

UNIVERSIDADE FEDERAL DE VIÇOSA

Estratégias de acolhimento versus evasão discente no primeiro ano: uma abordagem envolvendo cursos superiores na área de computação da Universidade Federal de Viçosa

Pedro Cardoso de Carvalho Mundim
Magister Scientiae

**FLORESTAL - MINAS GERAIS
2025**

PEDRO CARDOSO DE CARVALHO MUNDIM

Estratégias de acolhimento versus evasão discente no primeiro ano: uma abordagem envolvendo cursos superiores na área de computação da Universidade Federal de Viçosa

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

Orientadora: Thais R. de M. Braga Silva

Coorientadores: Gláucia Braga e Silva
Daniel Mendes Barbosa

Ficha catalográfica elaborada pela Biblioteca da Universidade Federal de Viçosa - Campus Florestal

T

M965e
2025
Mundim, Pedro Cardoso de Carvalho, 2000-
Estratégias de acolhimento versus evasão discente no primeiro ano: uma abordagem envolvendo cursos superiores na área de computação da Universidade Federal de Viçosa / Pedro Cardoso de Carvalho Mundim. – Florestal, MG, 2025.
1 dissertação eletrônica (79 f.): il. (algumas color.).

Orientador: Thais Regina de Moura Braga Silva.
Dissertação (mestrado) - Universidade Federal de Viçosa, Instituto de Ciências Exatas e Tecnológicas, 2025.

Referências bibliográficas: f. 73-79.

DOI: <https://doi.org/10.47328/ufvcaf.2025.023>

Modo de acesso: World Wide Web.

1. Computação. 2. Evasão universitária. 3. Aprendizado do computador. I. Silva, Thais Regina de Moura Braga, 1982-. II. Universidade Federal de Viçosa. Instituto de Ciências Exatas e Tecnológicas. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 23. ed. 004

PEDRO CARDOSO DE CARVALHO MUNDIM

Estratégias de acolhimento versus evasão discente no primeiro ano: uma abordagem envolvendo cursos superiores na área de computação da Universidade Federal de Viçosa

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

APROVADA: 5 de dezembro de 2025.

Assentimento:

Pedro Cardoso de Carvalho Mundim
Autor

Thais Regina de Moura Braga Silva
Orientadora

Essa dissertação foi assinada digitalmente pelo autor em 30/12/2025 às 20:55:18 e pela orientadora em 05/01/2026 às 13:01:11. As assinaturas têm validade legal, conforme o disposto na Medida Provisória 2.200-2/2001 e na Resolução nº 37/2012 do CONARQ. Para conferir a autenticidade, acesse <https://siadoc.ufv.br/validar-documento>. No campo 'Código de registro', informe o código **WWEY.TS1A.T7AA** e clique no botão 'Validar documento'.

Dedico este trabalho, em especial, aos meus pais, pelo apoio constante, aos amigos que compartilharam desta jornada e a todos que valorizam a importância do conhecimento e da educação.

AGRADECIMENTOS

Agradeço primeiramente a Deus por sempre estar comigo, dando forças para que eu não desistisse e por despertar em mim o interesse e a paixão pela área da educação. Agradeço à minha orientadora, Prof.^a Dr.^a Thais Regina de Moura Braga Silva, e aos meus coorientadores, Prof.^a Dr.^a Gláucia Braga e Silva e Prof. Dr. Daniel Mendes Barbosa, pela orientação, paciência e discussões enriquecedoras. Agradeço à Universidade Federal de Viçosa (UFV), aos professores do Programa de Pós-Graduação em Ciência da Computação (PPGCC) pela formação, aos colaboradores dos três campi da UFV, Vânia Valente, Renato Ruas e Benício Ramalho e a todos os coordenadores de curso que colaboraram para a conclusão da pesquisa.

Agradeço também à minha família, especialmente aos meus pais, Elizabeth Cardoso de Carvalho e Oscavu Alves Mundim, pelo amor, apoio e incentivo incondicional em todas as etapas desta jornada. Tudo que vivo hoje é fruto das condições que vocês me proporcionaram. À minha psicóloga Márcia Lelis e Silva, minha sincera gratidão pelo acolhimento, pela escuta atenta e pelo apoio fundamental ao longo dessa caminhada. Ao meu círculo de amigos da UFV, Otávio Gomes, Igor Lucas, Lucas Souza, Daniel Louzada, Douglas Diones, Marcos Biscotto, Gabriel Bonfim, Estela Batista, Paula Gibrim e Ludmila Souza, agradeço por estarem ao meu lado ao longo de toda essa jornada, oferecendo companheirismo, motivação e incentivo, especialmente nos momentos em que mais precisei. Aos amigos que, mesmo de longe, sempre estiveram presentes, Beatriz Braga, Beatriz Pinheiro, Anelyse Abbad e Antônio Abud, agradeço pelo incentivo constante e pela confiança em mim.

Por fim, agradeço ao Universo Narrado, em especial ao Felipe Guisoli e à Júnia, minha sincera gratidão por me acolher profissionalmente logo após o término da bolsa, proporcionando não apenas um emprego incrível, mas também um ambiente de trabalho inspirador e acolhedor. Essa oportunidade foi essencial para que eu pudesse dar continuidade e concluir este trabalho com dedicação e tranquilidade.

Este trabalho foi realizado com o apoio das seguintes agências de pesquisa brasileiras: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

"Se estiver se sentindo desmotivado ou sentindo que não é bom o suficiente, incendeie o seu coração. Enxugue as lágrimas e siga em frente. Quando se entristecer ou se acovardar, lembre-se de que o fluxo do tempo nunca para. Ele não vai te esperar enquanto você se afoga em tristeza."
(Kyojuro Rengoku)

RESUMO

MUNDIM, Pedro Cardoso de Carvalho, M.Sc., Universidade Federal de Viçosa, dezembro de 2025. **Estratégias de acolhimento versus evasão discente no primeiro ano: uma abordagem envolvendo cursos superiores na área de computação da Universidade Federal de Viçosa.** Orientadora: Thais Regina de Moura Braga Silva. Coorientadores: Glaucia Braga e Silva e Daniel Mendes Barbosa.

A evasão no ensino superior é um dos principais desafios enfrentados pelas instituições, com impactos acadêmicos, sociais e financeiros significativos. Entre as áreas de conhecimento, os cursos da área de computação merecem destaque por apresentarem elevadas taxas de evasão, em grande parte devido à dificuldade nos primeiros períodos, quando disciplinas introdutórias como Programação são ofertadas. Nos últimos anos, observa-se uma crescente utilização de técnicas de análise exploratória e de aprendizado de máquina para identificar padrões e prever a evasão, geralmente com base em atributos de dados pessoais, desempenho acadêmico e quantidade de reprovações. No entanto, esses atributos dependem de informações obtidas após certo tempo de curso, o que limita sua utilidade em ações preventivas. Dessa forma, torna-se relevante investigar fatores disponíveis desde o ingresso do estudante, como as estratégias de acolhimento, que surgem como alternativas promissoras, especialmente no primeiro ano do curso, mas ainda pouco exploradas na literatura. Assim, o objetivo desta pesquisa é avaliar se as estratégias de acolhimento adotadas nos cursos de computação da Universidade Federal de Viçosa (UFV) possuem impacto na evasão no primeiro ano. Os resultados indicaram que a adoção progressiva dessas estratégias esteve associada a reduções de até 20 pontos percentuais nas taxas médias de evasão. Além disso, a integração dos dados acadêmicos e sociodemográficos com os atributos sobre acolhimento elevou o desempenho de modelos de aprendizado de máquina, com destaque para a Regressão Logística e o Support Vector Machine (SVM). Esses resultados indicam que tais atributos podem funcionar como indicadores precoces de evasão e, ao mesmo tempo, sugerem uma possível relação com o fortalecimento do sentimento de pertencimento dos calouros, fornecendo suporte para intervenções mais direcionadas por parte das instituições de ensino.

Palavras-chave: evasão no ensino superior; aprendizado de máquina; estratégias de acolhimento; cursos de computação; evasão no primeiro ano

ABSTRACT

MUNDIM, Pedro Cardoso de Carvalho, M.Sc., Universidade Federal de Viçosa, December, 2025. **Onboarding strategies versus freshmen dropout: an approach involving higher education computing programs at the Federal University of Viçosa.** Adviser: Thais Regina de Moura Braga Silva. Co-advisers: Glaucia Braga e Silva and Daniel Mendes Barbosa.

Dropout in higher education remains one of the main challenges faced by institutions, with significant academic, social, and financial impacts. Among the different fields of knowledge, computing programs deserve special attention due to their high dropout rates, largely related to the difficulties encountered in the first semesters, when introductory courses such as Programming are offered. In recent years, there has been a growing use of exploratory analysis and machine learning techniques to identify patterns and predict dropout, usually based on attributes such as personal data, academic performance, and the number of course failures. However, these attributes depend on information collected only after some time in the program, which limits their usefulness in preventive actions. Therefore, it becomes relevant to investigate factors available since students' entry into the university, such as onboarding strategies, which have emerged as promising alternatives, especially in the first year, but are still little explored in the literature. The objective of this research is to evaluate whether the onboarding strategies adopted in the computing programs at the Federal University of Viçosa (UFV) have an impact on first-year dropout. The results indicated that the progressive adoption of these strategies was associated with reductions of up to 20 percentage points in the average dropout rates. Furthermore, the integration of academic and sociodemographic data with information on onboarding strategies improved the performance of machine learning models, especially for Logistic Regression and Support Vector Machine (SVM). These findings suggest that such attributes can serve as early indicators of dropout and may also be related to the strengthening of freshmen's sense of belonging, thus supporting targeted institutional interventions.

Keywords: dropout in higher education; machine learning; onboarding strategies; computing programs; first year dropout

LISTA DE FIGURAS

1.1	Metodologia.	15
2.1	Processo de seleção dos trabalhos.	21
2.2	Distribuição da quantidade de trabalhos por ano de publicação.	23
3.1	Adesão dos cursos dos três campi da UFV à pesquisa.	34
3.2	Mapeamento das estratégias de acolhimento da UFV.	37
3.3	Evolução cumulativa da adoção de estratégias por categoria.	37
3.4	Estratégias de acolhimento dos cursos da área de computação da UFV.	38
3.5	Evolução cumulativa das subcategorias das estratégias por campus nos cursos de computação da UFV.	39
3.6	Evasão média de calouros por campus e curso de computação da UFV.	41
4.1	Atributos da base de dados das estratégias de acolhimento.	52
4.2	Atributos da base de dados das estratégias de acolhimento dos cursos de computação da UFV.	54
4.3	Processo de preparação dos dados.	56
4.4	Quantidade de alunos em cada classe da variável alvo.	61
4.5	Evasão anual de calouros nos cursos de computação da UFV.	61
4.6	Frequência dos atributos entre os 10 mais importantes.	67

LISTA DE TABELAS

2.1	Comparativo entre os trabalhos relacionados e este MSL.	19
2.2	String de busca para o MSL.	20
2.3	Critérios de inclusão e critérios de exclusão de trabalhos.	20
2.4	Descrição dos 28 trabalhos selecionados para análise.	22
2.5	Distribuição dos cursos por trabalhos.	24
2.6	Distribuição dos trabalhos por abordagem metodológica.	24
2.7	Distribuição de trabalhos por cursos: área de computação.	25
2.8	Distribuição dos trabalhos por abordagem metodológica: área de computação.	26
2.10	Técnicas de ML dos trabalhos que envolvem a área de computação. . .	27
2.9	Dimensões dos atributos e identificadores de trabalhos.	28
3.1	Taxas de evasão anuais de calouros nos cursos de computação da UFV. .	40
4.1	Comparativo de abordagens entre os trabalhos relacionados e este estudo. .	49
4.2	Atributos da base de dados acadêmica disponibilizada.	51
4.3	Matriz de confusão.	59
4.4	Desempenho dos modelos para o curso de CdC - campus CAF.	62
4.5	Desempenho dos modelos para o curso de CdC - campus CAV.	63
4.6	Desempenho dos modelos para o curso de SI (Integral) - campus CRP. .	64
4.7	Desempenho dos modelos para o curso de SI (Noturno) - campus CRP. .	65
4.8	Posição dos dez atributos mais importantes por modelo e curso.	68

SUMÁRIO

1	INTRODUÇÃO GERAL	11
1.1	Questão de Pesquisa	12
1.2	Justificativa	13
1.3	Objetivo	13
1.4	Contribuições	14
1.5	Organização do Trabalho	14
2	EVASÃO EM CURSOS SUPERIORES NA ÁREA DE COMPUTAÇÃO: UM MAPEAMENTO SISTEMÁTICO DA LITERATURA	16
2.1	Introdução	16
2.2	Trabalhos Relacionados	17
2.3	Mapeamento Sistemático da Literatura	19
2.4	Resultados	23
2.4.1	QP1 - Qual é o panorama geral das pesquisas sobre evasão no ensino superior e, em particular, como essas pesquisas se relacionam com a área de computação?	23
2.4.2	QP2 - Como as abordagens metodológicas encontradas têm sido aplicadas para se analisar a evasão no ensino superior em cursos da área de computação?	26
2.5	Considerações Finais	27
3	ACOLHIMENTO VERSUS EVASÃO NO PRIMEIRO ANO: UMA ANÁLISE PANORÂMICA NA UFV COM ÊNFASE NOS CURSOS DE COMPUTAÇÃO	30
3.1	Introdução	31
3.2	Trabalhos Relacionados	32
3.3	Materiais e Métodos	33
3.3.1	Coleta de Dados	33
3.3.2	Análise de Dados	35
3.4	Resultados	36
3.4.1	Mapeamento das Estratégias de Acolhimento da UFV	36
3.4.2	Análise das Estratégias de Acolhimento dos Cursos de Computação da UFV	38
3.4.3	Análise das Estratégias de Acolhimento e da Evasão de Calouros nos Cursos de Computação da UFV	39
3.5	Considerações Finais	42
4	USO DE ESTRATÉGIAS DE ACOLHIMENTO COMO ATRIBUTOS PARA MODELOS DE PREVISÃO DE EVASÃO DISCENTE NO PRIMEIRO ANO EM CURSOS SUPERIORES DE COMPUTAÇÃO	44
4.1	Introdução	45
4.2	Trabalhos Relacionados	47
4.2.1	Uso de EDM e ML na Análise da Evasão	47
4.2.2	Uso de Estratégias de Acolhimento	48
4.2.3	Similaridades e Diferenças deste Trabalho	49

4.3	Materiais e Métodos	50
4.3.1	Entendimento do Problema e do Conjunto de Dados	50
4.3.2	Preparação dos Dados	53
	Filtragens	53
	Remoção de Atributos, Valores Nulos e Criação da Variável Alvo	54
	Enriquecimento da Base de Dados Acadêmicos e Engenharia de	
	Atributos	56
4.3.3	Modelagem	56
	Seleção de Atributos e Separação dos Dados	57
	Desbalanceamento dos Dados	58
4.3.4	Avaliação	59
4.4	Resultados	60
4.4.1	Análise Geral	60
4.4.2	Resultado por Curso	62
4.4.3	Análise da Importância dos Atributos	66
4.5	Considerações Finais	69
5	CONCLUSÃO GERAL	71
	REFERÊNCIAS BIBLIOGRÁFICAS	73

Capítulo 1

Introdução Geral

A evasão no ensino superior é um desafio que afeta tanto os estudantes quanto as instituições de ensino. De acordo com Coimbra et al. (2021), a evasão ocorre quando o estudante se desliga do curso antes de sua conclusão, podendo manifestar-se por diferentes formas, como abandono, cancelamento a pedido do aluno, cancelamento administrativo, transferência, falecimento, dentre outros. Suas consequências extrapolam o campo acadêmico, gerando impactos sociais, institucionais e financeiros consideráveis (Duran et al., 2023). Pesquisas como as de Flores et al. (2022) e Souza et al. (2025) evidenciam que esse fenômeno resulta de uma complexa interação entre fatores acadêmicos, socioeconômicos, institucionais, emocionais e relacionais, o que exige uma abordagem abrangente e integrada para sua compreensão e enfrentamento.

Na área de computação, a evasão tem se mostrado especialmente preocupante, uma vez que cursos dentro deste recorte apresentam desafios particulares, como elevada carga de disciplinas iniciais de programação e matemática, que costumam representar barreiras significativas para os ingressantes (Pachas et al., 2021).

Na Universidade Federal de Viçosa (UFV), este cenário também se manifesta nos três campi da instituição (Viçosa, Florestal e Rio Paranaíba), o que tem levado muitos de seus cursos, incluindo os da área de computação, sendo eles Ciência da Computação nos campi Viçosa e Florestal e Sistemas de Informação (Integral e Noturno) no campus Rio Paranaíba, a tentar compreender e mitigar as dificuldades enfrentadas pelos estudantes nos períodos iniciais. Uma das maneiras de se abordar essa questão tem sido por meio da aplicação de diversas estratégias de acolhimento, que envolvem orientação, apoio pedagógico, psicológico e social, especialmente nos períodos iniciais dos cursos, como propósito de facilitar a adaptação dos calouros e reduzir sua evasão (Silva et al., 2021b).

A necessidade dessas estratégias torna-se evidente ao se observar as taxas de evasão do primeiro ano (calouros) nos cursos de computação da UFV: 16,2% em Ciência da Computação (campus Viçosa), 22,3% em Ciência da Computação (campus Florestal), 25,4% em Sistemas de Informação - Integral (campus Rio Paranaíba) e 30,6% em Sistemas de Informação - Noturno (campus Rio Paranaíba). Para fins institucionais, a UFV considera como tipos de evasão¹ as situações classificadas como abandono (A), desligamento (D), falecimento (F), mudança de curso (M),

¹Conceitos Acadêmicos da UFV: <https://sre.caf.ufv.br/avaliacao-do-rendimento-academico-2/>

desligamento com pedido de reconsideração (R), transferência (T) e exclusão (X), que representam formas de saída definitiva do curso ou da universidade.

Diante do problema da evasão apresentado, tecnologias emergentes vêm se consolidando como aliadas no seu estudo (Santos et al., 2019). A mineração de dados educacionais (*Educational Data Mining* - EDM), por exemplo, tem se destacado por empregar técnicas analíticas sobre grandes volumes de dados educacionais com o objetivo de extrair padrões relevantes sobre a evasão (da Silva et al., 2019b). Além de descrever e diagnosticar causas, a EDM também tem sido aplicada à predição da evasão por meio de técnicas de aprendizado de máquina (*Machine Learning* - ML). Estudos como os de Santos et al. (2019) e Couto and Santana (2017) treinaram modelos de ML com algoritmos amplamente utilizados como Árvores de Decisão, *Random Forest* e Redes Bayesianas para identificar estudantes em risco de evasão, considerando variáveis como desempenho acadêmico, frequência e quantidade de reprovações. Os resultados indicaram que, no primeiro estudo, modelos baseados em Árvores de Decisão e *Random Forest* alcançaram acurácia média entre 66% e 72%. Já no segundo, modelos com Redes Bayesianas obtiveram acurácia de 86%.

As abordagens para o estudo da evasão no ensino superior variam desde análises exploratórias até métodos preditivos baseados em EDM e ML, geralmente focando em variáveis quantitativas de mensuração mais direta e objetiva, baseadas apenas em parâmetros do próprio estudante após algum tempo de curso, como o rendimento acadêmico e o número de reprovações. De modo geral, esses estudos analisam a evasão como um fenômeno amplo, sem distinção entre perfis ou momentos específicos da trajetória discente, o que limita a compreensão, por exemplo, sobre os fatores que afetam de forma particular os calouros.

Portanto, aspectos institucionais, como as estratégias de acolhimento, ainda carecem de maior atenção na literatura, especialmente em relação à evasão. Assim, surge uma lacuna na investigação sobre como e em que medida essas estratégias impactam na evasão do ensino superior, particularmente nos cursos da área de computação ofertados pela UFV em seus três campi.

1.1 Questão de Pesquisa

Diante da possível relevância do acolhimento institucional e da crescente utilização de técnicas de EDM e ML no estudo da evasão, surge a seguinte questão de pesquisa: "Como as estratégias de acolhimento implementadas pelos cursos da área de computação da Universidade Federal de Viçosa (UFV) podem contribuir, de forma mensurável, para a redução e predição da evasão durante o primeiro ano?"

1.2 Justificativa

A evasão em cursos de STEM (Ciência, Tecnologia, Engenharia e Matemática) é um fenômeno conhecido internacionalmente (Cavalcanti et al. (2025)), com taxas particularmente elevadas no primeiro ano, representando desafios significativos para a formação de profissionais qualificados e para o aproveitamento dos investimentos públicos em educação superior. Além disso, a evasão acarreta impactos sociais e econômicos (Máximo et al. (2024)), tanto para os estudantes, que enfrentam atrasos na formação e custos pessoais, quanto para a sociedade, que perde mão de obra especializada em áreas estratégicas (da Silveira and Tonini (2023)). Embora diversas ações de acolhimento tenham sido implementadas ao longo dos últimos anos, estudos anteriores têm se limitado a análises qualitativas, sem mensurar seus efeitos sobre a retenção dos ingressantes. Nesse contexto, o presente estudo se justifica pela necessidade de avaliar quantitativamente o impacto das estratégias de acolhimento nos cursos de computação da UFV, oferecendo evidências sobre sua eficácia na redução da evasão no primeiro ano e subsidiando políticas institucionais que promovam a permanência estudantil.

1.3 Objetivo

O objetivo geral desta pesquisa é realizar uma avaliação quantitativa do impacto das estratégias de acolhimento implementadas pelos cursos da área de computação da UFV na evasão no primeiro ano. Para atingir esse objetivo, mostra-se necessário os seguintes objetivos específicos:

- Mapear a literatura relacionada à evasão em cursos da área de computação, buscando identificar os principais fatores associados ao fenômeno, bem como as abordagens metodológicas, técnicas de análise e variáveis preditoras mais recorrentes nos estudos existentes.
- Construir um conjunto de dados enriquecido com informações acadêmicas, demográficas e institucionais dos estudantes dos cursos de computação da UFV, incorporando também dados referentes às estratégias de acolhimento implementadas, de modo a possibilitar análises exploratórias que revelem padrões relevantes.
- Desenvolver e avaliar modelos de ML voltados à previsão de evasão de calouros dos cursos da área de computação da UFV, utilizando novos atributos relacionados às estratégias de acolhimento.

1.4 Contribuições

- Artigo publicado: **“Evasão em cursos superiores na área de Computação: um Mapeamento Sistemático da Literatura”** (Mundim et al. (2025)), nos *Anais do XXXIII Workshop sobre Educação em Computação* (WEI 2025, Qualis A4).
- Artigo aceito para publicação: **"Acolhimento versus Evasão no primeiro ano: Uma Análise panorâmica na UFV com ênfase nos cursos de Computação"**, na *Revista Novas Tecnologias na Educação* (RENOTE, Qualis A4).
- Artigo submetido: **"Uso de Estratégias de Acolhimento como atributos para modelos de previsão de Evasão Discente no Primeiro Ano em cursos superiores de Computação"**, para a *Revista Brasileira de Informática na Educação* (RBIE, Qualis A4).

Os resultados deste estudo se materializam em três frentes diferentes. Inicialmente, destaca-se o mapeamento sistemático da literatura, o qual trouxe as principais metodologias utilizadas no estudo da evasão, bem como os atributos mais utilizados para tal fim. Em segundo lugar, tem-se o mapeamento das estratégias de acolhimento adotadas nos cursos de graduação da UFV, com foco nos cursos de computação, a partir de uma categorização que organiza e descreve as ações institucionais voltadas à integração e à permanência estudantil. Esse processo resultou, por fim, na construção de uma base de dados enriquecida, de uso exclusivo da instituição, que integra informações acadêmicas e institucionais a novos atributos relacionados ao acolhimento para uso em treinamentos de modelos de ML.

Os achados da pesquisa indicam que o uso contínuo e diversificado das estratégias de acolhimento parece estar associado à redução da evasão nos cursos da área de computação da UFV. Além disso, evidenciam que os novos atributos de acolhimento podem ser empregados em modelos preditivos de evasão, oferecendo a vantagem de estarem disponíveis desde o ingresso dos estudantes.

1.5 Organização do Trabalho

A organização desta dissertação segue o padrão estabelecido pela Comissão do Programa de Pós-Graduação em Ciência da Computação (PPGCC) da UFV, sendo estruturada em formato de coletânea composta por três artigos. Cada artigo corresponde a uma etapa da metodologia adotada (Figura 1.1), contribuindo de forma complementar para o alcance dos objetivos da pesquisa.

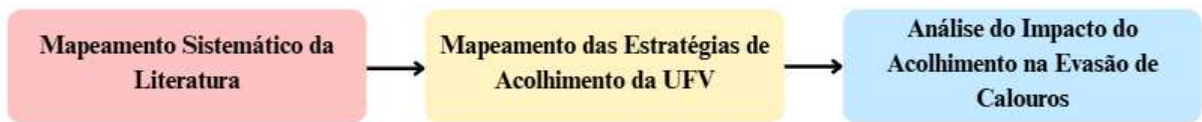


Figura 1.1: Metodologia.

O primeiro artigo, apresentado no Capítulo 2, corresponde à etapa inicial do estudo, na qual foi conduzido um Mapeamento Sistemático da Literatura (MSL). O objetivo dessa etapa foi identificar as principais metodologias, algoritmos e atributos utilizados em pesquisas sobre evasão no ensino superior, com ênfase em cursos da área de computação. A partir dessa revisão, foi possível delinear o estado da arte e reconhecer lacunas na literatura, orientando as etapas subsequentes da pesquisa.

