situação em que o modelo aprende padrões específicos e ruídos do conjunto de treino, perdendo capacidade de generalização para novos dados), mesmo quando muitas árvores são utilizadas.

Uma característica importante do Random Forest é o uso das chamadas amostras *out-of-bag* (OOB). Durante a criação de cada árvore, aproximadamente um terço das observações não é selecionado na amostra *bootstrap*. Esses exemplos que "ficam de fora" são chamados de OOB e funcionam como uma espécie de conjunto de validação interno. Assim, cada árvore dispõe de exemplos que não foram usados em seu treinamento e que permitem avaliar seu desempenho sem necessidade de um conjunto externo de teste. Esse mecanismo interno de validação é central para o cálculo da importância das variáveis.

(BREIMAN, 2001) descreve um procedimento simples e eficiente para medir a importância das variáveis (*feature importance*) usando diretamente as estimativas OOB. Depois que cada árvore é construída, os valores da variável m nos exemplos OOB são aleatoriamente permutados, isto é, embaralhados, produzindo uma versão "deturpada" dessa variável. Em seguida, esses exemplos modificados são passados novamente pela árvore, e a classificação obtida é comparada com aquela gerada quando os valores originais estavam intactos. Esse processo é repetido para cada variável, uma de cada vez. Ao final da construção de toda a floresta, compara-se a taxa de erro obtida com os dados originais com a taxa obtida quando cada variável foi artificialmente embaralhada. Segundo o autor, a importância de uma variável é dada pelo aumento percentual da taxa de erro OOB causado por essa permutação. Quanto maior o aumento no erro ao destruir a informação contida na variável, mais relevante ela é considerada para a predição do modelo.

Embora o Random Forest forneça uma estimativa interna de importância das variáveis por meio do aumento percentual do erro *out-of-bag* (OOB), esse mecanismo apresenta uma visão essencialmente global e agregada da relevância dos atributos. Para aprofundar a compreensão sobre como cada variável influencia as predições, tanto no nível global quanto no nível individual, é possível combinar o modelo Random Forest com a metodologia SHAP, potencializando a interpretabilidade. Conforme discutido por (WANG *et al.*, 2024) e fundamentado na teoria dos valores de Shapley, os valores SHAP quantificam a contribuição marginal de cada variável para a predição específica de uma instância, permitindo decompor a saída do modelo de forma aditiva e interpretável.

No contexto deste trabalho, após o treinamento do modelo Random Forest, os valores SHAP são calculados utilizando o TreeExplainer, um método de explicação otimizado especificamente para modelos baseados em árvores. Proposto por (LUNDBERG *et al.*, 2020), o TreeExplainer constitui um avanço significativo na interpretabilidade desses modelos ao permitir o cálculo exato de valores de Shapley em tempo polinomial, explorando diretamente a estrutura hierárquica das árvores de decisão que compõem o algoritmo.

Enquanto o cálculo tradicional de valores de Shapley exige considerar todas as possíveis combinações de atributos — um problema classicamente NP-difícil, inviável para modelos reais — o TreeExplainer reformula esse processo ao colapsar o somatório combinatorial em operações estruturadas sobre os caminhos e folhas das árvores. Dessa forma, o método substitui aproximações estocásticas por um algoritmo determinístico capaz de computar explicações exatas, consistentes e localmente fiéis, preservando integralmente as propriedades fundamentais dos valores de Shapley, como eficiência, simetria e monotonicidade.

Essa eficiência decorre da capacidade do algoritmo de rastrear, ao longo dos caminhos decisórios, a proporção de subconjuntos de atributos que fluem para cada folha, o que permite atribuir corretamente a contribuição marginal de cada variável para a predição. (LUNDBERG *et al.*, 2020) demonstram que, ao explorar essa estrutura, o TreeExplainer supera limitações de abordagens anteriores, como métodos heurísticos específicos para árvores (por exemplo, Saabas), que sofrem de inconsistência e distorcem a importância de variáveis conforme sua profundidade na árvore, e métodos modelo-agnósticos baseados em amostragem, que apresentam alta variância e custos computacionais elevados

Assim, o TreeExplainer oferece explicações precisas mesmo em modelos compostos por centenas de árvores, como florestas aleatórias, tornando possível interpretar de forma confiável a contribuição individual de cada atributo para cada previsão.

Além disso, enquanto as medidas tradicionais de importância do Random Forest fornecem apenas uma visão global das variáveis mais relevantes — frequentemente influenciada por métricas internas como ganho ou Gini — os valores SHAP calculados pelo TreeExplainer permitem compreender não apenas a magnitude, mas também o sentido (positivo ou negativo) e até interações locais entre variáveis em cada instância analisada. (LUNDBERG *et al.*, 2020) mostram que essa abordagem produz explicações que refletem com maior fidelidade o comportamento real do modelo e revelam padrões que métodos globais não conseguem capturar, como efeitos raros, interações específicas e subgrupos de dados com relações particulares entre atributos e predições

Dessa forma, a combinação entre a importância global tradicional e os valores SHAP — calculados de forma eficiente e consistente pelo TreeExplainer — fornece uma análise interpretativa mais completa e robusta: primeiro identifica-se quais variáveis são importantes para o modelo; em seguida, investiga-se como elas influenciam cada previsão individualmente.

Para tornar esse processo mais claro, o fluxo operacional pode ser representado pelo pseudocódigo a seguir, que descreve a integração entre Random Forest e SHAP utilizada neste estudo:

```
1. Carregar o dataset original.
2. Separar variáveis preditoras (X) e variável alvo (y).
3. Dividir os dados em treinamento e teste.
```

```
4. Treinar o Random Forest com n árvores.
5. Avaliar o desempenho utilizando amostras OOB.
6. Gerar importâncias tradicionais via permutação das variáveis.
7. Inicializar o TreeExplainer a partir do modelo treinado.
8. Calcular valores SHAP para o conjunto de teste.
9. Agregar valores SHAP para obter importância global média.
10. Visualizar importância com summary plots e listas rankeadas.
11. Produzir explicações individuais com gráficos específicos.
```

O código implementado segue exatamente esse fluxo, permitindo tanto a visualização global da importância média dos atributos — por meio dos gráficos *summary plot* e do ranqueamento numérico — quanto explicações individualizadas das predições, utilizando gráficos como o *force plot*. Adicionalmente, após identificar as oito variáveis mais relevantes segundo a média absoluta dos valores SHAP, realiza-se uma nova visualização focada apenas nesse subconjunto de atributos, facilitando a interpretação e destacando os padrões presentes nas predições do modelo.

Assim, a integração entre Random Forest e SHAP reforça a interpretabilidade do processo de modelagem, oferecendo uma análise robusta, transparente e alinhada às recomendações contemporâneas de Explainable Artificial Intelligence (XAI). Enquanto o Random Forest contribui com estimativas empíricas baseadas em perturbações das variáveis, os valores SHAP complementam essa informação ao explicitar a contribuição individual de cada atributo sobre a predição, garantindo uma visão mais detalhada, confiável e explicativa do comportamento do modelo.

2.4 TRABALHOS CORRELATOS

2.4.1 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; 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 aprendi-

zagem.

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.

2.4.2 Fatores determinantes no interesse e autopercepção em STEM

A seguir, apresenta-se um fichamento resumido dos principais trabalhos relacionados ao tema deste estudo, conforme a Tabela 1.

Tabela 1 – Fichamento de trabalhos relacionados

Fonte	**Descrição**
(MILLER *et al.*, 2024) (MASTER; MELTZOFF; CHERYAN, 2021) (MCGUIRE; HOFFMAN *et al.*, 2022) (MASTER; TANG *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; 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.
(OLIVEIRA MENEZES; 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.

Fonte	**Descrição**
(ZÚÑIGA-MEJÍAS; 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; 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; 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; 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; MARTÍ-GONZÁLEZ, 2025)	Mulheres universitárias relatam maiores dificuldades psicossociais, menor autonomia e mais problemas de saúde mental no processo de aprendizagem do que homens.
(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; 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.

