

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

ANNA CARLA FREIRE LUNA CAMPÊLO BASTOS

**PROCESSO DECISÓRIO DE GESTÃO DA EVASÃO NOS CURSOS
DE GRADUAÇÃO DA UNIVERSIDADE FEDERAL DO SUL DA BAHIA:
CONTRIBUIÇÕES DA ABORDAGEM DA MINERAÇÃO DE DADOS
EDUCACIONAIS**

CURITIBA

2024

ANNA CARLA FREIRE LUNA CAMPÊLO BASTOS

**PROCESSO DECISÓRIO DE GESTÃO DA EVASÃO NOS CURSOS DE
GRADUAÇÃO DA UNIVERSIDADE FEDERAL DO SUL DA BAHIA:
CONTRIBUIÇÕES DA ABORDAGEM DA MINERAÇÃO DE DADOS
EDUCACIONAIS**

**Dropout Management Decision-Making Process in Undergraduate Courses at
the Federal University of Southern Bahia: Contributions of the Educational
Data Mining Approach**

Dissertação apresentada como requisito para
obtenção do título de Mestre em Administração
Pública do Mestrado em Administração Pública
em Rede Nacional da Universidade Tecnológica
Federal do Paraná.

Orientador: Prof. Dr. Oséias Santos de Oliveira
(UTFPR)

Coorientador: Prof. Dr. Ari Melo Mariano (UNB)

CURITIBA

2024



[4.0 Internacional](https://creativecommons.org/licenses/by-nc-sa/4.0/)

Esta licença permite remixe, adaptação e criação a partir do trabalho, para fins não comerciais, desde que sejam atribuídos créditos ao(s) autor(es) e que licenciem as novas criações sob termos idênticos. Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.



**Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Curitiba**



ANNA CARLA FREIRE LUNA CAMPELO BASTOS

**PROCESSO DECISÓRIO DE GESTÃO DA EVASÃO NOS CURSOS DE GRADUAÇÃO DA UNIVERSIDADE
FEDERAL DO SUL DA BAHIA: CONTRIBUIÇÕES DA ABORDAGEM DA MINERAÇÃO DE DADOS
EDUCACIONAIS**

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Administração Pública da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Administração Pública.

Data de aprovação: 20 de Novembro de 2024

Dr. Oseias Santos De Oliveira, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Ari Melo Mariano, Doutorado - Universidade de Brasília (Unb)

Dra. Hilda Alberton De Carvalho, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Itzhak David Simao Kaveski, Doutorado - Universidade Federal de Mato Grosso do Sul (Ufms)

Dra. Jane Correa Alves Mendonca, Doutorado - Fundação Universidade Federal da Grande Dourados (Ufgd)

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 20/11/2024.

Esta dissertação é dedicada ao meu marido e filhos:
meu alicerce e minha determinação vêm de vocês.
Àqueles que já partiram, pelo legado que deixaram em
mim. Àqueles que virão, para que sigam em frente.

AGRADECIMENTOS

Há de se começar pelo princípio e não haveria outra pessoa que não fosse minha mãe (*in memoriam*) por sempre me dizer que o caminho do conhecimento era fundamental e que eu poderia sempre ir além.

Agradeço à família que construí, ao meu marido e companheiro de toda a vida e para todo o sempre, pelo ombro amigo, forte e seguro, sempre acreditando que eu chegaria lá.

Aos meus filhos, pela compreensão nas ausências, pelos carinhos nas noites intermináveis, pelo colo que me deram e pelos incontáveis olhares de admiração. Tudo foi visto e sentido intensamente.

Aos amigos, poucos, seletos e insubstituíveis, vocês extraem o melhor de mim e eu sou melhor por vocês.

Aos colegas de trabalho, aos colegas pesquisadores, à UTFPR e à UFSB pela oportunidade de realizar um sonho.

Aos colegas de mestrado, por dividirem as pedras do caminho, mas também as manhãs, tardes e noites, vocês me doaram dois preciosos anos de suas companhias e carregarei cada um de vocês na memória.

Ao meu orientador, pelo altruísmo e leveza com que compartilhou todo este conhecimento, saio desta experiência melhor que quando comecei.

Ao meu coorientador pelo ensinamento imensurável, sua generosidade não teve limites e sua mão firme e amiga me conduziram pelo caminho do desconhecido. E que jornada trilhamos! Que o destino nos reserve novas descobertas.

À espiritualidade amiga, que nunca desampara.

Nenhuma sociedade que esquece a arte de questionar pode esperar encontrar respostas para os problemas que a afligem.

(ZYGMENT BAUMAN)

RESUMO

BASTOS, Anna Carla Freire Luna Campêlo. **Processo Decisório de Gestão da Evasão nos Cursos de Graduação da Universidade Federal do Sul da Bahia: Contribuições da Abordagem da Mineração de Dados Educacionais**. 2024. 123 páginas. Dissertação (Mestrado Profissional em Administração Pública) - Universidade Tecnológica Federal do Paraná. Curitiba, 2024

A Educação é fundamental para a Administração Pública democrática, pois promove equidade, desenvolvimento econômico e inclusão social. No entanto, apesar dos avanços rumo à universalização, desafios complexos persistem, e a evasão se destaca como um grande obstáculo, demandando Políticas Públicas eficientes. Esta pesquisa tem como objetivo apresentar estratégias específicas para a gestão da evasão acadêmica nos cursos de graduação da Universidade Federal do Sul da Bahia (UFSB) a partir da utilização da abordagem de Mineração de Dados Educacionais, entregando um Relatório de Consultoria, com vistas ao enfrentamento desta problemática e a sugestão de práticas para melhoria da gestão da evasão existente na instituição. O método de pesquisa se pauta por uma abordagem quantitativa, de natureza aplicada e aliada à pesquisa exploratória explicativa. A partir de um estudo bibliométrico, composto pelo referencial teórico e revisão bibliográfica com aplicação da Teoria do Enfoque Meta-Analítico Consolidado (TEMAC) associada a uma pesquisa documental foram investigando dados secundários, quantitativos, com análise de Regressão Logística e Árvore de Decisão aplicadas ao estudo de caso. A amostra inicial é composta por dados secundários, com 11.626 linhas relacionadas aos cursos de graduação e 502.403 relacionadas as disciplinas, dos estudantes de graduação da Universidade Federal do Sul da Bahia quando 23 variáveis foram nomeadas. A amostra foi preparada observando a curva de normalidade, identificando dados faltantes e os outliers. Os resultados apontam a Regressão Logística realizada por curso, sendo destacadas como as principais variáveis o CR (β 0,362) e a Permanência (β 0,139) e por disciplinas, a relevância que recaiu sobre a Média Final (β 1,759) e a Carga Horária Total (β 0,007). A pesquisa foi aprofundada com a modelagem por Árvore de Decisão, apresentando acurácia de 86,7% e revelando que o fator mais relevante para prever a evasão na UFSB é o tempo de Permanência, seguido de Coeficiente de Rendimento (CR), Idade, Campus e Distância da residência em relação à universidade. Os resultados evidenciam tópicos que não foram encontrados em outras literaturas científicas sobre o tema da evasão. O propósito do trabalho foi completamente alcançado com elaboração do Produto Técnico-Tecnológicos, delimitado em um *Dashboard* do Observatório da Evasão apresentado sob o formato de um Relatório de Consultoria contendo propostas de intervenções e a sugestão de desenvolvimento de aplicativo.

Palavras-Chave: Gestão da Educação. Ensino Superior. Evasão. Mineração de Dados Educacionais.

ABSTRACT

BASTOS, Anna Carla Freire Luna Campêlo. **Dropout Management Decision-Making Process in Undergraduate Courses at the Federal University of Southern Bahia: Contributions of the educational Data Mining Approach.** 2024. 123 pages. Dissertation (Professional Master's in Public Administration) - Federal Technological University of Paraná. Curitiba, 2024.

Education is essential for democratic Public Administration, as it promotes equity, economic development, and social inclusion. However, despite advances towards universalization, complex challenges persist, and dropout is a major obstacle, demanding efficient Public Policies. This research aims to propose specific strategies for the management of academic dropout in undergraduate courses at the Federal University of Southern Bahia (UFSB) based on the use of the Educational Data Mining approach, delivering a Consulting Report, to address this problem and suggest practices to improve the management of dropout in the institution. The research method is based on a quantitative approach, of an applied nature and combined with explanatory exploratory research. Based on a bibliometric study, composed of the theoretical framework and bibliographic review with the application of the Theory of Consolidated Meta-Analytic Approach (TEMAC) associated with documentary research, secondary, quantitative data were investigated, with Logistic Regression and Decision Tree analysis applied to the case study. The initial sample is composed of secondary data, with 11,626 lines related to undergraduate courses and 502,403 related to disciplines, from undergraduate students at the Federal University of Southern Bahia when 23 variables were named. The sample was prepared by observing the normality curve and identifying missing data and outliers. The results indicate the Logistic Regression performed by course, highlighting as the main variables the CR (β 0.362) and Permanence (β 0.139) and by disciplines, the relevance that fell on the Final Average (β 1.759) and Total Workload (β 0.007). The research was deepened with Decision Tree modeling, presenting an accuracy of 86.7% and revealing that the most relevant factor in predicting dropout at UFSB is the length of Permanence, followed by Performance Coefficient (CR), Age, Campus, and Distance from residence to the university. The results highlight topics not found in other scientific literature of evasion. The purpose of the work was fully achieved with the preparation of the Technical-Technological Product, delimited in a Dashboard of the Evasion Observatory presented in the format of a Consulting Report containing intervention proposals and suggestions for application development.

Keywords: Dropout, University Education, Education Management, Educational Data Mining.

LISTA DE SIGLAS

ANDIFES – Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior

BI – *Business Intelligence*, Inteligência de Negócios

BN – *Bayesians Networks*

CC – Componente Curricular

CEP – Comitê de Ética em Pesquisa

CH – Carga Horária

CJA – *campus* Jorge Amado

CNN – *Convolutional Neural Networks*

COGRAD – Colégio de Pró-Reitores de Graduação das Instituições Federais de Educação Superior

CPA – Comissão Própria de Avaliação

CPF – *campus* Paulo Freire

CR – Coeficiente de Rendimento

CRISP-DM - *Cross Industry Standard Process for Data Mining*, Processo Padrão entre Indústrias para Mineração de Dados

CSC – *campus* Sosígenes Costa

DAG – *Directed Acyclic Graph*, Grafo Direcionado Acíclico

DPA – Diretoria de Percursos Acadêmicos

EDM – *Educational Data Mining*, Mineração de Dados Educacionais

FICA – Ferramentas de Identificação e Combate ao Abandono

FIS – *Fuzzy Inference System*

FORGRAD – Fórum de Pró-Reitores de Graduação

FORPLAD - Fórum Nacional de Pró-Reitores de Planejamento e de Administração das Instituições Federais de Ensino Superior