O segundo artigo, apresentado no Capítulo 3, refere-se à etapa de mapeamento institucional, cujo propósito foi identificar e categorizar as estratégias de acolhimento adotadas nos cursos de graduação da UFV. Essa etapa buscou organizar e descrever as ações voltadas à integração e permanência estudantil, especialmente no contexto dos cursos de computação, gerando uma base de dados consolidada a ser utilizada nas análises posteriores.

Por fim, o terceiro artigo, apresentado no Capítulo 4, corresponde à etapa analítica da pesquisa. Seu objetivo foi avaliar o impacto dos atributos de acolhimento na evasão de calouros dos cursos de computação da UFV, por meio da aplicação de técnicas de EDM e ML. Essa etapa permitiu investigar se a inclusão desses atributos possui impacto na evasão no primeiro ano e se contribui para melhorar o desempenho preditivo dos modelos.

Capítulo 2

Evasão em cursos superiores na área de Computação: Um Mapeamento Sistemático da Literatura

Mundim, P., Silva, T., e Silva, G. B., and Barbosa, D. (2025). Evasão em cursos superiores na área de computação: Um mapeamento sistemático da literatura. In *Anais do XXXIII Workshop sobre Educação em Computação*, pages 1055–1067, Porto Alegre, RS, Brasil. SBC.

Resumo: A evasão estudantil nos cursos superiores de computação afeta tanto os estudantes quanto as instituições, impactando a estrutura universitária, a trajetória acadêmica e o mercado de trabalho. Este artigo apresenta um Mapeamento Sistemático da Literatura (MSL) sobre essa questão, com base em 28 trabalhos encontrados nas bases SOL, ACM, Scopus e IEEE. Os resultados indicam que a área de computação é bem representativa nos estudos analisados (18/28) e que a análise exploratória e o uso de algoritmos de aprendizado de máquina foram as abordagens metodológicas predominantes nos estudos primários. Também foram identificados e categorizados os principais atributos relacionados à evasão e a tendência de desenvolver modelos para prever a evasão.

Abstract: Student dropout in higher education computer courses affects both students and institutions, impacting the university structure, academic trajectories, and the job market. This paper presents a Systematic Literature Mapping (SLM) on this issue, based on 28 studies found in the SOL, ACM, Scopus, and IEEE databases. The results indicate that the field of computing is well represented in the analyzed studies (18/28), and that exploratory analysis and the use of machine learning algorithms were the predominant methodologies in the primary studies. The main attributes related to dropout were identified, along with a trend toward developing models to predict dropout.

2.1 Introdução

A evasão no ensino superior impacta alunos e instituições em todo o mundo, gerando consequências sociais e financeiras que afetam a estrutura universitária e o percurso acadêmico dos discentes (da Cruz et al., 2023). Considerando a área de computação,

Duran et al. (2023) destacam que a evasão está relacionada a questões estruturais no sistema educacional, influenciadas por fatores acadêmicos, socioeconômicos, de etnia/raça, de gênero e emocionais, internos e externos às instituições. Analisar a evasão nesse contexto é essencial, dado seu histórico elevado e a crescente demanda por profissionais em um mercado em expansão (Menolli et al., 2020).

Este trabalho apresenta um Mapeamento Sistemático da Literatura (MSL), que tem como objetivo identificar e analisar pesquisas publicadas no período de 2018 a 2023 sobre a evasão no ensino superior, com foco em cursos da área de computação, destacando as abordagens metodológicas utilizadas e como foram aplicadas. A necessidade deste MSL está em, portanto, consolidar o conjunto de publicações encontrados na literatura, fornecendo às instituições e pesquisadores *insights* que contribuam para mitigar o problema.

Como principais resultados deste MSL, observou-se que as abordagens metodológicas utilizadas pelos trabalhos primários considerados neste contexto são a análise exploratória e a modelagem de algoritmos de aprendizado de máquina (*Machine Learning* - ML) para se tentar prever a evasão. A partir desta análise, foi possível ainda construir uma categorização dos atributos utilizados, composta por quatro dimensões: *Demográfico Socioeconômico e Pessoal; Conhecimento Prévio; Institucional; e Desempenho Acadêmico*.

O artigo está organizado de forma que: a Seção 2.2 traz os trabalhos relacionados encontrados na literatura; a Seção 2.3 descreve o protocolo para a condução do MSL; a Seção 2.4 apresenta os resultados; e, a Seção 2.5 traz as considerações finais.

2.2 Trabalhos Relacionados

Para identificar os trabalhos relacionados, foram utilizadas as bases SOL, ACM, Scopus e IEEE, usando os termos da Tabela 2.2¹. A seleção incluiu artigos de 2018 a 2023, em inglês e português, sobre evasão no ensino superior e que fossem Revisões Sistemáticas da Literatura (RSL) ou MSL, visando consolidar estudos sobre o tema.

Em Silva et al. (2023) foi realizada uma RSL com objetivo de identificar os principais fatores associados à predição de abandono e retenção no ensino superior utilizados por estudos primários. As bases de dados utilizadas pelos autores foram SCOPUS e IEEE (sem o uso de bases nacionais), tendo sido selecionados pela *string* de busca, 52 trabalhos para serem analisados. A *string* foi composta pelos termos: *student, undergraduate, predict, dropout, retention, attrition, metric, measurement, indicate, higher education e bachelor degree*. Os resultados revelaram 29 fatores utilizados pelos trabalhos, em que alguns dos mais citados foram média geral, gênero, curso, idade,

¹Foram utilizados exatamente os mesmos termos da *string* de busca definida para o MSL.

etnia, horário de curso e tipo de ingresso na universidade. Esses fatores foram classificados pelos autores em três grandes grupos: *Demográficos*, *Acadêmicos* e de *Aprendizagem*. Os fatores mais recorrentes em cada grupo foram, respectivamente, gênero, média geral e regularidade de estudo.

Os autores em [da Silva et al. \(2019b\)](#) apresentam um MSL que investiga a área de mineração de dados educacionais (*Educational Data Mining - EDM*), com foco na identificação de tendências e contribuições para estudos futuros. Os 357 trabalhos primários foram selecionados a partir das bases da ACM, IEEE e por consultas ao Google Acadêmico (sem o uso de bases nacionais), por meio de uma *string* de busca que possuía os termos mineração de dados educacionais e *educational data mining*. Eles foram categorizados em três critérios: *Foco de Pesquisa*, *Tipo de Contribuição* e *Tipo de Pesquisa*. Os resultados revelaram que a maioria dos estudos está centrada no desempenho acadêmico dos alunos e na identificação de fatores relacionados à evasão escolar, com foco na predição e relação entre eles.

O trabalho de [Oliveira and Barbosa \(2023\)](#) descreve uma RSL sobre o uso de modelos de aprendizado de máquina multiníveis para análise e previsão de evasão escolar. O objetivo dos autores foi oferecer uma visão abrangente sobre os fatores associados à evasão em diferentes níveis educacionais, além de investigar o uso de técnicas de inteligência artificial no aprimoramento e interpretação desses modelos. Foram utilizadas as bases de pesquisa ACM, IEEE, SCOPUS, Elsevier, Engineering Village e Web of Science (sem o uso de bases nacionais) e, no total, 36 estudos primários foram selecionados. Alguns termos de busca foram: *multilevel, hierarchical, nested, mixed, random coefficients, analysis, model, academic, student, dropout* e *evasion*. Os achados da pesquisa destacaram que o ambiente educacional mais pesquisado é o ensino médio, seguido pelos níveis secundário e superior. Os fatores de impacto para a evasão, na modelagem dos trabalhos primários, foram divididos em três níveis: *Estudantil*, *Escolar* e *Outros*. No primeiro, prevaleceram aspectos demográficos e socioeconômicos; no segundo, questões estruturais; e no terceiro os demais. Entre os métodos mais usados, sobressaíram a Regressão Linear e a Regressão Logística.

Em [Colpo et al. \(2020\)](#), os autores realizaram uma RSL sobre a previsão da evasão estudantil com técnicas de EDM, analisando trabalhos do Congresso Brasileiro de Informática na Educação (CBIE). Foram encontrados 104 trabalhos no período de 2006 - 2019, dos quais 23 foram para análise. Alguns termos utilizados para a seleção foram evasão e *dropout*. Os resultados indicaram que a maioria das pesquisas utiliza a classificação, destacando algoritmos de Árvores de Decisão implementados no Weka.

Já [dos Santos et al. \(2021b\)](#) analisaram 50 trabalhos nacionais e internacionais sobre EDM aplicados à evasão escolar, examinando ferramentas, algoritmos e bases de dados. Os termos da busca foram *prediction of students, school dropout, school retention, school failure, educational data mining, knowledge discovery, machine learning,*

institution e university. Os resultados indicam a predominância do Weka e do Scikit-learn e, entre os algoritmos mais recorrentes, destacam-se *Árvore de Decisão*, *Naive Bayes*, *Random Forest* e *Support Vector Machine (SVM)*, com os dois últimos apresentando melhor desempenho. As bases de dados analisadas possuem entre 501 e 5.000 registros, contendo atributos acadêmicos e fatores socioeconômicos.

A Tabela 2.1 compara os trabalhos discutidos com este MSL. Diferente dos estudos anteriores, que abrangem diversos níveis de ensino ou áreas não especificadas, este MSL identifica uma predominância de pesquisas voltadas para cursos da área de computação no ensino superior. Além disso, este trabalho considera um período mais recente, de 2018 a 2023, o que permite analisar tendências e padrões atuais, diferentemente, por exemplo, de [Silva et al. \(2023\)](#) e [de Oliveira and Barbosa \(2023\)](#), que analisaram períodos iniciados nas décadas de 1990. Por fim, os trabalhos de [da Silva et al. \(2019b\)](#), [de Oliveira and Barbosa \(2023\)](#), [Colpo et al. \(2020\)](#) e [dos Santos et al. \(2021b\)](#) estão restritos a um recorte de pesquisas primárias mais específicas, sendo, trabalhos sobre EDM e modelos de ML.

Tabela 2.1: Comparativo entre os trabalhos relacionados e este MSL.

Trabalho	Tipo	Nível de ensino	Curso(s)	Período	Qtd. de Trabalhos
Silva et al. (2023)	RSL	Superior	Não especificado	1993-2021	52
da Silva et al. (2019b)	MSL	Técnico e Superior	Não especificado	2014-2019	357
de Oliveira and Barbosa (2023)	RSL	Fundamental, Médio e Superior	Não especificado	1997-2022	36
Colpo et al. (2020)	RSL	Fundamental, Médio, Técnico e Superior	Não especificado	2006-2019	23
dos Santos et al. (2021b)	RSL	Fundamental, Médio, Técnico e Superior	Não especificado	2008-2020	50
Este trabalho	MSL	Superior	Área de computação	2018-2023	28

2.3 Mapeamento Sistemático da Literatura

A estrutura deste MSL foi definida em três fases – planejamento, condução e documentação – seguindo as diretrizes estabelecidas para estudos dessa natureza ([Kitchenham et al., 2007](#)). Esse delineamento permite organizar o estudo de forma sistemática, assegurando que cada etapa contribua para a construção de um panorama consistente e embasado sobre o tema investigado.

Na fase de planejamento, as seguintes questões de pesquisa (QP) foram definidas: **QP1** - Qual é o panorama geral das pesquisas sobre evasão no ensino superior e, em particular, como essas pesquisas se relacionam com a área de computação?; **QP2** - Como as abordagens metodológicas encontradas têm sido aplicadas para se analisar

a evasão no ensino superior em cursos da área de computação?

Em seguida, foi definida a *string* de busca (Tabela 2.2) a ser utilizada nas buscas avançadas das bases de pesquisa, composta por duas partes: termos sobre evasão e retenção; e termos referentes a tipos de Instituições de Ensino Superior (IESs).

Tabela 2.2: String de busca para o MSL.

	Palavras-Chave
Evasão/Retenção	(evasão, retenção, abandono, desistência, transferência, mudança de curso, dropout, attrition, abandonment, retention, failure, desertion, transfer, change of course, persistence)
Tipos de IESs	(universidade pública, universidade federal, instituto federal, instituição de ensino superior, public university, federal university, federal institute, higher education institution)
String de Busca	("evasão" OR "abandono" OR "desistência" OR "retenção" OR "transferência" OR "mudança de curso" OR "dropout" OR "attrition" OR "abandonment" OR "retention" OR "failure" OR "desertion" OR "transfer" OR "change of course" OR "persistence") AND ("universidade pública" OR "universidade federal" OR "instituto federal" OR "instituição de ensino superior" OR "public university" OR "federal university" OR "federal institute" OR "higher education institution")

Para execução da *string* de busca, foram consideradas quatro bases científicas: SOL², ACM Digital Library³, Scopus⁴ e IEEE Xplore⁵, por serem as principais bases de estudos científicos da área de computação. Por fim, para seleção dos estudos, a Tabela 2.3 apresenta os Critérios de Inclusão (CI) e de Exclusão (CE) definidos de modo a selecionar os trabalhos de maior qualidade e relevância para este estudo.

Tabela 2.3: Critérios de inclusão e critérios de exclusão de trabalhos.

Critérios de Inclusão
CI1 - Trabalhos que abordem evasão no ensino superior presencial, incluindo estudos tanto da área de computação quanto de outras áreas.
CI2 - Trabalhos primários.
Critérios de Exclusão
CE1 - Trabalhos duplicados.
CE2 - Trabalhos que não estejam em inglês ou português.
CE3 - Trabalhos publicados até 2022 que não foram citados nenhuma vez.
CE4 - Trabalhos publicados em veículos sem Qualis ou com Qualis C.

A fase de condução abrangeu a execução da *string* de busca em cada uma das bases citadas e a seleção dos trabalhos, a partir da aplicação dos CI e CE definidos,

²SOL: <https://sol.sbc.org.br/index.php/indice>

³ACM: <https://dl.acm.org/>

⁴Scopus: <https://www.scopus.com/>

⁵IEEE Xplore: <https://ieeexplore.ieee.org/Xplore/home.jsp>

conforme ilustra a Figura 2.1. Visando a seleção de trabalhos mais recentes, este MSL foi realizado nas bases de dados com a filtragem no intervalo de 2018 a 2023. Com a execução da *string* de busca nas bases, foram retornados 6158 trabalhos.

A execução das buscas ocorreu nos meses de novembro e dezembro de 2023. O processo de seleção dos estudos, incluindo triagem por títulos, palavras-chave, resumos e leitura completa, quando aplicável, foi conduzido pelo primeiro autor do trabalho, e as etapas de aplicação dos CI e CE foram revisados e validados pelos demais autores ao longo do processo, de modo a assegurar consistência nas decisões.

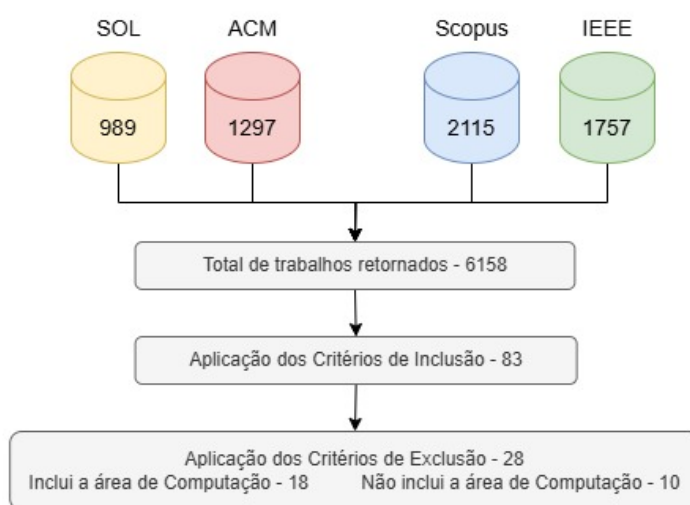


Figura 2.1: Processo de seleção dos trabalhos.

A partir dos 6158 trabalhos retornados, inicialmente, aplicaram-se os CI para identificar quais estudos estavam alinhados com o tema da evasão no ensino superior presencial, incluindo aqueles relacionados a cursos na área de computação, e verificar se eram trabalhos primários. Os critérios foram aplicados aos títulos, resumos e às palavras-chave dos trabalhos, resultando em 83 trabalhos selecionados.

Sobre os 83 trabalhos previamente selecionados, foram aplicados os CE, com os seguintes resultados: o CE1 eliminou dez trabalhos que eram duplicados; com o CE2 foram excluídos sete trabalhos que não estavam redigidos em inglês ou português; o CE3 descartou dez trabalhos que não haviam sido citados nenhuma vez, em que esta verificação foi feita através do Google Acadêmico, verificando um trabalho por vez; e por fim, o CE4 eliminou 28 trabalhos publicados em veículos sem classificação Qualis ou com classificação Qualis C. Dessa forma, após a aplicação dos CE, restaram 28 estudos, que foram considerados relevantes para análise. Vale destacar ainda que não houve nenhuma restrição em relação às localidades das instituições que foram estudadas nos trabalhos selecionados.

Por fim, na fase de documentação, os trabalhos selecionados foram sintetizados na Tabela 2.4, que apresenta cada um dos 28 estudos com um identificador, sua

citação bibliográfica e uma breve descrição. Após a leitura e categorização dos estudos, optou-se por organizá-los em dois grupos: (i) TR1 a TR18, que correspondem a trabalhos que investigam a evasão em diversos cursos, com foco predominante na área de computação; e (ii) TR19 a TR28, que abrangem estudos voltados à evasão em áreas do ensino superior presencial diferentes da área de computação. Essa categorização foi realizada com base no escopo temático identificado em cada estudo e teve como objetivo facilitar a análise comparativa entre pesquisas focadas na Computação e aquelas de caráter mais geral.

Tabela 2.4: Descrição dos 28 trabalhos selecionados para análise.

ID	Referência	Descrição
TR1	Duran et al. (2023)	Investigar fatores relacionados à intenção de abandono em programas de computação.
TR2	Santos et al. (2019)	Avaliar modelos de predição para identificar o perfil dos estudantes em risco de evasão.
TR3	Hoed et al. (2018)	Analisar a evasão dos cursos de graduação em Computação no Brasil.
TR4	Oliveira et al. (2020)	Analisar a evasão em três cursos de computação da UFG e seus fatores contribuintes.
TR5	Rocha et al. (2021)	Examinar os impactos sociais e econômicos da evasão e seus fatores contribuintes.
TR6	Costa et al. (2021)	Predizer alunos em risco de evasão com base nas notas dos 3 primeiros semestres.
TR7	Menolli et al. (2020)	Estudar a evasão nos cursos de sistema de informação com o uso de metodologia BI.
TR8	da Cruz et al. (2023)	Propor um método para analisar a evasão em Sistemas de Informação e cursos relacionados.
TR9	Opazo et al. (2021)	Comparar modelos de ML na previsão da evasão de alunos do primeiro ano de engenharia.
TR10	Marques et al. (2020)	Identificar os motivos da evasão no curso de Ciência da Computação da UFERSA.
TR11	Santos et al. (2021)	Utilizar técnicas de mineração de dados para prever evasão e formação de alunos.
TR12	dos Santos et al. (2021a)	Propor um modelo de Cadeia de Markov para analisar a probabilidade de evasão e conclusão.
TR13	do Carmo et al. (2022)	Investigar a evasão dos estudantes de um curso de Ciência da Computação da UDESC.
TR14	Viana et al. (2022)	Descrever um processo de MDE da UFPI para prever entre alunos evadidos e graduados.
TR15	Fukao et al. (2023)	Analisar estatísticas de evasão, motivações e avaliar o impacto da pandemia na evasão.
TR16	Vasconcelos and Andrade (2018)	Investigar as razões da evasão no curso de Licenciatura em Computação da UFRPE.
TR17	Carvalho et al. (2019)	Identificar os fatores relacionados à evasão dos calouros dos cursos de computação da UFAM.
TR18	Silva et al. (2021a)	Identificar os motivos de evasão dos alunos de Ciência da Computação da UFC.
TR19	Llauró et al. (2021)	Identificar estudantes em risco de evasão usando dados, questionários e percepção de tutores.
TR20	Santos et al. (2020)	Propor um modelo que integra árvores de decisão e algoritmos genéticos para classificar a evasão.
TR21	Nájera and Ortega (2022)	Desenvolver um modelo de ML para antecipar e reduzir a evasão no ensino superior.
TR22	da Silva et al. (2019a)	Propor três modelos de regressão para prever a evasão em Instituições de Ensino Superior.
TR23	Pachas et al. (2021)	Aplicar ML para identificar fatores da evasão estudantil e analisar o tempo até sua ocorrência.
TR24	de Brito et al. (2020)	Identificar atributos relevantes na evasão escolar utilizando técnicas de MDE e ML.
TR25	Colpo et al. (2021)	Identificar o perfil de evasão em cursos de graduação, utilizando técnicas de MDE e ML.
TR26	Kantorski et al. (2023)	Apresentar um processo de análise e previsão da evasão utilizando métodos de MDE e ML.
TR27	Carvalho et al. (2023)	Investigar modelos de MDE, focando no problema da evasão de estudantes no ensino superior.
TR28	Falcao et al. (2023)	Construir e avaliar um modelo de classificação para prever a evasão de alunos ingressantes.

2.4 Resultados

2.4.1 QP1 - Qual é o panorama geral das pesquisas sobre evasão no ensino superior e, em particular, como essas pesquisas se relacionam com a área de computação?

Para responder à primeira pergunta de pesquisa foram analisados os resultados referentes a todos os 28 trabalhos selecionados para as análises.

Na Figura 2.2 tem-se a distribuição dos anos de publicação dos 28 trabalhos (barras azuis), oferecendo um panorama da evolução das pesquisas sobre evasão no ensino superior entre 2018 e 2023. O maior número de publicações ocorreu em 2021, com nove trabalhos, possivelmente devido a fatores como a pandemia de COVID-19, que trouxe novos desafios à permanência estudantil. Nos anos seguintes, a distribuição foi mais equilibrada, com uma queda de seis trabalhos em 2022 e um aumento de três em 2023, indicando que o tema continua relevante.

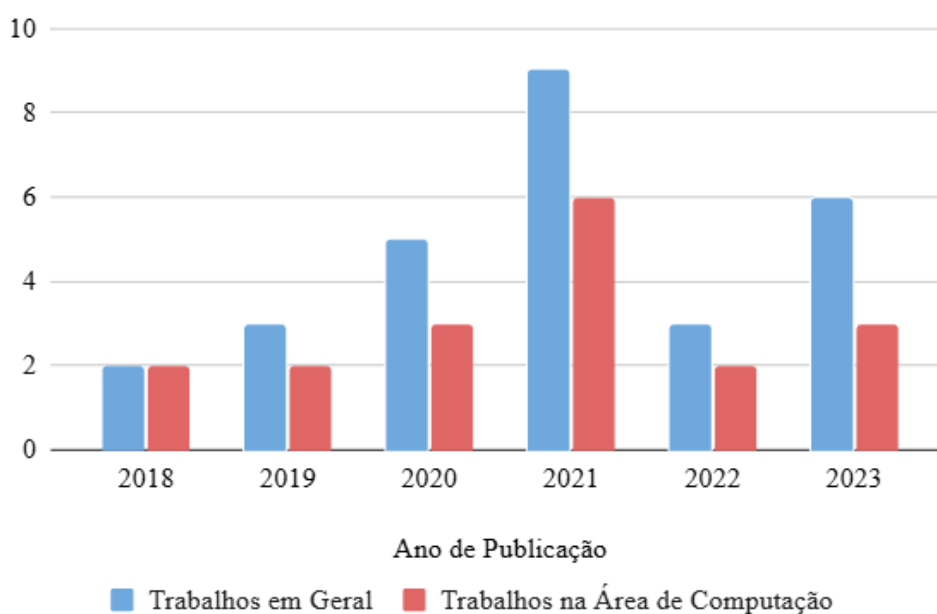


Figura 2.2: Distribuição da quantidade de trabalhos por ano de publicação.

A Tabela 2.5 apresenta a distribuição dos cursos analisados nos trabalhos. Observa-se uma predominância de estudos voltados para a área de computação, com destaque para Ciência da Computação, que lidera com treze trabalhos, seguida por Sistemas de Informação com seis. Esse padrão reflete a crescente preocupação com a evasão nos cursos de computação, que, apesar da alta demanda no mercado, enfrentam altas taxas de evasão devido à carga curricular rigorosa e à complexidade técnica. Em contraste, cursos como Artes, Arquitetura e Ciências Sociais são abordadas por apenas um trabalho cada, evidenciando a lacuna nas investigações

sobre evasão em contextos acadêmicos menos explorados. Isso destaca a necessidade de expandir as pesquisas para outras áreas.

Tabela 2.5: Distribuição dos cursos por trabalhos.