Fonte	**Descrição**
(MCGUIRE; MONZAVI *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.

3 DELINEAMENTO METODOLÓGICO

Este estudo adota uma abordagem quantitativa para investigar o fenômeno da da disparidade de gênero em áreas de STEM. 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.

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 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 e tecnologia. 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. A técnica escolhida para essa análise será a clusterização hierárquica, utilizando o método de Ward, que minimiza a variância intra-cluster. A distância entre os pontos será calculada utilizando a métrica de distância euclidiana, adequada para variáveis numéricas. ALém disso, as features selecionadas para a montagem dos clusters serão definidas na fase de análise de importância das variáveis, utilizando o modelo de Random Forest junto com os valores SHAP.

A definição do número ideal de grupos será feita com base no método do cotovelo (elbow method). Após a criação dos clusters, cada grupo será analisado em termos de suas características predominantes, permitindo a descrição dos perfis de estudantes identificados.

4 RESULTADOS

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.

No Apêndice B, encontra-se a estrutura de perguntas e opções de respostas das principais perguntas do questionário aplicado.

Para as etapas de modelagem e clusterização, foi realizado um tratamento prévio dos dados com o objetivo de garantir consistência, comparabilidade e alinhamento teórico com a literatura sobre interesse em STEM. Inicialmente, foram selecionadas apenas as variáveis diretamente relacionadas aos fatores identificados na revisão bibliográfica como influentes no interesse ou desinteresse por áreas STEM — tais como percepções de pertencimento, apoio familiar e escolar, histórico de interesse, experiências de exclusão e indicadores de estereótipos de gênero. Essa seleção teve como finalidade evitar que variáveis não relacionadas ao fenômeno investigado interferissem na formação dos clusters. Em seguida, as variáveis ordinais foram codificadas respeitando a ordem natural de suas categorias, enquanto os atributos nominais foram convertidos por meio de Label Encoding (técnica que substitui cada categoria por um número inteiro, permitindo que variáveis categóricas sejam utilizadas por algoritmos que exigem entradas numéricas). Após garantir que todos os valores estivessem em formato numérico, foi aplicado o método de normalização Min–Max em todas as colunas, exceto a variável alvo “gender”, pois já se encontrava com valores limitados à 0 e 1, de modo a preservar a proporcionalidade entre escalas e evitar distorções nas distâncias utilizadas pelos algoritmos de agrupamento. O conjunto final de dados, contendo apenas as variáveis selecionadas e devidamente tratadas, foi então utilizado para as análises e métodos de clusterização apresentados nos capítulos seguintes.

Figura 1 – Número de respondentes por gênero

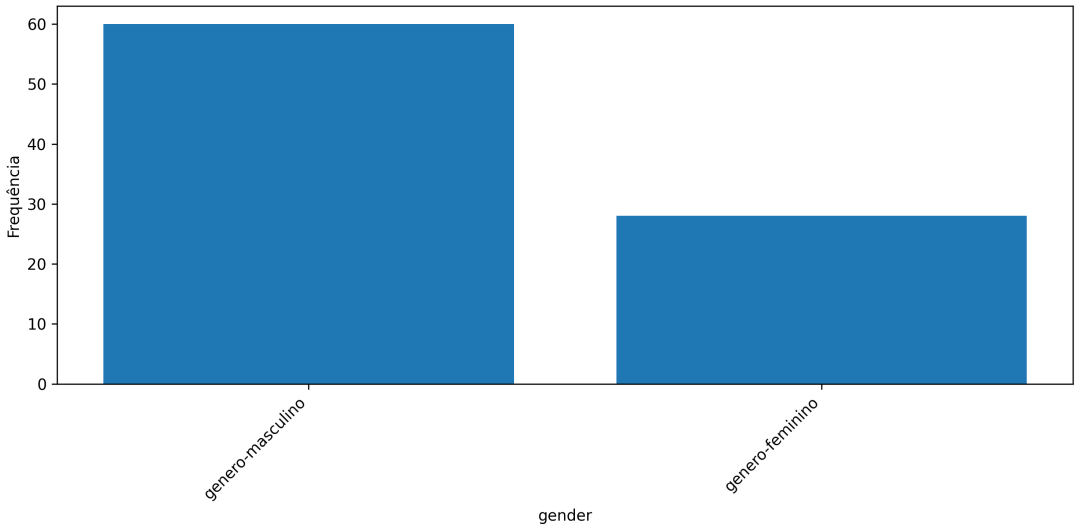


Figura 2 – Número de respondentes por curso

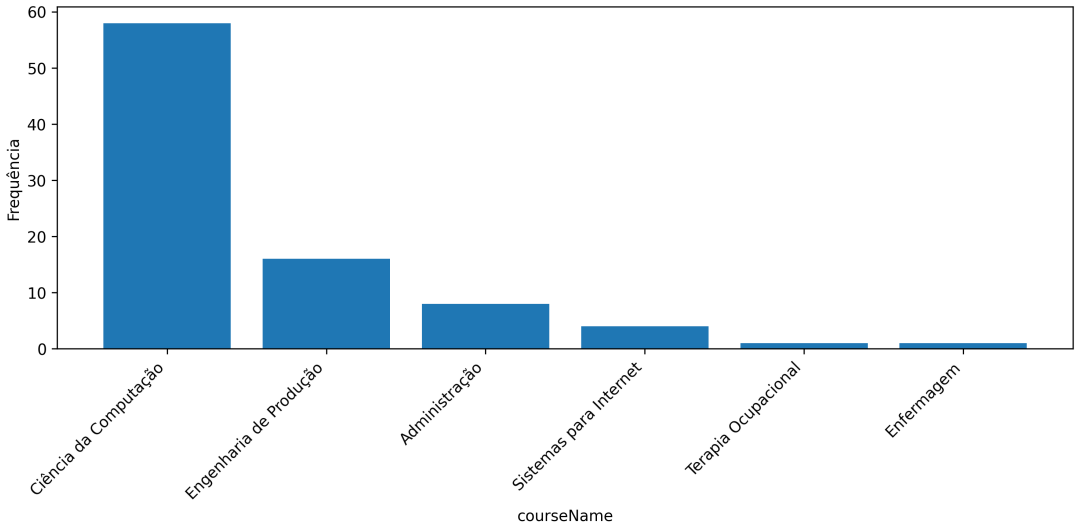
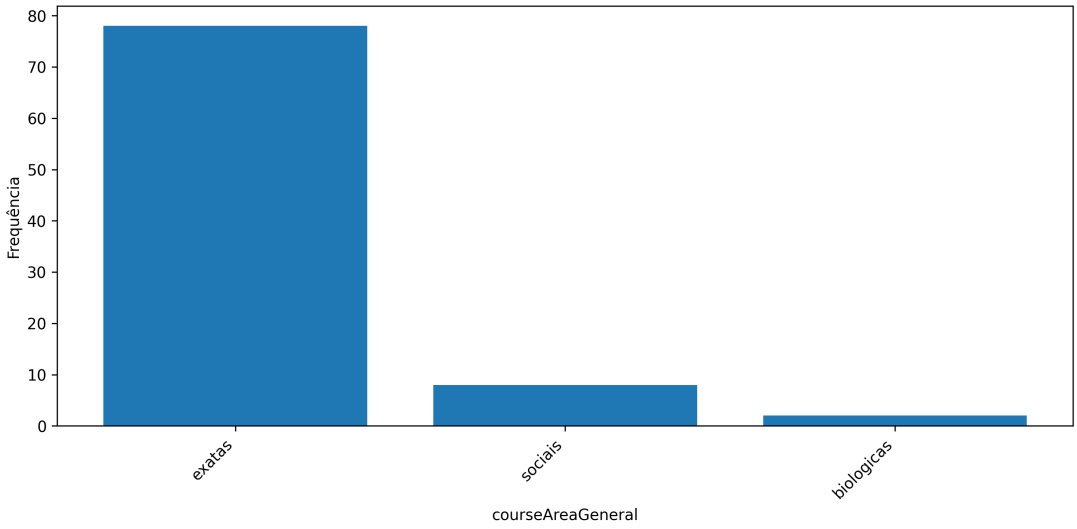


Figura 3 – Número de respondentes por área geral do curso



4.2 MODELO DE CLASSIFICAÇÃO E ANÁLISE DAS VARIÁVEIS MAIS IMPORTANTES

4.2.1 Análise com ambos os gêneros

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 foi carregado e a variável-alvo *gender* foi separada das demais variáveis explicativas selecionadas. Em seguida, os dados foram divididos em treino e teste, e um modelo de *Random Forest Classifier* foi ajustado com base nesse conjunto filtrado, com o objetivo de predizer a probabilidade de cada respondente pertencer ao gênero feminino.