IBGE – Instituto Brasileiro de Geografia e Estatística

ICERI – *International Conference of Education, Research and Innovation*

IES – Instituição de Educação Superior

INEP – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

IQR – Intervalo Interquartil

KDD – *Knowledge Discovery from Data*

K-NN – *K-Nearest Neighbors*

LDB - Lei de Diretrizes e Bases da Educação Nacional

OMS – Organização Mundial da Saúde

PNAES – Plano Nacional de Assistência Estudantil

PROAF – Pró-Reitoria de Ações Afirmativas

PROFIAP – Mestrado Profissional em Administração Pública

PROGEAC – Pró-Reitoria de Gestão Acadêmica

PROPLAN – Pró-Reitoria de Planejamento

ProUni – Programa Universidade para Todos

PTT – Produto Técnico-Tecnológico

RB – Redes Bayesianas

SEI – Superintendência de Estudos Econômicos e Sociais da Bahia

SISU – Sistema de Seleção Unificada

SVM – *Support Vector Machines*

TDIC – Tecnologias Digitais de Informação e Comunicação

TEMAC – Teoria do Enfoque Meta-Analítico Consolidado

UFFS – Universidade Federal da Fronteira Sul

UFSB – Universidade Federal do Sul da Bahia

UTFPR - Universidade Tecnológica Federal do Paraná

VN – Verdadeiro Negativo

VP – Verdadeiro Positivo

WOS – *Web of Science*

LISTA DE FIGURAS

Figura 1: Modelo CRISP-DM.....	43
Figura 2: Mapa de coocorrência	47
Figura 3: Publicações por ano indexadas em <i>Web of Science</i>	48
Figura 4: Autores que mais publicaram	50
Figura 5: Autores mais citados	51
Figura 6: Países que mais publicaram.....	52
Figura 7: Agências financiadoras.....	53
Figura 8: Áreas de pesquisa	53
Figura 9: Revistas que mais publicaram	54
Figura 10: Universidades que mais publicam	55
Figura 11: Palavras mais frequentes nos títulos dos artigos.....	56
Figura 12: Mapa de cocitação.....	57
Figura 13: Mapa de <i>coupling</i>	58
Figura 14: Modelo integrador.....	64
Figura 15: Taxonomia da pesquisa.....	67
Figura 16: Mapa Mesorregião Sul Baiano	70
Figura 17: Estudantes da UFSB	71
Figura 18: Plataforma digital <i>dashboard</i>	79
Figura 19: Limpeza de dados da planilha de Cursos	81
Figura 20: Índice de desempenho acadêmico global – CR.....	82
Figura 21: Limpeza de dados da planilha de Disciplinas	83
Figura 22: Boxplot de Médias	84
Figura 23: Regressão Logística por Cursos.....	89
Figura 24: Regressão Logística por Disciplinas	92
Figura 25: Tratamento de dados	95
Figura 26: Tratamento de dados Blocos 01 a 05	96
Figura 27: Tratamento de dados Blocos 06 a 09	98
Figura 28: Árvore de Decisão	99
Figura 29: Árvore de Decisão identificada	100
Figura 30: Raíz da Árvore de Decisão, percurso I a IV.....	101

LISTA DE FIGURAS

Figura 31: Primeiro Caminho, percurso V a X	102
Figura 32: Primeiro Caminho, percurso XI a XX.....	103
Figura 33: Segundo Caminho, percurso XXI a XXII.....	104
Figura 34: Segundo Caminho, percurso XXIII a XXVIII	105
Figura 35: Segundo Caminho, percurso XXIX a XXXIX.....	106
Figura 36: Segundo Caminho, percurso XL a XLIV	107
Figura 37: <i>Dashboard</i> da Evasão	113
Figura 38: Evasão na UFSB	115
Figura 39: Aplicativo de Acompanhamento Acadêmico.....	116

LISTA DE QUADROS

Quadro 1: Estudos brasileiros sobre evasão	36
Quadro 2: Etapas do CRISP-DM.....	43
Quadro 3: Mapa de busca	46
Quadro 4: Artigos mais citados.....	49
Quadro 5: Artigos selecionados.....	60
Quadro 6: Desenho da pesquisa	69
Quadro 7: Dados dos alunos por Curso	77
Quadro 8: Dados dos alunos por Disciplina.....	78
Quadro 9: Atribuição de novas etiquetas.....	81
Quadro 10: Enriquecimento dos dados.....	85
Quadro 11: Recomendações de uso do Produto Técnico Tecnológico	117

LISTA DE TABELAS

Tabela 1: Cálculo de Kappa	59
Tabela 2: Descrição da amostra	73
Tabela 3: Regressão Logística por Cursos.....	88
Tabela 4: Regressão Logística por Disciplina.....	91
Tabela 5: Matriz de Confusão.....	108

SUMÁRIO

1	INTRODUÇÃO	16
1.1	Problema de Pesquisa	18
1.2	Objetivo Geral.....	18
1.3	Objetivos Específicos	18
1.4	Justificativa.....	19
1.5	Estrutura da Dissertação	21
2	REFERENCIAL TEÓRICO	23
2.1	A Administração Pública e as Políticas Educacionais	23
2.2	A Problemática da Evasão: Um Desafio Global, Brasileiro e da UFSB	27
2.3	As Variáveis Críticas no Estudo da Evasão	33
2.4	A Gestão da Evasão e a Inteligência de Dados.....	36
3	EVASÃO NO ENSINO SUPERIOR: ESTADO DA ARTE	45
3.1	Preparação da Pesquisa	45
3.2	Apresentação e Interrelação de Dados.....	46
3.3	Detalhamento, Modelo Integrador e Validação por Evidências	56
4	MÉTODOLOGIA DA PESQUISA	66
4.1	Tipo de Pesquisa	66
4.2	Desenho da Pesquisa	68
4.3	Local da Pesquisa	70
4.4	Objeto da Pesquisa	72
4.5	Coleta e Tratamento de Dados	72
5	ANÁLISE SITUACIONAL	75
5.1	Compreensão do Negócio	75
5.2	Compreensão dos Dados	77
5.3	Preparação dos Dados.....	80
5.4	Modelagem.....	86
5.4.1	Análise por Regressão Logística.....	87
5.4.2	Análise por Árvore de Decisão.....	94
5.4.2.1	Análise dos Nós e Folhas da Árvore de Decisão	94
5.5	Avaliação.....	108
5.5.1	Matriz de Confusão.....	108
5.6	Implementação	110

5.6.1	Implicações do Modelo Proposto e Apresentação do <i>Dashboard</i>	110
6	CONSIDERAÇÕES FINAIS	120
	REFERÊNCIAS	124

1 INTRODUÇÃO

Os princípios fundamentais da Administração Pública envolvem o planejamento, a organização, a direção e o controle de recursos, visando o alcance de metas previamente estabelecidas (Motta, 2001 e 2003). Neste contexto, é fundamental o reconhecimento de que as atividades da Administração Pública são conduzidas pelo Estado com o objetivo de atender ao interesse público (Di Pietro, 2010). Estes conceitos formam a base para o estudo e a compreensão da administração no setor público, com ênfase sobre o compromisso com o bem-estar coletivo e a eficiência na gestão dos recursos públicos.

A Administração Pública brasileira, atualmente, apresenta características predominantemente gerenciais, uma abordagem que busca aplicar princípios e práticas da gestão privada ao setor público, visando maior eficiência, eficácia e responsabilidade na prestação de serviços públicos. Um dos pilares fundamentais da Administração Pública democrática é a educação e, para atender à formação cidadã, é necessário que ela seja ampla, acessível e de qualidade. Sob esta perspectiva, muito se avançou localmente, seja guiado pelas normativas construídas ou pelas políticas públicas implementadas.

O desenvolvimento de políticas públicas educacionais de qualidade é de suma importância para a promoção da equidade, desenvolvimento econômico e inclusão social e estas devem ser regularmente reavaliadas considerando sua elaboração, estabelecimento, implementação e acompanhamento, sendo importante avaliar o histórico destas políticas para projetar o futuro considerando as necessidades emergentes do público a quem ela se destina (Valente, 2020). Estas políticas visam, notadamente, a ampliação do acesso à educação em todos os níveis, para todo os cidadãos.

Apesar da regulação e do financiamento serem formas amplamente utilizadas como instrumento de controle pela Administração Pública, outras ferramentas podem ser utilizadas para garantir que instituições públicas (universidades inclusas) se mantenham adequadas às métricas de desempenho e missões estabelecidas, adicionando ao campo as Teorias Organizacionais (Misra, 2024), construindo interesse social para conquistar apoio e legitimidade. O controle instrumental e a

perspectiva da legitimidade institucional podem ser combinados em arranjos de implementação de políticas mais abrangentes (Misra, 2024).

Ainda, na métrica da gestão gerencial, fundamentada em características como descentralização, competição e incentivo (Dunleavy, 2006), foram também introduzidos índices de avaliação para mensurar o avanço na qualidade do ensino, tanto na perspectiva dos educandos quanto dos educadores e das instituições de ensino. Neste sentido, a Governança Colaborativa traz formadores de opinião (*stakeholders*) públicos e privados para discutir em fóruns com as agências públicas, no intuito de formar decisões consensuais (Ansell, 2008).

Apesar dos avanços constatados no desenvolvimento das políticas educacionais brasileiras, os desafios ainda são enormes, profundos e complexos. Uma vez que se ampliou significativamente o acesso à educação, é primordial que sejam direcionados esforços à permanência e conclusão das etapas do ensino, o que ainda é um problema significativo.

Altas taxas de evasão e repetência, especialmente no Ensino Fundamental e no Ensino Médio, indicam problemas na gestão educacional e na adequação dos métodos de ensino às necessidades dos alunos. No Ensino Superior, o abandono deixa posto que a garantia de acesso não é suficiente para que o estudante conclua esta etapa, impactando a instituição e os investimentos feitos para a permanência deste estudante, a sociedade em geral e o indivíduo (Santos, 2020).

Muitos estudos têm tentado explicar a evasão no Ensino Superior, podendo ser citada a pesquisa realizada por Santos (2019), na Universidade Federal de Sergipe, com dados de 23.690 estudantes do Departamento de Ciências da Computação a partir de Árvore de Decisão e Floresta Randômica onde se constatou que os maiores índices de evasão ocorrem no segundo e terceiro ano de curso. Por outro lado, Preto e Gutierrez-Pachas (2023), encontraram que alunos que possuem maior índice de aprovação, evadem menos, assim como programas de ensino de base, como a abordagem STEM que reflete a adaptação das escolas às novas demandas globais, que exigem uma preparação mais integrada nas áreas de Ciência, Tecnologia, Engenharia e Matemática e que contribuem para a diminuição do índice de evasão. Os autores também forneceram indícios de que as notas são decisivas no processo de evasão, assim como as horas de ausência em um curso.

A Universidade Federal do Sul da Bahia também enfrenta a evasão em seus cursos de graduação, com um índice de 33,42% de evasão (UFSB, 2023). Os

controles realizados pela Comissão Própria de Avaliação (CPA) e pela Diretoria de Percursos Acadêmicos (DPA) servem de base de dados para a tomada de decisão da gestão e são periodicamente divulgados para livre acesso pela Pró-Reitoria de Gestão Acadêmica (PROGEAC) e Pró-Reitoria de Planejamento (PROPLAN), no Painel Valor Público, no site oficial da instituição (UFSB 2023).

1.1 Problema de Pesquisa

A evasão acadêmica é um desafio persistente no contexto do Ensino Superior brasileiro, afetando negativamente a formação dos estudantes e o uso eficiente dos recursos educacionais. Sendo assim, a questão de pesquisa que orienta este estudo é assim apresentada: quais os fatores relevantes para o enfrentamento da evasão nos cursos de graduação da Universidade Federal do Sul da Bahia (UFSB), potencializando os resultados destas ações para a qualificação da gestão pública educacional?

1.2 Objetivo Geral

O objetivo geral deste trabalho consiste em apresentar estratégias específicas para a gestão da evasão acadêmica nos cursos de graduação da Universidade Federal do Sul da Bahia (UFSB) a partir da utilização da abordagem de Mineração de Dados Educacionais, entregando um Relatório de Consultoria, com vistas ao enfrentamento desta problemática e a sugestão de práticas para melhoria da gestão da evasão existente na instituição.

1.3 Objetivos Específicos

Para que o objetivo geral desta pesquisa seja alcançado propõem-se os seguintes objetivos específicos:

1. Situar a problemática da evasão no Ensino Superior público, contemplando os macros cenários da Educação, Políticas Públicas e Gestão da Educação, e também da Administração Pública.

2. Identificar as variáveis que mais impactam na evasão no Ensino Superior contidas na literatura científica de alto impacto;
3. Propor um modelo para explicar a evasão dos cursos de graduação da Universidade Federal do Sul da Bahia, utilizando as técnicas de Mineração de Dados Educacionais;
4. Validar o modelo proposto.

1.4 Justificativa

A Educação sempre foi uma prioridade em minha formação familiar. Filha única, tive muitas oportunidades de estudo, sempre acompanhada de grandes cobranças por bons desempenhos. Aprendi, em especial com minha mãe, que a educação era o caminho da independência e da autossuficiência num contexto em que nossa família morava distante das oportunidades educacionais. Ainda adolescente, problemas de saúde me levaram à algumas restrições físicas e foi, mais uma vez, nos estudos que canalizei as minhas energias.