Curso	Trabalhos	Total
Ciência da Computação	TR1, TR2, TR3, TR4, TR5, TR6, TR10, TR11, TR13, TR14, TR15, TR17, TR18	13
Sistemas de Informação	TR1, TR2, TR3, TR4, TR7, TR8	6
Engenharia da Computação	TR1, TR2, TR3, TR17	4
Engenharia de Software	TR1, TR3, TR4, TR17	4
Licenciatura em Computação	TR3, TR16	2
Engenharias não especificadas	TR19, TR21	2
Engenharia de Informática	TR9	1
Informática	TR15	1
Engenharias Bioinformática	TR9	1
Engenharia Civil	TR9	1
Engenharia Industrial	TR9	1
Engenharia Mecatrônica	TR9	1
Engenharia de Minas	TR9	1
Engenharia Mecânica	TR9	1
Artes	TR19	1
Negócio	TR19	1
Arquitetura	TR19	1
Ciências Sociais	TR21	1
Ciências Administrativas	TR21	1
Química	TR28	1
Física	TR28	1
Ciências Biológicas	TR28	1
Nutrição	TR28	1
Letras (Português Espanhol)	TR28	1

Já a Tabela 2.6 mostra a distribuição dos trabalhos conforme as abordagens metodológicas utilizadas na análise dos cursos. A análise exploratória se destaca como a metodologia predominante, empregada em todos os 28 trabalhos revisados, sendo amplamente adotada como uma abordagem inicial para compreender os dados e identificar padrões de evasão no ensino superior. Dentre as técnicas de análise exploratória, a EDM foi a mais utilizada. Já o ML aparece em 17 trabalhos, refletindo o crescente interesse por métodos de análise e predição da evasão.

Tabela 2.6: Distribuição dos trabalhos por abordagem metodológica.

Abordagem Metodológica	Trabalhos	Total
Análise Exploratória	TR1 - TR28	28
Aprendizado de Máquina	TR2, TR6, TR8, TR9, TR10, TR11, TR14, TR17, TR20, TR21, TR22, TR23, TR24, TR25, TR26, TR27, TR28	17

Para concluir a QP1, foi feito um recorte focado nos 18 trabalhos sobre evasão na área de computação. Na Figura 2.2, as barras vermelhas evidenciam a distribuição

desses estudos ao longo dos anos. Observa-se um crescimento expressivo em 2021, ano em que foram publicados seis trabalhos, o maior volume registrado no período. Em 2018, todos os dois estudos identificados tratavam de cursos de computação, enquanto, entre 2019 e 2021, os trabalhos dessa área representaram aproximadamente 66% das publicações do período. Embora 2022 apresente uma redução, a proporção de estudos voltados à Computação ainda se mantém relevante, voltando a crescer em 2023. Assim, as barras vermelhas mostram que, apesar de oscilações anuais, a produção científica sobre evasão especificamente em cursos de computação tem sido predominante e consistente ao longo do período analisado.

A Tabela 2.7 apresenta a distribuição dos cursos da área de computação analisados nos 18 trabalhos. Como comentado, destacam-se os cursos de Ciência da Computação e Sistemas de Informação, que apareceram em treze e seis trabalhos respectivamente, seguidos por Engenharia da Computação e Engenharia de Software, com quatro trabalhos cada, e Licenciatura em Computação, com dois trabalhos. A predominância de Ciência da Computação e Sistemas de Informação pode estar relacionada ao fato de que, no contexto brasileiro, esses cursos possuem maior número de ofertas registradas na plataforma do MEC, o que tende a gerar mais estudos nacionais sobre o tema. Além disso, Ciência da Computação possui um ciclo básico com disciplinas como Cálculo, Física e Programação, conhecidas pelos altos índices de reprovação, um fator que pode influenciar diretamente a evasão.

Tabela 2.7: Distribuição de trabalhos por cursos: área de computação.

Curso	Trabalhos	Total
Ciência da Computação	TR1, TR2, TR3, TR4, TR5, TR6, TR10, TR11, TR13, TR14, TR15, TR17, TR18	13
Sistemas de Informação	TR1, TR2, TR3, TR4, TR7, TR8	6
Engenharia da Computação	TR1, TR2, TR3, TR17	4
Engenharia de Software	TR1, TR3, TR4, TR17	4
Licenciatura em Computação	TR3, TR16	2
Engenharia de Informática	TR9	1

No que se refere à abordagem metodológica de análise dos dados, os trabalhos que abordam evasão em cursos da área de computação seguem o mesmo padrão dos demais, conforme a Tabela 2.8. Todos os 18 trabalhos utilizaram análise exploratória como abordagem inicial para compreensão dos dados e identificação de padrões de evasão, sendo que nove deles aplicaram EDM. Em contraste, o uso de métodos de análise e predição de evasão por meio de ML foi explorado em cerca de 44% dos trabalhos da área de computação (8/18), enquanto nos trabalhos de outras áreas esse uso foi de 90% (9/10).

Tabela 2.8: Distribuição dos trabalhos por abordagem metodológica: área de computação.

Abordagem Metodológica	Trabalhos	Total
Análise Exploratória	TR1 - TR18	18
Aprendizado de Máquina	TR2, TR6, TR8, TR9, TR10, TR11, TR14, TR17	8

2.4.2 QP2 - Como as abordagens metodológicas encontradas têm sido aplicadas para se analisar a evasão no ensino superior em cursos da área de computação?

Para responder à QP2, foram analisadas as abordagens metodológicas e consolidados os atributos extraídos das bases de dados dos trabalhos selecionados da área de computação.

A Tabela 2.9 apresenta os 62 atributos identificados nos trabalhos da área de computação, que utilizam análises exploratórias ou modelos de predição de evasão no ensino superior por meio de algoritmos de ML. Esses atributos estão organizados em quatro dimensões principais: *Demográfico Socioeconômico e Pessoal; Conhecimento Prévio; Institucional; e Desempenho Acadêmico*. Essa categorização é uma das principais contribuições deste trabalho, pois organiza de maneira sistemática os elementos usados para analisar a evasão nos cursos da área de computação. Ao facilitar a identificação de padrões e a compreensão das variáveis mais relevantes, ela oferece uma base para estudos futuros e também proporciona uma visão geral dos atributos em cada dimensão, que podem ser combinados e explorados em investigações subsequentes.

Ao comparar a categorização deste MSL com os trabalhos relacionados, observa-se que tanto [Silva et al. \(2023\)](#) quanto [de Oliveira and Barbosa \(2023\)](#) utilizaram dimensões semelhantes. O primeiro estudo agrupou 29 atributos nas categorias *Demográficas, Acadêmicas* e de *Aprendizagem*. Já o segundo identificou 15 atributos, distribuídos em três dimensões: *Estudantil, Escolar* e *Outros*. No entanto, como este MSL se concentra em trabalhos da área de computação, uma área de grande relevância atualmente, a análise resultou em um número bem maior: 62 atributos. Essa variedade justificou o uso das quatro dimensões mencionadas para organizá-los de forma eficaz.

Essa categorização revela a diversidade de atributos analisados, abrangendo desde características pessoais e sociais dos alunos até fatores institucionais e de desempenho acadêmico. Também é possível observar quais atributos estão mais relacionados em cada dimensão. Na categoria *Demográfico Socioeconômico e Pessoal*, os mais frequentes foram gênero e idade, em que apareceram em dez e seis trabalhos respectivamente, seguidos por cidade / local de residência em cinco

trabalhos. Na categoria *Conhecimento Prévio*, o conhecimento em matemática e informática se destacou em seis trabalhos. Em *Institucional*, os atributos mais recorrentes foram o curso, a estrutura do curso e a forma de ingresso, em sete, seis e cinco trabalhos. Por fim, em *Desempenho Acadêmico*, destacaram-se a situação de evasão, ano de admissão e número de reprovações, em nove, cinco e cinco trabalhos.

Outra consolidação produzida neste trabalho foi a Tabela 2.10 que apresenta os algoritmos de ML utilizados nos trabalhos analisados que envolvem a área de computação. Ao comparar este MSL com o estudo de Oliveira and Barbosa (2023), observam-se algoritmos em comum, como *Support Vector Machine (SVM)*, *Random Forest*, *Árvores de Decisão* e *Regressão Logística*. No entanto, este trabalho não encontrou a *Regressão Linear*, a qual foi utilizada no trabalho citado, mas apresentou *Redes Neurais*, *K-Nearest Neighbors (KNN)*, *Naive Bayes*, entre outros algoritmos.

Tabela 2.10: Técnicas de ML dos trabalhos que envolvem a área de computação.

Algoritmo	Trabalhos	Total
<i>Random Forest</i>	TR2, TR6, TR9, TR11, TR14, TR17	6
<i>Árvores de Decisão</i>	TR2, TR6, TR9, TR11, TR14	5
<i>Support Vector Machine (SVM)</i>	TR2, TR9, TR14, TR17	4
<i>K-Nearest Neighbors (KNN)</i>	TR2, TR8, TR9, TR14	4
<i>Regressão Logística</i>	TR6, TR9, TR17	3
<i>Naive Bayes</i>	TR2, TR9, TR14	3
<i>Redes Neurais</i>	TR2, TR9	2
<i>K-Means</i>	TR10	1
<i>Extra Trees (ET)</i>	TR14	1
<i>Multiayer Perception (MLP)</i>	TR14	1

2.5 Considerações Finais

Este artigo apresentou um MSL sobre evasão no ensino superior, com foco em cursos da área de computação. Os resultados indicaram que as principais abordagens metodológicas envolvem análise exploratória com uso de EDM e a aplicação de algoritmos de ML. Além disso, 18 dos 28 trabalhos analisados focaram em cursos da área de computação. Por fim, foi possível identificar e categorizar os principais atributos relacionados à evasão, além de observar a tendência crescente de desenvolvimento de modelos de predição de evasão.

As principais limitações e ameaças à validade deste MSL dizem respeito aos termos utilizados na *string* de busca, aos CE adotados e ao processo de seleção dos estudos. A *string* de busca foi composta majoritariamente por termos associados ao

Tabela 2.9: Dimensões dos atributos e identificadores de trabalhos.

Demográfico, Socioeconômico e Pessoal		
Atributo	Trabalhos	Total
Gênero	TR1, TR4, TR6, TR7, TR8, TR9, TR14, TR15, TR17, TR18	10
Idade	TR1, TR6, TR8, TR9, TR14, TR18	6
Cidade / Local de residência	TR5, TR8, TR9, TR10, TR14	5
Raça / Etnia	TR1, TR7, TR10, TR14	4
Região / Região de Origem (do país)	TR7, TR9, TR10, TR14	4
Apoio Estudantil	TR5, TR18	2
Estado de residência	TR7, TR14	2
Estado Civil	TR14, TR17	2
Deslocamento / Tipo de Transporte	TR17, TR18	2
Estatuto Socioeconômico	TR1, TR17	2
Naturalidade	TR6	1
Nacionalidade	TR7	1
Relação Conjugal dos pais	TR10	1
Tipo de moradia (alugada / própria)	TR17	1
Quantidade de Filhos	TR17	1
Acesso à internet / computador	TR17	1
Impossibilidade de Trabalhar e Estudar	TR18	1
Problemas de Saúde Mental	TR18	1
Dificuldade de Foco e de Organização	TR18	1
Falta de Motivação	TR18	1
Conhecimento Prévio		
Atributo	Trabalhos	Total
Conhecimento Matemático / Conhecimento em Informática	TR3, TR9, TR10, TR16, TR17, TR18	6
Escola / Grau Acadêmico (Público / Privado)	TR8, TR9, TR14	3
Preferência / Escolha do curso	TR9, TR17	2
Graduação Prévia	TR14, TR17	2
Pontuação no ensino médio	TR9	1
Pontuação nos testes de Admissão Universitária	TR9	1
Posição em escala de notas no ensino médio	TR9	1
Pontuação em Linguagem	TR9	1
Pontuação em Teste Nacional	TR9	1
Institucional		
Atributo	Trabalhos	Total
Curso	TR1, TR4, TR5, TR11, TR15, TR17, TR18	7
Estrutura do curso / Disciplinas	TR4, TR5, TR11, TR13, TR16, TR18	6
Forma de Ingresso	TR10, TR14, TR15, TR17, TR18	5
Tipo de instituição (pública ou privada)	TR6, TR7	2
Corpo docente (nome ou ID)	TR4, TR18	2
Grade Curricular Desatualizada	TR5, TR18	2
Mercado de trabalho	TR5, TR18	2
Estrutura Física da Instituição	TR5, TR18	2
Recursos pedagógicos	TR5	1
Estrutura Administrativa da Instituição	TR5	1
Acompanhamento discente	TR5	1
Falta de interdisciplinaridade	TR5	1
Modalidade (integral, diurno, noturno)	TR7	1
Método de Ensino (Presencial / EAD)	TR7	1
Trabalho Teórico (em horas)	TR8	1
Trabalho Prático (em horas)	TR8	1
Universidade (nome ou ID)	TR9	1
ID / Matrícula	TR9	1
Carga horária obrigatória	TR18	1
Atuação da Coordenação do curso	TR18	1
Desempenho Acadêmico		
Atributo	Trabalhos	Total
Situação de Evasão	TR6, TR7, TR8, TR9, TR11, TR12, TR13, TR15, TR17	9
Ano de Admissão	TR4, TR9, TR14, TR15, TR17	5
Quantidade de Reprovações	TR4, TR8, TR13, TR14, TR18	5
Quantidade de Aprovações	TR4, TR8	2
Período de Admissão	TR11, TR14	2
Problemas Vocacionais / Falta de afinidade com o Curso	TR3, TR18	2
Integração social / Engajamento	TR5, TR18	2
Período do curso	TR7, TR8	2
Frequência	TR13, TR14	2
Período de Saída	TR11	1
Facilidade com Programação	TR3	1
Média nos três primeiros períodos	TR6	1
Primeiros Meses	TR8	1
Disciplina (Concluída / Trancada / Cancelada)	TR14	1
Média da nota dos Aprovados	TR14	1
Média da nota dos Reprovados	TR14	1

contexto brasileiro, o que, embora tenha recuperado alguns estudos internacionais, pode ter limitado a identificação de trabalhos globais relevantes devido à ausência de expressões mais amplas e amplamente utilizadas em pesquisas internacionais (como *higher education*, *college* ou *university* sem qualificadores). Além disso, o critério CE3 excluiu trabalhos publicados até 2022 que não haviam recebido citações, buscando evitar a inclusão de estudos de baixa visibilidade. Entretanto, essa decisão pode ter eliminado pesquisas recentes, de nicho ou que ainda não tiveram tempo para acumular impacto, introduzindo um viés em direção a estudos mais consolidados. Por fim, a triagem inicial baseou-se em títulos, resumos e palavras-chave, sem a realização sistemática da leitura completa de todos os textos elegíveis, o que pode ter levado à exclusão prematura de trabalhos relevantes cujos elementos textuais não refletiam adequadamente seu conteúdo.

Como contribuições deste estudo, destacam-se: o panorama geral das pesquisas sobre evasão no ensino superior, com recorte para a área de computação, destacando as abordagens metodológicas aplicadas; e a categorização dos atributos analisados em: *Demográfico Socioeconômico e Pessoal; Conhecimento Prévio; Institucional; e Desempenho Acadêmico*, que fornece uma base para estudos futuros relacionados ao desenvolvimento de modelos de predição de evasão. Além disso, pesquisas futuras podem incluir levantamentos junto às instituições de ensino superior sobre estratégias para mitigar a evasão, permitindo a avaliação de seus impactos e a identificação de novos atributos relevantes.

Capítulo 3

Acolhimento Versus Evasão no Primeiro Ano: Uma Análise Panorâmica na UFV com Ênfase nos Cursos de Computação

Resumo: Entender e tentar mitigar a evasão nos cursos de computação, especialmente no primeiro ano, tem sido um desafio constante para as instituições de ensino superior. Este estudo apresenta um levantamento das estratégias de acolhimento praticadas em cursos de graduação da Universidade Federal de Viçosa (UFV), que possui três campi, com o intuito de realizar uma análise preliminar, com recorte para cursos da área de computação, sobre o impacto destas estratégias nas taxas de evasão de estudantes do primeiro ano. A partir dos resultados, foi possível observar que, com a implementação progressiva das estratégias, houve uma diminuição da taxa média de evasão nos cursos da área de computação, com destaque para dois cursos, localizados no campus Florestal e no campus Rio Paranaíba, onde houve queda, respectivamente, de 20 e 10 pontos percentuais.

Palavras-chave: Primeiro Ano; Evasão nos Cursos de Computação; Estratégias de Acolhimento.

Abstract: Understanding and attempting to mitigate dropout in computing programs, especially during the first year, has been a constant challenge for higher education institutions. This study presents a survey of the onboarding strategies implemented in undergraduate programs at the Federal University of Viçosa (UFV) across its three campuses. The aim is to conduct a preliminary analysis, focusing on computing programs, to assess the impact of these strategies on first-year student dropout rates. Based on the results, it was possible to observe that, with the progressive implementation of these strategies, there was a reduction in the average dropout rate in the computing programs, with emphasis on two programs, located at the Florestal and Rio Paranaíba campuses, which showed decreases of 20 and 10 percentage points, respectively.

Keywords: First Year; Dropout in Computing Programs; Onboarding Strategies.

3.1 Introdução

A evasão no ensino superior é um fenômeno complexo e multifatorial que impacta a sociedade como um todo (Marques et al., 2019). No Brasil, esse problema tem sido amplamente discutido, especialmente em cursos com alta demanda e altos índices de desistência, como acontece na área de computação (Menolli et al., 2020). De acordo com Duran et al. (2023), a evasão em cursos de computação está frequentemente associada à dificuldade de adaptação ao ritmo acadêmico, a falta de identificação com o curso e a carência de suporte institucional. Segundo Pachas et al. (2021), ela ocorre, principalmente, nos quatro primeiros períodos de curso, onde disciplinas introdutórias como Programação são lecionadas (dos Santos Silva et al., 2020).

Dentre os esforços institucionais, individuais e coletivos, de se tentar reduzir a evasão, alguns estudos têm destacado a importância de se utilizar, por exemplo, estratégias de acolhimento nos anos iniciais dos cursos. Silva et al. (2021b), Graça and Nascimento (2021) e de Carvalho (2019) destacam que estratégias como mentorias, programas de nivelamento e integração entre calouros e veteranos podem desempenhar um papel crucial na redução da evasão.

Na Universidade Federal de Viçosa (UFV), altas taxas de evasão no primeiro ano dos cursos da área de computação também se fazem presentes. Na tentativa de mitigar esse fenômeno, algumas iniciativas institucionais têm sido adotadas. Em um trabalho anterior Mundim et al. (2024), os autores avaliaram o impacto de duas estratégias de acolhimento voltadas a alunos do primeiro ano do curso de Ciência da Computação da UFV-Florestal: uma baseada em gamificação e outra em grupos de estudos em programação. Os resultados indicaram impacto positivo no desempenho acadêmico e no engajamento dos participantes. No entanto, o estudo foi realizado localmente, sem considerar o contexto institucional ou a existência de ações similares em outros cursos da área de computação nos demais campi da UFV.

Este trabalho realiza um levantamento das estratégias de acolhimento praticadas nos cursos de graduação da UFV, em seus três campi, com recorte para os cursos da área de computação, de forma a analisar preliminarmente o impacto dessas estratégias sobre a evasão no primeiro ano. Para isso, o estudo foi guiado pelas seguintes questões de pesquisa (QP): **QP1** - Qual o panorama geral das estratégias de acolhimento adotadas pelos cursos da UFV?; **QP2** - Quais as estratégias de acolhimento adotadas pelos cursos de computação da UFV?; **QP3** - Como as estratégias de acolhimento se relacionam com a evasão de calouros nos cursos de computação da UFV?

Dentre as contribuições deste trabalho, destacam-se: (i) criação de um conjunto de dados referente às estratégias de acolhimento de 41 cursos de graduação nos três campi da UFV; (ii) agrupamento das estratégias em três categorias: *Recepção, Apoio*

ao *Discente e Engajamento e Motivação*; (iii) panorama de distribuição das estratégias e suas categorias; (iv) análises preliminares do impacto das estratégias sobre a evasão nos cursos da área de computação, com base em dados institucionais.

O artigo está organizado da seguinte forma: a Seção 3.2 apresenta os trabalhos relacionados ao uso de estratégias de acolhimento no primeiro ano; na Seção 3.3, encontram-se descritos os materiais e métodos utilizados na condução da pesquisa; a Seção 3.4 apresenta e discute os resultados obtidos com o levantamento e o mapeamento dos dados sobre estratégias e a análise dos mesmos em conjunto com dados institucionais sobre evasão; e, por fim, as considerações finais são apresentadas na Seção 3.5.

3.2 Trabalhos Relacionados

O estudo de [Holanda et al. \(2022b\)](#) sobre a evasão de calouros, no contexto dos cursos de computação, mostra que o primeiro ano da graduação é um dos maiores desafios para os calouros na Universidade de Brasília (UnB). Os resultados indicaram que as dificuldades enfrentadas, tais como barreiras impostas pelas disciplinas iniciais, como a programação, e lacunas na formação educacional prévia ao ingresso à universidade, nesse período, são alguns dos principais fatores que levam muitos estudantes à evasão.

Diversos estudos têm proposto estratégias de acolhimento para mitigar a evasão de calouros em cursos de computação. Na UnB, por exemplo, o Departamento de Ciência da Computação (CIC) implementou três iniciativas. Em [Holanda et al. \(2021\)](#), os autores apresentam o **Programa de Acolhimento dos Calouros**, realizado virtualmente em 2020. O programa incluiu encontros sobre o departamento, palestras sobre dificuldades do início do curso e interação com ex-alunos, recebendo avaliações bastante positivas.

Já [Holanda et al. \(2022a\)](#) descrevem a implementação das **Monitorias** em disciplinas iniciais, também em 2020. Essa iniciativa ofereceu plantões de dúvidas, atendimentos agendados e aulas extras, sendo reconhecida pelos alunos como um espaço de suporte e colaboração. Por fim, [Holanda et al. \(2022b\)](#) apresentam a disciplina **Introdução ao Ambiente Universitário da Computação (IAUC)**, criada em 2021. Estruturada em núcleos de informação, instrução, motivação e interação, a disciplina incluiu desde dicas de estudo até maratonas e jogos de integração.

Essas três iniciativas da UnB contaram com *feedbacks* positivos dos estudantes, destacando a interação com ex-alunos como um fator que promoveu maior engajamento e contribuiu para a adaptação acadêmica.

Por sua vez, [Pinto et al. \(2020\)](#) detalham o planejamento, execução e avaliação do evento **Semana do Calouro**, promovido pelo Programa de Educação Tutorial em

Computação (PETComp) da Universidade Federal do Maranhão (UFMA). Este evento tem como principal objetivo acolher e orientar os calouros do curso de Ciência da Computação, apresentando-lhes diversas oportunidades acadêmicas e incentivando sua permanência e integração na UFMA. A metodologia incluiu o planejamento de atividades como tira-dúvidas, dinâmicas de grupo, visitas a laboratórios e uma palestra sobre um tema atual em computação. Os resultados foram a melhoria da ambientação dos calouros, a facilitação da compreensão sobre o curso e a redução potencial da evasão.

As estratégias de acolhimento não se restringem à Computação. Em [Brito and Ciampi \(2023\)](#), os autores descrevem a *Out of the Box Week*, semana de dinâmicas e exercícios voltados a calouros de engenharia, que recebeu avaliações positivas e mostrou eficácia na retenção. Já [Graça and Nascimento \(2021\)](#) analisam o **Projeto Mentoria** na Universidade Tiradentes, envolvendo diversos cursos, e apontam o fortalecimento da relação aluno-instituição, melhor adaptação e redução da evasão. Outros estudos, como [Rios et al. \(2021\)](#) e [Theobald et al. \(2020\)](#), também confirmam a relevância do acolhimento na permanência estudantil em diferentes áreas.

Diferentemente dos estudos anteriores, que trataram de estratégias isoladas, esta pesquisa adota uma abordagem mais ampla, mapeando as ações de acolhimento nos cursos dos três campi da UFV, com ênfase nos cursos de computação, e classificando-as em categorias e subcategorias. Além disso, outro diferencial deste estudo é a análise da possível relação dessas estratégias com a evasão de calouros, comparando taxas antes e depois de sua implementação.

3.3 Materiais e Métodos

Esta seção apresenta a coleta e análise dos dados dos discentes e das estratégias de acolhimento da UFV, incluindo métodos, ferramentas e recursos computacionais.

3.3.1 Coleta de Dados

A extração dos dados dos discentes foi realizada a partir da base de dados acadêmicos da UFV, que reúne diversos atributos essenciais dos alunos, como ano de nascimento, gênero, campus, curso, período de ingresso e saída, situação (evasão ou matrícula), nota do ENEM, raça, número de reprovações e coeficiente de rendimento.

O conjunto de dados inicial, filtrado para abranger apenas os cursos de computação, contava com 3.509 calouros no período de 2000 a 2024. Após a limpeza dos registros de entrada e saída, considerando apenas aqueles válidos para análise, o total foi reduzido para 3.310 alunos. O ano de 2024, entretanto, foi excluído, uma vez que seus dados ainda não estavam completos no momento da disponibilização da

base, resultando na amostra final de 3.112 calouros. Cabe destacar que os intervalos temporais podem variar entre os cursos, de acordo com seus anos de criação.

Já as informações referentes às estratégias de acolhimento foram coletadas por meio de entrevistas com as coordenações dos cursos dos três campi da UFV, a partir de um roteiro com perguntas¹ elaboradas especificamente para este trabalho. Os coordenadores que não puderam participar das entrevistas tiveram a oportunidade de responder manualmente as perguntas no Google Docs, ampliando a cobertura dos cursos respondentes. Esta coleta permitiu a construção de uma base de dados das estratégias de acolhimento para os cursos da UFV, inicialmente com os seguintes atributos: campus, área, curso, duração, período de realização, ano de início, tipo da estratégia (extensão ou ensino), grupos participantes, executor (coordenação ou professores), público-alvo e idealização (pela UFV ou pelo curso).

A coleta ocorreu entre abril e novembro de 2024, de forma remota e presencial nos campi de Florestal (CAF) e Viçosa (CAV), e totalmente remota no campus de Rio Paranaíba (CRP). A participação dos cursos variou entre os campi, como apresentado na Figura 3.1, sendo possível obter respostas de todos os 10 cursos de Florestal (100%), 8 de 10 cursos de Rio Paranaíba (80%) e 23 de 47 cursos de Viçosa (49%).