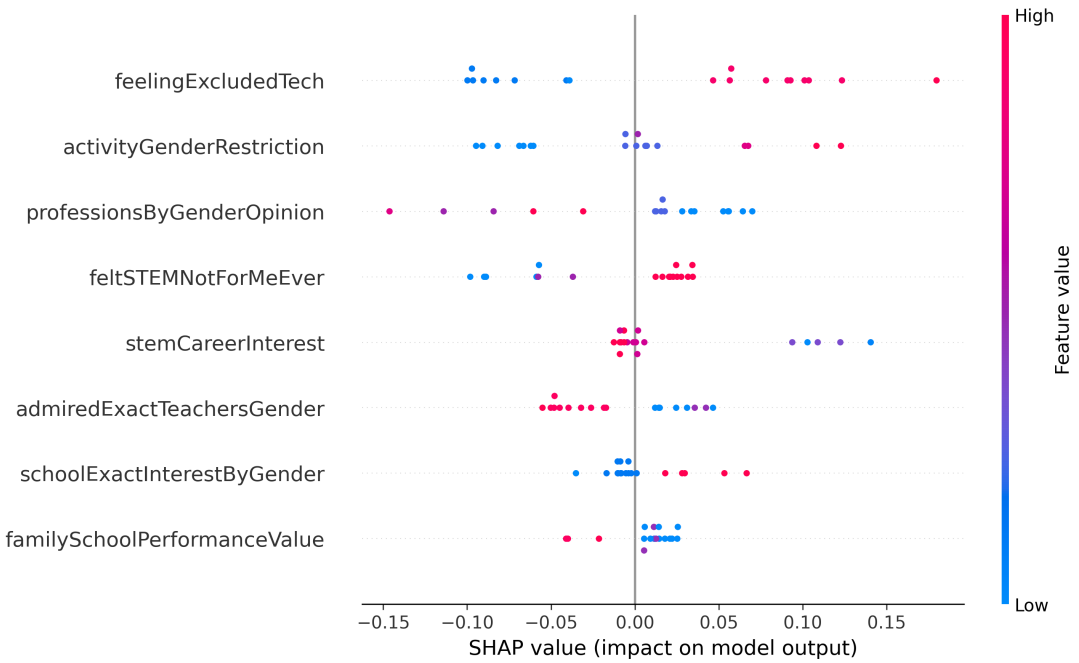
Uma vez treinado, o modelo foi interpretado por meio dos valores SHAP aplicados sobre as amostras de teste, permitindo identificar de forma transparente como cada um desses fatores relacionados ao interesse por STEM contribui para as predições da variável *gender*. No gráfico de SHAP resultante, valores mais vermelhos indicam maior probabilidade de o respondente ser do gênero feminino, enquanto valores mais azuis indicam menor probabilidade, ou seja, maior associação com o gênero masculino.

A análise inicial foi conduzida utilizando a base completa de respondentes, sem separação por gênero. O gráfico de SHAP oferece uma visão global dos principais fatores

que distinguem respondentes do gênero feminino e masculino, permitindo observar tanto a relevância relativa de cada variável quanto a direção de seus efeitos.

No Apêndice C, encontra-se a tabela de tradução das features utilizadas neste gráfico, facilitando a compreensão dos termos técnicos empregados.

Figura 4 – Gráfico de SHAP para o modelo de classificação de gênero



4.2.1.1 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 mais vermelhos indicam maior probabilidade de o respondente ser do gênero feminino, enquanto pontos mais azuis indicam maior probabilidade de pertencer ao gênero masculino. Essa distinção visual permite compreender a direção das relações entre as variáveis explicativas e a probabilidade de o modelo predizer um determinado gênero.

4.2.1.2 Sentimento de Exclusão em Tecnologia (*feelingExcludedTech*)

Esta foi a variável mais relevante para distinguir gêneros. Valores mais altos de *feelingExcludedTech* estão fortemente associados ao gênero feminino, indicando que meninas relatam maior sensação de exclusão em áreas de tecnologia. Valores baixos da variável tendem a associar-se ao público masculino, o que reforça achados da literatura sobre experiências diferenciadas de inclusão em áreas de tecnologia.

4.2.1.3 Restrição de Atividades por Motivo de Gênero (*activityGenderRestriction*)

Valores altos indicam que a pessoa já deixou de realizar atividades por questões de gênero. Observa-se maior frequência de respostas elevadas entre meninas, embora alguns meninos também reportem valores altos, possivelmente por interpretação mais ampla da pergunta. Essa variável destaca diferenças de percepção sobre barreiras sociais relacionadas a gênero.

4.2.1.4 Opinião sobre Profissões por Gênero (*professionsByGenderOpinion*)

Valores altos refletem maior concordância com a ideia de que algumas profissões são mais adequadas a um gênero. Valores elevados foram mais associados a meninos, enquanto valores baixos foram mais frequentes entre meninas. Isso indica que meninas tendem a rejeitar estereótipos profissionais com maior intensidade.

4.2.1.5 Percepção de que STEM Nunca Foi Para Mim (*feltSTEMNotForMeEver*)

Valores altos dessa variável indicam que o respondente já sentiu que STEM não era para ele/ela. No gráfico, observa-se que respostas elevadas estão concentradas no gênero feminino, sugerindo que meninas apresentam maior tendência a se afastar de STEM em algum momento, reforçando questões de pertencimento e autoconfiança.

4.2.1.6 Interesse em Carreira STEM (*stemCareerInterest*)

Esta variável mostra a disposição em seguir carreiras em STEM. Curiosamente, valores médios ou baixos aparecem mais associados a meninas, enquanto meninos tendem a indicar interesse mais intenso. Essa diferença pode refletir fatores de autoconfiança ou percepção de oportunidade em STEM.

4.2.1.7 Admiração por Professores (*admiredExactTeachersGender*)

Valores mais altos indicam que o respondente admira professores de um gênero específico. Pontos vermelhos (maior probabilidade de feminino) concentram-se em respostas indicando admiração por professores do gênero feminino ou ambos, sugerindo que meninas tendem a reconhecer figuras de autoridade em diferentes gêneros, enquanto meninos apresentam tendencia a admirarem professores de exatas do gênero masculino ou nenhum professor.

4.2.1.8 Interesse Escolar por Gênero (*schoolExactInterestByGender*)

Esta variável mede a percepção dos estudantes em relação a quem possui maior interesse em certas matérias conforme o gênero. Meninas (pontos vermelhos) tendem a

relatar que meninos possuem mais interesse ou não veem diferença, enquanto os meninos tendem a relatar que meninas possuem mais interesse nessas disciplinas.

4.2.1.9 Percepção do Desempenho Escolar Familiar (*familySchoolPerformanceValue*)

Valores mais altos dessa variável indicam maior valorização familiar do desempenho escolar. Muitas das meninas tendem a apresentar valores mais baixos dessa variável, indicando que sentem menor valorização familiar em relação ao desempenho escolar, enquanto meninos tendem a apresentar valores mais altos, sugerindo maior reconhecimento familiar.