A proposta apresentada para esta pesquisa, no âmbito do Mestrado Profissional em Administração Pública (PROFIAP) executado em parceria entre a Universidade Tecnológica Federal do Paraná (UTFPR) e a Universidade Federal do Sul da Bahia (UFSB), pode suscitar contribuições significativas ao estudo da gestão pública, especialmente no que se refere ao enfrentamento da problemática da evasão existente no Ensino Superior, buscando, para isto, o entendimento do campo de estudo e a identificação de técnicas e ferramentas que possam auxiliar no processo decisório, aplicando-as em dados reais extraídos da Instituição de Ensino Superior (IES) pesquisada – a Universidade Federal do Sul da Bahia, uma universidade criada recentemente, inserida em um contexto socioeconômico marcado pela diversidade cultural e inclusão, numa região historicamente carente de instituições públicas de Ensino Superior.

A análise de dados resultante deste estudo fornece subsídios para o desenvolvimento de práticas e estratégias específicas, alinhadas às características da UFSB, visando aprimorar a gestão das ações de enfrentamento à evasão, o que inclui a apresentação de um Produto Técnico-Tecnológico (PTT) em formato de Relatório

de Consultoria apresentando uma proposta de ações de maior impacto para a gestão da evasão nos cursos de graduação da UFSB e ajustes nas políticas institucionais.

A garantia do acesso ao conhecimento é um princípio humanitário e norteador das ações educacionais. Muito já se caminhou na questão do acesso ao Ensino Superior brasileiro com a aquisição de políticas públicas como o Sistema de Seleção Unificada (SISU) e o Programa Universidade para Todos (ProUni) e muito há ainda por fazer relativo à permanência e conclusão dos cursos escolhidos, mas esta não é a única dificuldade.

Associada às questões acadêmicas, a gestão da educação pública também pode ser afetada pelas restrições orçamentárias, de infraestrutura e modernização, contratação e qualificação dos profissionais envolvidos na gestão acadêmica e nos processos de ensino, além das particularidades inerentes a cenários e públicos tão distintos quanto encontrados nas diferentes regiões do país.

Assim, tratar a evasão acadêmica é olhar para o indivíduo, aquele que carrega sonhos e expectativas, perspectivas de uma melhora na condição de vida própria e familiar e para a construção social mais igualitária. Isto requer, também um olhar sobre a eficiência do sistema de gestão, sobre a aplicação eficaz de recursos, sobre a análise dos impactos socioeconômicos de políticas públicas e, sobretudo, sobre a capacidade de planejamento e implementação destas políticas.

A modernização que se vê necessária e urgente não diz respeito apenas à sala de aula, aos recursos didáticos ou ao formato de ensino (presencial, à distância ou híbrido). Este avanço precisa alcançar também a gestão da educação com pesquisa e inovação. Por este motivo a utilização da Mineração de Dados Educacionais se insere como ferramenta moderna, robusta e imparcial, de extrema relevância para o auxílio nos processos decisórios de ampla repercussão.

Assim, esta proposta busca uma análise da evasão acadêmica na UFSB, ao mesmo tempo em que a apresentação de contribuições práticas e tangíveis para a gestão educacional, visando aprimorar a qualidade do Ensino Superior e o fortalecimento da efetividade das políticas institucionais. A abordagem de Mineração de Dados Educacionais é evidenciada, nesta pesquisa, como uma estratégia muito significativa para enfrentar esta problemática de maneira orientada e embasada em evidências, uma vez que a análise das múltiplas variáveis e dos impactos da evasão acadêmica pode ser um passo importante em direção aos resultados mais eficientes que se almeja na Administração Pública, contribuindo para o desenvolvimento do país.

Outrossim, é oportuno situar a relevância deste estudo justificado pela escassez de pesquisas sobre evasão acadêmica no contexto brasileiro, especificamente na região Sul da Bahia. Para além da produção limitada sobre o tema em nível nacional este estudo revela que não foram encontradas análises específicas realizadas na Bahia ou na sua região sul, o que reforça a necessidade de uma investigação que aborde esta problemática localmente. Diante disto, este trabalho pode favorecer o preenchimento desta lacuna e contribuir para o desenvolvimento de estratégias eficazes que possam ser aplicadas ao contexto regional da Universidade Federal do Sul da Bahia (UFSB), oferecendo um diagnóstico mais preciso e soluções adaptadas à realidade da instituição.

1.5 Estrutura da Dissertação

Esta dissertação está estruturada a partir de quatro capítulos. O primeiro capítulo é introdutório, sendo composto por uma apresentação do tema, o problema de pesquisa, o objetivo geral e os objetivos específicos do trabalho, além da justificativa da pesquisa e estrutura da dissertação.

O segundo capítulo traz o referencial teórico, subdividido nas temáticas: administração pública e as políticas públicas de educação, os panoramas da evasão, as variáveis críticas no estudo da evasão e a gestão da evasão na perspectiva da Mineração de Dados. Estes conceitos são debatidos à luz das produções de autores que se dedicam às temáticas, conforme estudos destacados em períodos mais recentes.

O terceiro capítulo insere o Estado da Arte, e oferece uma visão abrangente e atualizada do que já foi estudado e publicado sobre evasão no Ensino Superior, fornecendo uma base sólida para a análise apresentada nesta pesquisa, identificando lacunas na literatura e estabelecendo o contexto teórico e conceitual para a investigação que é conduzida utilizando-se a Teoria do Enfoque Meta-Analítico Consolidado (TEMAC) por Mariano e Rocha (2017).

O quarto capítulo tem seu enfoque na metodologia e contém a descrição da pesquisa a realizada: um estudo de caso baseado em análise quantitativa dos dados acadêmicos da evasão nos cursos de graduação da UFSB, sob a ótica da utilização da EDM (*Educational Data Mining*, ou Mineração de Dados Educacionais) para propor

ações pontuais de alto impacto na gestão da evasão. Para isto o capítulo engloba a contextualização da instituição analisada, os elementos que são estudados, a descrição das fontes de dados utilizadas e como estes foram coletados. Este capítulo aborda ainda a técnica de Mineração de Dados Educacionais empregada e a justificativa para esta escolha em relação à análise da evasão acadêmica, procedimentos de coleta e preparação de dados, identificação de padrões e geração de resultados, apresentando uma discussão sobre a exploração dos dados e sua interpretação à luz das técnicas de EDM, apontando as métricas e indicadores que são utilizados nesta avaliação.

O quinto capítulo aborda os resultados apurados das análises de Regressão Logística e Árvore de Decisão e sua relação com a literatura levantada nos capítulos dois e três de acordo com o modelo do CRISP-DM. Com base neste modelo adotado, foram desenvolvidos seis tópicos, contemplando todas as etapas do método: Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelagem (englobando a análise por regressão Logística e por Árvore de Decisão), Avaliação (Matriz de Confusão) e a Implementação, onde discute-se as implicações do modelo proposto. Neste ponto do capítulo, é descrita a construção do Produto Técnico-Tecnológico, um *dashboard* intitulado Observatório da Evasão que será entregue à Universidade Federal do Sul da Bahia e recomendação de desenvolvimento futuro de um aplicativo para *smartphone* baseado nos resultados da Regressão Logística e da Árvore de Decisão. Este capítulo se encerra com as considerações finais, limitações e futuras linhas de pesquisa, trazendo os principais achados e o fechamento do trabalho.

2 REFERENCIAL TEÓRICO

O presente capítulo insere uma análise abrangente sobre a interseção entre a Administração Pública e as políticas educacionais, oferecendo uma perspectiva contextualizada para compreender os desafios enfrentados na gestão da evasão acadêmica.

Inicialmente, são exploradas as dinâmicas da Administração Pública, com foco nas políticas educacionais que permeiam o cenário global e nacional, delineando o papel crítico da gestão pública na condução efetiva da educação.

A discussão se estende aos panoramas da evasão acadêmica em níveis mundial, nacional (considerando a realidade brasileira) e local, com especial atenção para a Universidade Federal do Sul da Bahia (UFSB). Posteriormente, são identificadas as variáveis críticas no estudo da evasão, investigando fatores que influenciam diretamente a permanência dos estudantes nos cursos de graduação apontados pela literatura.

Por fim, o capítulo aborda a gestão da evasão sob a perspectiva inovadora da mineração de dados, destacando como esta abordagem pode revelar percepções valiosas para o aprimoramento das estratégias de retenção e promoção do sucesso acadêmico, oferecendo uma base teórica abrangente para a compreensão e investigação do fenômeno da evasão em diversos contextos educacionais.

2.1 A Administração Pública e as Políticas Educacionais

A Administração Pública brasileira passou por uma série de reformas e contrarreformas, numa sucessão de disputas entre a burocracia e suas características centralizadoras, corporativas e de Estado máximo e novas formas de gestão, descentralizadas, ágeis, empreendedoras e eficientes (Alves, 2021). A era do conhecimento e da informação, a partir de 1950, e suas grandes mudanças tecnológicas, sociais e econômicas, também exigiram das organizações (públicas e privadas) transformações significativas em suas formas de atuar e mensurar seu desempenho (Guimarães, 2000).

A Administração Pública passou por uma significativa reforma com a implementação do Estado Novo, fundamentado nos princípios do mérito e da impessoalidade, em contraposição às antigas práticas patrimonialistas (Farah, 2011). Com isto, as organizações públicas adotam padrões de gestão até então utilizados no ambiente privado e passam a buscar eficiência e qualidade na prestação dos serviços públicos (Guimarães, 2000).

Tanto o tema das políticas públicas quanto dos processos decisórios começam a ser explorados dentro do campo da Administração a partir dos anos de 1950, ainda voltados para os estudos da ciência política e diferenciados da Administração Pública. Esta aproximação, com os campos influenciando-se mutuamente, ocorre por volta do fim dos anos 1960, adicionando às competências do administrador público a elaboração das políticas públicas (Farah, 2011).

Hoje, a Administração Pública volta-se à descentralização, maior flexibilidade na forma de desempenhar as atividades e conseqüente menor grau de padronização, responsabilização (*accountability*) com a manutenção de núcleos estratégicos menores e grupos operacionais maiores, gestões por contrato e maiores níveis de autonomia, com métricas pré-estabelecidas de resultados. Ao passo em que as organizações dependem de profissionais capacitados para conferirem a flexibilidade e a velocidade decisória demandada, estes profissionais precisam satisfazer um novo conjunto de habilidades exigidas, buscando das próprias organizações a capacitação necessária (Tamada, 2022).

Na Educação Pública há também a necessidade de qualificação, o que exige um olhar sobre questões como condições adversas à saúde dos professores, incorporação abrupta das Tecnologias Digitais de Informação e Comunicação (TDIC) ao campo educacional durante a pandemia da Covid-19, singularidades da educação inclusiva, efeitos das políticas educacionais e precarização do trabalho docente têm sido temas de trabalhos acadêmicos no Brasil, conforme evidencia Assis (2023) em pesquisa bibliográfica analisando 286 registros extraídos da *Web of Science*, no período de 2008 a 2023.

A Educação, no Brasil, é regulamentada por uma série de leis e diretrizes que visam orientar a organização, o financiamento e a qualidade do sistema educacional. A Constituição Federal (BRASIL, 1988), em seu Artigo 205, define a educação como um direito de todos e um dever do Estado e da família, estabelecendo como princípios e objetivos fundamentais a igualdade de condições para o acesso e permanência na

escola, a liberdade de aprender, ensinar, pesquisar e divulgar o pensamento, a arte e o saber, o pluralismo de ideias e de concepções pedagógicas, a gestão democrática do ensino público, a garantia de padrão de qualidade, a valorização dos profissionais da educação e a gestão democrática do ensino público.

Assim, a educação tem como um de seus principais regramentos a Lei de Diretrizes e Bases da Educação Nacional (LDB), Lei nº 9.394/96 (BRASIL, 1996), que define os princípios e as normas para a organização da educação, a estruturação dos currículos, a formação de professores, o financiamento da educação, entre outros aspectos. Quanto à filosofia/teoria da educação que fundamenta estas diretrizes, o sistema educacional brasileiro é influenciado por diversas correntes filosóficas e teorias da educação, incluindo abordagens como o construtivismo, o humanismo e o socio construtivismo.

A pedagogia brasileira também é influenciada por diversos teóricos da educação, com destaque a Anísio Teixeira (1900-1971) que com seu entusiasmo pela Escola Nova e com uma abordagem moderna e progressista, centrada no aluno, na universalidade e na gratuidade da educação, influenciou as políticas educacionais que muito contribuiu para o desenvolvimento da educação pública no Brasil (Alves, 2021).