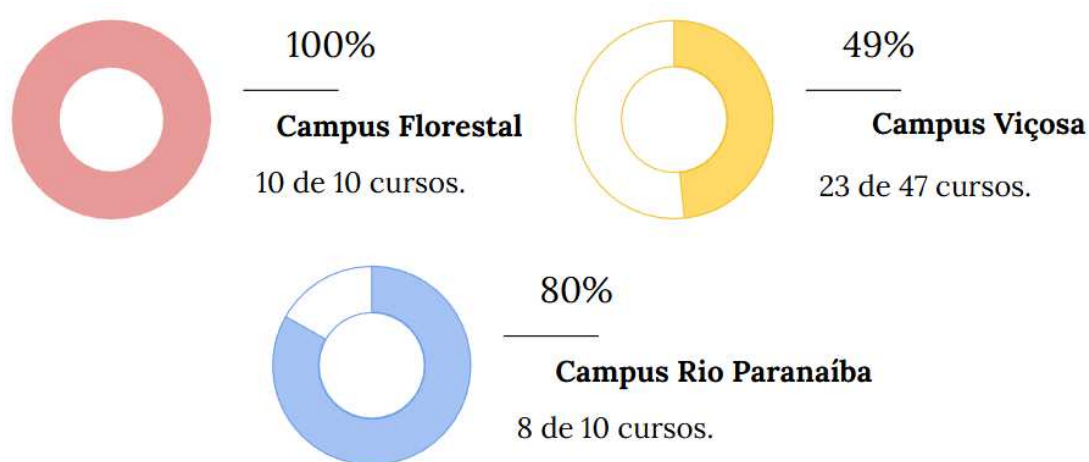


Figura 3.1: Adesão dos cursos dos três campi da UFV à pesquisa.

No campus Viçosa, a coleta presencial foi determinante: sem essa presença, apenas 8 cursos teriam respondido às entrevistas, enquanto com a aplicação presencial foram obtidas 23 respostas. Já em Rio Paranaíba, a coleta também se mostrou difícil, sendo necessária a intervenção direta da diretora geral do campus e de outros funcionários-chave para viabilizar a participação dos cursos. Esse processo evidenciou que a coleta de informações demandou adaptações e esforços adicionais para garantir maior cobertura e representatividade dos dados.

É importante destacar que esta pesquisa não utiliza dados identificáveis de

¹Perguntas: <https://github.com/carvalhopedro22/Perguntas-das-entrevistas>

indivíduos, pois todos os dados dos discentes são anônimos, sem possibilidade de identificação pessoal. De forma semelhante, as informações sobre as estratégias, coletadas por entrevistas com coordenadores, referem-se exclusivamente às práticas institucionais dos cursos, sem vinculação a dados pessoais dos participantes.

3.3.2 Análise de Dados

Com base nos dados levantados pelas entrevistas, as estratégias da UFV foram organizadas em três categorias principais: *Recepção*, *Apoio ao Discente* e *Engajamento e Motivação*. A UFV foi selecionada como campo de estudo por se tratar de uma universidade federal multicampi, com oferta diversificada de cursos e estratégias de acolhimento já implementadas. A criação das categorias foi realizada utilizando-se um método comumente empregado na análise de dados qualitativos, o *open coding* (codificação aberta) (Miles et al., 2014), que consiste na leitura sistemática dos dados e na atribuição de códigos para identificar padrões e agrupar informações por similaridade. Como resultado, tem-se as três categorias mencionadas.

As estratégias de *Recepção* ocorrem na primeira semana do primeiro período e incluem **Palestras Informativas** que fornecem aos ingressantes informações sobre o curso e a universidade, além de **Encontros com Docentes**, **Encontros com Veteranos** e outros grupos que promovem uma ambientação inicial. Algumas ações específicas são o *Campus Tour*, que apresenta laboratórios e setores do campus, a **Trilha Ecológica**, voltada para calouros de Ciências Biológicas, permitindo uma introdução prática a conceitos do curso, e as **Confraternizações**, com lanches durante os dias de recepção.

Já as estratégias de *Apoio ao Discente* ocorrem ao longo do primeiro período e incluem o **Apadrinhamento de Calouros**, no qual veteranos auxiliam os calouros, bem como o **Acompanhamento Acadêmico**, individual ou em grupo, realizado por docentes. Outra prática comum é a **Disciplina de Introdução**, que apresenta o curso e a universidade.

Por fim, as estratégias de *Engajamento e Motivação* são as mais estruturadas e incluem a **Gamificação**, **Grupos de Estudos**, **Desenvolvimento Supervisionado de Projetos** e **Visitas Técnicas**, as quais também ocorrem durante o primeiro período.

Com a categorização concluída, os dados puderam ser analisados por meio da linguagem de programação Python na plataforma Jupyter, com o uso das bibliotecas Pandas², NumPy³ e Seaborn⁴. Para os cursos de computação, foram extraídas as taxas anuais de evasão dos calouros, as taxas médias de evasão de acordo com os anos de implementações das estratégias e o total de ingressantes e evadidos por ano.

²Pandas: <https://pandas.pydata.org/docs/>

³NumPy: <https://numpy.org/>

⁴Seaborn: <https://seaborn.pydata.org/>

No total, dos 67 cursos que compõem os três campi da UFV, 41 participaram da entrevista (Figura 3.1). Neste artigo, foram analisados especificamente todos os cursos de computação da UFV, sendo eles Ciência da Computação (CdC) nos campi CAF e CAV e Sistemas de Informação (SI), Integral e Noturno, no campus CRP.

É importante salientar que, para as análises dos calouros, foram considerados apenas os ingressantes no primeiro período de cada ano (denotados por /1), pois as estratégias de acolhimento são realizadas exclusivamente nesses períodos. Dessa forma, os alunos que ingressaram no segundo período do ano (/2) foram desconsiderados da contagem de calouros. Entretanto, a evasão foi analisada considerando todas as saídas possíveis no primeiro ano, incluindo, portanto, tanto o primeiro quanto o segundo período.

3.4 Resultados

Nesta seção são apresentados os resultados obtidos a partir da análise das estratégias de acolhimento coletadas, considerando inicialmente todos os cursos da UFV que responderam à pesquisa, e, em seguida, com foco específico na área de computação. A análise aborda a diversidade e a temporalidade das estratégias, bem como a possibilidade de correlacioná-las com as taxas de evasão observadas.

3.4.1 Mapeamento das Estratégias de Acolhimento da UFV

Para responder à primeira questão de pesquisa (QP1 – Qual o panorama geral das estratégias de acolhimento adotadas pelos cursos da UFV?), a Figura 3.2 traz a categorização, desenvolvida neste trabalho, para as estratégias de acolhimento adotadas pelos cursos da UFV que participaram desta pesquisa. Os campos exibidos na Figura 3.2 são: ID – usado para abreviar cada estratégia de cada categoria (por exemplo, *Recepção* é *R*, e a primeira estratégia de *Recepção* é *R1*; o mesmo vale para as demais); Descrição – nome da estratégia; Qtd – quantidade de cursos que utiliza a estratégia; e Duração – duração da aplicação da estratégia pelo curso.

Há uma variação significativa no que diz respeito à quantidade de itens (subcategorias) dentro de cada categoria estabelecida pelo mapeamento, sendo a categoria *Recepção* a mais ampla neste sentido, o que indica a tendência dos cursos em realizar acolhimento principalmente no momento de chegada dos calouros.

Além disso, é muito diverso o número de cursos que adotam cada subcategoria. Na categoria *Recepção*, algumas subcategorias, como **Encontro com Docentes** e **Encontro com Veteranos**, são utilizadas por todos os cursos, enquanto a **Trilha Ecológica** consta em apenas um curso (Ciências Biológicas - CAF), devido à sua especificidade.

Recepção (R)				Apoio ao Discente (AD)			
ID	Descrição	Qtd	Duração	ID	Descrição	Qtd	Duração
R1	Palestras Informativas	11	Semanal	AD1	Apadrinhamento de Calouros (entre discentes)	1	Semestral
R2	Encontro com Comissão Coordenadora	4	Semanal	AD2	Acompanhamento Acadêmico Individual	6	Semestral
R3	Encontro com Docentes	41	Semanal	AD3	Acompanhamento Acadêmico em Grupo	1	Semestral
R4	Encontro com Representação Discente	2	Semanal	AD4	Disciplina de Introdução	34	Semestral
R5	Encontro com Veteranos	41	Semanal	Engajamento e Motivação (EM)			
R6	Encontro com Alunos de Pós-graduação	11	Semanal	ID	Descrição	Qtd	Duração
R7	Encontro com Egressos	19	Semanal	EM1	Estratégias de Gamificação	1	Semestral
R8	Encontro com Empresa Júnior	3	Semanal	EM2	Grupos de Estudos	1	Semestral
R9	Confraternização	9	Semanal	EM3	Desenvolvimento Supervisionado de Projetos	1	Semestral
R10	Campus Tour	11	Semanal	EM4	Visitas Técnicas	3	Semestral
R11	Trilha Ecológica	1	Semanal				

Figura 3.2: Mapeamento das estratégias de acolhimento da UFV.

Em *Apoio ao Discente*, a **Disciplina de Introdução** é a prática mais consolidada, adotada por 34 cursos (83% dos respondentes). Já ações como o **Apadrinhamento de Calouros** e o **Acompanhamento Acadêmico em Grupo** estão presentes em apenas um curso cada (Engenharia Civil - CRP e Educação Infantil - CAV, respectivamente).

Já em *Engajamento e Motivação*, a baixa adesão às subcategorias pode estar relacionada à maior complexidade de implementação, uma vez que essas estratégias demandam mais recursos e planejamento. Observa-se que **Gamificação** e **Grupos de Estudos** são exclusivas de um curso (Ciência da Computação - CAF), assim como o **Desenvolvimento Supervisionado de Projetos** (Engenharia Civil - CAV).

Os gráficos da Figura 3.3 mostram a evolução acumulada da adoção de estratégias de cada uma das categorias ao longo do tempo, abrangendo todos os campi e cursos participantes, o que permite analisar o ritmo de implementação de cada uma delas.

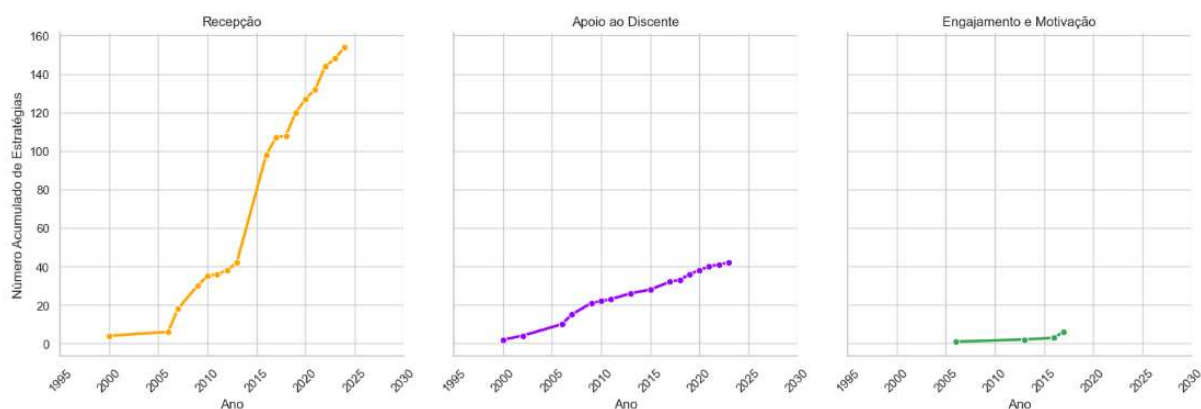


Figura 3.3: Evolução cumulativa da adoção de estratégias por categoria.

Observa-se que a categoria *Recepção* registra um aumento expressivo a partir de 2013, o que pode indicar uma preferência e facilidade de adoção mais imediata deste tipo de estratégia pelos cursos em geral. Além disso, as estratégias desta categoria foram, de fato, aquelas mais incentivadas e reforçadas pela própria administração da universidade. Em contraste, a categoria *Apoio ao Discente* mostra um crescimento contínuo, mas sem picos significativos de expansão. Por fim, a categoria *Engajamento e Motivação* apresenta o menor crescimento e quantidade de estratégias. A adoção deste tipo de acolhimento normalmente requer grande alinhamento com o curso, necessitando apoio significativo do corpo docente e dos próprios discentes.

3.4.2 Análise das Estratégias de Acolhimento dos Cursos de Computação da UFV

Respondendo à segunda questão de pesquisa (QP2 - Quais as estratégias de acolhimento adotadas pelos cursos de computação da UFV?), a Figura 3.4 apresenta um panorama das estratégias de acolhimento implementadas pelos cursos de computação da UFV. Embora compartilhem a mesma área de conhecimento, eles adotaram abordagens variadas de acolhimento para calouros ao longo dos anos.

Curso	Campus	Ação	Estratégia(s)	Ano de Início
CdC	CAF	Disciplina de Introdução	AD4	2015
		Semana de Acolhimento	R2, R3, R5, R6, R7	2016
		Lord of Florestal	EM1	2017
		MinasCoders	EM2	2017
		Orientação Acadêmica	AD2	2019
CdC	CAV	Disciplina de Introdução	AD4	2007
		Evento de Boas Vindas	R1, R3, R5, R7, R9	2016
SI	CRP	Palestras de Recepção ao Curso	R1, R3, R5	2022

Figura 3.4: Estratégias de acolhimento dos cursos da área de computação da UFV.

No CAF, o curso de CdC, criado em 2012, estruturou um conjunto diversificado de ações, abrangendo 47% das subcategorias, com destaque para a **Semana de Acolhimento**, em que ocorrem diversos encontros com a comunidade acadêmica, iniciados em 2016, além da implementação da **Disciplina de Introdução** iniciada em 2015 e da **Orientação Acadêmica** com **Acompanhamento Acadêmico Individual** pelos docentes implantada em 2019. Além disso, é o único curso de computação que possui estratégias da categoria de *Engajamento e Motivação*, sendo o *Lord of Florestal*

uma estratégia de **Gamificação**, com um jogo de cartas e duelos, e o *MinasCoders* uma estratégia de **Grupos de Estudos** para programação voltada para as meninas, porém com a participação de alguns meninos.

Já no CAV, que abriga o curso mais antigo da lista (iniciado em 1986), percebe-se que algumas estratégias surgiram mais cedo, como a **Disciplina de Introdução** implantada em 2007, ao passo que, mais recentemente, também foram implementados encontros com a comunidade acadêmica. O curso conta com estratégias em 32% das subcategorias. No CRP, por sua vez, o curso de SI, criado em 2008, apresenta um conjunto mais conciso de estratégias, abrangendo 16% das subcategorias, tendo iniciado suas ações de recepção apenas em 2022 com as **Palestras de Recepção ao Curso**, nas quais ocorrem **Encontros com Docentes e Veteranos**. É importante ressaltar que, embora o curso de SI tenha turnos integral e noturno, as estratégias de acolhimento são idênticas. Além disso, **Encontros com Docentes e Veteranos** foram as únicas subcategorias comuns a todos os cursos.

Por fim, o gráfico da Figura 3.5 ilustra a evolução cronológica da implementação das estratégias nos cursos de computação da UFV. Observa-se que, no CAF, a curva apresenta um crescimento acelerado em um curto período, sugerindo uma adoção concentrada de subcategorias em poucos anos. Já no CAV, a curva tem um crescimento mais linear e distribuído ao longo do tempo, indicando uma implementação gradual. Entretanto, no CRP, as subcategorias foram iniciadas significativamente mais tarde que os demais.

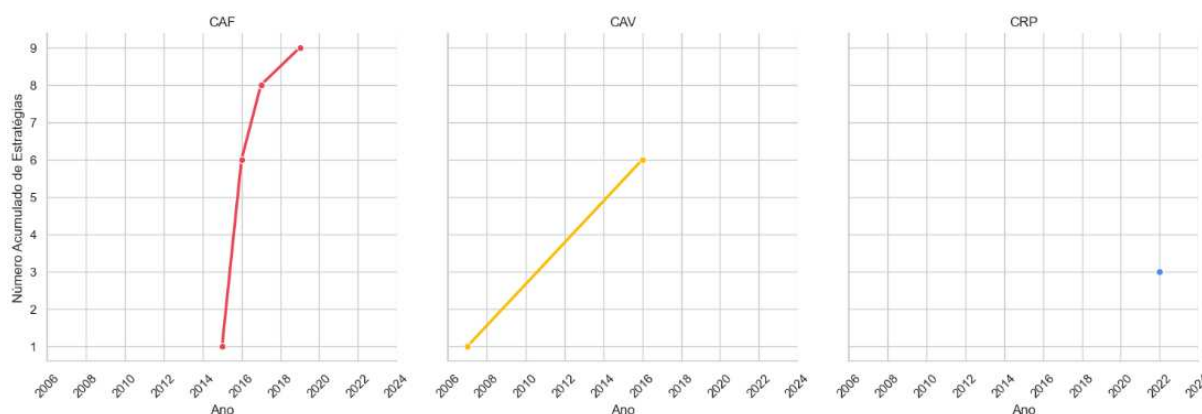


Figura 3.5: Evolução cumulativa das subcategorias das estratégias por campus nos cursos de computação da UFV.

3.4.3 Análise das Estratégias de Acolhimento e da Evasão de Calouros nos Cursos de Computação da UFV

Com relação à terceira questão de pesquisa (QP3 - Como as estratégias de acolhimento se relacionam com a evasão de calouros nos cursos de computação da UFV?), a Tabela

3.1 apresenta as taxas de evasão anuais dos calouros nos cursos de computação da UFV. As marcações em cores indicam os anos de implementação das estratégias em cada curso, permitindo a comparação das taxas antes e depois de sua adoção.

É importante ressaltar que, neste estudo, a evasão é considerada não apenas como a saída definitiva do aluno da universidade, mas também inclui situações como mudança de curso e transferência. Outras situações que a UFV considera como tipos de evasão⁵ são o abandono (A), desligamento (D), falecimento (F), desligamento com pedido de reconsideração (R) e exclusão (X), abrangendo diferentes formas de interrupção ou redirecionamento do percurso acadêmico.

Tabela 3.1: Taxas de evasão anuais de calouros nos cursos de computação da UFV.

ANO	Ciência da Computação (CAF)		Ciência da Computação (CAV)		S. Informação Integral (CRP)		S. Informação Noturno (CRP)	
	Ingressantes /1	Evasão no 1º Ano	Ingressantes /1	Evasão no 1º Ano	Ingressantes /1	Evasão no 1º Ano	Ingressantes /1	Evasão no 1º Ano
2000	-	-	40	7 (17,5%)	-	-	-	-
2001	-	-	40	2 (5,0%)	-	-	-	-
2002	-	-	40	4 (10,0%)	-	-	-	-
2003	-	-	46	3 (6,5%)	-	-	-	-
2004	-	-	43	5 (11,7%)	-	-	-	-
2005	-	-	44	2 (4,6%)	-	-	-	-
2006	-	-	42	1 (2,4%)	-	-	-	-
2007	-	-	40	2 (5,0%)	-	-	-	-
2008	-	-	43	1 (2,3%)	-	-	-	-
2009	-	-	42	4 (9,5%)	24	5 (20,8%)	27	0 (0,0%)
2010	-	-	42	4 (9,5%)	49	10 (20,4%)	50	7 (14,0%)
2011	-	-	42	6 (14,2%)	32	9 (28,1%)	49	8 (16,3%)
2012	50	16 (32,0%)	46	6 (13,0%)	49	17 (34,6%)	50	14 (28,0%)
2013	51	22 (43,1%)	54	11 (20,4%)	48	17 (35,4%)	49	19 (38,7%)
2014	56	15 (26,8%)	47	4 (8,5%)	54	20 (37,1%)	53	19 (35,9%)
2015	47	22 (46,8%)	45	13 (28,9%)	48	13 (27,1%)	52	17 (32,7%)
2016	48	10 (20,8%)	53	10 (18,9%)	56	16 (28,6%)	55	21 (38,2%)
2017	55	8 (14,5%)	52	11 (21,1%)	51	17 (33,3%)	52	17 (32,7%)
2018	47	7 (14,9%)	47	8 (17,1%)	57	10 (17,6%)	58	18 (31,0%)
2019	50	10 (20,0%)	47	6 (12,8%)	43	9 (21,0%)	50	15 (30,0%)
2020	50	2 (4,0%)	40	2 (5,0%)	43	1 (2,3%)	39	1 (2,6%)
2021	50	12 (24,0%)	40	5 (12,5%)	50	16 (32,0%)	50	21 (42,0%)
2022	56	10 (17,8%)	44	4 (9,1%)	49	6 (12,3%)	49	12 (24,4%)
2023	54	5 (9,3%)	44	12 (27,3%)	50	10 (20,0%)	49	17 (34,7%)

Em CdC do CAF, os índices de evasão no primeiro ano diminuíram a partir de 2017, passando a se manter abaixo dos patamares observados nos anos anteriores, em que alguns deles chegaram a superar 30% de evasão. Essa redução coincide com o início da aplicação do *Lord of Florestal* e do *MinasCoders*. Entretanto, como pode ser observado, não houve um declínio constante, nem mesmo uma estabilidade em determinado nível, uma vez que outros fatores, para além do acolhimento, também interferem na evasão [Mundim et al. \(2024\)](#).

Já em CdC do CAV, os índices de evasão, que eram inicialmente baixos, de modo geral aumentaram a partir de 2011. As estratégias de acolhimento adotadas foram a **Disciplina de Introdução** (2007) e o **Evento de Boas Vindas** (2016) com os encontros com a comunidade acadêmica. Após 2016, observa-se uma redução da evasão, mas com algumas oscilações em 2021 e 2023.

Por fim, o CRP registrou os maiores índices de evasão dos cursos analisados. Em SI, as estratégias incluem as **Palestras de Recepção ao Curso com Encontros com**

⁵Conceitos Acadêmicos da UFV: <https://sre.caf.ufv.br/avaliacao-do-rendimento-academico-2/>

Docentes e Veteranos, implementadas apenas em 2022, o que pode ser uma justificativa para a elevada evasão nos anos anteriores, em especial no curso noturno.

A Figura 3.6, por sua vez, apresenta as taxas médias de evasão anuais dos alunos calouros por campus e curso, a partir de 2012. A escolha desse período se justifica pela necessidade de garantir o mesmo intervalo de tempo para as análises em todos os cursos, sendo que foi apenas após esta data que os três cursos considerados já haviam sido criados. Os marcadores da Figura 3.6 indicam o ano em que cada categoria de estratégia passou a ser adotada em cada campus. Por exemplo, em 2012 há “*Sem Estratégias*” para os cursos de CAF e ambos de CRP, enquanto que em CAV já havia estratégias da categoria *AD*. O mesmo princípio se aplica aos demais marcadores.

Para o cálculo das taxas médias anuais de evasão exibidas no gráfico, considera-se, por exemplo, que para CdC do CAF a média referente ao marcador “*Sem Estratégias*” abrange o período de 2012 a 2014, conforme apresentado na Tabela 3.1. No marcador seguinte da curva do CAF, correspondente à categoria *AD*, a média é calculada a partir do ano de implementação dessa categoria (2015 até o ano final). Esse procedimento se repete para os demais marcadores, sempre calculando a média do ano inicial de vigência da nova categoria até o último ano da amostra.

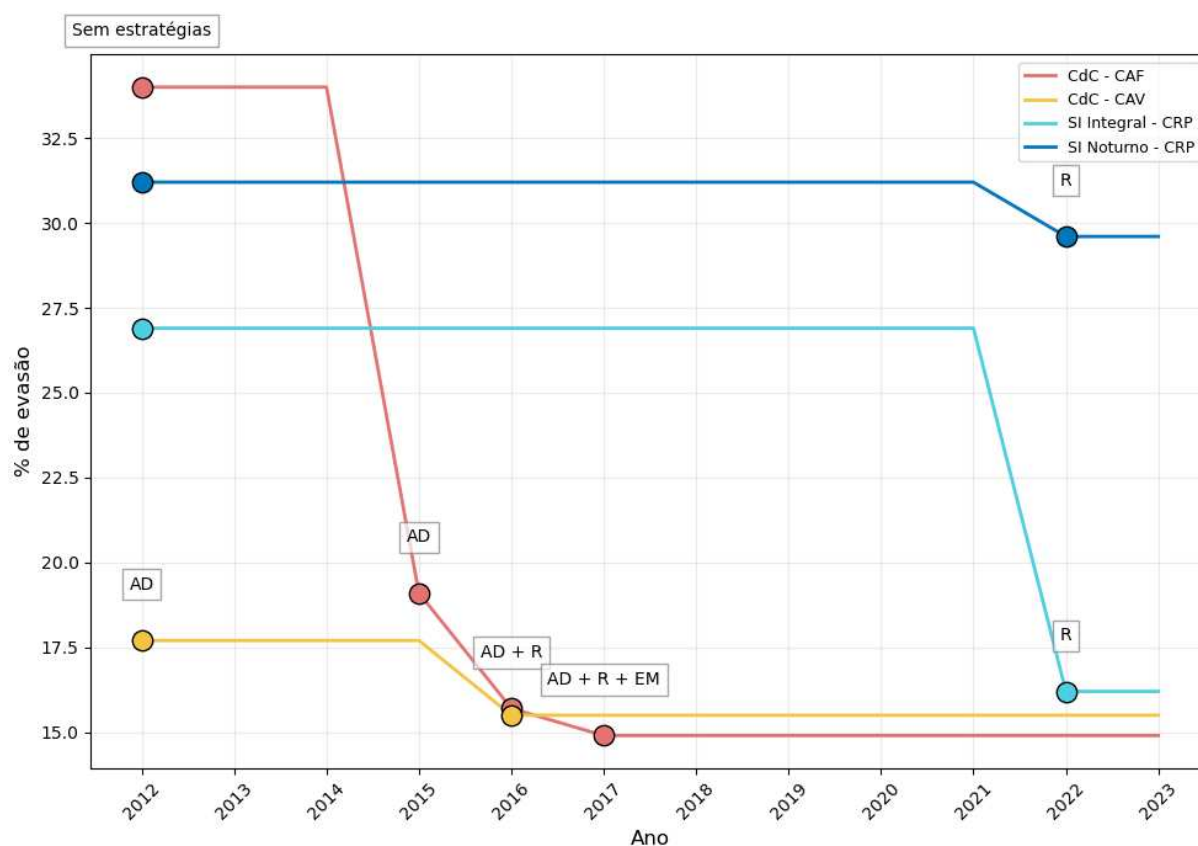


Figura 3.6: Evasão média de calouros por campus e curso de computação da UFV.