4.2.1.10 Resumo

A análise das oito variáveis selecionadas mostra um padrão consistente: características relacionadas à exclusão percebida, restrições por gênero, experiências passadas e incentivo familiar estão fortemente associadas ao gênero feminino. Esse padrão reforça evidências da literatura sobre a influência de fatores sociais e psicológicos na decisão de meninas e meninos de se engajar em carreiras STEM.

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 quanto a pessoa consegue se imaginar em uma carreira em áreas de 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.

4.2.2.1 Diversidade dos Gráficos e Diferença no Volume de Respostas

Inicialmente, é válido destacar que 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.

4.2.2.2 Similaridades Entre os Modelos: Interesse Prévio em STEM

Nos dois modelos analisados, tanto para mulheres quanto para homens, o interesse prévio em STEM destaca-se como uma das variáveis mais influentes na previsão da capacidade do estudante de se visualizar trabalhando em carreiras de STEM. Essa relação é intuitiva, uma vez que estudantes que demonstram afinidade e gosto pela área tendem naturalmente a imaginar-se atuando em carreiras de STEM, pois já possuem maior familiaridade, motivação e identificação com esses conteúdos. 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 alvo (capacidade de se visualizar atuando em uma carreira STEM), tanto para meninas quanto para meninos, sugerindo que práticas pedagógicas de apoio podem desempenhar papel relevante na motivação dos estudantes.

4.2.2.3 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 autopercepção das mulheres. 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.

4.2.2.4 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 que conseguem se enxergar em carreiras em STEM tendem a reproduzir menos estereótipos de gênero, enquanto aqueles que não conseguem se visualizar trabalhando em STEM ou apresentam pouco envolvimento com a área manifestam maior adesão a esses estereótipos.

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

Outro ponto relevante é que a variável *feelingExcludedTech* aparece entre as mais importantes no modelo masculino, mas não figura entre as principais no modelo feminino. Isso ocorre porque a grande maioria das meninas — mesmo aquelas que afirmam gostar de STEM — relata algum nível de sentimento de exclusão em contextos tecnológicos. Esse padrão já havia se destacado na análise conjunta, em que o sentimento de exclusão foi a principal variável para diferenciar gênero. Dessa forma, dentro do grupo feminino, a variável apresenta baixa variabilidade, o que reduz sua capacidade de discriminar quais estudantes conseguem se imaginar atuando em STEM. Entre os meninos, porém, essa percepção varia de forma mais expressiva, fazendo com que *feelingExcludedTech* contribua de maneira mais significativa para o modelo masculino.

4.2.2.6 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 conseguem se enxergar atuando em áreas de STEM tendem a admirar professores de exatas do gênero feminino ou de ambos (valores mais baixos), enquanto aquelas que não conseguem se imaginar com muita clareza nessas áreas tendem a admirar professores do gênero masculino ou nenhum professor (valores mais altos).

Figura 5 – Gráfico de SHAP para o modelo de classificação sobre visão de futuro em STEM - Feminino

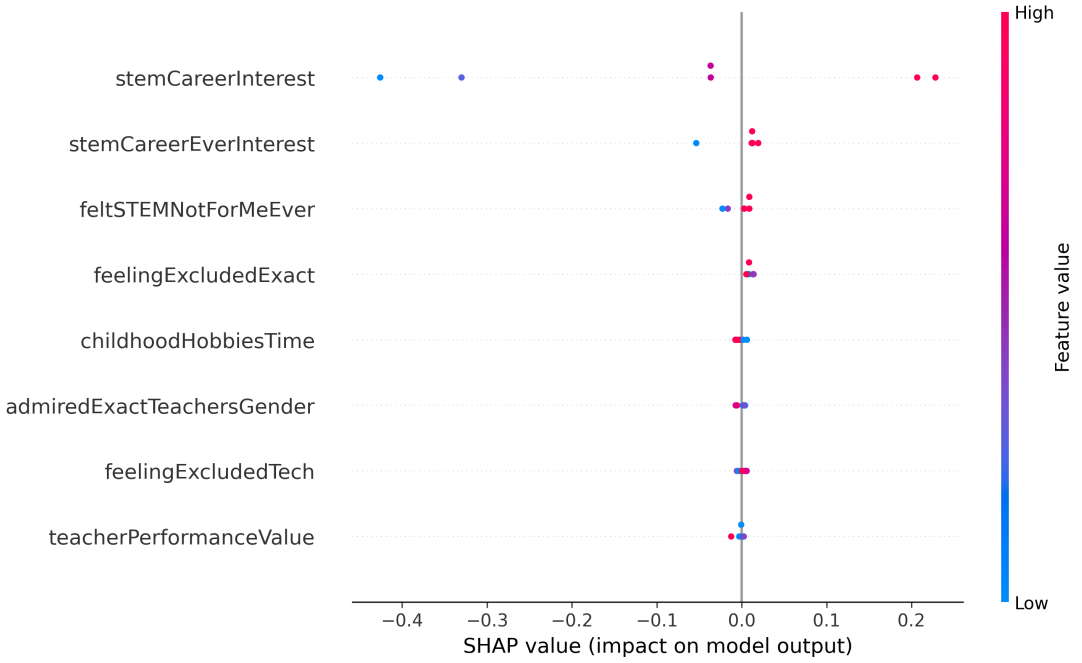
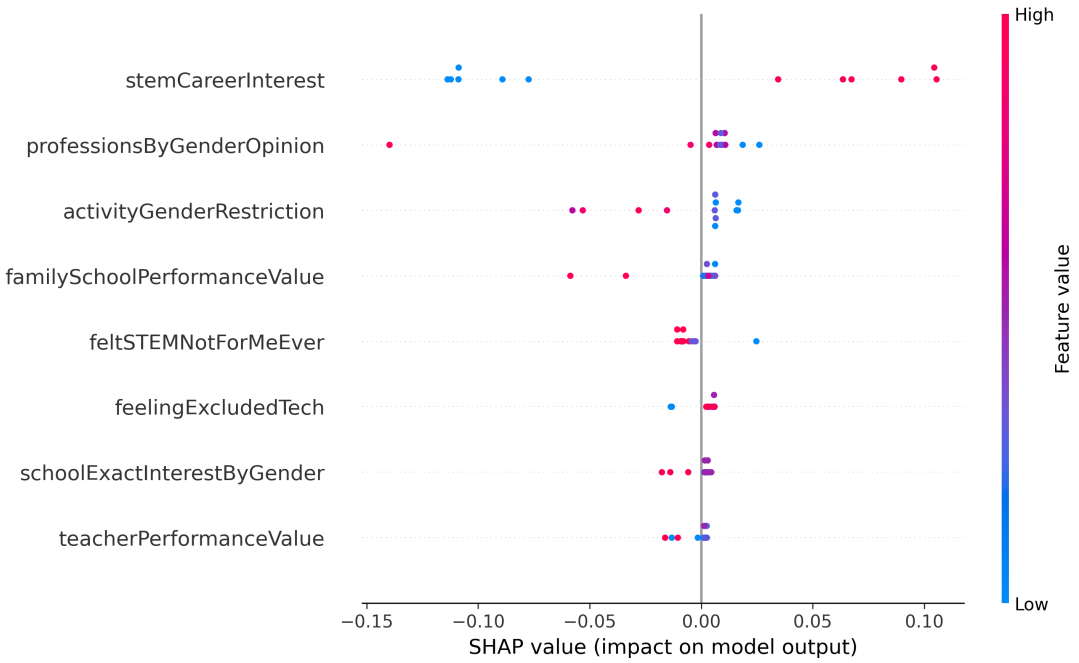


Figura 6 – Gráfico de SHAP para o modelo de classificação sobre visão de futuro em STEM - Masculino



4.3 FORMAÇÃO E INTERPRETAÇÃO DOS GRUPOS

Após a identificação das oito variáveis mais relevantes pelo modelo Random Forest com SHAP na análise geral, procedeu-se à formação de grupos (*clusters*) utilizando a técnica de *clustering* hierárquico. O objetivo desta etapa foi verificar se os respondentes apresentavam padrões semelhantes de comportamento ou percepção, e se esses padrões permitiam a formação de perfis distintos de estudantes.