Ainda, no rol dos pensadores cabe destacar o importante legado de Paulo Freire (1921-1997), que defendeu uma educação libertadora e crítica, pautada em uma educação popular. Tanto Anísio Teixeira quanto Paulo Freire deixaram um legado duradouro nas políticas educacionais brasileiras, influenciando a forma como a educação é concebida e implementada no país, uma vez que suas ideias continuam a inspirar debates e práticas educacionais até os dias atuais.

As políticas educacionais atuais, sob a perspectiva apresentada por Libâneo (2012), trazem três pontos críticos que merecem reflexão: a concepção restrita de aprendizagem, a importância do papel do ensino e a necessidade de uma escola inclusiva e igualitária.

Em relação à concepção restrita de aprendizagem, as políticas educacionais adotadas nas últimas décadas seguiram uma abordagem neoliberal que reduziu a aprendizagem a um mero instrumento com o propósito da aquisição de competências básicas para a sobrevivência social. Esta visão limitada desconsidera o aspecto cultural e cognitivo da aprendizagem, desvalorizando o papel do ensino na promoção do desenvolvimento pleno dos indivíduos. Ao restringir a aprendizagem a noções

mínimas, as escolas são enfraquecidas, professores são mal preparados e desmotivados, o que prejudica o processo educacional como um todo.

A partir da teoria histórico-cultural, baseada nas contribuições de Vygotsky (2008) e seus seguidores, pode ser enfatizado o papel da escola em proporcionar aos alunos a apropriação da cultura e da ciência acumuladas historicamente. Segundo esta visão, a escola é uma importante instância de democratização social e inclusão, sendo sua tarefa básica promover a atividade de aprendizagem dos alunos.

Assim, surge a necessidade de uma instituição educacional que promova a igualdade e a inclusão, reconhecendo a diversidade como uma condição essencial para a integração de todos os alunos, mas que só pode ser alcançada se a diversidade cultural for acolhida.

Motta (2003) traz um breve recorte do cenário escolar, onde a autogestão pedagógica se diferencia da autogestão institucional, enfatizando que a autogestão social passa, necessariamente, pelas questões de educação politécnica e polivalente (aprendizagem não autoritária).

Esta diversidade de pensamentos e avaliações mostra a complexidade do tema Gestão Pública da Educação numa Administração Pública de dimensões continentais e perfis sócio, econômico e culturais tão diversos, como é o caso do Brasil.

Para traçar um panorama da realidade do Ensino Superior brasileiro este estudo pauta-se nos dados disponíveis no Portal do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2021) sobre a educação brasileira e a evolução da educação superior pública, em especial.

Utilizando como marco temporal inicial o ano de 2010, o cenário educacional apresentava 278 Instituições de Educação Superior (IES) públicas, das quais 99 eram federais, 108 estaduais e 71 municipais. Estas instituições englobavam um universo de 1.643.298 estudantes, dos quais 938.656 pertencem à rede federal. Em 2020 existiam 304 IES, das quais 118 são federais, 129 estaduais e 57 municipais, abraçando 1.956.352 estudantes, dos quais 1.254.080 são discentes federais (INEP, 2021).

Assim, mantendo a perspectiva da dimensão da diversidade encontrada num campo tão vasto quanto o da Educação, tanto na perspectiva humana quanto na gerencial, aplicado a regiões com características sociais, econômicas e culturais tão distintas, associada às mais diversas teorias de aprendizado e de gestão pública, entende-se que é necessário mais que representatividade ou inclusão,

descentralização ou centralização, inovação ou conservadorismo. É preciso buscar um método decisório que contemple um amplo espectro de nuances de maneira objetiva e otimizada, atendendo ao maior público possível.

A gestão pública da Educação Superior é um desafio complexo – no que tange lidar com as problemáticas existentes - e de suma importância para o desenvolvimento educacional e social de um país. No contexto brasileiro, a administração e aprimoramento das instituições de Ensino Superior públicas têm sido objeto de constante debate e pesquisa, à medida que se busca aprimorar o acesso e a qualidade do ensino.

Os principais problemas da educação pública superior são multifacetados e incluem questões como a falta de recursos financeiros, a infraestrutura precária, a carência de políticas educacionais eficazes, utilização das TDIC (Valente, 2020) e a alta taxa de evasão, que impacta negativamente a conclusão dos cursos.

Um dos maiores problemas enfrentados na educação em todos os seus níveis (fundamental, médio e superior) é, sem dúvidas, a evasão. Muito se avançou nas políticas públicas de acesso à educação, ampliando o número de instituições de ensino, de vagas disponibilizadas, garantindo o acesso de minorias por meio das políticas de quotas, mas permanecer no ambiente educacional ainda é um desafio de muitos e a conclusão uma etapa de ensino, ainda hoje, pode ser difícil.

2.2 A Problemática da Evasão: Um Desafio Global, Brasileiro e da UFSB

A evasão é um problema que afeta as instituições de ensino em forma e intensidade diversas em todo o mundo, ao passo em que emerge como uma problemática que transcende fronteiras, ao mesmo tempo em que apresenta nuances singulares no contexto brasileiro.

Os países que mais têm se dedicado à publicação de estudos sobre evasão são a Espanha, Itália e Alemanha. Mesmo em um cenário de países desenvolvidos, a evasão ainda é um problema de impacto social e financeiro relevantes (Aina, 2022, Rodriguez-Gomes, 2012, Contini, 2018, Vignoles, 2009). Os estudos sobre estas temáticas têm se multiplicado e mantido a tendência de alta, retratando a relevância das discussões realizadas, conforme dados extraídos da plataforma de pesquisa *Web*

of Science, mantendo uma produção anual entre 21 e 26 artigos nos últimos 05 anos (2019 a 2023).

Há décadas a taxa de evasão tem sido quantificada e utilizada como indicador de performance, mas o tema também traz como cenário secundário as questões dos custos da educação superior, do uso ineficiente destes recursos e das expectativas (alcançadas ou não) dos estudantes (Rodriguez-Gomez, 2012). É preciso considerar, ainda, que a elevada taxa de evasão de estudantes em Instituições de Ensino Superior não apenas compromete os esforços em direção à democratização do conhecimento, mas também desafia a capacidade dos sistemas educacionais em preparar cidadãos qualificados para os desafios contemporâneos.

No cenário internacional, a evasão acadêmica representa um fenômeno multifacetado. Diferenças econômicas, culturais e estruturais influenciam diretamente nas razões que levam os estudantes a abandonarem seus cursos. A pressão financeira, a falta de preparo acadêmico, a ausência de suporte emocional e a desconexão entre as expectativas do Ensino Superior e a realidade do mercado de trabalho são alguns dos fatores recorrentes que permeiam este panorama global (Aina, 2022; Rodriguez-Gomez, 2012, Contini, 2018; Vignoles, 2009; Behr, 2020; Montmarquette, 2001).

A evasão é um fenômeno complexo, que se encontra difundido em todo o globo e um apanhado geral não demonstraria as nuances entre instituições, mesmo dentro do mesmo país. Mesmo o suporte da literatura socioeconômica não esclarece se o abandono representa uma perda de tempo e de investimentos financeiros ou seria uma etapa inevitável para alguns estudantes (Aina, 2022).

A Espanha tem uma política de acesso mais equitativa em relação ao perfil econômico do ingressante (Rodriguez-Gomez, 2012). Estudos espanhóis apontam que índices de evasão podem variar amplamente em relação à escolha do curso, devido ao tipo de disciplinas, cultura de ensino-aprendizagem, proporção de estudantes, apoio estudantil e pela satisfação, motivação e qualidade acadêmica, mas não apontavam o perfil sociodemográfico como determinante, sendo que 50% da evasão ocorre no primeiro ano. A pouca ou nenhuma conexão do curso com o mercado de trabalho também é uma variável relevante (Rodriguez-Gomez, 2012).

Ainda no contexto europeu, ficou comprovado que estudantes menos favorecidos economicamente tinham tanto menores chances de matrícula no Ensino Superior quanto maiores possibilidades de evadir no decorrer do curso ou terminá-lo

fora do prazo padrão. Estas desvantagens acumulativas mostram-se bastante significativas, variando em função das experiências escolares anteriores (Contini, 2018). Este pensamento corrobora com Vignoles (2009), para quem o perfil socioeconômico é o grande diferencial nos fatores de evasão.

Em complemento ao cenário europeu, na Alemanha apenas 50% dos estudantes de origem na classe trabalhadora acessam à Educação Superior e as variáveis acadêmicas do Ensino Médio aparecem como críticas, assim como fatores subjetivos tais como satisfação e autoavaliação (Behr, 2020).

No Canadá há um ano de preparação entre o ensino médio e o universitário que se propõe a apresentar aos estudantes tanto as possibilidades de mercado quanto o universo acadêmico, onde se considera também as questões envolvidas tanto no acesso ao ensino superior quanto aos recursos que deixam de ser recebidos durante o período acadêmico (Montmarquette, 2001).

Em estudo realizado na África do Sul, restou evidenciado que muito se avançou no acesso ao Ensino Superior, mas a alta evasão e a baixa taxa de conclusão da graduação ainda permanecem como problemas graves, onde o país figurou, em 2005, dentre os que apresentam menor índice de graduação, com taxa de apenas 15% (Sommer, 2011).

No contexto latino-americano um estudo realizado no Chile proporciona uma visão complementar ao panorama inicialmente apresentado, trazendo a perspectiva de um país em desenvolvimento, relatando a necessidade de estudos que considerem estas diferentes contextualizações e apontando consequências, tais como altas perdas de recursos estatais e institucionais, menores oportunidades para os estudantes e suas famílias, diminuição da produtividade do país (Palácios, 2021).

No âmbito brasileiro, a evasão acadêmica se manifesta de maneira peculiar, amplificando desafios já existentes. O acesso ampliado ao Ensino Superior nas últimas décadas trouxe consigo a necessidade de lidar com a diversidade de perfis acadêmicos e socioeconômicos dos estudantes. As desigualdades regionais, as carências estruturais nas instituições de ensino e as dificuldades de integração no ambiente universitário também se somam às razões que impulsionam a saída prematura de muitos alunos.

No conjunto de mais de dois milhões de discentes das 304 IES brasileiras, é complexo apontar um índice médio nacional para a evasão no Ensino Superior quando

o INEP fornece dados que variam em uma mesma instituição, de acordo com o curso analisado, de 1,6% a 87,5% (INEP, 2022).

Para Santos (2022) não há consenso sobre o conceito nem quanto ao cálculo da evasão uma vez que as variáveis críticas mapeadas, em geral configuradas em fatores socioeconômicos, acadêmicos e individuais, não promovem os resultados esperados para as políticas públicas de enfrentamento da evasão. O autor suscita que a permanência dos estudantes depende de condições efetivas que vão além do alcance da gestão da educação, tal como melhor condição financeira familiar.

É certo que num problema de alta complexidade nem todos os fatores estarão sob controle de uma das partes envolvidas, assim como é limitante indicar um fator externo como determinante de uma questão multivariada. Apontam em outra direção os estudos de Nierotka (2023), desenvolvidos na Universidade Federal da Fronteira Sul (UFFS), cujo estudo de caso realizado com dados de 1.391 estudantes ingressantes revela que o fator socioeconômico não é uma variável crítica na predição da evasão.

A comparação de diferentes estudos dentro de cenários nacionais permite uma confirmação inicial da necessidade de englobar diferentes fatores na construção de um processo de evasão como uma forma mais eficiente para esta predição que o apontamento de uma causa fundamental para o evento.

As ações brasileiras sobre a evasão contam com discussões em ambientes de gestão da educação como o Fórum de Pró-Reitores de Graduação (Forgrad) e o Colégio de Pró-Reitores de Graduação de Graduação das Instituições Federais de Educação Superior (Cograd) e agem por políticas públicas de iniciativa do Estado como o Plano Nacional de Assistência Estudantil (BRASIL, 2010) criada pela Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior (Andifes) e o Programa Bolsa Permanência (BRASIL, 2013).

Urge, portanto, uma abordagem integrada e proativa para lidar com esta problemática. Investimentos em políticas de permanência estudantil, aprimoramento do aconselhamento acadêmico, avaliações constantes dos métodos de ensino e a criação de programas de orientação profissional são algumas das medidas que podem contribuir significativamente para reverter o quadro de evasão.