É possível observar que, com a implementação progressiva das estratégias, houve uma diminuição da taxa média de evasão. No CAF, por exemplo, a evasão decaiu

cerca de 20 pontos percentuais, entre os alunos que passaram pelas três categorias de estratégias em comparação aos que não tiveram nenhuma estratégia. No CRP (SI Integral), essa redução foi de aproximadamente 10 pontos percentuais, enquanto que CdC do CAV e SI Noturno do CRP tiveram as menores reduções. Esse resultado reforça a hipótese de que estratégias combinadas nas três categorias podem contribuir para a redução da evasão.

3.5 Considerações Finais

Este estudo apresentou um levantamento das estratégias de acolhimento praticadas nos cursos de graduação da UFV, em seus três campi. A partir de dados coletados junto às coordenações dos cursos da instituição, houve um agrupamento das estratégias em três categorias: *Recepção*, *Apoio ao Discente* e *Engajamento e Motivação*. Assim, o trabalho forneceu um panorama de distribuição das estratégias em cada campus. Por fim, a partir de um recorte para cursos da área de computação, foram apresentados resultados de uma análise preliminar, cruzando dados das estratégias com dados institucionais para investigar o impacto das mesmas sobre as taxas de evasão de calouros.

Os resultados indicaram que, no campus Florestal têm sido praticadas estratégias das três categorias com um crescimento acelerado em um curto período. No campus Viçosa, houve um crescimento mais linear e, no campus Rio Paranaíba, a adoção de estratégias teve um início mais recente.

No que diz respeito ao impacto das estratégias de acolhimento sobre taxas de evasão nos cursos da área de computação, os resultados preliminares evidenciaram, para a amostra considerada, que a implementação progressiva das estratégias contribui para a redução da taxa média de evasão. No campus Florestal, foi observado maior impacto com a taxa de evasão decaindo cerca de 20 pontos percentuais entre os alunos que passaram pelas três categorias de estratégias. Importante destacar que esse foi o único campus com adoção de estratégias da categoria *Engajamento e Motivação*. Nos outros dois campi, também houve redução, com segundo maior impacto observado no campus Rio Paranaíba (SI Integral), com redução de aproximadamente 10 pontos percentuais.

Em termos de ameaças à validade, destaca-se que a categorização e subcategorização das estratégias de acolhimento foi conduzida pelo primeiro autor do trabalho, de forma manual e intuitiva. No entanto, para mitigar possíveis inconsistências e agrupamentos equivocados, esse procedimento foi refinado com checagens iterativas de validação realizadas pelos demais autores. Além disso, é importante considerar que, devido à pandemia da COVID-19, em 2020 a universidade flexibilizou a obrigatoriedade da matrícula dos alunos, o que pode ter

enviesado as taxas de evasão e retenção analisadas nesse ano. Adicionalmente, como não há registros individuais ou listas de presença que permitam verificar a participação efetiva dos estudantes nas estratégias, assumiu-se que todos os calouros ingressantes após a implementação de uma estratégia foram igualmente expostos a ela, o que pode não refletir a realidade.

No que se refere às limitações do estudo, destaca-se que, embora pertençam à mesma área do conhecimento, os cursos de computação apresentam particularidades curriculares relevantes. Diferenças na estrutura dos cursos, como a maior carga de disciplinas de matemática no curso de CdC em comparação ao curso de SI, podem impactar o desempenho acadêmico e a permanência dos estudantes. Outro aspecto importante refere-se ao turno de oferta dos cursos, especialmente o curso de SI noturno, cujos estudantes podem frequentemente conciliar trabalho e estudo, o que pode influenciar tanto a participação nas estratégias de acolhimento quanto as taxas de evasão observadas. Outras limitações incluem a ausência de análises causais e possíveis riscos de viés nas entrevistas realizadas com coordenadores de curso.

Nesse contexto, este trabalho contribui para a literatura ao preencher a lacuna relacionada à ausência de estudos quantitativos e correlacionais que analisem a relação entre estratégias de acolhimento e os índices de evasão, especialmente em cursos da área de computação de uma universidade pública brasileira. Ao adotar uma abordagem quantitativa, o estudo investiga a associação entre a implementação dessas estratégias e a evasão de estudantes ingressantes, oferecendo evidências que podem subsidiar o planejamento e a adaptação de ações de acolhimento em outros contextos institucionais.

Como trabalhos futuros, pretende-se aprofundar as análises para investigar o impacto individual e combinado das estratégias em cada uma das três categorias, ainda em termos de evasão no primeiro ano; considerar a aplicação de novas análises estatísticas que possam investigar relações causais entre o acolhimento e a redução da evasão; e analisar de forma específica os cursos oferecidos no turno noturno, os quais podem introduzir desafios adicionais à permanência dos alunos no curso.

Capítulo 4

Uso de Estratégias de Acolhimento como Atributos para Modelos de Previsão de Evasão Discente no Primeiro Ano em Cursos Superiores de Computação

Resumo: A evasão no ensino superior permanece como um dos principais desafios enfrentados pelas instituições de ensino, gerando impactos acadêmicos, sociais e financeiros. Este estudo tem como objetivo avaliar se atributos relacionados a estratégias de acolhimento têm impacto sobre a evasão de calouros em cursos da área de computação da Universidade Federal de Viçosa (UFV). Para isso, foram construídas e analisadas duas bases de dados: uma disponibilizada pela instituição, composta apenas pelos atributos acadêmicos e sociodemográficos, e outra enriquecida com atributos específicos sobre as estratégias de acolhimento desenvolvidas por cada curso. Cinco algoritmos de aprendizado de máquina foram aplicados a ambas as bases para a realização de análises segmentadas por curso. Os resultados mostraram que a inclusão dos atributos de acolhimento contribuiu positivamente, de maneira geral, para o aumento do desempenho preditivo dos modelos. Embora essas variáveis ainda não figurem entre as mais importantes na maioria dos casos, sendo superadas por atributos acadêmicos como CRA e número de reprovações, seu papel como indicadores precoces é promissor, visto que, diferentemente dos atributos acadêmicos, as informações sobre acolhimento estão disponíveis desde o ingresso dos estudantes.

Palavras-chave: Evasão no Ensino Superior; Estratégias de Acolhimento; Área de Computação; Algoritmos; Aprendizado de Máquina; Atributos de Acolhimento; Desempenho Preditivo.

Abstract: Dropout in higher education remains one of the main challenges faced by educational institutions, generating academic, social, and financial impacts. This study aims to assess whether attributes related to freshmen onboarding strategies influence the dropout of first-year students in computer courses at Federal University of Viçosa (UFV). To this end, two datasets were built and analyzed: one provided by the institution, composed only of academic and sociodemographic attributes, and another enriched with specific attributes related to the onboarding strategies developed by each course. Five machine learning algorithms were applied to both datasets to conduct analyses segmented by course. The results showed that

the inclusion of the onboarding attributes generally contributed positively to the increase in the predictive performance of the models. Although these variables are still not among the most important in most cases, being surpassed by academic attributes such as GPA and the number of failed subjects, their role as early indicators is promising, given that, unlike academic attributes, information on welcome strategies is available from the moment students enter the institution.

Keywords: Dropout in Higher Education; Onboarding Strategies; Computer Courses; Machine Learning; Algorithms; Onboarding Attributes; Predictive Performance.

4.1 Introdução

A evasão no ensino superior é um fenômeno complexo e multifatorial que impacta os estudantes mas também as próprias instituições (da Cruz et al., 2023). Esse fenômeno está presente no contexto dos cursos da área de computação, conhecidos por sua alta complexidade e elevados índices de desistência (Menolli et al., 2020). Conforme Duran et al. (2023), a evasão em cursos dessa área está comumente relacionada a obstáculos como a dificuldade de acompanhar o ritmo das atividades acadêmicas, a ausência de afinidade com o curso escolhido e a insuficiência de apoio institucional. Nesse contexto, a evasão tende a se concentrar nos primeiros períodos dos cursos, fase em que os estudantes estão mais vulneráveis a dificuldades de adaptação e enfrentam disciplinas com maior taxa de reprovação, como as de programação, dada a complexidade de seus conteúdos iniciais e a demanda por raciocínio lógico e abstrato (Pachas et al., 2021; dos Santos Silva et al., 2020).

Como forma de entender o problema e os principais fatores causadores da evasão, diversas pesquisas como as de Couto and Santana (2017) e do Nascimento et al. (2024) têm explorado o uso da mineração de dados educacionais (*Educational Data Mining* - EDM) como ferramenta para analisar padrões e fatores associados ao abandono escolar. A EDM permite extrair conhecimento útil a partir de grandes volumes de dados educacionais, possibilitando a identificação de perfis de risco e o mapeamento de variáveis que influenciam a permanência ou evasão dos estudantes (da Silva et al., 2019b). De forma complementar, técnicas de aprendizado de máquina (*Machine Learning* - ML) vêm sendo amplamente utilizadas na predição da evasão, dada sua capacidade de lidar com dados complexos e identificar padrões não triviais (de Brito et al., 2020). Modelos como Árvores de Decisão e Florestas Aleatórias têm alcançado bons resultados na detecção de estudantes em situação de risco de abandono (Santos et al., 2019; Costa et al., 2021). Além disso, a integração entre EDM e ML potencializa a acurácia das previsões, permitindo intervenções mais

precisas e oportunas por parte das instituições de ensino (Kantorski et al., 2023).

A evasão nos anos iniciais também tem sido investigada sob a perspectiva das ações que têm sido aplicadas nas instituições como forma de mitigar o problema. Alguns estudos têm destacado a importância de iniciativas institucionais de acolhimento, especialmente, durante a fase inicial dos cursos. Silva et al. (2021b), Graça and Nascimento (2021) e de Carvalho (2019) ressaltam que ações como mentorias, programas de nivelamento e atividades de integração entre calouros e veteranos podem desempenhar papel crucial na retenção estudantil, em especial dos alunos ingressantes. No entanto, ainda não foram encontrados trabalhos na literatura que avaliem o impacto dessas estratégias a partir de abordagens quantitativas e preditivas, em especial no contexto dos cursos de computação. Os estudos conhecidos tem se limitado ao uso de atributos distribuídos em quatro dimensões principais: Demográfico, Socioeconômico e Pessoal; Conhecimento Prévio; Institucional; e Desempenho Acadêmico (Mundim et al., 2025).

Na Universidade Federal de Viçosa (UFV), cursos superiores da área de computação, distribuídos em seus três campi, têm enfrentado desafios para reter seus estudantes, especialmente nos períodos iniciais, quando são ministradas disciplinas introdutórias, como Programação e Estruturas de Dados, frequentemente associadas a elevadas taxas de reprovação. Na tentativa de mitigar o problema, esses cursos têm implementado estratégias de acolhimento voltadas à integração e ao apoio aos calouros. Em um trabalho anterior (Mundim et al., 2024), os autores apresentaram algumas estratégias de acolhimento de um dos cursos de computação da UFV, sendo uma de gamificação e outra grupos de estudos em programação. Os resultados indicaram um impacto positivo no desempenho acadêmico e no engajamento dos participantes. No entanto, os efeitos dessas ações foram avaliados apenas de forma exploratória em um dos campi da instituição, desconsiderando as estratégias adotadas por outros cursos da área e em outros campi.

Dessa forma, o objetivo deste estudo é investigar o uso de novos atributos relacionados às estratégias de acolhimento praticadas nos cursos de computação da UFV, de forma combinada com atributos comumente utilizados nos modelos preditivos de evasão, buscando avaliar tanto o panorama da evasão do primeiro ano quanto o impacto da incorporação desses novos atributos no desempenho dos modelos e a importância de cada um deles. O estudo foi guiado pelas seguintes questões de pesquisa (QP): **QP1** - Qual a situação geral da evasão, com ênfase nos calouros, nos cursos de computação da UFV?; **QP2** - Como os modelos preditivos de evasão se comportam quando atributos acadêmicos são combinados com atributos relacionados a estratégias de acolhimento?; e **QP3** - Qual a importância dos atributos relacionados a estratégias de acolhimento nos modelos preditivos de evasão de calouros nos cursos de computação da UFV?

Este estudo mostra-se relevante e traz contribuições importantes para a compreensão da evasão nos cursos da área de computação da UFV, ao investigar o impacto do uso de novos atributos relacionados a estratégias de acolhimento no comportamento dos modelos preditivos de evasão de calouros. Ao contrário de atributos como número de reprovações, coeficiente de rendimento acadêmico (CRA), raça, entre outros, que se destacam nos modelos de predição de evasão, mas trazem previsões com base no passado (a evasão já aconteceu) (Lakkaraju et al., 2015), atributos relacionados a estratégias de acolhimento podem contribuir para prever futuras evasões, fornecendo indicadores importantes para que as instituições possam fortalecer estratégias já existentes e fomentar novas ações como forma de mitigar a evasão no primeiro ano.

Além disso, ao integrar dados institucionais com técnicas de EDM e ML, a pesquisa fornece evidências quantitativas sobre o impacto das estratégias de acolhimento na trajetória dos calouros, ampliando a compreensão sobre sua efetividade. Por fim, o trabalho apresenta uma metodologia replicável, que pode ser adaptada por outras instituições de ensino superior interessadas em avaliar evasão discente e aprimorar suas próprias estratégias de acolhimento.

O artigo está organizado da seguinte forma: a Seção 4.2 apresenta os trabalhos relacionados ao uso de EDM, ML e estratégias de acolhimento no estudo da evasão; a Seção 4.3 descreve a metodologia utilizada na pesquisa; a Seção 4.4 discute os resultados obtidos na execução dos modelos de ML com e sem a inclusão dos atributos sobre estratégias de acolhimento, além da análise de importância dos atributos; e, por fim, a Seção 4.5 apresenta as considerações finais.

4.2 Trabalhos Relacionados

4.2.1 Uso de EDM e ML na Análise da Evasão

O estudo de Duran et al. (2023) faz uso de técnicas de EDM para investigar fatores que influenciam o abandono em cursos de computação, como raça/etnia, gênero e status socioeconômico, além de elementos que motivam a permanência. A análise exploratória foi feita a partir de dados de uma pesquisa online com 3.193 alunos. Os resultados apontam que boas perspectivas salariais e de emprego são os principais fatores de retenção, enquanto o excesso de teoria e dificuldades em programação e matemática motivam o abandono. Embora gênero, raça/etnia e status socioeconômico tenham pouco impacto, a predominância masculina, o assédio e a dificuldade dos cursos afetam mais as mulheres.

Em de Jesus et al. (2021), os autores buscaram prever o risco de evasão no curso de Licenciatura em Computação da Universidade do Estado do Amazonas (UEA),

com foco nas dificuldades dos alunos em disciplinas de programação. A hipótese era que os históricos acadêmicos poderiam indicar alunos com maior propensão à evasão. Utilizando uma metodologia de EDM e Redes Neurais de múltiplas camadas, foi construído um modelo preditivo que alcançou 98% de precisão, demonstrando o potencial do uso de ML na identificação de riscos de evasão.

A pesquisa de [Oliveira and Medeiros \(2024\)](#) desenvolveu e avaliou um modelo preditivo para identificar estudantes com alta propensão à evasão, utilizando dados de autoavaliações institucionais semestrais da Universidade Federal da Paraíba (UFPB). Com base na metodologia *CRoss-Industry Standard Process for Educational Data Mining* (CRISP-EDM) ([Ramos et al., 2020](#)), os autores aplicaram técnicas de classificação como *Árvore de Decisão*, *Random Forest* e *Support Vector Machine* (SVM). Os modelos foram avaliados por métricas como acurácia, precisão, revocação e F1-score, com destaque para a precisão de 91,72%. O estudo também revelou que cerca de 59% dos alunos ativos admitidos a partir de 2017 apresentaram alta probabilidade de evasão.

4.2.2 Uso de Estratégias de Acolhimento

Para combater a evasão de calouros, algumas instituições adotam estratégias de acolhimento, como disciplinas introdutórias e semanas de recepção. Na UnB, o Departamento de Ciência da Computação (CIC) implementou três ações distintas. A primeira, descrita por [Holanda et al. \(2021\)](#), é o **Programa de Acolhimento dos Calouros**, realizado virtualmente no primeiro semestre de 2020. O programa foi estruturado em uma série de encontros, iniciando com a apresentação do departamento e das diferentes formas de apoio oferecidas pela UnB. Posteriormente, os calouros assistiram a palestras sobre os principais desafios enfrentados no primeiro ano da graduação. Nos encontros finais, ex-alunos participaram, promovendo uma interação enriquecedora com os ingressantes. A partir do segundo encontro, foram aplicadas avaliações das atividades, cujos *feedbacks* indicaram uma recepção bastante positiva do programa como um todo.

A segunda experiência é relatada por [Holanda et al. \(2022a\)](#), que aborda o desenvolvimento de **Monitorias** em disciplinas introdutórias do CIC, implementadas no segundo semestre de 2020. Essas monitorias envolveram plantões de dúvidas, atendimentos individuais agendados, monitoramento constante e aulas realizadas aos sábados. Ao final do semestre, uma pesquisa de avaliação foi aplicada aos participantes, cujos resultados revelaram que a maioria considerou a experiência positiva, destacando o papel fundamental das monitorias na criação de um ambiente de suporte e colaboração.

Por fim, [Holanda et al. \(2022b\)](#) apresenta a criação da disciplina **Introdução ao**

Ambiente Universitário da Computação (IAUC), implementada em 2021 nos cursos do CIC com o propósito de reduzir a evasão. A disciplina foi organizada em quatro núcleos temáticos: informação, instrução, motivação e interação. O núcleo de informação introduzia os alunos à universidade e ao curso; o de instrução oferecia orientações e estratégias de estudo; o de motivação incluía relatos de professores e ex-alunos, além da apresentação de projetos e uma maratona de programação; e o de interação promovia dinâmicas e jogos entre os calouros para estimular a integração. Os *feedbacks* foram positivos, com destaque especial para os encontros com ex-alunos, que obtiveram as melhores avaliações e um nível maior de engajamento.

Já na Universidade Federal do Maranhão (UFMA), [Pinto et al. \(2020\)](#) mostraram o planejamento, execução e avaliação da **Semana do Calouro**, promovida pelo Programa de Educação Tutorial em Computação (PETComp). O evento visa acolher e orientar os calouros de Ciência da Computação, oferecendo informações sobre oportunidades acadêmicas e promovendo sua integração à universidade. As atividades incluíram sessões de tira-dúvidas, dinâmicas, visitas a laboratórios e palestras sobre temas atuais da área. Os resultados indicaram melhor ambientação dos alunos, maior compreensão do curso e indícios de redução da evasão.

4.2.3 Similaridades e Diferenças deste Trabalho

Diferentemente dos estudos anteriores, que abordam EDM e ML (separadamente ou em conjunto), ou se restringem à análise qualitativa de estratégias de acolhimento, esta pesquisa propõe integrar essas três abordagens. A Tabela 4.1 resume as principais diferenças metodológicas entre este estudo e os trabalhos relacionados.

Tabela 4.1: Comparativo de abordagens entre os trabalhos relacionados e este estudo.

Trabalho	ML	EDM	Acolhimento	Atributos Acolhimento	Comparação ML
Duran et al. (2023)	X	✓	X	X	X
de Jesus et al. (2021)	✓	✓	X	X	X
Oliveira and Medeiros (2024)	✓	✓	X	X	✓
Holanda et al. (2021)	X	X	✓	X	X
Holanda et al. (2022a)	X	X	✓	X	X
Holanda et al. (2022b)	X	X	✓	X	X
Pinto et al. (2020)	X	X	✓	X	X
Mundim et al. (2024)	X	✓	✓	X	X
Este trabalho	✓	✓	✓	✓	✓

Os trabalhos analisados apresentam abordagens complementares, mas com focos distintos. Observa-se que a maioria das pesquisas utiliza técnicas de EDM e ML para prever ou compreender a evasão estudantil, como em [de Jesus et al. \(2021\)](#) e [Oliveira and Medeiros \(2024\)](#). Entretanto, apenas os estudos voltados ao contexto da UnB ([Holanda et al., 2021, 2022a,b](#)) e da UFMA ([Pinto et al., 2020](#)) exploram estratégias de acolhimento voltadas à permanência de calouros, sem, contudo, relacioná-las a

modelos quantitativos de predição. Além disso, nenhum dos trabalhos anteriores investigou atributos específicos de acolhimento como variáveis explicativas nos modelos de ML. Nesse contexto, este trabalho diferencia-se por integrar EDM, ML e variáveis relacionadas ao acolhimento em um mesmo estudo, permitindo avaliar o impacto dessas estratégias na evasão e comparar o desempenho de diferentes modelos de predição.

4.3 Materiais e Métodos

Esta seção apresenta a metodologia da coleta, tratamento e análise dos dados dos discentes e das estratégias de acolhimento dos cursos de computação da UFV. Será apresentado o processo de modelagem e de avaliação dos modelos de ML, incluindo os métodos, ferramentas e recursos computacionais utilizados.

Com o intuito de atingir os objetivos propostos na Seção 4.1, foi conduzido um processo de EDM e aplicação de algoritmos de ML no contexto educacional. Para orientar esse processo, adotou-se a metodologia CRISP-EDM, proposta por [Ramos et al. \(2020\)](#), que é uma adaptação da metodologia CRISP-DM, porém especificamente voltada para o domínio educacional. As etapas da CRISP-EDM, em ordem, são: Entendimento do Problema; Entendimento do conjunto de Dados; Preparação dos Dados; Modelagem; Avaliação e; Implementação da Solução. Ressalta-se que a última etapa não foi contemplada nesta pesquisa, visto que o objetivo deste trabalho é apenas avaliar se os atributos relacionados às estratégias de acolhimento impactam na evasão dos calouros de cursos de computação.

4.3.1 Entendimento do Problema e do Conjunto de Dados

As duas primeiras etapas do processo são entender o problema, o qual foi abordado na Seção 4.1, e o domínio da aplicação do conjunto de dados, ou seja, o contexto dos dados educacionais que foram minerados para o estudo da evasão.

Nesse sentido, o primeiro passo prático foi a extração dos dados dos discentes, a qual foi realizada a partir da base de dados acadêmicos da UFV, que reúne diversos atributos essenciais dos alunos de todos os cursos da instituição. A Tabela 4.2 apresenta esses atributos, seus tipos e uma breve descrição de cada um deles.

Uma observação importante reside no **A12**, o qual mostra todos os conceitos acadêmicos da UFV. A instituição considera como evasão¹ as situações classificadas como abandono (A), desligamento (D), falecimento (F), mudança de curso (M),

¹Código identificador (conceito) da situação do aluno (A12): <https://sre.caf.ufv.br/avaliacao-do-rendimento-academico-2/>

Tabela 4.2: Atributos da base de dados acadêmica disponibilizada.

ID	Atributo	Tipo	Descrição
A1	Identificador	Numérico	Identificação numérica dos alunos (não é a matrícula).
A2	Ano de Nascimento	Numérico	Ano de nascimento de cada aluno.
A3	Gênero	Catégorico Nominal	Gênero do aluno, definido por M ou F.
A4	Campus	Catégorico Nominal	Campus da instituição em que o aluno ingressou.
A5	Identificador de Curso	Numérico	Identificador numérico do curso de um determinado campus.
A6	Curso	Catégorico Nominal	Nome do curso em que o aluno está matriculado.
A7	Estado de Nascimento	Catégorico Nominal	Estado de nascimento de cada aluno.
A8	Região de Nascimento	Catégorico Nominal	Região (do Brasil) de nascimento de cada aluno.
A9	Município de Nascimento	Catégorico Nominal	Município de nascimento de cada aluno.
A10	Período de Admissão	Temporal	Período de ingresso do aluno na instituição.
A11	Período de Saída	Temporal	Período de saída do aluno da instituição.
A12	Código da Situação do Aluno	Catégorico Nominal	Código identificador (conceito) da situação do aluno.
A13	Situação do Aluno	Catégorico Nominal	Situação do aluno escrita por extenso de acordo com o código.
A14	Situação Agrupada	Catégorico Nominal	Agrupamento das diversas situações do aluno em apenas quatro situações gerais: evasão, matriculado, conclusão e retenção.
A15	Modalidade Inscrita pelo ENEM	Numérico	Modalidade do ENEM utilizada para ingresso.
A16	Nota do ENEM	Numérico	Nota do ENEM de cada aluno.
A17	Número de Reprovações	Numérico	Número total de reprovações de cada aluno.
A18	Raça	Catégorico Nominal	Raça de cada aluno.
A19	Coeficiente de Rendimento Acumulado	Numérico	CRA de cada aluno.
A20	Município de Endereço	Catégorico Nominal	Município em que o aluno mora atualmente.
A21	Estado de Endereço	Catégorico Nominal	Estado em que o aluno mora atualmente.

desligamento com pedido de reconsideração (R), transferência (T) e exclusão (X), que representam formas de saída definitiva do curso ou da universidade.