As oito *features* utilizadas como base para essa etapa foram: *feelingExcludedTech*, *activityGenderRestriction*, *professionsByGenderOpinion*, *feltSTEMNotForMeEver*, *stemCareerInterest*, *admiredExactTeachersGender*, *schoolExactInterestByGender* e *familySchoolPerformanceValue*. Essas variáveis foram selecionadas por representarem os fatores mais importantes para diferenciação entre os gêneros na análise anterior, o que justificou sua utilização como descritores dos possíveis perfis.

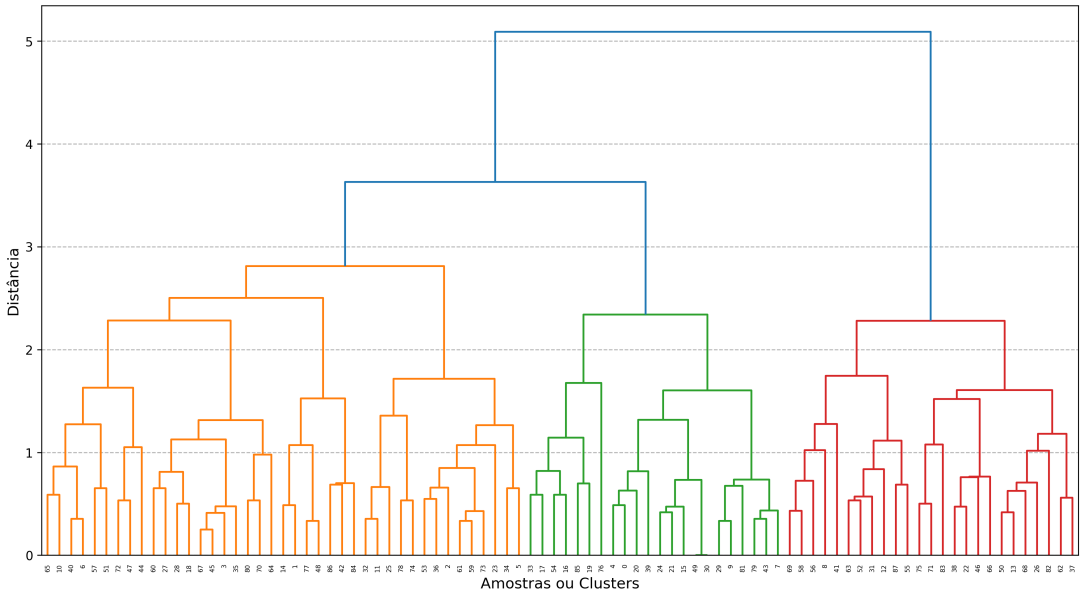
4.3.1 Construção da Estrutura Hierárquica

Para a formação dos grupos foi aplicado o método hierárquico aglomerativo com ligação de Ward, que busca minimizar a variância interna durante a junção dos grupos. O dendrograma resultante, apresentado na Figura 7, permite visualizar a forma como as amostras se agrupam ao longo do processo de fusão, bem como as distâncias às quais as junções ocorrem.

Ao observar o dendrograma, nota-se a presença de dois grandes blocos claramente separados, o que já indica a possibilidade de existência de dois grupos bem definidos

entre os respondentes. A separação visual observada no gráfico sugere que os estudantes apresentam dois padrões predominantes de respostas nas variáveis analisadas.

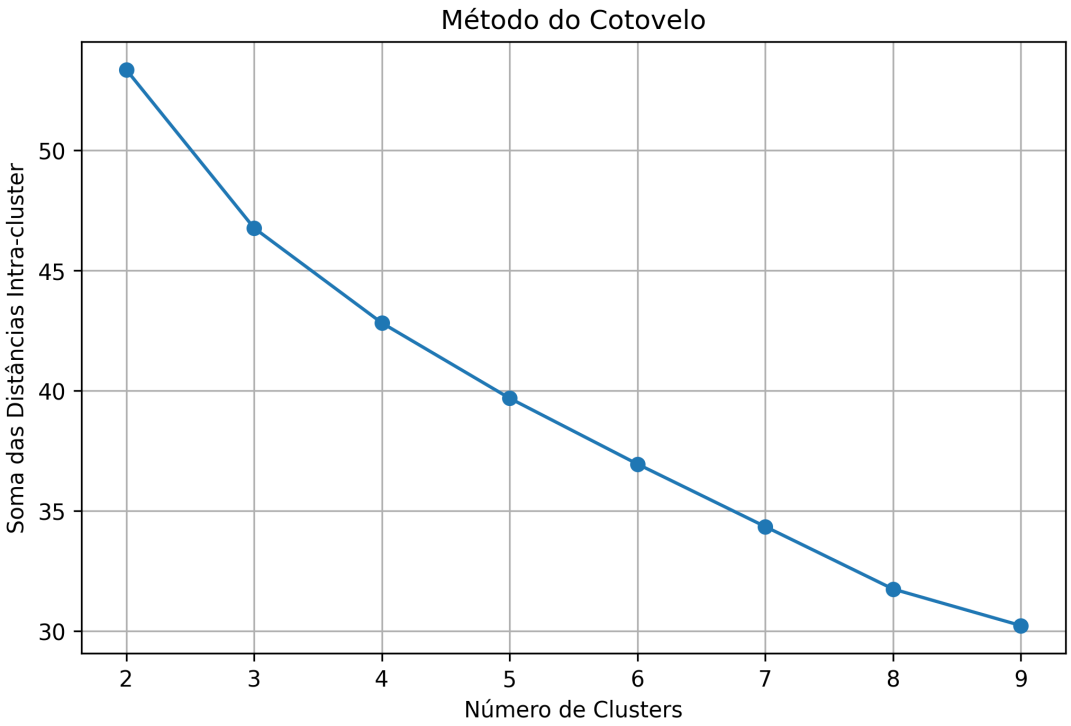
Figura 7 – Dendrograma da Análise de Clusters Hierárquica



4.3.2 Determinação do Número Ideal de Clusters

Para corroborar a indicação visual do dendrograma, aplicou-se o método do cotovelo, conforme a Figura 8. A análise da soma das variâncias intra-grupo revelou uma queda acentuada ao se passar de um para dois clusters, seguida de uma redução significativamente menor a partir de três grupos. Esse comportamento caracteriza o "cotovelo" clássico no ponto correspondente a dois clusters.

Figura 8 – Análise do Cotovelo para Determinação do Número de Clusters



4.3.3 Estrutura e Tamanho dos Clusters

Os dois grupos formados apresentam tamanhos distintos: o **Cluster 1** com 25 participantes e o **Cluster 2** com 63 participantes. A distribuição de gênero indica uma diferença marcante entre os grupos: o Cluster 1 possui maioria masculina (21 homens e apenas 4 mulher), enquanto o Cluster 2 apresenta uma composição mais equilibrada (24 mulheres e 39 homens). Esse desequilíbrio sugere, desde o início, que os clusters podem refletir diferenças de percepção associadas ao gênero e potencialmente relacionadas às barreiras ou estímulos vinculados ao interesse por STEM.

4.3.4 Interpretação dos Perfis Identificados

A Tabela no Apêndice A (resultante dos cruzamentos das variáveis com os clusters) permite visualizar as características de cada grupo.

Cluster 1: Perfil com maior interesse e engajamento em STEM

O Cluster 1, mais numeroso, apresenta os seguintes padrões:

- Menor percepção de exclusão (*feelingExcludedTech*) e menor influência de restrições por gênero (*activityGenderRestriction*);

- Menor concordância com estereótipos de gênero em profissões (*professionsByGenderOpinion*) e apreciação mais equilibrada de professores (*admiredExactTeachersGender*);
- Maior histórico de sentimento de STEM "para mim" (*feltSTEMNotForMeEver* concentra-se em níveis mais altos, indicando que muitos se sentiram aptos à área) e maior interesse declarado em seguir carreira em STEM (*stemCareerInterest*);
- Maior envolvimento em atividades escolares (*schoolExactInterestByGender*) e percepção mais positiva do desempenho familiar (*familySchoolPerformanceValue*).