Além disto, é imperativo que se estabeleça uma comunicação efetiva entre as Instituições de Ensino Superior e o mercado de trabalho, a fim de alinhar as

expectativas dos acadêmicos com as demandas da sociedade. A promoção de estágios, parcerias com empresas e a valorização de competências transversais são estratégias que podem não apenas manter os universitários engajados, mas também prepará-los para uma transição mais suave para a vida profissional.

Em suma, os panoramas da evasão acadêmica, tanto no mundo quanto no Brasil, exigem uma análise profunda e a implementação de soluções abrangentes. Somente a partir de um comprometimento conjunto entre instituições de ensino, governos e sociedade civil será possível construir um sistema educacional superior mais inclusivo, equitativo e capaz de preparar os futuros profissionais para os desafios dos tempos atuais.

Após traçar um breve panorama da problemática da evasão no contexto global e brasileiro, o foco desta discussão volta-se ao cenário da Universidade Federal do Sul da Bahia (UFSB), por ser este o lócus onde a presente pesquisa se estrutura.

A Universidade Federal do Sul da Bahia (UFSB), lócus onde esta pesquisa é desenvolvida, é uma das IES mais novas do país, criada em 2013, com uma personalidade marcada pela inclusão social representada em sua altíssima adesão a políticas de quotas, também integra este cenário, enfrentando os inúmeros desafios relacionados à gestão da educação de nível superior, sendo um dos principais a evasão de estudantes em seus cursos de graduação. O desafio da evasão na UFSB, portanto, está intrinsecamente ligado à sua missão de proporcionar uma educação de qualidade, acessível e alinhada às demandas da sociedade contemporânea.

No contexto da UFSB, a evasão se torna ainda mais premente, afetando a trajetória acadêmica de muitos estudantes sócio e economicamente vulneráveis e, conseqüentemente, o cumprimento de sua missão institucional e que tem impactado significativamente o ambiente acadêmico e o percurso dos discentes.

A evasão, no contexto da UFSB (2023), é caracterizada pela saída prematura de discentes antes de concluírem seus cursos de graduação. A UFSB, desde sua criação, tem enfrentado desafios significativos relacionados à evasão acadêmica, uma questão que não pode ser negligenciada dada sua repercussão nos objetivos educacionais e na trajetória dos acadêmicos. Diversos fatores convergem para a complexidade desta problemática nesta IES pública, demandando uma análise crítica e ações assertivas. Este fenômeno pode ser influenciado por uma variedade de

fatores, incluindo questões socioeconômicas, falta de preparo acadêmico, dificuldades de adaptação à vida universitária, problemas de saúde mental e, em alguns casos, a incompatibilidade entre as expectativas dos estudantes e a realidade dos cursos oferecidos.

Primeiramente, é inegável que as desigualdades socioeconômicas e regionais, tão presentes no cenário brasileiro, exercem uma influência acentuada na dinâmica da evasão na UFSB. A diversidade de perfis acadêmicos dos estudantes, provenientes de diferentes realidades e contextos educacionais, coloca desafios extras para a instituição na adequação de suas práticas pedagógicas e na oferta de suporte necessário para a permanência destes acadêmicos.

Na UFSB também estão presentes desafios estruturais, como a falta de recursos financeiros, a infraestrutura insuficiente e a carência de políticas educacionais efetivas para lidar com a evasão. Ainda, a falta de programas estruturados de orientação acadêmica e profissional pode ser percebida como outra lacuna que contribui para a evasão na UFSB.

A ausência de um suporte efetivo na escolha de carreiras, no acompanhamento psicopedagógico e na identificação precoce de possíveis dificuldades acadêmicas deixa os acadêmicos expostos a maiores vulnerabilidades, aumentando a probabilidade de desistência. Os cursos de graduação da UFSB abrangem uma variedade de áreas, desde ciências humanas até ciências exatas e da saúde, o que torna o enfrentamento da evasão uma tarefa complexa que demanda estratégias específicas para cada campo de estudo.

Os índices de evasão também variam amplamente em função do curso e período analisado, o que torna um desafio delicado a análise dos dados e também a proposição de ações para gerenciá-la (INEP, 2023).

A evasão não apenas impacta a vida dos estudantes, mas também pode comprometer a eficiência do sistema educacional e a missão da UFSB de promover o desenvolvimento social, econômico e cultural da região, gerando desperdício de recursos públicos e comprometendo a qualidade do ensino (Santos, 2020).

Como medida para gestão da evasão a UFSB implantou algumas ações, como o Programa de Monitorias, e o Programa de Tutorias em implantação em 2023. Programas assim são utilizados em muitas universidades brasileiras e estrangeiras, com o objetivo final de monitorar, reduzir e prevenir a evasão universitária e o atraso na conclusão dos cursos, a exemplo da Itália, onde o programa de tutorias é

estabelecido em lei desde 1990, oferecendo serviços de apoio e orientação aos alunos (Da Re, 2011).

Existem muitas variáveis que foram mapeadas na literatura científica e cada abordagem dedica-se a um grupo delas, variando entre dados socioeconômicos, índices de desempenho acadêmico, fatores culturais e perspectivas pessoais (Castro-Lopez, 2022; Aina, 2013 e 2020; Contini, 2018; Vilorio, 2019). Estas variáveis possuem diferentes graus de influência a depender do local, curso, país, conforme discussão efetivada no próximo tópico

2.3 As Variáveis Críticas no Estudo da Evasão

Apesar das revisões bibliográficas apresentarem, com frequência, as variáveis acadêmicas, demográficas e financeiras como pontos críticos para a evasão, não há um consenso definitivo sobre quais são as mais relevantes para um estudo sobre o tema. Há apenas uma concordância quanto à multiplicidade de fatores envolvidos na questão.

Em Montmarquette (2001) as variáveis consideradas como determinantes para a evasão são as habilidades e interesses do estudante e, como variável ambiental, é considerado o número de alunos nos cursos obrigatórios de primeiro ano também como relevante. Dentro do contexto canadense, o autor faz uma análise da evasão com ênfase no viés econômico e nas escolhas dos ingressantes, sobretudo no primeiro ano da graduação.

Pesquisas realizadas na Espanha apontam que existe uma interdependência entre as variáveis analisadas no estudo da evasão, dizendo-se que é necessária uma abordagem holística e o emprego de soluções que considerem múltiplos fatores e o peso de cada um deles no resultado final (Castro-Lopez, 2022).

Para Aina (2022), a evasão estudantil depende de um mix de fatores individuais, institucionais e econômicos. Estes fatores são mediados pela aptidão do discente em adaptar-se ao sistema acadêmico. Seu estudo da evasão na Itália integra a abordagem econômica com a análise de custos e benefícios associada à uma abordagem sociológica complementar, ampliando a investigação além do individual, para os contextos sociais e institucionais.

Em sua abordagem inicial Aina (2013) traz um estudo mais restrito às variáveis de contexto socioeconômico familiar e seus efeitos nos modelos de investimento em ensino superior das famílias, em âmbito nacional. Seus estudos, realizados em 2020, mostram que variáveis relacionais - tais como nível de integração com o sistema acadêmico, de identificação com atitudes, valores e padrões, extensão e qualidade das relações desenvolvidas com a instituição e seus pares - são consideradas pela autora como críticos para a maioria dos estudos dentro da abordagem sociológica. Nesta análise, estratégias de orientação estudantil antes da entrada na universidade, ações de integração acadêmica e social, e intervenções de aconselhamento, mentoria e tutorias depois do ingresso são apontadas como relevantes para a gestão da evasão (Aina, 2022).

No mesmo cenário geográfico, Contini (2018) aponta como variáveis críticas o contexto educacional e econômico familiar, dados demográficos, características escolares anteriores, características da universidade e do programa, além de variáveis contextuais, tendo Vilorio (2019) com a mesma indicação de fatores socioeconômicos e de escores anteriores.

Estudos realizados na Inglaterra também tem foco no contexto econômico das variáveis da evasão indicando que o acesso independe destes fatores, mas que a permanência e conclusão da graduação estão fortemente relacionados às variáveis deste gênero. Foram utilizados também informações acadêmicas longitudinais (da idade entre 11 e 18 anos de cada estudante incluso no estudo) e socioculturais, além dos dados universitários como disciplina cursada e instituição escolhidas de todos os estudantes de primeira graduação no Reino Unido (Vignoles, 2009), com o objetivo de identificar apenas se os estudantes permaneciam no Ensino Superior de um ano para outro, na mesma instituição e se grupos de alunos com contextos socioeconômicos diferentes apresentavam taxas de evasão distintas, sendo que este estudo analisou 121.827 casos de 120 instituições diferentes.

A análise do citado estudo concluiu que, à época, havia significante menor evasão dentro de perfis socioeconômicos mais elevados e de vizinhanças com maior nível de escolaridade, num contexto em que a evasão total girava em torno de 6% e uma forte correlação dos índices acadêmicos nas idades mais próximas ao ingresso e as probabilidades de evasão. O estudo também apontou uma menor evasão dentro de minorias étnicas, corroborando com outros trabalhos da época. No contexto britânico, o aspecto socioeconômico já não tem mais peso no ingresso ao ensino

superior para estudantes com diferentes históricos financeiros e desempenhos escolares semelhantes, mas ainda persiste como fator relevante para a evasão no cenário universitário (Vignoles, 2009).

Em Portugal, estudos realizados na Universidade de Aveiro apontam sete variáveis críticas para monitoramento de risco de evasão: taxas de sucesso acadêmicos abaixo de 50%, inadimplência, não obtenção de bolsa de estudos, nota de entrada, assiduidade, baixo nível de satisfação com a vida e experiências negativas (Ferreira, 2020).

Em outra perspectiva, grupos de variáveis como notas do ensino médio, índices de pobreza, quintil econômico e notas do primeiro e segundo anos do Ensino Superior mesclam fatores acadêmicos e econômicos e são utilizados por Palácios (2021), no contexto chileno, numa universidade onde 50% dos ingressantes vêm do quintil mais pobre do país e apenas 5% vêm do quintil mais rico. Neste mesmo contexto geográfico, Vallejos (2017) faz uma seleção de variáveis mais próxima das relatadas por Contini (2018), como indicado anteriormente.

Por isto a importância dada às variáveis de origem econômica, indicadores financeiros, origem geográfica e performance escolar e universitária. No Chile, as taxas de evasão variam entre 11% na área de saúde a 44% nas engenharias e 50% dos alunos matriculados não terminam a graduação no mesmo curso que iniciaram.

Estudos brasileiros (Quadro 1), no período entre 2019 e 2023, apontam uma correlação temporal entre a criação de políticas públicas e órgãos gestores de educação e a diminuição nos índices de evasão, mas não comprovam esta relação direta. Mesmo que os estudos analisem as diminuições destes índices nos anos consecutivos à adoção de políticas públicas na área educacional não conseguem identificar elementos diretos que possam ser apontados como uma relação de causa e efeito, dada a complexidade e variedade de fatores envolvidos (Santos, 2020). Os estudos sobre gestão da evasão com utilização da Mineração de Dados no Brasil ainda são poucos, mas já há experimentação de alguns algoritmos com este propósito e discussões iniciais sobre sua precisão e quais são mais indicados ao contexto nacional (Santos, 2019).

Quadro 1. Estudos brasileiros sobre evasão, no período entre 2019 e 2023.

NOME DO ESTUDO	AUTORES	TIPO DE ESTUDO	VARIÁVEIS ENCONTRADAS
Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout.	SANTOS, K. J. D. O., MENEZES, A. G., de CARVALHO, A. B., & MONTESCO, C. A. (2019)	Artigo Estudo de Caso	Evasão predominante em 4ª e 6º semestres, amostra 23.690 estudantes. Utilizado Árvore de Decisão e Floresta Randômica.
Diagnosis of Dropout and Strategies for Retention of University Students: from Theory to Practice in Times of Pandemic.	Dos SANTOS, J.P., SANTANA, A.D.; JUNGER, A.P.; THIMOTEO, A.C.D. (2023)	Artigo Estudo de Caso	Organização e fatores emocionais como variáveis críticas. Análise de educação à distância.
Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic, Socioeconomic, and Equity Factors through Machine Learning and Survival Analysis Methods in the Latin American Context.	GUTIERREZ-PACHAS, D.A., GARCIA-ZANABRIA, G., CUADROS-VARGAS, E., CAMARA-CHAVEZ, G., GOMEZ-NIETO (2023)	Artigo Análise Exploratória	Análise de variáveis demográficas, socioeconômicas e acadêmicas de 13.696 estudantes durante o período de 2008 a 2020.
A comparative study of WHO and WHEN prediction approaches for early identification of university students at dropout risk.	GUTIERREZ-PACHAS, D.A., GARCIA-ZANABRIA, G., CUADROS-VARGAS, E., CAMARA-CHAVEZ, G., GOMEZ-NIETO, E. (2021)	Artigo Estudo de Caso	Análise do perfil do estudante que evade e predição do momento crítico da evasão.
Dropout Management in Brazilian Educational Policies: from Undergraduate to Stricto Sensu Postgraduate.	SANTOS JÚNIOR, J. D. S., MAGALHÃES, A. M. D. S., & REAL, G. C. M. (2020).	Artigo Análise Exploratória	Análise quanti-qualitativa, documental e bibliográfica do processo de gestão da evasão a partir das políticas públicas de educação.