Em seguida, foram coletadas, por meio de entrevistas realizadas com as coordenações dos cursos dos três campi da UFV, as informações relacionadas às estratégias de acolhimento, utilizando um roteiro de perguntas². Para os coordenadores que não puderam participar das entrevistas, foi disponibilizada a opção de responder ao questionário por escrito, por meio de um documento no Google Docs, o que possibilitou ampliar a quantidade de cursos participantes.

Essa coleta ocorreu entre abril e novembro de 2024, de forma remota ou presencial nos três campi. A participação variou consideravelmente entre eles: no campus Florestal (CAF), foi possível obter respostas de todos os 10 cursos (100%); no campus Rio Paranaíba (CRP), obteve-se retorno de 8 dos 10 cursos (80%); e no campus Viçosa (CAV), 23 dos 47 cursos (49%).

No CAV, a atuação presencial foi decisiva, sem ela, apenas 8 cursos teriam participado. No CRP, o processo também exigiu esforços adicionais, incluindo a intervenção direta da diretora geral e de outros funcionários-chave para viabilizar as respostas. Esses desafios reforçam que a coleta de informações não foi trivial, demandando adaptações e mobilização ativa para ampliar a cobertura e garantir maior representatividade dos dados.

²Perguntas: <https://github.com/carvalhopedro22/Perguntas-das-entrevistas>

A partir das respostas, foi compilada uma base única com as estratégias adotadas, organizadas na Figura 4.1 em três categorias, em que cada estratégia corresponde a um atributo. Essas categorias foram definidas pela aplicação do método de codificação aberta (*open coding*) (Miles et al., 2014). Esse método tem como finalidade organizar e resumir os dados por meio da identificação de padrões e similaridades. Como resultado, foram obtidas as categorias da Figura 4.1.

Recepção (R)		Apoio ao Discente (AD)	
ID	Atributo	ID	Atributo
R1	Palestras Informativas	AD1	Apadrinhamento de Calouros (entre discentes)
R2	Encontro com Comissão Coordenadora	AD2	Acompanhamento Acadêmico Individual
R3	Encontro com Docentes	AD3	Acompanhamento Acadêmico em Grupo
R4	Encontro com Representação Discente	AD4	Disciplina de Introdução
R5	Encontro com Veteranos		
R6	Encontro com Alunos de Pós-graduação	Engajamento e Motivação (EM)	
R7	Encontro com Egressos	ID	Atributo
R8	Encontro com Empresa Júnior	EM1	Estratégias de Gamificação
R9	Confraternização	EM2	Grupos de Estudos
R10	Campus Tour	EM3	Desenvolvimento Supervisionado de Projetos
R11	Trilha Ecológica	EM4	Visitas Técnicas

Figura 4.1: Atributos da base de dados das estratégias de acolhimento.

As estratégias de *Recepção* concentram-se principalmente na primeira semana de aula, esclarecendo aspectos fundamentais do curso e da universidade, além de promoverem encontros com docentes, veteranos, egressos e demais membros da comunidade acadêmica, favorecendo a ambientação inicial. As estratégias de *Apoio ao Discente*, por sua vez, desenvolvem-se ao longo do primeiro período letivo e têm como objetivo facilitar a adaptação dos calouros, por meio da apresentação do curso e da integração com os colegas. Já as estratégias de *Engajamento e Motivação* são mais estruturadas e contínuas, voltadas à promoção da interação entre calouros, veteranos e professores ao longo de todo o primeiro período.

Por fim, cabe destacar que esta investigação se baseia exclusivamente em dados acadêmicos anonimizados, impossibilitando qualquer forma de identificação individual dos estudantes. Do mesmo modo, os dados referentes às estratégias de acolhimento dizem respeito apenas a práticas institucionais implementadas pelos cursos, sem qualquer associação com informações pessoais.

4.3.2 Preparação dos Dados

A terceira etapa consiste na preparação e análise exploratória dos dados. Ela foi dividida nas fases: Filtragens, Remoção de Variáveis, de Valores Nulos e Criação da Variável Alvo, Enriquecimento da base de dados acadêmicos (disponibilizada pela instituição) e Engenharia de Atributos. Para isso, foi utilizada a linguagem de programação Python³, na plataforma Jupyter⁴, com as bibliotecas Pandas⁵ e NumPy⁶, que são fundamentais no processamento e manipulação dos dados.

Filtragens

A base de dados acadêmicos disponibilizada pela instituição utilizada neste trabalho abrange o período de 2000 a 2024 e todos os cursos da UFV. No entanto, como o escopo da pesquisa abrange apenas os cursos de computação, foi feito um recorte específico, considerando apenas os cursos dessa área, sendo eles Ciência da Computação (CdC) nos campi CAF e CAV e Sistemas de Informação (SI), integral (Int) e noturno (Not) no campus CRP. Após essa filtragem inicial, a base passou a conter 3509 alunos.

Na etapa seguinte, realizou-se uma análise exploratória com foco na evasão dos calouros dos cursos da área de computação, ou seja, considerando apenas as evasões que ocorreram no mesmo ano de entrada. Dessa forma, foram incluídos somente os estudantes que ingressaram no primeiro semestre de cada ano (identificados por /1), uma vez que as estratégias de acolhimento são aplicadas apenas nesse período. Por isso, os alunos que começaram no segundo semestre (/2) foram excluídos da contagem de calouros. Entretanto, a evasão foi analisada considerando todas as saídas possíveis no primeiro ano, incluindo, portanto, tanto o primeiro quanto o segundo período. Após esse refinamento, o conjunto de dados passou a incluir 3310 alunos.

Além disso, embora o ano de 2024 conste na base, seus dados foram desconsiderados nas análises, pois ainda não estavam completos no momento da extração. Com essa exclusão, obteve-se a amostra composta por 3112 calouros.

Por fim, como esses cursos foram criados em anos distintos, estabeleceu-se um ponto de corte comum em 2012, correspondente ao ano de início do curso mais recente, o que garante homogeneidade na série histórica analisada e evita que os anos anteriores apresentem valores zerados para cursos que ainda não existiam. Após aplicar esse corte temporal, a base final considerada neste estudo passou a conter 2377 alunos.

Assim como ocorreu com os dados acadêmicos, as informações sobre as estratégias de acolhimento abrangeram diversos cursos dos três campi da UFV. No

³Python: <https://www.python.org/>

⁴Jupyter: <https://jupyter.org/>

⁵Pandas: <https://pandas.pydata.org/docs/>

⁶NumPy: <https://numpy.org/>

total, 41 dos 67 cursos participaram da entrevista. No entanto, considerando que o foco deste trabalho está nos cursos da área de computação, foi aplicada uma filtragem nessa base para manter apenas os registros correspondentes a esses cursos, para os quais foram consideradas 11 estratégias distintas. A Figura 4.2 apresenta quais atributos referentes às estratégias de acolhimento permaneceram e o ano em que elas iniciaram, o qual foi importante no momento de enriquecer a base de dados acadêmicos. Ainda assim, a realização das entrevistas com um número expressivo de cursos mostrou-se fundamental, pois permitiu levantar um conjunto amplo de informações, contribuindo para a identificação e categorização das estratégias de acolhimento nas três categorias citadas anteriormente.

<i>Curso</i>	<i>Campus</i>	<i>Estratégia (Atributo)</i>	<i>Ano de Início</i>
CdC	CAF	AD4	2015
		R2, R3, R5, R6, R7	2016
		EM1	2017
		EM2	2017
		AD2	2019
CdC	CAV	AD4	2007
		R1, R3, R5, R7, R9	2016
SI	CRP	R1, R3, R5	2022

Figura 4.2: Atributos da base de dados das estratégias de acolhimento dos cursos de computação da UFV.

Remoção de Atributos, Valores Nulos e Criação da Variável Alvo

Durante o processo de preparação dos dados, alguns atributos listados na Tabela 4.2 foram removidos por diferentes razões relacionadas à sua relevância para os modelos de ML. A exclusão desses atributos teve como objetivo tornar a base de dados mais consistente e adequada às análises.

Primeiramente, alguns atributos foram descartados por conterem apenas valores numéricos aleatórios que não representavam nenhuma informação útil para o contexto da análise. É o caso do atributo identificador (A1), cujo valor era gerado unicamente para fins internos da base de dados.

Além disso, alguns atributos foram eliminados por não apresentarem viabilidade ou utilidade prática na transformação dos dados para aplicação em modelos de

aprendizado de máquina. Entre os principais motivos estão a elevada quantidade de valores distintos, a presença de informações duplicadas ou a perda de sentido dos atributos após a segmentação da base por curso. É o caso dos atributos campus (A4), identificador de curso (A5), curso (A6), município de nascimento (A9), período de admissão (A10), período de saída (A11), código da situação do aluno (A12), situação do aluno (A13) e situação agrupada (A14).

No caso específico dos atributos A4 e A6, por exemplo, sua utilização se torna redundante, uma vez que a base, ao ser analisada, foi segmentada exatamente de acordo com os valores que esses atributos representam (campus e curso). Já para o A10, como o preenchimento das novas colunas de estratégias foi definido a partir do respectivo ano de criação, sua informação acabou se tornando redundante. Isso ocorre porque a participação do aluno em determinada estratégia está diretamente condicionada ao ano em que ela foi implementada, que por sua vez coincide com o ano (período) de ingresso do próprio aluno.

Também foram removidos os atributos que apresentavam alto grau de incompletude, o que comprometeria a qualidade das análises. Essa foi a justificativa para a exclusão do município de endereço (A20) e estado de endereço (A21).

Por fim, foi criado o atributo alvo dos modelos, denominado **evasão primeiro ano**. Para sua definição, realizou-se uma filtragem com base nos atributos que indicam os períodos de admissão e de saída dos alunos. Assim, foram classificados como casos de **evasão** aqueles em que o período de saída ocorreu no mesmo ano do ingresso e como **não evasão** todos os demais casos.

Dessa forma, a base de dados acadêmicos final que foi utilizada pelos modelos e enriquecida com os atributos de acolhimento da Figura 4.2 passou a conter, além do atributo alvo, outros nove atributos: ano de nascimento (A2), gênero (A3), estado de nascimento (A7), região de nascimento (A8), modalidade inscrita pelo ENEM (A15), nota do ENEM (A16), número de reprovações (A17), raça (A18), coeficiente de rendimento acumulado (A19).

Após as filtrações, remoção dos atributos e criação do atributo alvo, foi realizada a limpeza dos registros com valores nulos. Como as filtrações anteriores já haviam reduzido significativamente a presença de nulos, essa etapa final gerou um impacto mínimo sobre o volume total de dados (de 2377 para 2142 alunos). A Figura 4.3 ilustra todo esse processo de preparação, desde a base fornecida pela instituição até a obtenção da amostra final. Por fim, é importante destacar que, após o tratamento, os 2142 alunos foram segmentados por campus, de modo que as análises passaram a utilizar quatro conjuntos de dados distintos, um para cada curso de computação.

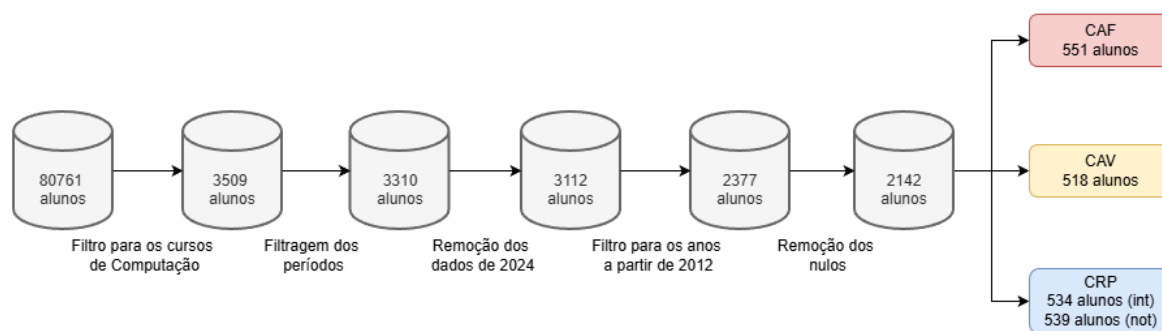


Figura 4.3: Processo de preparação dos dados.

Enriquecimento da Base de Dados Acadêmicos e Engenharia de Atributos

O processo de enriquecimento da base de dados disponibilizada pela instituição consistiu na inserção dos atributos relacionados às estratégias de acolhimento mapeadas nos cursos de computação (Figura 4.2). Cada uma das 11 estratégias foi associada ao respectivo ano de implementação, permitindo cruzar essa informação com o ano de ingresso de cada discente. Para operacionalizar essa informação, realizou-se diretamente nessa etapa uma codificação binária (*one-hot encoding*), na qual cada estratégia foi representada como uma variável com valores “Sim” (1) ou “Não” (0), indicando se o estudante teve ou não provável exposição à estratégia.

Além disso, esse enriquecimento foi aplicado separadamente nas quatro bases correspondentes aos cursos de computação da UFV. Nos cursos em que determinada estratégia não é adotada, a coluna correspondente permanece com valores zerados, independentemente do ano de ingresso dos alunos. Esse procedimento garante que todas as bases resultantes mantenham a mesma estrutura de atributos.

A etapa final de preparação dos dados envolveu complementar esse processo de Engenharia de Atributos para os demais campos da base após todas as filtragens realizadas. Basicamente, os atributos categóricos e temporais foram convertidos em valores numéricos, de modo a viabilizar um melhor desempenho dos modelos. A abordagem utilizada foi a utilização de dicionários de mapeamento (*map*), nos quais cada categoria foi associada a um valor inteiro correspondente. Essa transformação buscou garantir a consistência dos dados e permitir que os algoritmos de aprendizado pudessem processar as variáveis de forma eficiente, sem perda de significado semântico.

4.3.3 Modelagem

Nesta etapa, foram definidos os algoritmos de ML para construir os modelos preditivos da evasão no primeiro ano, sendo eles: Árvore de Decisão (*Decision Tree - DT*), Floresta Aleatória (*Random Forest - RF*), Regressão Logística (*Logistic Regression -*

LR), *Naive Bayes* (NB) e *Support Vector Machine* (SVM). A escolha desses algoritmos se baseou na sua ampla utilização em estudos da literatura, sendo recorrentes em trabalhos voltados à predição de evasão (Mundim et al., 2025). Todos os modelos foram implementados por meio da biblioteca Scikit-learn⁷, na linguagem Python.

A DT é um algoritmo que utiliza uma estrutura hierárquica de decisões simples para realizar tarefas de classificação ou regressão, sendo bastante interpretável e fácil de visualizar. Já a RF combina várias árvores de decisão para produzir previsões mais robustas e precisas, sendo eficaz especialmente em bases de dados com muitas variáveis. O algoritmo LR, por sua vez, é amplamente utilizado em tarefas de classificação binária, que estima a probabilidade de uma instância pertencer a uma determinada classe com base em uma combinação linear dos atributos de entrada. O NB é um algoritmo probabilístico baseado no Teorema de Bayes, que assume independência entre os atributos e costuma apresentar bom desempenho em classificações rápidas com grandes volumes de dados. Por fim, o SVM busca encontrar um limite ótimo (hiperplano) que separe as classes no conjunto de dados, sendo adequado para problemas complexos e conjuntos de pequeno a médio porte (Oliveira and Medeiros, 2024; Opazo et al., 2021).

Seleção de Atributos e Separação dos Dados

Todos os atributos originais disponíveis após a preparação foram utilizados para se gerar os modelos a partir dos cinco algoritmos, tanto antes quanto após o enriquecimento da base com os dados das estratégias de acolhimento. Essa abordagem foi adotada em razão da quantidade reduzida de atributos (9 atributos acadêmicos, 11 atributos de acolhimento e o atributo alvo).

Relativo à separação dos dados, uma maneira simples e amplamente adotada para avaliar a capacidade de generalização de um modelo de aprendizado de máquina é por meio da divisão em conjuntos de treinamento e teste. Neste trabalho, a divisão dos dados foi feita da seguinte forma: 80% dos dados foram reservados para o treinamento e 20% para o teste. Com essa abordagem, torna-se possível mensurar o desempenho do modelo em dados que não foram utilizados durante o processo de aprendizado, obtendo uma estimativa do seu erro em dados desconhecidos. No entanto, uma de suas limitações é que não há garantias de que a divisão realizada represente adequadamente a distribuição geral da base de dados, o que pode comprometer a validade da avaliação (Géron, 2019).

Para contornar essas limitações e realizar uma busca por combinações de hiperparâmetros que resultem no melhor desempenho dos modelos, foi utilizada neste trabalho a técnica de validação cruzada com *GridSearchCV*⁸ da biblioteca

⁷Scikit-learn: <https://scikit-learn.org/stable/>

⁸GridSearchCV: https://scikit-learn.org/stable/modules/generated/sklearn.model_

Scikit-learn. Esse método realiza uma varredura exaustiva sobre um espaço definido de hiperparâmetros, combinando isso com validação cruzada *k-fold* para garantir avaliações mais robustas. A cada combinação de hiperparâmetros, o modelo é treinado e validado em diferentes divisões dos dados, utilizando os *folds* como conjuntos de validação alternados. Ao final do processo, a combinação que apresenta melhor desempenho médio é selecionada. Essa abordagem, além de ajudar a reduzir o risco de *overfitting*, permite uma escolha mais confiável dos hiperparâmetros ideais com base em dados não vistos durante o treinamento.

Desbalanceamento dos Dados

Um aspecto importante considerado durante a tarefa de classificação foi o desbalanceamento das classes no conjunto de treinamento. Em todos os campi analisados, a quantidade de registros da classe negativa (não evasão) superou a da classe positiva (evasão). Especificamente, a classe majoritária foi aproximadamente 3,7 vezes maior no CAF, 4,8 vezes maior no CAV, 3,1 vezes maior no CRP Integral e 2,1 vezes maior no CRP Noturno.

Muitos algoritmos de ML partem do pressuposto de que os conjuntos de dados de treinamento apresentam uma distribuição equilibrada entre as classes, ou seja, que há uma proporção semelhante de exemplos para cada categoria a ser prevista. Contudo, esse equilíbrio raramente é encontrado em cenários reais. É comum que uma das classes possua um número significativamente maior de instâncias em comparação à outra, caracterizando o problema do desbalanceamento (Oliveira and Medeiros, 2024). Essa situação é bastante frequente em problemas de classificação e pode dificultar o desempenho dos modelos, prejudicando sua capacidade de identificar corretamente os casos da classe minoritária.

Dessa forma, durante o desenvolvimento dos modelos, foi avaliada a aplicação de técnicas de balanceamento para tentar reduzir o impacto do desbalanceamento entre as classes. Entre elas, testou-se a *Synthetic Minority Over-sampling Technique* (SMOTE), que gera exemplos sintéticos da classe minoritária com base nas características dos registros existentes, e a *Random Undersampling* (RUS), que busca equilibrar as classes por meio da redução aleatória de exemplos da classe majoritária. No entanto, após a realização de testes preliminares, observou-se que os modelos treinados com dados balanceados por essas técnicas não apresentaram ganhos significativos nas métricas de desempenho quando comparados aos modelos desenvolvidos com os dados originais desbalanceados.

Essa análise foi conduzida por meio da comparação direta entre os modelos treinados com os dados originais e aqueles treinados após o balanceamento. Para

cada técnica, avaliou-se se houve melhora ou piora nos valores das métricas. Em geral, observou-se que o número de pioras superou o de melhorias, e que as variações obtidas foram pequenas, indicando ausência de ganhos consistentes decorrentes do balanceamento. Assim, optou-se por não incluir as análises com SMOTE e RUS na apresentação dos resultados finais, priorizando a interpretação com base nos cenários mais próximos da realidade institucional.

4.3.4 Avaliação

Para avaliar o desempenho dos modelos, foram utilizadas algumas métricas de avaliação, sendo elas: **acurácia**, **precisão**, **revocação**, **especificidade**, **F1-score** e **AUC**. Os modelos foram avaliados em dois cenários distintos: (1) com a base obtida apenas a partir dos dados acadêmicos, denominada neste trabalho como "original", e (2) com a mesma base enriquecida com atributos relacionados às estratégias de acolhimento dos cursos de computação.

Ao analisar os resultados de um algoritmo, as métricas estão associadas ao número de previsões realizadas, as quais geralmente são representadas por uma matriz chamada matriz de confusão (Tabela 4.3), em que o "Sim" corresponde ao aluno que evadiu no primeiro ano e o "Não" ao que não evadiu no primeiro ano.

Tabela 4.3: Matriz de confusão.

		Detectada	
		Sim	Não
Real	Sim	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

A **acurácia** é calculada pela razão entre o número de previsões corretas (VP + VN) e o total de previsões (VP + VN + FP + FN). Ela representa a proporção de acertos do modelo em relação a todas as instâncias avaliadas, sendo uma medida geral de desempenho (Junior et al., 2022).

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (4.1)$$

Já a **precisão** é calculada pela divisão das previsões verdadeiras positivas (VP) pela soma das previsões verdadeiras e falsas positivas (VP + FP). Ou seja, ela mede a proporção de instâncias classificadas como positivas que realmente pertencem à classe positiva (Junior et al., 2022).

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (4.2)$$

A **revocação**, por sua vez, é calculada pela divisão das predições verdadeiras positivas (VP) pela soma das predições verdadeiras positivas e falsas negativas (VP + FN). Ela indica a capacidade do modelo de identificar corretamente as instâncias da classe positiva, ou seja, a proporção de positivos que foram corretamente reconhecidos (Junior et al., 2022).

$$\text{Revocação} = \frac{VP}{VP + FN} \quad (4.3)$$

A **especificidade** é calculada pela divisão das predições verdadeiras negativas (VN) pela soma das predições verdadeiras negativas e falsas positivas (VN + FP), ou seja, ela mede a proporção de instâncias negativas que foram corretamente identificadas como tal (Junior et al., 2022).

$$\text{Especificidade} = \frac{VN}{VN + FP} \quad (4.4)$$

O **F1-Score** é calculado a partir da média harmônica entre a Precisão e a Revocação. Ela busca um equilíbrio entre essas métricas, sendo especialmente útil em cenários com dados desbalanceados, pois considera tanto os falsos positivos quanto os falsos negativos (Junior et al., 2022).

$$\text{F1-Score} = \frac{2 \cdot (\text{Precisão} \cdot \text{Revocação})}{\text{Precisão} + \text{Revocação}} \quad (4.5)$$

Por fim, a **AUC** representa a Área sob a Curva ROC e indica a capacidade do modelo em distinguir entre as classes. Quanto mais próxima de 1, melhor o desempenho do modelo. A AUC resume o desempenho do classificador, considerando a taxa de verdadeiros positivos (Revocação) e a taxa de falsos positivos (1 - Especificidade) (Junior et al., 2022).

4.4 Resultados

Nesta seção, são apresentados os resultados dos modelos preditivos aplicados antes e depois da inclusão dos atributos de acolhimento, bem como a discussão sobre os atributos mais relevantes.

4.4.1 Análise Geral

Para responder à primeira questão de pesquisa (QP1 - Qual a situação geral da evasão, com ênfase nos calouros, nos cursos de computação da UFV?), a Figura 4.4 apresenta a quantidade de calouros evadidos e não evadidos, em relação ao total de alunos, por curso da área de computação da UFV, no intervalo de 2012 a 2023. Destaca-se que o

curso de CdC do campus CAV apresenta a menor taxa de evasão de calouros (inferior a 20%) e o curso de SI do Campus CRP (noturno) apresenta a maior (superior a 30%).

Curso	Campus	Total de Alunos	Não Evasão - 1º ano	Evasão - 1º ano	Evasão - 1º ano (%)
CdC	CAF	551	433	118	21,4%
CdC	CAV	518	429	89	17,2%
SI (Int)	CRP	534	402	132	24,7%
SI (Not)	CRP	539	366	173	32,1%

Figura 4.4: Quantidade de alunos em cada classe da variável alvo.

Já a Figura 4.5 mostra a quantidade de calouros evadidos por ano de admissão, sendo que a Figura 4.5a traz o total acumulado e a Figura 4.5b detalha os valores para cada curso de computação.