Esse perfil indica maior engajamento, interesse consolidado em STEM e menor percepção de barreiras ou estereótipos de gênero, compatível com o equilíbrio de gênero observado no cluster.

Cluster 2: Perfil mais sensível à exclusão e menor engajamento em STEM

O Cluster 2 reúne predominantemente mulheres e apresenta as seguintes características:

- Valores elevados em *feelingExcludedTech* e *feltSTEMNotForMeEver*, indicando maior percepção de exclusão em ambientes de tecnologia e maior restrição de atividades por gênero;
- Maior concordância com estereótipos de profissões (*professionsByGenderOpinion*) e certa valorização de professores do gênero predominante (*admiredExactTeachersGender*);
- Menor interesse prévio por STEM (*feltSTEMNotForMeEver* indica que poucos se sentiram excluídos permanentemente, mas ainda existe percepção de barreira) e menor interesse declarado em seguir carreira em STEM (*stemCareerInterest* apresenta concentração em níveis intermediários);
- Menor envolvimento em atividades escolares de interesse por gênero (*schoolExactInterestByGender*) e percepção intermediária de desempenho familiar (*familySchoolPerformanceValue*).

Esse perfil sugere um grupo com maior sensibilidade a barreiras de gênero e percepções de exclusão em STEM, com engajamento intermediário, refletindo a predominância feminina no cluster.

4.3.5 Síntese Interpretativa dos Grupos

Os resultados confirmam que a segmentação em dois clusters reflete diferenças estruturais relevantes entre os estudantes, especialmente no que diz respeito a interesse por STEM, percepção de exclusão e internalização de estereótipos de gênero.

O **Cluster 2** é caracterizado por maior sensibilidade a barreiras de gênero e menor engajamento em STEM, proporcionalmente (considerando a quantidade de respondentes de cada gênero) com maior presença feminina, enquanto o **Cluster 1** demonstra maior motivação, interesse consolidado e menor percepção de restrições de gênero, proporcionalmente com maior presença masculina.

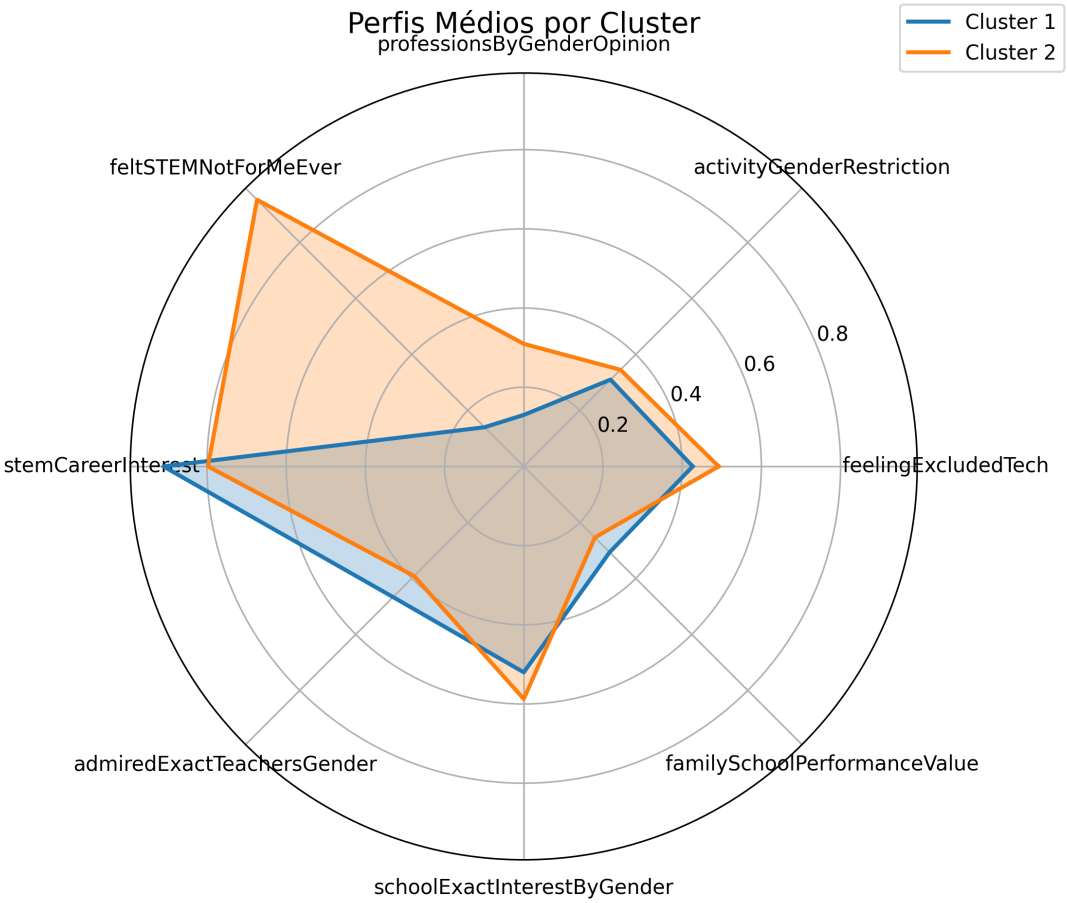
4.3.6 Visualização Comparativa dos Perfis por Cluster

A Figura 9 apresenta um gráfico do tipo radar com as médias normalizadas das oito *features* utilizadas na clusterização. O gráfico evidencia as divergências entre os clusters:

- **Cluster 1**: valores mais altos em *stemCareerInterest*, *schoolExactInterestByGender* e *familySchoolPerformanceValue*, evidenciando maior interesse em STEM, maior engajamento escolar e suporte familiar mais consistente.
- **Cluster 2**: valores mais altos em *feelingExcludedTech*, *feltSTEMNotForMeEver* e *professionsByGenderOpinion*, indicando maior percepção de exclusão e estereótipos de gênero;

O gráfico permite compreender de forma intuitiva a estrutura e as diferenças qualitativas entre os grupos, reforçando os padrões observados nas análises numéricas e na segmentação hierárquica.

Figura 9 – Gráfico de radar comparativo entre os clusters



5 CONCLUSÕES

Este trabalho partiu da observação de um padrão de distribuição de gênero nas diferentes áreas do conhecimento: embora as mulheres sejam maioria no Ensino Superior brasileiro, sua presença permanece concentrada em determinadas áreas, com subrepresentação nas carreiras de STEM. A motivação central foi compreender os fatores que influenciam a escolha de curso superior e identificar perfis de estudantes capazes de explicar, parcial ou totalmente, essa desigualdade de representatividade. Buscou-se, com isso, fornecer subsídios empíricos que possam orientar políticas e práticas de incentivo ao ingresso e à permanência das mulheres em áreas científicas e tecnológicas.

Em alinhamento a essa motivação, os objetivos propostos foram atendidos de forma parcial e coerente com o escopo da pesquisa: elaborou-se e aplicou-se um instrumento de coleta de dados em instituições de ensino superior da região do Sertão Central de Pernambuco; realizou-se análise exploratória dos dados; treinou-se um modelo supervisionado explicável (Random Forest + interpretação por valores SHAP) para identificar variáveis associadas a diferenças de gênero e interesse por STEM; e empregou-se uma técnica de clusterização hierárquica (método de Ward) para mapear perfis de estudantes a partir das principais variáveis identificadas.

Metodologicamente, a pesquisa combinou técnicas quantitativas clássicas e modernas: processamento e codificação das respostas do questionário, análise exploratória, modelagem supervisionada com Random Forest e interpretação de importância e direção de efeito via SHAP, seguida de clusterização hierárquica com validação visual pelo dendrograma e método do cotovelo. Essa combinação permitiu não apenas identificar quais variáveis se associam com maior força às diferenças por gênero, mas também descrever agrupamentos de respondentes com perfis similares.