Fonte: Própria (2024). Extraída de *Web of Science*.

Estas variáveis serão testadas no capítulo cinco, para estabelecer uma relação entre os achados da literatura científica explorada e as variáveis críticas identificadas no estudo de caso da Universidade Federal do Sul da Bahia, a partir das técnicas de mineração de dados.

2.4 A Gestão da Evasão e a Inteligência de Dados

A mineração de dados é o processo de descobrir padrões, relações e informações úteis ou não triviais a partir de grandes conjuntos de dados. Ela utiliza técnicas de análise estatística, aprendizado de máquina e inteligência artificial para

explorar e extrair conhecimento relevante a partir de dados, auxiliando na tomada de decisões, previsões e identificação de tendências.

O primeiro estudo identificado sobre a utilização da mineração de dados para auxiliar na compressão do comportamento de estudantes e identificar a população de risco para evasão utilizava uma modelagem de dados baseada nas correntes de Markov, com o objetivo de aprimorar as políticas para redução dos índices de evasão (Massa, 1999).

Um estudo realizado por Agrusti (2019) considerando 73 artigos sobre evasão nas universidades com aplicação de técnicas de mineração de dados, coletados das bases de dados Scopus e Web of Science, aponta que as técnicas mais utilizadas para estes estudos são a Árvore de Decisão (67%), a classificação Bayesiana (49%) e as Redes Neurais (40%), visto que muitas abordagens optam pela utilização de mais de uma técnica na mesma análise. Assim, torna-se oportuno apresentar uma breve descrição das técnicas mais difundidas, conforme análise de Agrusti (2019):

1. **Árvore de Decisão:** é um algoritmo de aprendizado de máquina supervisionado usado para tarefas de classificação e regressão. Ele é uma técnica popular devido à sua interpretabilidade e à capacidade de lidar com conjuntos de dados complexos. A Árvore de Decisão divide recursivamente o conjunto de dados em subconjuntos mais homogêneos com base em características específicas, permitindo a tomada de decisões de forma hierárquica. O algoritmo de Árvore de Decisão segue um processo de construção hierárquico, no qual, em cada nó, a árvore faz uma decisão com base em um atributo específico. Cada nó representa um teste sobre um atributo, e cada ramo representa o resultado deste teste. Os nós folha da árvore contêm as classes de saída para tarefas de classificação ou valores de previsão para tarefas de regressão. Seus principais conceitos são a raiz, ou nó inicial da árvore, nó interno, que representa um teste em um atributo, nó folha, que representa a classe de saída (para classificação) ou o valor previsto (para regressão) e os ramos, que são as conexões entre os nós e representam as decisões tomadas com base nos testes dos atributos. Suas vantagens são a compreensão fácil, comum a lógica de decisão fácil de entender e interpretar, a capacidade de lidar com dados

categoricos e numericos com pouca necessidade de pré-processamento dos dados. Ajustar a complexidade da árvore (usando parâmetros como profundidade máxima ou número mínimo de amostras por folha) e técnicas como poda de árvores são estratégias comuns para evitar o sobre ajuste e melhorar o desempenho do modelo. O algoritmo de Árvore de Decisão é frequentemente usado em conjunto com métodos ensemble, como Florestas Aleatórias ou *Gradient Boosting*, para melhorar ainda mais o desempenho.

2. Classificações Bayesianas e Redes Bayesianas: As Redes Bayesianas (BNs) são modelos probabilísticos gráficos que representam as relações entre variáveis aleatórias por meio de um grafo acíclico direcionado (DAG, do inglês *Directed Acyclic Graph*). Estes modelos são úteis para representar incertezas e inferir sobre as relações causais entre diferentes variáveis em um domínio. Seus princípios básicos são Grafo Direcionado Acíclico (DAG), que representa as variáveis e suas dependências probabilísticas. Cada nó no grafo representa uma variável aleatória, e as arestas indicam relações direcionadas entre estas variáveis. Nós e Arestas Probabilísticas, onde cada nó na rede está associado a uma distribuição de probabilidade condicional que expressa a probabilidade da variável associada, dada a probabilidade das variáveis parentes (variáveis que têm uma aresta direcionada para o nó em questão). Regra da Cadeia: probabilidade conjunta de todas as variáveis em um DAG, pode ser decomposta em uma série de probabilidades condicionais multiplicadas, conforme indicado pela regra da cadeia da probabilidade. As BNs são eficientes para realizar inferência probabilística. Dada uma evidência observada em um conjunto de variáveis, o objetivo é calcular a distribuição de probabilidade para as variáveis não observadas. Suas maiores vantagens são a representação explícita de relações causais e dependências probabilísticas, sua eficiência em lidar com incertezas e variáveis ocultas e a facilidade de interpretação. Em contrapartida pode ser difícil determinar a estrutura adequada do grafo em domínios

complexos, além da dificuldade de estimar parâmetros em conjunto de dados limitados.

3. **Redes Neurais:** As redes neurais são modelos computacionais inspirados no funcionamento do cérebro humano. Uma rede neural consiste em camadas de unidades chamadas neurônios, cada uma das quais está conectada a neurônios em camadas adjacentes. As informações fluem por meio da rede, e os pesos associados às conexões entre neurônios são ajustados durante o treinamento para realizar tarefas específicas. Sua arquitetura é composta pelas camadas de entrada – que recebe as entradas no sistema -, ocultas, que realizam as transformações intermediárias nas entradas e as camadas de saída da rede. O treinamento de uma rede neural envolve a apresentação de dados de treinamento à rede, ajustando iterativamente os pesos das conexões para minimizar a diferença entre as saídas previstas e as saídas desejadas. O algoritmo de retro propagação (*backpropagation*) é comumente usado para ajustar os pesos durante o treinamento. Cada neurônio em uma rede neural utiliza uma função de ativação para determinar sua saída com base na soma ponderada das entradas. Seus pontos fortes são o reconhecimento de padrões, processamento de linguagem natural, visão computacional e previsão de séries temporais.

4. **Support Vector Machine:** Os *Support Vector Machines* (SVM) são algoritmos de aprendizado de máquina supervisionado utilizados para classificação e regressão. O objetivo é encontrar um hiperplano de decisão que separe os dados em classes (no caso de classificação) ou ajuste uma função para prever valores contínuos (no caso de regressão). O SVM procura um hiperplano que maximize a margem entre as classes. A margem é a distância entre o hiperplano e os pontos mais próximos de cada classe, chamados de vetores de suporte. A ideia é escolher o hiperplano que generaliza bem para novos dados. Em muitos casos, os dados não são linearmente separáveis. O SVM utiliza o "*kernel trick*" para mapear os dados para um espaço de maior dimensionalidade, onde podem se tornar linearmente separáveis. Exemplos de kernels incluem

o kernel linear, o kernel polinomial e o kernel radial (RBF). O SVM é capaz de classificar textos e imagens, reconhecer padrões e realizar previsões, sendo eficiente em espaços de alta dimensionalidade e eficaz mesmo em conjuntos de dados com poucas amostras. Porém, é sensível a outliers, sendo necessário um ajuste cuidadoso dos parâmetros.

5. *K-Nearest Neighbour*: O algoritmo *k-Nearest Neighbors* (k-NN) é um método de classificação e regressão utilizado em aprendizado de máquina supervisionado. Ele é conhecido como um "método baseado em instância" porque faz uso direto das instâncias de treinamento para fazer previsões para novos dados. O princípio fundamental do k-NN é bastante simples. Para classificar ou prever o valor de uma nova instância, o algoritmo procura pelos k vizinhos mais próximos desta instância nos dados de treinamento. A classe ou o valor da nova instância é então determinado pela maioria (para classificação) ou média (para regressão) dos rótulos ou valores dos k vizinhos mais próximos. A escolha do valor de k é crucial. Valores pequenos de k (por exemplo, 1 ou 3) podem tornar o modelo sensível a ruídos, enquanto valores grandes de k podem suavizar demais as fronteiras de decisão. A seleção de k depende da natureza do conjunto de dados e deve ser escolhida por meio de técnicas de validação cruzada ou outras estratégias de ajuste de hiper parâmetros. Este algoritmo é sensível a atributos irrelevantes ou redundantes e também à escala de atributos. Pode ser computacionalmente caro, especialmente em conjuntos de dados grandes.

6. *Regressão Logística*: A Regressão Logística é um método estatístico utilizado para modelar a relação entre uma variável dependente binária (que assume dois valores, geralmente 0 e 1) e uma ou mais variáveis independentes. Ao contrário da Regressão Linear, que é usada para prever valores contínuos, a Regressão Logística é especialmente adequada para problemas de classificação binária, onde o objetivo é prever a probabilidade de uma observação pertencer a uma determinada classe. A Regressão Logística modela a relação entre as variáveis

independentes e a probabilidade de um evento ocorrer. A função logística é aplicada para transformar a saída linear em uma probabilidade, limitada entre 0 e 1 e, por isto, é frequentemente utilizada em problemas de classificação binária, como prever se um estudante evadiu ou não. Os coeficientes estimados na Regressão Logística representam a mudança logarítmica nas probabilidades da variável dependente para uma mudança unitária nas variáveis independentes, o que facilita a interpretação. São limitações deste sistema a sua presunção de uma relação linear entre as variáveis e a transformação logística, sendo apropriado considerar técnicas mais avançadas caso as relações não sejam lineares. Também é sensível a outliers, em especial em amostras pequenas, impactando significativamente os coeficientes. Ainda assim, a Regressão Logística é uma ferramenta poderosa e amplamente utilizada, especialmente em contextos de classificação binária. No entanto, é importante entender suas limitações e considerar alternativas, como modelos de aprendizado de máquina mais complexos, quando necessário.

Embora estas sejam as técnicas mais utilizadas, não há uma hegemonia na aplicação de um algoritmo específico. Em Vitoria (2019), são utilizados a Árvore de Decisão, as Redes Neurais e as Redes de Classificação Bayesianas, que também são empregadas por Vallejos (2017).

Para Mezzini (2019) as Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNN) são um ponto divisor de águas no uso da Inteligência Artificial, utilizando dezenas ou centenas de camadas neurais convolucionais proporcionando, assim, uma estrutura de rede profunda que têm produzido modelos altamente precisos. O modelo foi aplicado aos dados de 6.000 estudantes da Universidade Roma Tre com resultados considerados muito promissores.

Estudos da Universidade de Aveiro, Portugal, utilizam uma ferramenta própria denominada Ferramentas de Identificação e Combate ao Abandono (FICA), com o objetivo de facultar acesso aos dados críticos para gestão da evasão por diferentes stakeholders, baseando o desenvolvimento desta ferramenta na acessibilidade dos dados, permitindo uma exploração visual dos dados gerados, obtendo respostas de forma rápida e interativa (Ferreira, 2020).

Para Castro-Lopez (2022), a aplicação de modelo difuso permite diminuir as incertezas inerentes ao comportamento humano de forma mais eficaz, processando conceitos linguísticos que são típicos da linguagem natural usada em alguns dados e convertendo-os em descrições mais reais e objetivas dos problemas e suas soluções. Seu estudo utiliza o Sistema de Inferência Difuso (*Fuzzy Inference System – FIS*) de forma adaptada para ampliar o diagnóstico da intenção de permanência na universidade.

Em artigos ainda mais atuais as propostas tecnológicas são apontadas como o caminho para lidar com a evasão universitária, num estudo que analisa oito modelos preditivos num banco de dados de 4.365 acadêmicos da Universidade Nacional de Moquegua, no Peru, cujo resultado indica a Floresta Randômica como o modelo mais preciso e robusto, com 97% de acerto (Flores, 2022), mesmo algoritmo aplicado em Behr (2020).