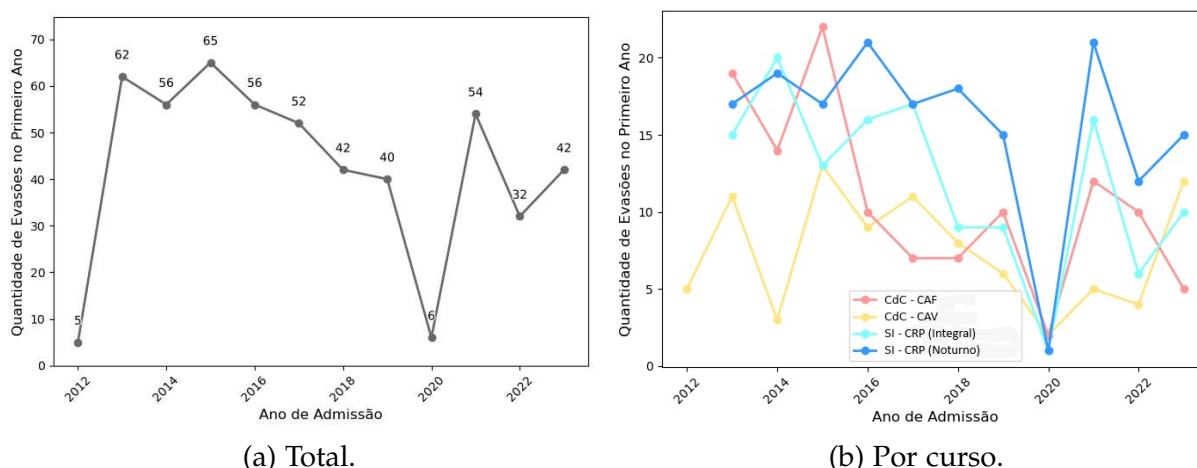


Figura 4.5: Evasão anual de calouros nos cursos de computação da UFV.

Nota-se, pela Figura 4.5a, que as taxas de evasão em 2012 aparecem relativamente baixas devido ao processo de preparação dos dados, que resultou na eliminação de registros de evasão daquele ano. No caso do curso do campus CAF, por exemplo, havia 16 evasões registradas nesse ano, mas todas foram descartadas por apresentarem atributos importantes preenchidos como nulos. Por consequência, em 2013 observa-se um aumento expressivo, seguido por relativa estabilidade até 2015, período em que a média foi de 61 evasões e a mediana de 62. Já entre 2016 e 2019 verifica-se uma redução nesse patamar, com média de 47,5 e mediana de 47, indicando uma queda consistente em relação ao triênio anterior. Em 2020, em função da pandemia da COVID-19, a universidade flexibilizou a obrigatoriedade de matrícula, o que pode ter gerado viés nas taxas de evasão. Parte das desistências pode ter sido registrada apenas em 2021, contribuindo para o pico observado nesse ano. Por fim, a partir de 2023, nota-se uma nova tendência de alta nas taxas.

A Figura 4.5b, por sua vez, evidencia que o curso de SI Noturno manteve taxas de evasão superiores aos demais cursos na maioria dos anos, ao passo que o curso de CdC do campus CAV apresentou, em geral, as menores. Observa-se ainda uma tendência de forte redução das taxas para o curso de CdC do campus CAF, a partir de 2015. Outro resultado interessante é que, em 2023, a tendência de alta observada no primeiro gráfico não é generalizada, uma vez que o curso de CdC do campus CAF, isoladamente, apresentou queda.

4.4.2 Resultado por Curso

Esta subseção apresenta os resultados dos modelos preditivos aplicados tanto à base de dados disponibilizada pela instituição quanto à base enriquecida com os atributos de acolhimento, com os resultados separados por curso, com o objetivo de responder à segunda questão de pesquisa (QP2 - Como os modelos preditivos de evasão se comportam quando atributos acadêmicos são combinados com atributos relacionados a estratégias de acolhimento?). As células em verde nas Tabelas 4.4, 4.5, 4.6 e 4.7 indicam uma melhora no desempenho da métrica após o enriquecimento da base, enquanto as células em vermelho uma piora. As demais células indicam que o desempenho se manteve. Já as células em destaque com retângulos representam os maiores valores obtidos em cada métrica.

A Tabela 4.4 apresenta o desempenho dos modelos de classificação aplicados especificamente ao curso de CdC do campus CAF.

Tabela 4.4: Desempenho dos modelos para o curso de CdC - campus CAF.

Algoritmo	Acurácia		Precisão		Revocação		Especificidade		F1-Score		AUC	
	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida
DT	0,86	0,87	0,88	0,88	0,52	0,56	0,98	0,98	0,65	0,68	0,90	0,88
RF	0,91	0,90	1,00	1,00	0,63	0,59	1,00	1,00	0,77	0,74	0,97	0,93
LR	0,92	0,93	0,95	1,00	0,70	0,70	0,99	1,00	0,81	0,83	0,97	0,99
NB	0,89	0,89	0,89	0,94	0,63	0,59	0,98	0,99	0,74	0,73	0,95	0,94
SVM	0,90	0,92	0,90	0,95	0,67	0,70	0,98	0,99	0,77	0,81	0,98	0,99

Observa-se que a base enriquecida contribuiu de forma geral para o aumento do desempenho dos modelos, sobretudo em métricas associadas ao equilíbrio entre acerto e erro. Destaca-se, em especial, a LR, que alcançou os maiores valores de revocação (0,70) e F1-Score (0,83) com a base enriquecida, superando os demais modelos. Outro ponto positivo foi o desempenho do SVM, que também obteve ganho consistente em todas as métricas, atingindo revocação de 0,70 e F1-Score de 0,81, valores próximos aos da LR. Esses dois modelos demonstram que a inclusão de atributos de acolhimento favoreceu algoritmos mais sensíveis a padrões lineares ou de margens bem definidas, o que se mostra particularmente relevante, já que a revocação é a métrica mais crítica no problema de predição de evasão, por indicar maior capacidade de identificar corretamente os estudantes que de fato evadiram.

Por outro lado, modelos como RF e NB apresentaram queda em revocação com a base enriquecida (de 0,63 para 0,59 em ambos). Isso sugere que, para esses algoritmos, os novos atributos podem ter introduzido ruído, dificultando a correta detecção de estudantes em risco. A DT, por sua vez, apesar de apresentar melhora em relação à base original, obteve os menores valores de revocação e F1-Score (0,56 e 0,68, respectivamente). Esses resultados explicam a redução no F1-Score observada em alguns casos, já que a perda de revocação compromete o equilíbrio com a precisão que, no geral, manteve-se bastante alta. Ainda assim, a inclusão dos atributos de acolhimento revelou-se valiosa, contribuindo para aprimorar o desempenho global dos modelos.

De modo geral, observa-se que a acurácia permanece elevada em todos os modelos, mas ela não é tão informativa neste cenário, já que o desafio maior está em prever os casos minoritários de evasão. Assim, a análise evidencia que a LR e o SVM se mostram mais promissores para o contexto do curso de CdC no campus CAF, conciliando boa precisão com níveis mais altos de revocação e F1-Score.

A Tabela 4.5, por sua vez, apresenta os resultados dos modelos aplicados ao curso de CdC do campus CAV.

Tabela 4.5: Desempenho dos modelos para o curso de CdC - campus CAV.

Algoritmo	Acurácia		Precisão		Revocação		Especificidade		F1-Score		AUC	
	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida
DT	0,85	0,84	0,53	0,52	0,89	0,83	0,84	0,84	0,67	0,64	0,91	0,89
RF	0,88	0,92	0,65	0,86	0,72	0,67	0,92	0,98	0,68	0,75	0,93	0,92
LR	0,88	0,90	0,71	0,79	0,56	0,61	0,95	0,97	0,62	0,69	0,95	0,95
NB	0,90	0,90	0,90	0,90	0,50	0,50	0,99	0,99	0,64	0,64	0,88	0,88
SVM	0,88	0,89	0,69	0,77	0,61	0,56	0,94	0,97	0,65	0,65	0,95	0,95

Os resultados apresentam um comportamento distinto em relação ao observado no campus CAF. Embora a base enriquecida tenha contribuído para ganhos em acurácia, precisão e especificidade em quase todos os modelos, os valores de revocação apresentaram melhora em apenas um caso.

A DT foi o modelo com maior destaque em revocação deste curso, alcançando 0,89 com a base disponibilizada pela instituição e 0,83 com a base enriquecida. Apesar da leve queda após o enriquecimento, continuou sendo o maior valor na métrica. Entretanto, sua precisão foi baixa e, conseqüentemente, o mesmo ocorreu para o F1-Score. Esse resultado indica que o modelo passou a identificar mais estudantes em risco de evasão, mas à custa de um número maior de falsos positivos, o que comprometeu o equilíbrio geral do desempenho.

O RF, por sua vez, alcançou precisão de 0,86 e F1-Score de 0,75 com a base enriquecida. Esses valores indicam maior equilíbrio entre acertos e erros, mostrando-se particularmente eficaz na identificação correta dos não evadidos (especificidade = 0,98). No entanto, esse ganho ocorreu às custas da revocação, que caiu de 0,72 para 0,67, reduzindo a sensibilidade do modelo em identificar os alunos

em risco de evasão. Para o SVM, observa-se um comportamento semelhante, mas com um F1-Score menor. A queda na revocação pode estar ligada à sensibilidade desses modelos ao balanceamento entre as classes, que os levou a priorizar a classe majoritária (não evadidos) e, assim, reduziram a métrica ao deixar de identificar parte dos alunos em risco.

A LR apresentou comportamento mais estável, com aumento de revocação (de 0,56 para 0,61) e crescimento no F1-Score (de 0,62 para 0,69). Esse resultado, embora não seja o mais alto, sugere que o modelo conseguiu se beneficiar dos novos atributos sem perder tanta sensibilidade, equilibrando razoavelmente bem precisão e revocação. Por outro lado, o NB manteve resultados praticamente inalterados, com revocação em 0,50 e F1-Score em 0,64, evidenciando dificuldade em capturar os padrões de evasão mesmo após o enriquecimento da base. Além disso, esse valor de revocação é o mais baixo entre todos os modelos do campus CAV.

De modo geral, os resultados do campus CAV evidenciam que a base enriquecida trouxe ganhos relevantes em acurácia e precisão, mas não garantiu avanços consistentes na revocação. O contraste entre os modelos reforça que há um *trade-off* importante entre aumentar a sensibilidade, como ocorreu na DT, e manter o equilíbrio entre acertos e erros, como na LR.

Já a Tabela 4.6 mostra o desempenho dos modelos de classificação para o curso de SI (Integral) do campus CRP.

Tabela 4.6: Desempenho dos modelos para o curso de SI (Integral) - campus CRP.

Algoritmo	Acurácia		Precisão		Revocação		Especificidade		F1-Score		AUC	
	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida
DT	0,93	0,93	0,96	0,96	0,79	0,79	0,99	0,99	0,87	0,87	0,97	0,97
RF	0,92	0,93	0,96	1,00	0,76	0,76	0,99	1,00	0,85	0,87	0,99	0,99
LR	0,93	0,91	0,96	0,96	0,79	0,74	0,99	0,99	0,87	0,83	0,99	0,98
NB	0,94	0,96	0,97	0,97	0,85	0,91	0,99	0,99	0,91	0,94	0,98	0,99
SVM	0,94	0,94	0,94	0,94	0,88	0,88	0,97	0,97	0,91	0,91	0,98	0,98

Observa-se um desempenho consistentemente elevado em todos os modelos, com métricas próximas ou superiores a 0,90 em quase todos os casos. Isso indica que, para esse curso, ambas as bases já fornecem informações suficientemente discriminativas para os algoritmos, o que reduz as diferenças entre eles.

Os modelos que mais se beneficiaram do enriquecimento foram o RF e o NB. O NB apresentou melhora em sua revocação, a qual foi de 0,85 para 0,91 e o F1-Score de 0,91 para 0,94, atingindo os maiores valores entre todos os algoritmos nesse curso. Esse resultado é relevante porque sugere que os novos atributos auxiliaram especialmente em um modelo probabilístico mais simples, aumentando sua capacidade de identificar corretamente os alunos evadidos sem perda de precisão.

O SVM manteve resultados muito altos e estáveis, com revocação de 0,88 e F1-Score de 0,91, destacando-se como outro modelo confiável. Já a LR apresentou pequena redução de desempenho após o enriquecimento da base (revocação de 0,79 para 0,74

e F1-Score de 0,87 para 0,83), o que pode indicar alguma sensibilidade ao acréscimo dos novos atributos.

De forma geral, o cenário do curso SI (integral) do campus CRP difere dos anteriores por mostrar desempenhos muito próximos entre os modelos, todos com boa capacidade de prever a evasão. Nesse contexto, a discussão se desloca menos para o ganho absoluto trazido pelo enriquecimento e mais para destacar a robustez dos modelos e o fato de que, mesmo com pequenas variações, todos alcançam níveis elevados de revocação e F1-Score. Um possível motivo para essa estabilidade é que o curso apresenta um número reduzido de estratégias de acolhimento, o que naturalmente limitou o impacto esperado sobre as métricas de desempenho.

Por fim, a Tabela 4.7 apresenta o desempenho dos modelos para o curso de SI (Noturno) do campus CRP.

Tabela 4.7: Desempenho dos modelos para o curso de SI (Noturno) - campus CRP.

Algoritmo	Acurácia		Precisão		Revocação		Especificidade		F1-Score		AUC	
	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida	Original	Enriquecida
DT	0,86	0,82	0,83	0,76	0,77	0,74	0,91	0,87	0,80	0,75	0,87	0,81
RF	0,84	0,84	0,76	0,79	0,82	0,77	0,86	0,88	0,79	0,78	0,92	0,93
LR	0,81	0,82	0,79	0,81	0,67	0,67	0,90	0,91	0,72	0,73	0,91	0,91
NB	0,81	0,81	0,76	0,76	0,72	0,72	0,87	0,87	0,74	0,74	0,91	0,91
SVM	0,84	0,85	0,82	0,87	0,72	0,69	0,91	0,94	0,77	0,77	0,91	0,92

Os resultados, em comparação com os demais cenários, apresentaram métricas em geral inferiores e com maior oscilação entre os modelos.

O modelo DT foi o que mais perdeu desempenho após o enriquecimento, com queda em todas as métricas. Isso indica que os novos atributos podem ter introduzido ruído ou aumentado a complexidade do balanceamento entre as classes de forma negativa para esse algoritmo.

Por outro lado, a LR apresentou leve ganho, aumentando a precisão e mantendo estável a revocação (0,67), o que resultou em pequena melhora no F1-Score (de 0,72 para 0,73), sendo um caso em que o enriquecimento trouxe benefício, ainda que limitado. Já o RF e o SVM apresentaram um cenário misto: embora tenham melhorado em precisão e especificidade, sofreram redução na revocação, comportamento semelhante ao observado em alguns modelos dos cursos anteriores.

De forma geral, nota-se que os modelos do curso noturno não apresentaram melhora nos níveis de revocação, comprometendo também os valores de F1-Score. Isso indica que, neste contexto, a base enriquecida não foi suficiente para aumentar a sensibilidade dos algoritmos, apontando para a necessidade de explorar futuramente atributos mais diretamente relacionados às condições específicas do curso noturno.

Em síntese, a inclusão dos atributos de acolhimento impactou o desempenho dos modelos de maneira heterogênea, mas com tendência positiva em vários cenários. Algoritmos como LR, NB e SVM, mais robustos a variáveis binárias, foram os que mais se beneficiaram, apresentando ganhos consistentes em diversas métricas.

Embora alguns modelos baseados em árvores, como o DT e RF, tenham mostrado algumas quedas em revocação e F1-Score em alguns casos, essas limitações se devem em parte à representação simplificada dos acolhimentos como variáveis binárias (0/1), que pode não capturar toda a complexidade do fenômeno. Ainda assim, os resultados indicam que tais atributos funcionam como variáveis complementares relevantes, capazes de enriquecer a previsão da evasão em determinados contextos, reforçando seu potencial de contribuição para análises preditivas mais precisas.

4.4.3 Análise da Importância dos Atributos

Para a identificação dos atributos mais importantes em cada modelo, foi utilizada a análise de importância de variáveis (*feature importance*), empregando metodologias específicas de acordo com a natureza de cada algoritmo, com o objetivo de responder à terceira questão de pesquisa (QP3 - Qual a importância dos atributos relacionados a estratégias de acolhimento nos modelos preditivos de evasão de calouros nos cursos de computação da UFV?).

Nos modelos baseados em árvores (DT e RF), a importância das *features* foi determinada com base na redução da impureza (Índice de Gini) nos nós das árvores, refletindo o quanto cada variável contribui para a divisão dos dados. Na LR e no SVM, a relevância das variáveis foi avaliada a partir do módulo dos coeficientes estimados, considerando que maiores coeficientes indicam maior impacto na previsão da classe-alvo. No NB, utilizou-se a medida de informação mútua (*mutual information*) entre cada *feature* e a variável-alvo, quantificando a dependência estatística entre elas. Em todos os casos, o cálculo da importância foi realizado exclusivamente sobre os dados de treinamento, e os atributos foram posteriormente ordenados de acordo com seu grau de relevância para o modelo.

Na análise, optou-se por apresentar os 10 atributos mais importantes da base enriquecida, a qual contempla 20 atributos no total. Para a contagem das frequências dos atributos, foram consideradas as execuções de cada um dos cinco modelos aplicados a cada um dos quatro cursos de computação da UFV, totalizando 20 execuções. É importante observar que todos os 20 atributos, incluindo as estratégias de acolhimento, estão presentes em todas as execuções dos modelos. Nos cursos em que determinada estratégia não é aplicada, sua coluna permanece zerada, mas ainda assim alguns algoritmos podem atribuir importância não nula a essas variáveis, seja por particularidades do cálculo de importância ou por correlações indiretas. Dessa forma, pode ocorrer de uma estratégia inexistente em um curso aparecer entre os dez atributos mais relevantes.

A Figura 4.6 apresenta a frequência com que cada atributo apareceu entre os dez mais importantes, separando entre atributos acadêmicos originais (4.6a) e atributos

de acolhimento (4.6b), o que permite visualizar esse comportamento. Como foram realizadas 20 execuções e em cada uma foram selecionados 10 atributos, os gráficos reúnem, ao todo, 200 registros de atributos.

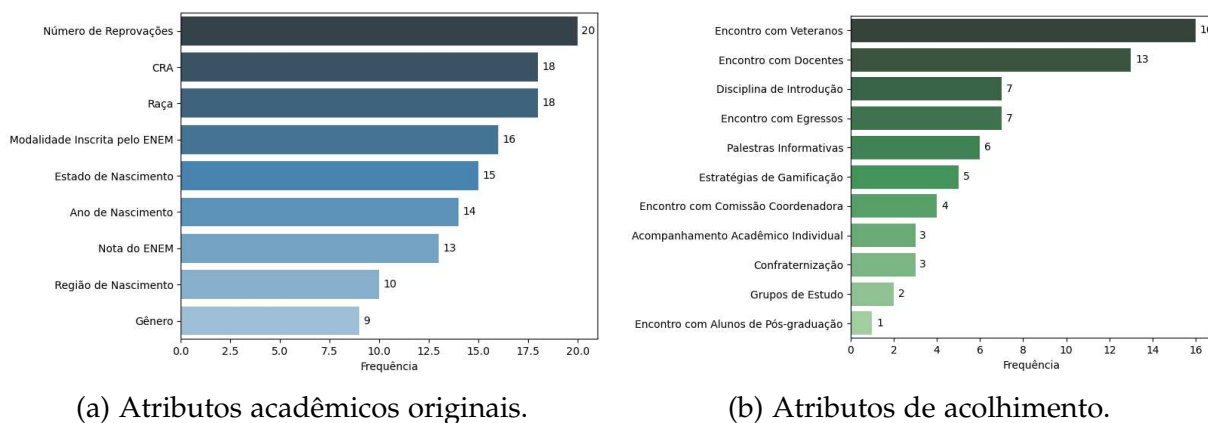


Figura 4.6: Frequência dos atributos entre os 10 mais importantes.

No primeiro gráfico, observa-se que o atributo **número de reprovações** foi o mais recorrente, aparecendo nas 20 execuções, seguido por **CRA** e **raça**, em 18, e por **modalidade inscrita pelo ENEM**, em 16. No caso dos estudantes que não evadiram, tanto o CRA quanto o número de reprovações correspondem aos valores acumulados ao longo de todo o curso, enquanto para os estudantes que evadiram esses atributos refletem apenas o desempenho registrado até o momento da saída. Esses resultados reforçam que características ligadas ao desempenho acadêmico e ao histórico escolar dos alunos desempenham papel central na análise da evasão.

O segundo gráfico, por sua vez, evidencia que as estratégias de **Encontro com Veteranos**, **Encontro com Docentes**, **Disciplina de Introdução** e **Encontro com Egressos** estiveram entre os atributos mais relevantes para a previsão nos modelos, destacando a importância das interações iniciais entre calouros e a comunidade acadêmica como fator potencial de redução da evasão.

Por fim, a Tabela 4.8 apresenta a distribuição dos 10 atributos mais relevantes identificados em cada um dos modelos gerados com a base de dados enriquecida. Diferentemente dos gráficos anteriores, os resultados estão agora organizados por curso e distribuídos conforme a posição de importância ocupada por cada atributo.

Ao analisar a Tabela, observa-se que, dentre os 200 atributos registrados, 67 correspondem a estratégias de acolhimento (células em azul), representando 33,5% do total. Ou seja, aproximadamente um terço dos atributos entre os 10 mais relevantes está relacionado a acolhimento.

Observa-se que o campus CAF apresenta a maior quantidade de atributos de acolhimento entre os 10 mais importantes, com 25 dos 67 possíveis (37,3%). Entre os mais recorrentes estão o **Encontro com Veteranos**, o **Encontro com Docentes** e a **Disciplina de Introdução**. O modelo que mais evidenciou a relevância dessas

Tabela 4.8: Posição dos dez atributos mais importantes por modelo e curso.

Atributo	CdC - campus CAF					CdC - campus CAV					SI (Int) - campus CRP					SI (Not) - campus CRP				
	DT	RF	LR	NB	SVM	DT	RF	LR	NB	SVM	DT	RF	LR	NB	SVM	DT	RF	LR	NB	SVM
Ano de Nascimento	3º	5º	-	9º	5º	4º	4º	-	-	4º	4º	4º	-	10º	3º	4º	4º	-	-	7º
Gênero	-	-	-	-	-	10º	8º	10º	-	-	10º	8º	3º	-	6º	-	8º	10º	-	-
Estado de Nascimento	-	-	4º	-	7º	6º	7º	-	10º	7º	7º	9º	6º	8º	7º	7º	6º	-	-	6º
Região de Nascimento	-	-	-	-	-	8º	9º	1º	9º	5º	-	-	8º	9º	5º	-	-	9º	10º	-
Modalidade Inscrita pelo ENEM	8º	6º	-	4º	-	5º	6º	3º	5º	3º	6º	6º	10º	-	7º	6º	6º	8º	-	5º
Nota do ENEM	4º	3º	-	-	6º	3º	3º	-	-	6º	3º	3º	-	8º	-	3º	3º	-	7º	4º
Número de Reprovações	2º	2º	3º	2º	2º	2º	2º	2º	2º	1º	2º	2º	1º	2º	2º	2º	2º	1º	2º	2º
Raça	6º	4º	7º	6º	-	7º	5º	9º	4º	-	5º	5º	2º	3º	4º	5º	5º	2º	3º	3º
Coefficiente de Rendimento Acumulado	1º	1º	-	1º	1º	1º	1º	-	1º	2º	1º	1º	7º	1º	1º	1º	1º	7º	1º	1º
Disciplina de Introdução	-	-	1º	10º	4º	9º	-	-	8º	-	-	-	-	-	-	10º	-	-	6º	-
Encontro com Comissão Coordenadora	-	8º	10º	3º	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8º	-
Encontro com Docentes	10º	9º	9º	7º	-	-	-	4º	-	-	9º	9º	4º	-	10º	-	10º	3º	9º	10º
Encontro com Veteranos	9º	7º	8º	8º	10º	-	-	6º	3º	8º	-	10º	5º	7º	9º	8º	9º	4º	-	9º
Encontro com Alunos de Pós-graduação	-	-	-	-	8º	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Encontro com Egressos	-	10º	-	5º	9º	-	-	7º	-	10º	-	-	-	4º	-	-	-	-	4º	-
Estratégias de Gamificação	7º	-	6º	-	-	-	-	-	6º	-	-	-	-	5º	-	9º	-	-	-	-
Grupos de Estudo	-	-	5º	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5º	-
Acompanhamento Acadêmico Individual	5º	-	2º	-	3º	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Palestras Informativas	-	-	-	-	-	-	10º	5º	-	-	8º	-	6º	-	-	-	-	5º	-	8º
Confraternização	-	-	-	-	-	-	-	8º	7º	9º	-	-	-	-	-	-	-	-	-	-

estratégias foi a LR, na qual sete dos 10 atributos mais importantes correspondem a estratégias de acolhimento. Destacam-se, ainda, a **Disciplina de Introdução** e o **Acompanhamento Acadêmico Individual**, que ocuparam, respectivamente, o primeiro e segundo lugares no ranking de importância do modelo.

No campus CAV, verificou-se que o **Encontro com Veteranos** e a **Confraternização** estiveram presentes em três dos cinco modelos avaliados. Os modelos com maior número de estratégias entre os 10 atributos mais relevantes foram a LR e o NB, cada um com quatro atributos de acolhimento. Neste campus, 14 dos 67 atributos de acolhimento (20,9%) figuraram entre os mais importantes.

Já no campus CRP, nos turnos integral e noturno, os atributos **Encontro com Docentes** e **Encontro com Veteranos** apareceram em oito dos 10 modelos. Os modelos da LR e NB destacaram-se pelo maior número de estratégias, sendo que, no turno noturno, metade dos 10 atributos mais importantes correspondia a atributos de acolhimento. O turno integral apresentou 12 dos 67 atributos (17,9%) como estratégias de acolhimento, enquanto o turno noturno contou com 16 (23,9%).

Esses resultados sugerem que iniciativas como apresentação do curso, interação com membros da comunidade acadêmica e atividades coletivas podem exercer influência positiva sobre o sentimento de pertencimento, contribuindo para a retenção e a redução da evasão de calouros.