Os principais achados podem ser sumarizados do seguinte modo:

- A percepção de exclusão em ambientes de tecnologia (*feelingExcludedTech*) emergiu como a variável mais relevante para distinguir perfis de gênero. Valores mais altos dessa variável estão fortemente associados ao público feminino, apontando para um problema de percepção e pertencimento que pode limitar a intenção de seguir carreiras tecnológicas.

- A ocorrência de restrição de atividades por motivo de gênero e a concordância com estereótipos sobre profissões também se mostraram importantes para diferenciação entre os grupos, indicando que normas e representações sociais continuam a influenciar escolhas e autopercepções.

- Incentivo familiar para os estudos e interesse prévio por disciplinas de exatas apareceram como fatores positivos associados ao maior interesse em seguir carreira em STEM.

- A análise por cluster revelou dois perfis principais: o Cluster 1 (predominantemente masculino na amostra) caracterizado por menor identificação com STEM, maior

concordância com estereótipos e menor incentivo familiar; e o Cluster 2 (mais heterogêneo) com maior interesse prévio em exatas, maior incentivo familiar e menor internalização de estereótipos.

Apesar das contribuições, a pesquisa tem limitações relevantes que merecem ser destacadas. A amostra é relativamente pequena (n = 88), com distribuição de gênero e composição institucional que limitam a generalização dos resultados para outras regiões ou para populações maiores. Algumas variáveis apresentaram baixa variabilidade dentro de subgrupos (por exemplo, sentimento de exclusão muito presente entre as mulheres), o que reduz sua capacidade discriminatória em análises estratificadas. Finalmente, escolhas de codificação e normalização de categorias, necessárias para a modelagem, podem inserir arbitrariedades que influenciam interpretações numéricas e comparações.

Em termos de contribuição, o estudo agrega: (i) um levantamento empírico localizado sobre percepções de gênero e interesse em STEM no município de Salgueiro–PE; (ii) a aplicação integrada de modelos explicáveis (SHAP) e técnicas de clusterização para mapear perfis de estudantes, ampliando a compreensão sobre como fatores psicossociais e contextuais se organizam; e (iii) o desenvolvimento (conforme proposto nas seções anteriores) de uma plataforma de divulgação de pesquisadoras brasileiras, que busca aumentar a visibilidade de modelos de referência e reduzir barreiras de representação.

5.1 TRABALHOS FUTUROS

Para trabalhos futuros, recomenda-se: ampliar a amostra e diversificar a abrangência geográfica para aumentar a robustez e a generalização dos achados; realizar desenhos longitudinais que permitam observar trajetórias e efeitos temporais (por exemplo, observando se intervenções de mentoria modificam intenção e permanência em STEM); incorporar métodos mistos, incluindo entrevistas qualitativas com estudantes e familiares, para aprofundar a compreensão das causas subjacentes aos sentimentos de exclusão e às dinâmicas de incentivo familiar; testar intervenções experimentais (programas de mentoria, oficinas com modelos femininos em STEM, mudanças curriculares) e avaliar seu impacto; e explorar outras técnicas de agrupamento e validação (por exemplo, LPA, clustering baseado em mistura ou validação em amostras externas).

Conclui-se que, embora não exista solução única para a sub-representação feminina em STEM, a combinação de políticas institucionais (mais modelos de referência, práticas pedagógicas acolhedoras), ações familiares e intervenções locais informadas por evidência empírica pode contribuir para reduzir barreiras percebidas e ampliar o sentido de pertencimento. Espera-se que os resultados aqui apresentados sirvam como ponto de partida para ações contextuais no Sertão Central de Pernambuco e como base para pesquisas ampliadas sobre o tema.

REFERÊNCIAS

ARSHAD, Aleena; ARSHAD, Arooj. 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. DOI: https://doi.org/10.52053/jpap.v5i4.317. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/josi.12655.

BOHRNSTEDT, George W. *et al.* Mathematics Motivation and Mathematics Performance: Does Gender Play a Role? **AERA Open**, v. 10, n. 1, p. 1–17, 2024. DOI: 10.1177/23328584241298272. Disponível em: https://doi.org/10.1177/23328584241298272.

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

BREIMAN, Leo. Random Forests. **Machine Learning**, Kluwer Academic Publishers, v. 45, n. 1, p. 5–32, 2001. Official publication; PDF attached in this work.

CIÊNCIA DA COMPUTAÇÃO, Bacharelado em. **Projeto Pedagógico do Curso (PPC): Bacharelado em Ciência da Computação**. [*S.l.*: *s.n.*], 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, Olga Victoria; MALDONADO, Darío; SÁNCHEZ, Fabio. 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. DOI: 10.1016/j.ijedudev.2022.102591. Disponível em: https://doi.org/10.1016/j.ijedudev.2022.102591.

ELVIRA-ZORZO, María Natividad; GANDARILLAS, Miguel Ángel; MARTÍ-GONZÁLEZ, Mariacarla. 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. DOI: 10.3390/socsci14020071. Disponível em: https://doi.org/10.3390/socsci14020071.

EMRAN, Ameer *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. DOI: 10.1007/s11191-020-00103-z. Disponível em: https://doi.org/10.1007/s11191-020-00103-z.

EUROPEAN DATA PROTECTION SUPERVISOR. **EDPS TechDispatch on Explainable Artificial Intelligence**. [*S.l.*: *s.n.*], 2021. https://edps.europa.eu. Placeholder entry — replace with the authoritative reference/URL.

FEIGE, Paulina *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, e0317837, 2025. DOI: 10.1371/journal.pone.0317837. Disponível em: https://doi.org/10.1371/journal.pone.0317837.

GENCEL-AUGUSTO, Jovanka *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. DOI: 10.3322/caac.21875. Disponível em: https://doi.org/10.3322/caac.21875.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining: Concepts and Techniques**. 3rd. [*S.l.*]: Morgan Kaufmann, 2011.

HSIEH, Tzu-Ying; SIMPKINS, Sandra 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. DOI: 10.1002/jad.12059. Disponível em: https://doi.org/10.1002/jad.12059.

JUNIOR, José Carlos Guimarães *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. DOI: 10.56083/RCV3N8-061.

LIU, Rui; LIU, Chang; HE, Peng. Chinese Grades 1–9 Students' Views of the Nature of Science: Do They Differ by Grade Level, Gender, and Parents' Occupation? **Science & Education**, 2024. DOI: 10.1007/s11191-024-00519-x. Disponível em: https://doi.org/10.1007/s11191-024-00519-x.

LUNDBERG, Scott M. *et al.* From local explanations to global understanding with explainable AI for trees. **Nature Machine Intelligence**, v. 2, n. 1, p. 56–67, 2020. Open access. DOI: 10.1038/s42256-019-0138-9.

MARTINS, Dayane Diniz *et al.* Clusterização do perfil de adolescentes escolares com predisposição ao uso de substância psicoativas. **Research, Society and Development**, v. 10, n. 2, e37510212528, 2021. Publicado: 19/02/2021. DOI: 10.33448/rsd-v10i2.12528.

MASTER, Allison; MELTZOFF, Andrew N.; CHERYAN, Sapna. 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, e2100030118, 2021. DOI: 10.1073/pnas.2100030118. Disponível em: https://doi.org/10.1073/pnas.2100030118.

MASTER, Allison; TANG, Daijiazi *et al.* Gender equity and motivational readiness for computational thinking in early childhood. **Early Childhood Research Quarterly**,

v. 64, p. 242–254, 2023. DOI: 10.1016/j.ecresq.2023.03.004. Disponível em: https://doi.org/10.1016/j.ecresq.2023.03.004.