O estudo mais recente identificado, aponta a utilização de múltiplas técnicas de mineração de dados numa base de dados robusta, incluindo uma verificação com amostra adicional superior a 10.000 dados, aplicando a Rede Neural, Floresta Randômica e logit, obtendo resultados com alta precisão (77%) e sensibilidade em desistências previstas (69%), mesmo utilizando apenas dados acadêmicos da universidade (Ortiz-Lozano, 2023). Este estudo demonstra a evolução do campo de estudo, dos resultados obtidos e das técnicas utilizadas.

Cabe destacar, inicialmente, que para esta pesquisa será utilizada a Regressão Logística para identificar as variáveis críticas para o estudo da evasão na Universidade Federal do Sul da Bahia, numa análise por cursos de graduação e também por disciplinas, e uma análise por Árvore de Decisão para entender de que forma estas variáveis se sobrepõe para alcançar os resultados observados, permitindo a visualização e interpretação dos dados, subsidiando as tomadas de decisão.

A indicação da opção pela Regressão Logística é importante, neste momento de discussão teórica, pois oportuniza o entendimento acerca da organização do ciclo de vida dos dados coletados, quando este estudo adotou o modelo CRISP-DM (*Cross Indutor Standard Processo Data Mining*) ou Processo Padrão entre Indústrias para Mineração de Dados. Este processo propõe um método composto por seis fases para desenvolver modelos preditivos de falhas e soluções a partir da análise de informações e dados.

As fases que constituem este modelo de processamento de dados são: Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implementação, conforme Figura 1 abaixo.

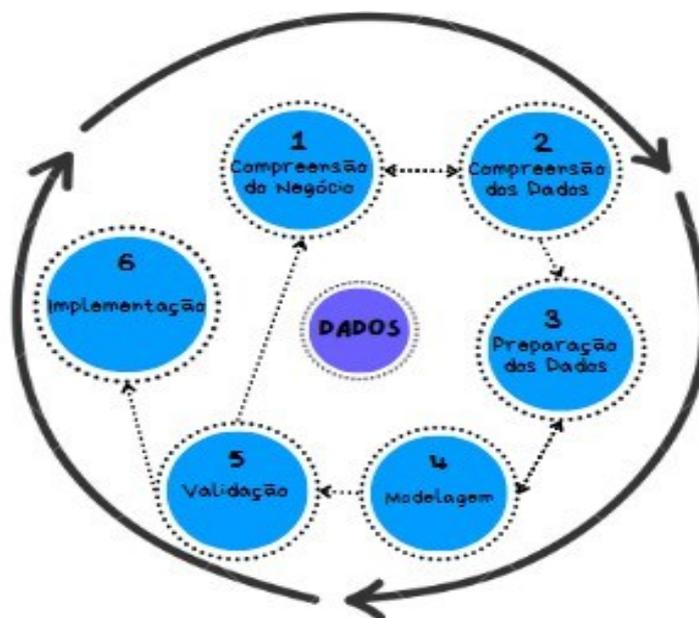


Figura 1 – Modelo CRISP-DM.

Fonte: Própria (2024).

O Quadro 2, sintetiza as seis etapas do modelo de processamento do CRISP-DM e suas definições, de acordo com os estudos de Salt (2021):

Quadro 2. Etapas do CRISP-DM

DETALHAMENTO CONFORME PROPOSIÇÃO DE SALTZ (2021)	
Compreensão do Negócio	A Compreensão do Negócio consiste na etapa de identificação e entendimento dos objetivos a serem alcançados pela pesquisa e as ações, requisitos, riscos, benefícios e ferramentas necessárias para isto, analisando o cenário atual em que está inserida, permitindo a elaboração de um plano de estudo detalhado.
Compreensão dos Dados	A fase seguinte, Compreensão dos Dados, se concentra na identificação, coleta e análise dos dados relevantes, diferenciando as informações que contribuem para o entendimento da pesquisa e para o alcance dos objetivos estabelecidos. São apontadas como tarefas desta etapa a coleta dos dados iniciais, sua descrição, exploração e apuração de qualidade, produzindo relatórios destas entregas.
Preparação dos Dados	Selecionados e coletados os dados, inicia-se a terceira etapa, a Preparação dos Dados, onde este conjunto de informações é tratado e seleções são feitas, delimitando o universo dos dados que serão submetidos à modelagem. Os critérios de inclusão/exclusão de dados devem ser definidos e descritos e todas as manipulações devem ser apontadas, como criação de atributos, mesclagem e formatação de dados
Modelagem	A Modelagem é a etapa de construção e avaliação dos modelos e experimentação de técnicas variadas, buscando um modelo que seja o mais adequado. São tarefas desta fase a seleção das técnicas de modelagem,

	criação do desenho de teste até a construção do modelo final e sua avaliação, sempre gerando relatórios
Validação	Seguindo o processo, avança-se para a fase de Avaliação dos resultados do modelo desenvolvido, considerando os critérios interessantes para a pesquisa, a revisão do trabalho construído e o planejamento das próximas ações. É uma etapa de validação de resultados e correção de possíveis erros. Os relatórios desta fase são voltados para a avaliação dos resultados obtidos na perspectiva dos parâmetros estabelecidos nos objetivos
Implementação	Todo este processo culmina na etapa de Implementação, desde o seu planejamento, passando pelo desenvolvimento de um método de monitoramento e manutenção do modelo em funcionamento, à produção do relatório final e revisão da pesquisa

Fonte: Própria (2024). Extraída de Saltz, 2021.

Com base na descrição das etapas do CRISP-DM, é possível perceber a importância de seguir um processo estruturado e iterativo para a realização de projetos de mineração de dados. Este modelo oferece uma metodologia clara para abordar problemas complexos, desde a compreensão inicial do domínio até a implementação de soluções práticas. Ao seguir estas etapas, a pesquisa garante maior rigor metodológico e a geração de insights relevantes, especialmente no contexto da evasão acadêmica analisada neste trabalho.

A aplicação destas etapas ao estudo dos dados da UFSB permite uma análise robusta, contribuindo para a identificação de padrões e possíveis intervenções para a gestão da evasão, alinhando-se aos objetivos propostos, o que será efetivamente desenvolvido no capítulo da análise de dados e dos resultados.

3 EVASÃO NO ENSINO SUPERIOR: ESTADO DA ARTE

Neste capítulo é explorada uma visão abrangente e atualizada sobre o campo a ser estudado: a evasão no Ensino Superior. Para isto, é preciso estabelecer aqui o entendimento sobre o termo utilizado e seu uso no campo de pesquisa. Tinto (1975) já alertava para a importância da definição do termo “evasão”, atribuindo à esta falta o fracasso de pesquisas anteriores. A não definição do campo e de modelos teóricos permitiu que comportamentos de abandono com características muito distintas fossem agrupadas sob a mesma nomenclatura, dificultando significativamente a construção de políticas voltadas para a educação superior.

Para a construção deste capítulo, nenhuma definição específica foi selecionada para que o estudo englobasse todos os trabalhos que abordam o tema da evasão sob alguma perspectiva, construindo uma figura abrangente da temática escolhida.

Para esta revisão sistemática bibliográfica foi utilizada a Teoria do Enfoque Meta-Analítico Consolidado (TEMAC) de Mariano & Rocha (2017), composta pelas etapas de preparação da pesquisa bibliográfica, apresentação e interrelação dos dados e detalhamento dos artigos mapeados, modelo integrador e validação por evidência que serão apresentados nos tópicos a seguir, para fundamentar a escolha das referências bibliográficas.

3.1 Preparação da Pesquisa

A busca foi realizada na base de dados *Web of Science* (WoS) devido a sua confiabilidade e amplitude e os dados foram salvos em planilhas de Excel para análises posteriores.

Foram utilizadas, inicialmente, as palavras-chaves “*university dropout*” e “*data mining*” ou “*data science*”, com todos os campos selecionados. Foram encontradas 31.174 publicações para estes parâmetros, levando a perceber que houve um equívoco na formulação da busca, onde a plataforma ofereceu todas as ocorrências para “*university dropout*” e “*data mining*”, como um termo único, ou todas as ocorrências para “*data science*”, um campo muito maior que a evasão. Assim, uma nova busca foi realizada utilizando apenas o termo “*university dropout*”, em todos os

campos. A nova busca retornou 164 publicações e foi adotada em sua integralidade, de acordo com a busca 2 do Quadro 3.

Quadro 3. Mapa de busca

Busca	String de busca	Período	Tipo de documento	Áreas	Total de registros
1	"University Dropout" AND "Data Mining" OR "Data Science"	Todos os anos	Todos	Todas	31.174
2	"University Dropout"	Todos os anos	Todos	Todas	164

Fonte: Própria (2023).

Assim, uma nova busca foi realizada utilizando apenas o termo "university dropout", em todos os campos. A nova busca retornou 164 publicações e foi adotada em sua integralidade, de acordo com a busca 2 do quadro 2, acima.

3.2 Apresentação e Interrelação de Dados

A partir do mapeamento realizado foi possível a identificação do registro mais antigo em relação ao tema da evasão universitária, sendo este a artigo de Massa e Puliafito (1999) "*An application of data mining to the problem of the University students' dropout using Markov chains*", onde é apresentada uma nova aplicação da mineração de dados à problemática da evasão no Ensino Superior, utilizando uma nova modelagem baseada na cadeia de Markov para extrair informações dos dados sobre comportamento estudantil.

É possível verificar no mapa de coocorrência de palavras-chave as diferentes tendências de abordagem dentro da temática, conforme Figura 2. Em 2018 as pesquisas giravam em torno de "education", "stress" e "college students" numa abordagem ainda voltada ao comportamento estudantil.

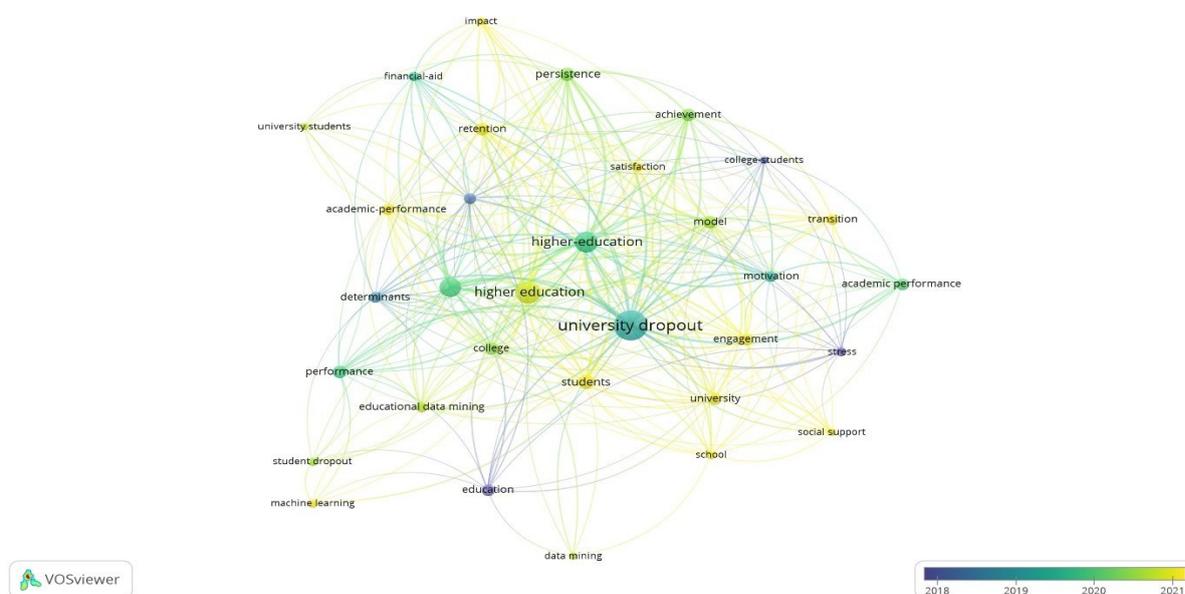


Figura 2 – Mapa de cocorrência

Fonte – Própria (2023). Criada a partir do VOSviewer 1.6.18

Em 2019 este foco se direciona a termos como “*University dropout*”, “*higher-education*”, “*motivation*” e “*performance*”, indicando uma preocupação maior com desempenho no ensino superior. Em 2020 termos como “*academic performance*” direcionando-se para “*educational data mining*” e “*student dropout*” mostram a introdução da mineração de dados neste campo. Em 2021 as temáticas apontam para “*machine learning*”, “*retention*” e “*academic-performance*” com maior participação das tecnologias.