Além das estratégias de acolhimento, outros atributos frequentemente identificados entre os mais relevantes, como visto também nos gráficos anteriores, estão relacionados ao desempenho acadêmico, como **CRA** e **Número de Reprovações**. Por refletirem o desempenho acumulado ao longo do curso, esses atributos são conhecidos na literatura como variáveis de resultado, ou pós-evento, já que só estão disponíveis após o aluno cursar ao menos um período. Em modelos

preditivos, seu uso exige cautela, pois envolvem informações que ainda não estariam acessíveis no momento da previsão. Por isso, são popularmente chamados de “profetas do acontecido”, pois antecipam comportamentos com base em dados que surgem apenas após eventos críticos, como reprovações ou queda no rendimento.

Apesar de apresentarem alto poder preditivo, essas variáveis devem ser complementadas com informações conhecidas desde o início da trajetória acadêmica, de modo a favorecer ações preventivas mais eficazes. Nesse sentido, as variáveis relacionadas ao acolhimento são identificáveis logo no início do curso, antes de qualquer desempenho concreto, reforçando seu valor como preditores preventivos e úteis para intervenções precoces na redução da evasão do primeiro ano.

Por fim, a análise das métricas apresentadas na subseção anterior evidencia uma relação consistente entre a utilização de atributos de acolhimento e o desempenho dos modelos. Nos casos em que os algoritmos atribuíram maior importância a essas estratégias, especialmente a LR e SVM, que destacaram múltiplos atributos de acolhimento entre os 10 mais relevantes, observou-se uma tendência de melhores resultados nas métricas de forma geral. Assim, os atributos de acolhimento surgem como variáveis preventivas que impactam diretamente a performance dos modelos.

4.5 Considerações Finais

Este trabalho apresentou resultados da investigação do impacto do uso de novos atributos relacionados às estratégias de acolhimento na evasão de calouros em cursos da área de computação da UFV por meio de uma abordagem preditiva com modelos de aprendizado de máquina.

De modo geral, os resultados evidenciaram diferenças significativas nas taxas anuais de evasão entre cursos e campi, com variações influenciadas por fatores como a pandemia. A inclusão dos novos atributos de acolhimento nos modelos preditivos mostrou impactos heterogêneos, porém majoritariamente positivos, especialmente em algoritmos mais sensíveis a variáveis binárias, como LR, NB e SVM, que apresentaram ganhos em métricas de desempenho como a revocação. Embora os modelos baseados em árvores tenham exibido oscilações em algumas métricas, os atributos de acolhimento se mostraram relevantes como variáveis complementares, capazes de enriquecer a previsão da evasão em determinados contextos.

Ainda que variáveis acadêmicas, como o CRA e o número de reprovações, tenham se destacado como mais preditivas, alguns atributos de acolhimento alcançaram posições de destaque, reforçando seu potencial como indicadores complementares e precoces.

Esses resultados apontam, para o contexto avaliado da UFV e para seus respectivos cursos da área de computação, a importância dos novos atributos relacionados ao

acolhimento nos modelos preditivos de evasão, no primeiro ano.

Dentre as limitações deste estudo, destaca-se o processo de preenchimento da base enriquecida: como não há listas de chamada ou registros individuais de participação nas estratégias de acolhimento, assumiu-se que todos os calouros ingressantes após a implementação de uma estratégia foram igualmente expostos a ela, o que pode não refletir a realidade. Além disso, algumas estratégias inexistentes em determinados cursos resultaram em colunas totalmente zeradas, o que pode ter influenciado as métricas de importância de atributos em alguns modelos.

Outra limitação relevante refere-se ao tratamento de valores nulos. Optou-se pela remoção dos registros incompletos, o que, embora tenha reduzido pouco a base final, pode ter descartado informações potencialmente úteis. Além disso, os modelos finais não ficaram balanceados devido à comparação dos resultados originais com os resultados obtidos pelas duas técnicas de balanceamento apresentadas.

Como trabalhos futuros, pretende-se ampliar o estudo para cursos de outras áreas, a fim de investigar como diferentes perfis de estratégias de acolhimento influenciam a evasão em contextos acadêmicos distintos. Além disso, planeja-se explorar métodos alternativos de tratamento de valores ausentes, utilizar outras técnicas de balanceamento de dados e avaliar as versões balanceadas dos modelos, estudar a revocação de forma mais profunda e analisar a robustez das predições sem atributos de desempenho acadêmico que possuem alto poder preditivo e dependem de eventos críticos.

Por fim, uma etapa promissora consiste em enriquecer ainda mais a base utilizando dados externos, como informações socioeconômicas e territoriais do IBGE, incluindo cidade de origem, distância ao campus e indicadores demográficos, o que pode aprimorar a capacidade preditiva e interpretativa dos modelos.

Capítulo 5

Conclusão Geral

Esta pesquisa buscou avaliar quantitativamente o impacto das estratégias de acolhimento adotadas nos cursos de computação da UFV na evasão do primeiro ano, por meio de uma abordagem que integrou três eixos complementares: (i) um mapeamento sistemático da literatura, (ii) um levantamento e análise preliminar das estratégias de acolhimento adotadas em diferentes campi da instituição, com ênfase na área de computação e (iii) a avaliação preditiva do impacto dessas estratégias por meio de modelos de aprendizado de máquina (*Machine Learning* - ML).

O primeiro estudo forneceu um panorama da produção científica relacionada à evasão, revelando a predominância de metodologias baseadas em análises exploratórias e modelos preditivos, sobretudo na área de computação. Além disso, destacou a relevância de atributos acadêmicos, institucionais, socioeconômicos e pessoais, sistematizando-os em quatro categorias que subsidiam futuras pesquisas.

O segundo estudo ampliou o olhar para a realidade institucional, mapeando as estratégias de acolhimento adotadas nos cursos de graduação dos três campi da UFV e realizando análises preliminares para os cursos da área de computação. Foi possível identificar diferenças no ritmo e na diversidade de adoção dessas práticas e, ainda, verificar indícios de que sua implementação progressiva pode estar associada à redução das taxas de evasão, especialmente quando contempladas estratégias ligadas ao engajamento e motivação estudantil. Apesar das limitações do estudo, os resultados sugerem que iniciativas institucionais bem estruturadas têm potencial de contribuir para a permanência dos calouros.

Por fim, o terceiro estudo explorou a capacidade preditiva das estratégias de acolhimento em modelos de ML. Embora variáveis acadêmicas tenham se mostrado mais robustas, alguns atributos relacionados às estratégias emergiram como preditores relevantes, sobretudo por estarem disponíveis já no ingresso do estudante. Esse resultado reforça a utilidade de tais atributos como indicadores complementares em sistemas de predição, possibilitando diagnósticos mais precoces e ações preventivas personalizadas, assim como demonstra que tais estratégias podem realmente impactar as taxas de evasão dos calouros.

De forma integrada, os três estudos contribuem para o entendimento de que a evasão é um fenômeno multifatorial, exigindo abordagens que combinem análise exploratória e modelos preditivos. Os resultados obtidos permitem responder diretamente à questão de pesquisa proposta: as estratégias de acolhimento

implementadas nos cursos de Computação da UFV contribuem de forma mensurável para a redução e predição da evasão no primeiro ano. Observou-se que cursos com maior diversidade de estratégias apresentam, em média, menores taxas de evasão, e que essas ações se destacam como variáveis relevantes em modelos de ML, estando entre os dez atributos mais importantes em um terço das 20 execuções realizadas.

Como perspectivas futuras, destaca-se a ampliação das análises para outros cursos e áreas do conhecimento, bem como a incorporação de abordagens causais que permitam avaliar, de forma mais robusta, a relação entre estratégias de acolhimento e evasão. Também se mostra promissora a inclusão de variáveis externas, como informações socioeconômicas e territoriais provenientes de bases públicas (por exemplo, IBGE), possibilitando avaliar efeitos associados à cidade de origem, distância até o campus e características demográficas dos estudantes.

Outra linha importante consiste em revisar o tratamento dado aos valores ausentes, explorando métodos alternativos ao descarte de registros, de forma a reduzir eventuais perdas de informação. Adicionalmente, recomenda-se investigar versões balanceadas dos modelos de ML utilizando outras técnicas de balanceamento para avaliar impactos sobre desempenho e estabilidade, bem como analisar a robustez das predições na ausência de atributos altamente preditivos, como *CRA* e *Número de Reprovações*.

Do ponto de vista metodológico, futuras pesquisas podem expandir o foco temporal da evasão: enquanto o presente trabalho se restringiu ao primeiro ano, investigações que considerem a evasão ao longo de todo o curso podem revelar padrões distintos de permanência e oferecer subsídios mais abrangentes para intervenções institucionais. Por fim, sugere-se examinar em maior profundidade as medidas de revocação (*recall*) dos modelos, de modo a compreender melhor sua capacidade de identificar corretamente estudantes em risco e orientar estratégias de intervenção mais eficazes.

Referências Bibliográficas

- Brito, C. R. and Ciampi, M. M. (2023). Engineering program “freshmen motivation week”. In *Proceedings of the Frontiers in Education Conference (FIE)*, pages 1–5, New York, NY, United States. IEEE.
- Carvalho, C. S., Mattos, J. C., and Aguiar, M. S. (2023). Avaliação da interpretabilidade de modelos por meio da clusterização de explicações no contexto da predição de evasão no ensino superior. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1191–1201, Porto Alegre, RS, Brasil. SBC.
- Carvalho, L., Santos, A., Nakamura, F., and Oliveira, E. (2019). Detecção precoce de evasão em cursos de graduação presencial em computação: um estudo preliminar. In *Anais do XXVII Workshop Sobre Educação Em Computação (WEI)*, pages 233–243, Porto Alegre, RS, Brasil. SBC.
- Cavalcanti, L., Freitas, N., Marques, E., Lima, S., Mello, R. F., Cristino, C. T., Alves, G., et al. (2025). Avaliação do impacto do gênero e da assistência estudantil sobre a sobrevivência dos estudantes em cursos de stem. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 731–745, Porto Alegre, RS, Brasil. SBC.
- Coimbra, C. L., Silva, L. B. e., and Costa, N. C. D. (2021). A evasão na educação superior: definições e trajetórias. *Educação e Pesquisa*, 47:e228764.
- Colpo, M. P., Primo, T. T., and de Aguiar, M. S. (2021). Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 873–884, Porto Alegre, RS, Brasil. SBC.
- Colpo, M. P., Primo, T. T., Pernas, A. M., and Cechinel, C. (2020). Mineração de dados educacionais na previsão de evasão: uma rsl sob a perspectiva do congresso brasileiro de informática na educação. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1102–1111, Porto Alegre, RS, Brasil. SBC.
- Costa, A. G., Mattos, J. C., Primo, T. T., Cechinel, C., and Muñoz, R. (2021). Model for prediction of student dropout in a computer science course. In *Proceedings of the 2021 XVI Latin American Conference on Learning Technologies (LACLO)*, pages 137–143, New York, NY, United States. IEEE.
- Couto, D. and Santana, A. (2017). Mineração de dados educacionais aplicada à identificação de variáveis associadas à evasão e retenção. In *Anais do II Congresso sobre Tecnologia na Educação (Ctrl+E)*, pages 333–344, Porto Alegre, RS, Brasil. SBC.

- da Cruz, R. C., Juliano, R. C., Monteiro Souza, F. C., and Correa Souza, A. C. (2023). A score approach to identify the risk of students dropout: an experiment with information systems course. In *Anais do XIX Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 120–127, Porto Alegre, RS, Brasil. SBC.
- da Silva, P. M., Lima, M. N., Soares, W. L., Silva, I. R., Fagundes, R. A. d. A., and De Souza, F. F. (2019a). Ensemble regression models applied to dropout in higher education. In *Proceedings of the 2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 120–125, New York, NY, United States. IEEE.
- da Silva, V. R. L. B., de Albuquerque Silva, F., and Burégio, V. (2019b). Characterizing educational data mining. In *Proceedings of the 2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–5, New York, NY, United States. IEEE.
- da Silveira, A. C. J. and Tonini, A. M. (2023). Evasão discente em cursos de engenharia de computação e o trabalho em TIC – tecnologia da informação e comunicação. *Revista Contexto & Educação*, 38(120):e10770.
- de Brito, B. C. P., de Mello, R. F. L., and Alves, G. (2020). Identificação de atributos relevantes na evasão no ensino superior público brasileiro. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1032–1041, Porto Alegre, RS, Brasil. SBC.
- de Carvalho, A. C. B. D. (2019). Semana de integração de calouros: uma prática de acolhimento. *Brazilian Journal of Development (BJD)*, 5(7):8811–8820.
- de Jesus, H. O., Rodriguez, L. C., and Junior, A. d. O. C. (2021). Predição de evasão escolar na licenciatura em computação. *Revista Brasileira de Informática na Educação (RBIE)*, 29:255–272.
- de Oliveira, M. M. and Barbosa, E. F. (2023). Multilevel modeling for the analysis and prediction of school dropout: a systematic review. In *Proceedings of the 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 103–112, New York, NY, United States. IEEE.
- do Carmo, Ê. P., Gasparini, I., and de Oliveira, E. H. T. (2022). Identificação de trajetórias de aprendizagem em um curso de graduação e sua relação com a evasão escolar. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 323–333, Porto Alegre, RS, Brasil. SBC.
- do Nascimento, F. F., de Oliveira Dantas, L. C., de Castro, A. F., and Queiroz, P. G. G. (2024). Técnicas de mineração de dados e aprendizado de máquina aplicados à

- evasão estudantil: um mapeamento sistemático da literatura. *Revista Brasileira de Informática na Educação (RBIE)*, 32:270–294.
- dos Santos, J. F., de Sousa, J. D. A., Mello, R. F., Cristino, C. T., and Alves, G. (2021a). Um modelo para análise do impacto da retenção e evasão no ensino superior utilizando cadeias de markov absorventes. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 813–823, Porto Alegre, RS, Brasil. SBC.
- dos Santos, V. H. B., Saraiva, D. V., and de Oliveira, C. T. (2021b). Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1196–1210, Porto Alegre, RS, Brasil. SBC.
- dos Santos Silva, D., Tamayo, S. C., Pessoa, M., Pires, F., Oliveira, D. B. F., de Oliveira, E. H. T., and de Carvalho, L. S. G. (2020). Minerando dados de um juiz on-line para prever a evasão de estudantes em disciplinas introdutórias de programação. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1343–1352, Porto Alegre, RS, Brasil. SBC.
- Duran, R., Bim, S. A., Gimenes, I., Ribeiro, L., and Correia, R. C. M. (2023). Potential factors for retention and intent to drop-out in brazilian computing programs. *ACM Transactions on Computing Education (TOCE)*, 23(3):1–33.
- Falcao, A. P., Villwock, R., and Miloca, S. A. (2023). Análise de dados pré-universidade para prever a evasão de alunos ingressantes em uma instituição de ensino superior. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1293–1304, Porto Alegre, RS, Brasil. SBC.
- Flores, V., Heras, S., and Julián, V. (2022). A new methodological framework for project design to analyse and prevent students from dropping out of higher education. *Electronics*, 11(18):2902.
- Fukao, A. T., Colanzi, T. E., Martimiano, L. A., and Feltrim, V. D. (2023). Estudo sobre evasão nos cursos de computação da universidade estadual de maringá. In *Anais do Simpósio Brasileiro de Educação em Computação (EduComp)*, pages 86–96, Porto Alegre, RS, Brasil. SBC.
- Géron, A. (2019). *Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow*. Alta Books.
- Graça, J. S. D. and Nascimento, E. F. V.-B. C. d. (2021). Acolhimento de calouros na universidade tiradentes: um olhar no projeto mentoria. In *Anais do XV Colóquio Internacional Educação e Contemporaneidade (EDUCON)*, São Cristóvão, SE, Brasil. Universidade Federal de Sergipe.

- Hoed, R. M., Ladeira, M., and Leite, L. L. (2018). Influence of algorithmic abstraction and mathematical knowledge on rates of dropout from computing degree courses. *Journal of the Brazilian Computer Society (JBACS)*, 24(1):1–16.
- Holanda, M., Castanho, C. D., Bandeira, I. N., and Da Silva, D. (2022a). Relato de experiência da monitoria da disciplina primeira linguagem de programação do departamento de ciência da computação da universidade de Brasília. In *Anais do XXX Workshop sobre Educação em Computação (WEI)*, pages 13–25, Porto Alegre, RS, Brasil. SBC.
- Holanda, M., Mandelli, M., Ishikawa, E., and da Silva, D. (2021). Um relato de experiência do acolhimento de novos alunos do departamento de ciência da computação da universidade de Brasília. In *Anais do XXIX Workshop sobre Educação em Computação (WEI)*, pages 151–160, Porto Alegre, RS, Brasil. SBC.
- Holanda, M., Mandelli, M., Ishikawa, E., and Da Silva, D. (2022b). Introdução ao ambiente universitário na computação: Uma disciplina para o acolhimento dos estudantes no departamento de ciência da computação da universidade de Brasília. *Revista Brasileira de Informática na Educação (RBIE)*, 30:494–518.
- Junior, G. B. V., Lima, B. N., Pereira, A. A., Rodrigues, M. F., Oliveira, J. R. L., Sílio, L. F., Carvalho, A. S., Ferreira, H. R., and Passos, R. P. (2022). Métricas utilizadas para avaliar a eficiência de classificadores em algoritmos inteligentes. *Revista CPAQV—Centro de Pesquisas Avançadas em Qualidade de Vida | Vol*, 14(2):2.
- Kantorski, G. Z., Martins, R. Z., Balejo, A., and Frick, M. (2023). Mineração de dados educacionais para predição da evasão em cursos de graduação presenciais no ensino superior. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1133–1142, Porto Alegre, RS, Brasil. SBC.
- Kitchenham, B., Charters, S., et al. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical report, Department of Computer Science, University of Durham.
- Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., and Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1909–1918, New York, NY, United States. ACM.
- Llauró, A., Fonseca, D., Villegas, E., Aláez, M., and Romero, S. (2021). Educational data mining application for improving the academic tutorial sessions, and the reduction of early dropout in undergraduate students. In *Ninth International*

- Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21)*, pages 212–218, New York, NY, United States. ACM.
- Marques, L. T., De Castro, A. F., Marques, B. T., Silva, J. C. P., and Queiroz, P. G. G. (2019). Mineração de dados auxiliando na descoberta das causas da evasão escolar: Um mapeamento sistemático da literatura. *Revista Novas Tecnologias na Educação (RENOTE)*, 17(3):194–203.
- Marques, L. T., Marques, B. T., Rocha, R. S., Chaves, L., de Castro, A. F., Queiroz, P. G. G., et al. (2020). Evasão acadêmica e suas causas em cursos de bacharelado em ciência da computação: Um estudo de caso na ufersa. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1042–1051, Porto Alegre, RS, Brasil. SBC.
- Menolli, A., Horita, F., Dias, J. J. L., and Coelho, R. (2020). Bi-based methodology for analyzing higher education: A case study of dropout phenomenon in information systems courses. In *Anais do XVI Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 1–8, Porto Alegre, RS, Brasil. SBC.
- Miles, M. B., Huberman, A. M., and Saldaña, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook*. Sage, Thousand Oaks, CA, United States, 3 edition.
- Mundim, P., Barbosa, D., Silva, G., and Silva, T. (2024). Impacto da gamificação e de grupos de apoio no ensino-aprendizagem de programação em um curso de ciência da computação. In *Anais do XXXII workshop sobre educação em computação (WEI)*, pages 419–430, Porto Alegre, RS, Brasil. SBC.
- Mundim, P., Silva, T., e Silva, G. B., and Barbosa, D. (2025). Evasão em cursos superiores na área de computação: Um mapeamento sistemático da literatura. In *Anais do XXXIII workshop sobre educação em computação (WEI)*, pages 1055–1067, Porto Alegre, RS, Brasil. SBC.
- Máximo, R. d. O., Carvalho, L., Lopes, J. E. F., and Ribeiro, K. C. d. S. (2024). Evasão e retenção no ensino superior: fatores socioeconômicos regionais e gestão de custos como determinantes da trajetória acadêmica. *Revista de Gestão e Secretariado (GESEC)*, 15(3):e3615.
- Nájera, A. B. U. and Ortega, L. A. M. (2022). Predictive model for taking decision to prevent university dropout. *IJIMAI*, 7(4):205–213.
- Oliveira, J. L., Ambrósio, A. P., Silva, U., Brancher, J., and Franco, J. J. (2020). Undergraduate students' effectiveness in an institution with high dropout index. In *Proceedings of the 2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–7, New York, NY, United States. IEEE.

- Oliveira, R. d. S. and Medeiros, F. P. A. d. (2024). Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação. *Revista Brasileira de Informática na Educação (RBIE)*, 32:1–21.
- Opazo, D., Moreno, S., Álvarez-Miranda, E., and Pereira, J. (2021). Analysis of first-year university student dropout through machine learning models: A comparison between universities. *Mathematics*, 9(20):2599.
- Pachas, D. A. G., Garcia-Zanabria, G., Cuadros-Vargas, A. J., Camara-Chavez, G., Poco, J., and Gomez-Nieto, E. (2021). A comparative study of who and when prediction approaches for early identification of university students at dropout risk. In *Proceedings of the XLVII Latin American Computing Conference (CLEI)*, pages 1–10, New York, NY, United States. IEEE.
- Pinto, D., Chagas, E., Costa, G., Nunes, K., Anchieta, R., Rocha, S., Paiva, A., and Rivero, L. (2020). De veteranos para os novatos: Avaliação das atividades de recepção aos calouros organizado pelo grupo petcomp. In *Anais do XXVIII workshop sobre educação em computação (WEI)*, pages 61–65, Porto Alegre, RS, Brasil. SBC.
- Ramos, J. L. C., Rodrigues, R. L., Silva, J. C. S., and de Oliveira, P. L. S. (2020). Crisp-edm: uma proposta de adaptação do modelo crisp-dm para mineração de dados educacionais. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1092–1101, Porto Alegre, RS, Brasil. SBC.
- Rios, I. C., Santos, C. D. V. d., Fernandes, E. M. d. O., Pacheco, M. K. O., Fernandes, M. T. d. A., and Vital, P. F. (2021). Mentoria de acolhimento para alunos ingressantes no curso de medicina. *Revista Brasileira de Educação Médica (RBEM)*, 45(suppl 1):e111.
- Rocha, R., Vanderlei, I., Araujo, J., et al. (2021). A model of actions to reduce evasion in higher technology courses: An experience report. In *Proceedings of the 2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6, New York, NY, United States. IEEE.
- Santos, C. H. D., de Lima Martins, S., and Plastino, A. (2021). É possível prever evasão com base apenas no desempenho acadêmico? In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 792–802, Porto Alegre, RS, Brasil. SBC.
- Santos, G., Belloze, K. T., Tarrataca, L., Haddad, D. B., Bordignon, A. L., and Brandão, D. N. (2020). Evolvedtree: Analyzing student dropout in universities. In *Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 173–178, New York, NY, United States. IEEE.

- Santos, K. J. d. O., Menezes, A. G., de Carvalho, A. B., and Montesco, C. A. (2019). Supervised learning in the context of educational data mining to avoid university students dropout. In *Proceedings of the 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, volume 2161, pages 207–208, New York, NY, United States. IEEE.
- Silva, E. C., Freitas, S., Soares Ramos, C., Muniz De Menezes, A. E., and Rodrigues De Araujo, L. K. S. (2023). A systematic review of the factors that impact the prediction of retention and dropout in higher education. In *Proceedings of the 2023 Hawaii International Conference on System Sciences (HICSS-56)*, pages 1271–1280, New York, NY, United States. IEEE.
- Silva, R. A. d. S., AF, B. B., Maria de Fátima, P. F., de Sousa Santos, I., and Andrade, R. M. (2021a). Evasão em computação na ufc sob a perspectiva dos alunos. In *Anais do XXIX Workshop sobre Educação em Computação (WEI)*, pages 338–347, Porto Alegre, RS, Brasil. SBC.
- Silva, T. R. d. M. B., Braga, G., and Henriques, M. T. A. (2021b). Mentoria em programação: aprendendo ao ensinar e ensinando ao aprender. In *Proceedings of the XV Women in Information Technology (WIT)*, pages 310–314, Porto Alegre, RS, Brasil. SBC.
- Souza, L. M., Mundim, P. C., Barbosa, D. M., Braga, G., and de MB Silva, T. R. (2025). Aplicação de uma metodologia de previsão de evasão: um estudo de caso no campus ufv-florestal. In *Anais do XXXIII Workshop sobre Educação em Computação (WEI)*, pages 290–301, Porto Alegre, RS, Brasil. SBC.
- Theobald, A. P. P., Moreira, A. F., Rodrigues, J. P., Brandão, K. F., de Moraes Costa, M. C., and de Oliveira Guimarães, R. S. (2020). A importância da recepção de calouros e o impacto na formação acadêmica. o caso de sucesso da escola de engenharia da ufmg: o engenharia recebe. In *Anais do Congresso Brasileiro de Educação em Engenharia (COBENGE)*, Belo Horizonte, MG, Brasil. UFMG.
- Vasconcelos, V. and Andrade, E. (2018). Análise da evasão de alunos na licenciatura em computação. In *Anais do XXVI Workshop sobre Educação em Computação (WEI)*, pages 28–35, Porto Alegre, RS, Brasil. SBC.
- Viana, F. S., Santana, A. M., and Rabêlo, R. d. A. L. (2022). Avaliação de classificadores para predição de evasão no ensino superior utilizando janela semestral. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 908–919, Porto Alegre, RS, Brasil. SBC.