MCGUIRE, Luke; HOFFMAN, Adam J. *et al.* Gender stereotypes and peer selection in STEM domains among children and adolescents. **Sex Roles**, v. 87, p. 455–470, 2022. DOI: 10.1007/s11199-022-01327-9. Disponível em: https://doi.org/10.1007/s11199-022-01327-9.

MCGUIRE, Luke; MONZAVI, Tina *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. DOI: 10.3389/fpsyg.2021.503237. Disponível em: https://doi.org/10.3389/fpsyg.2021.503237.

MILLER, David 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. DOI: 10.1037/bul0000456. Disponível em: https://doi.org/10.1037/bul0000456.

MORALES, Devon X.; GRINESKI, Sara E.; COLLINS, Timothy 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. DOI: 10.1111/nyas.14602. Disponível em: https://doi.org/10.1111/nyas.14602.

OLIVEIRA, Ana Luíza Matos 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. DOI: 10.1590/S1414-40772021000100013.

OLIVEIRA, Pamella Letícia Silva 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. DOI: 10.5753/rbie.2022.2508.

OLIVEIRA MENEZES, Suzy Kamylla de; SANTOS, Mario Diego Ferreira 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. DOI: 10.5753/RBIE.2021.29.0.456. Disponível em: https://doi.org/10.5753/RBIE.2021.29.0.456.

OPPS, Zachary; YADAV, Aman. Who Belongs in Computer Science? *In:* PROCEEDINGS of the 53rd ACM Technical Symposium on Computer Science Education (SIGCSE '22). Providence, RI, USA: ACM, 2022. P. 383. DOI: 10.1145/3478431.3499301. Disponível em: https://doi.org/10.1145/3478431.3499301.

PARK, Jeongeun *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. DOI: 10.1111/josi.12655. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/josi.12655.

PUGLIESE, Gustavo. STEM Education – um panorama e sua relação com a educação brasileira. **Curriculo sem Fronteiras**, v. 20, mar. 2020. DOI: 10.35786/1645-1384.v20.n1.12.

RANDRIAMIHAMISON, Nathanaël; VIALANEIX, Nathalie; NEUVIAL, Pierre. Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints. **Journal of Classification**, v. 38, p. 363–389, 2021. Open access under CC BY 4.0. DOI: 10.1007/s00357-020-09377-y.

SCHEL, Janina; DRECHSEL, Barbara. 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. DOI: 10.3389/fpsyg.2025.1527438.

SHI, Jie *et al.* A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. **EURASIP Journal on Wireless Communications and Networking**, v. 2021, n. 174, p. 1–20, 2021. Open access. DOI: 10.1186/s13638-021-01910-w.

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

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

TELLHED, Una; BJÖRKLUND, Fredrik; STRAND, Kalle Kallio. Tech-Savvy Men and Caring Women: Middle School Students' Gender Stereotypes Predict Interest in Tech-Education. **Sex Roles**, v. 88, p. 307–325, 2023. DOI: 10.1007/s11199-023-01353-1. Disponível em: https://doi.org/10.1007/s11199-023-01353-1.

WANG, Nan *et al.* On the use of explainable AI for susceptibility modeling: Examining the spatial pattern of SHAP values. **Geoscience Frontiers**, v. 15, n. 1, p. 101800, 2024. Open access under CC BY-NC-ND 4.0. DOI: 10.1016/j.gsf.2024.101800.

ZÚÑIGA-MEJÍAS, Vanessa; HUINCAHUE, Jaime. Gender stereotypes in STEM: a systemic review of studies conducted at primary and secondary school. **Educação e Pesquisa**, v. 50, e258677, 2024. DOI: 10.1590/S1678-4634202450258677. Disponível em: https://doi.org/10.1590/S1678-4634202450258677.

Apêndices

APÊNDICE A – ESTRUTURA DAS PERGUNTAS

Tabela 2 – Fichamento das questões das features identificadas pelas análises com Random Forest e SHAP values

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

Questionamento	**Tipo**	**Opções Simplificadas**	**Feature**
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
Você já sentiu, em algum momento, que as áreas de ciência, tecnologia, engenharia ou matemática (STEM) não eram para você?	Seleção única	Sim; Não; Nunca pensei sobre;	feltSTEMNotForMeEver
Os professores das ciências exatas que já admirei, em sua maioria, são:	Seleção única	Masculino; Feminino; Ambos; Nenhum;	admiredExactTeachersGender
Na sua experiência escolar, quem você percebia demonstrar mais interesse por disciplinas da área de exatas (como matemática, física, química ou tecnologia)?	Seleção única	Meninas; Meninos; Não observo diferença;	schoolExactInterestByGender
De que forma sua família valorizava seu desempenho escolar durante sua trajetória na educação básica?	Seleção única	Muito; Moderadamente; Eventualmente; Não demonstrava interesse; Não valorizava	familySchoolPerformanceValue
Você já teve interesse em seguir uma carreira na área de STEM (Ciência, Tecnologia, Engenharia ou Matemática)?	Seleção única	Sim; Não; Não tenho certeza;	stemCareerEverInterest
Me sinto excluído(a) em conversas sobre assuntos de exatas (física, matemática, química)	Seleção única	Escala likert	feelingExcludedExact

APÊNDICE B – TABELAS DE RESULTADOS DOS CLUSTERS GERAL

Tabela 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 – 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 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 6 – Tabela de tradução da variável feltSTEMNotForMeEver

Valor Original	Codificado	Normalizado
nao-nunca-senti	0	0.00
nunca-pensei	1	0.50
sim-ja-senti	2	1.00

Tabela 7 – 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 8 – Tabela de tradução da variável admiredExactTeachersGender

Valor Original	**Codificado**	**Normalizado**
ambos	0	0.00
feminino	1	1.00
masculino	2	2.00
nenhum	3	3.00

Tabela 9 – Tabela de tradução da variável schoolExactInterestByGender

Valor Original	**Codificado**	**Normalizado**
meninas	0	0.00
meninos	1	1.00
nao-diferenca	2	2.00

Tabela 10 – Tabela de tradução da variável familySchoolPerformanceValue

Valor Original	**Codificado**	**Normalizado**
nao-demonstrava-interesse	0	0.00
valorizava-eventualmente	1	0.33
valorizava-moderadamente	2	0.67
valorizava-muito	3	1.00

Tabela 11 – Tabela de tradução da variável feelingExcludedExact

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 12 – Tabela de tradução da variável stemCareerEverInterest

Valor Original	**Codificado**	**Normalizado**
nao	0	0.00
nao-tenho-certeza	1	0.50
sim	2	1.00

Tabela 13 – Tabela de tradução da variável childhoodHobbiesTime

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

Tabela 14 – 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 15 – Tabela de tradução da variável preCollegeExactInterestLevel

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

Tabela 16 – Tabela de tradução da variável teacherPerformanceValue

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

APÊNDICE C – TABELAS DE TRADUÇÃO DE FEATURES

Tabela 17 – Tabela de tradução da variável gender

Valor Original	**Codificado**
genero-feminino	0
genero-masculino	1

Tabela 18 – 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 19 – 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 20 – 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 21 – Tabela de tradução da variável feltSTEMNotForMeEver

Valor Original	**Codificado**	**Normalizado**
nao-nunca-senti	0	0.00
nunca-pensei	1	0.50
sim-ja-senti	2	1.00

Tabela 22 – 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 23 – Tabela de tradução da variável admiredExactTeachersGender

Valor Original	**Codificado**	**Normalizado**
ambos	0	0.00
feminino	1	1.00
masculino	2	2.00
nenhum	3	3.00

Tabela 24 – Tabela de tradução da variável schoolExactInterestByGender

Valor Original	**Codificado**	**Normalizado**
meninas	0	0.00
meninos	1	1.00
nao-diferenca	2	2.00

Tabela 25 – Tabela de tradução da variável familySchoolPerformanceValue

Valor Original	**Codificado**	**Normalizado**
nao-demonstrava-interesse	0	0.00
valorizava-eventualmente	1	0.33
valorizava-moderadamente	2	0.67
valorizava-muito	3	1.00