Sobre a quantidade de artigos há uma clara tendência de aumento (Figura 3), com aumento superior a 100% no número de publicações de 2017 para 2018, repetindo o mesmo feito em 2019. Os anos seguintes reportam uma estabilidade, mantendo-se o alto número de publicações dos anos anteriores.

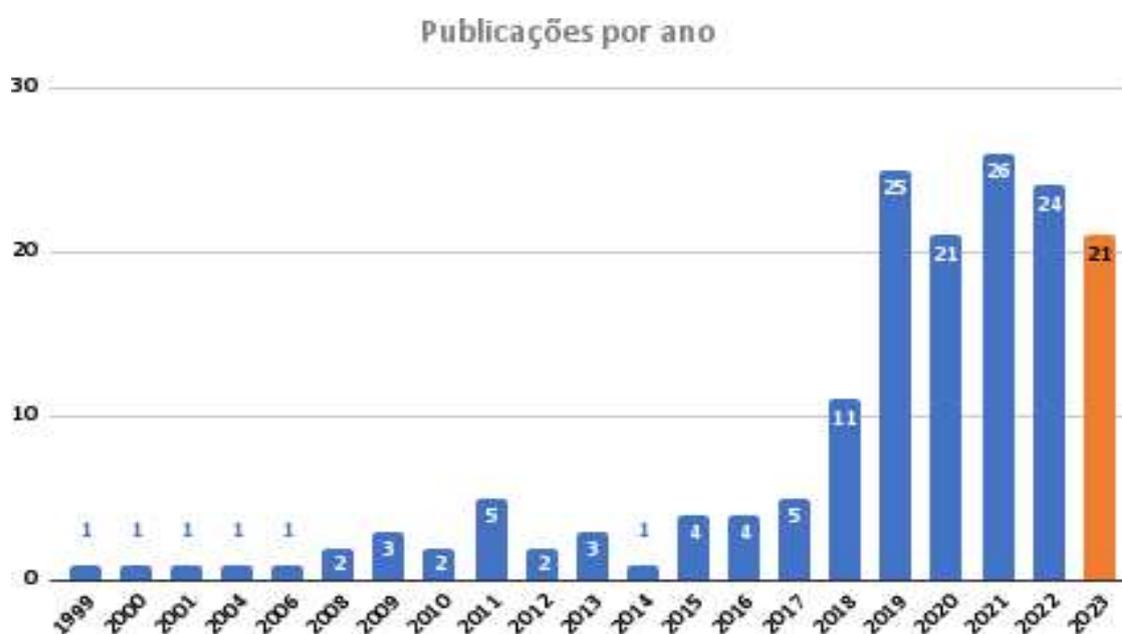


Figura 3 – Publicações por ano indexadas em *Web of Science*.

Fonte – Própria (2023). Extraída de *Web of Science*.

É possível verificar que mesmo nos anos relativos à pandemia da Covid-19 o número de publicações da área manteve a média. A pandemia da Covid-19 foi desencadeada pela propagação do vírus SARS-CoV-2 e teve início no final de 2019 na cidade de Wuhan, na China. Rapidamente, o vírus se disseminou globalmente, resultando em uma pandemia declarada pela Organização Mundial da Saúde (OMS) em março de 2020. Os impactos da pandemia na produção científica acadêmica foram profundos e dicotômicos: ao passo em que as comunidades científicas de todo o mundo reuniam esforços para produzir soluções científicas, o isolamento e restrições necessárias à época também dificultaram o prosseguimento dos estudos de algumas áreas. No campo objeto deste estudo, as produções se mantiveram e os anos subsequentes, como o de 2023, já confirma esta tendência com 21 registros.

Do período apresentado, são destacados, neste momento, os seis artigos mais citados, com ampla margem para os dois mais citados (Quadro 4).

Quadro 4. Artigos mais citados

Autor(es)	Título do Artigo	Contribuições	Citações
Montmarquette, C; Mahseredijan, S; Houle, R (2001)	The determinants of university dropouts: a bivariate probability model with sample selection	Analisa determinantes da evasão universitária com modelo Probit Bivariado com viés de seletividade, considerando interesses e habilidades estudantis.	77
Aina, Carmen (2013)	Parental background and university dropout in Italy	Analisa taxas de ingresso e evasão observando condições culturais e financeiras familiares para investimento educacional.	69
Vignoles, Anna F.; Powdthavee, Nattavudh (2009)	The Socioeconomic Gap in University Dropouts	Análise do índice de evasão após o primeiro ano nas universidades de estudantes economicamente menos favorecidos.	39
Lovon Cueva, Marco Antonio; Cisneros Terrones, Sandra Amelia (2020)	Impact of virtual classes on the university students in the context of COVID-19 quarantine: The case of the PUCP	Analisa o impacto das aulas virtuais na saúde mental de estudantes universitários no recorte da quarentena da Covid-19.	33
Contini, Dalit; Cugnata, Federica; Scagni, Andrea (2018)	Social selection in higher education. Enrolment, dropout and timely degree attainment in Italy	Análise das probabilidades de conclusão do ensino superior para jovens oriundos de meios desfavorecidos.	32
Di Pietro, Giorgio (2006)	Regional labour market conditions and university dropout rates: Evidence from Italy	Análise das determinantes da evasão universitária usando modelo Probit Bivariado com seleção de amostra.	32

Fonte: Própria (2023). Extraída de *Web of Science*.

Os trabalhos mais citados fazem, em geral, uma análise dos fatores socioeconômicos e culturais que favorecem a evasão, à exceção do trabalho de Cueva e Terrones (2020), que trata do impacto das aulas virtuais na saúde mental dos estudantes universitários durante a pandemia da Covid-19. Dois destes trabalhos, (Di Pietro, 2006 e Montmarquette, 2001) utilizam o modelo Probit Bivariado com viés de seletividade, tornando relevante o entendimento da escolha desta ferramenta, sendo pertinente a dois autores dos mais citados no campo.

Os autores que mais publicaram foram Bernado e Galve-Gonzalez, com sete e seis publicações, respectivamente (Figura 4).

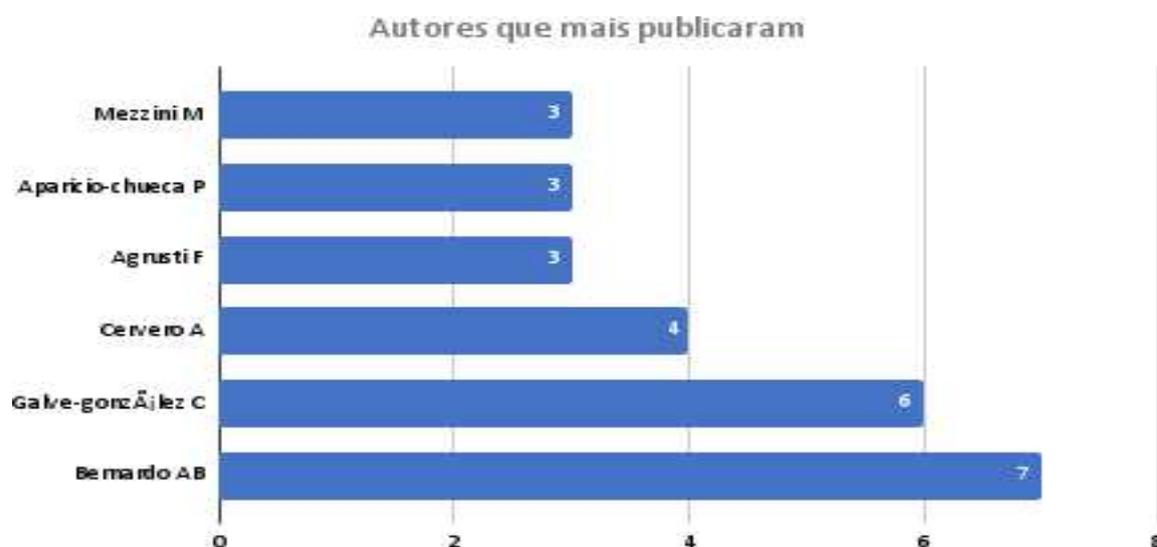


Figura 4 – Autores que mais publicaram

Fonte: Própria (2023). Extraída de *Web of Science*.

Dos sete artigos publicados por Bernardo (2019, 2020, 2021, 2021, 2022, 2022, 2023), seis foram feitos em colaboração com Galve (2020, 2021, 2021, 2022, 2022, 2023), e tratam da análise de fatores que levam a evasão. Percebe-se das publicações a ênfase na multiplicidade destes fatores e na impossibilidade de apontar uma causa única para o fenômeno do abandono. Nos trabalhos são utilizadas técnicas de análise diferentes, prevalecendo como constante a correlação das causas de evasão sobre o método de predição.

Quanto aos autores mais citados, suas abordagens incluem a aplicação de modelagens de variáveis dependentes, análise de impactos e probabilidades (Figura 5).

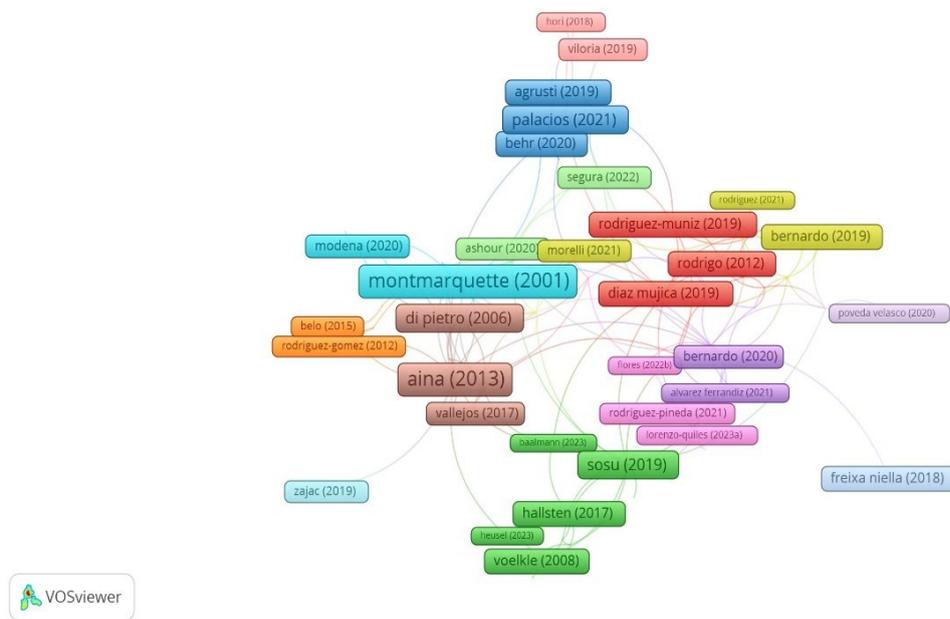


Figura 5 – Autores mais citados

Fonte: Própria (2023). Extraída de *Web of Science*.

É possível observar a conformidade entre os autores mais citados e os artigos mais citados, no entanto, Aina e Montmarquette utilizam perspectivas diferentes para estudar a evasão em seus campos de pesquisa, a primeira abordando questões sociais, familiares e econômicas e o segundo concentrando-se em questões individuais como habilidades e interesses do estudante.

A distribuição dos trabalhos por região geográfica mostra a predominância da Espanha (47), seguida por Itália (22) e Alemanha. O Brasil ocupa a oitava posição junto com o México, ambos com 06 publicações (Figura 6).



Figura 6 – Países que mais publicaram

Fonte: Própria (2023). Extraída de *Web of Science*.

É possível visualizar a hegemonia europeia na produção de trabalhos sobre evasão com ampla margem para Espanha e Itália, mas também se percebe que, fora do contexto europeu, são os países que mais têm se dedicado a este tema.

Sobre as agências financiadoras, foram identificadas 75 investidoras das quais 68 promoveram um estudo. A agência com maior número de incentivos é a alemã *Federal Ministry of Education and Research* (Bundesministerium für Bildung und Forschung, BMBF), seguida do governo da Espanha e União Europeia (Figura 7).



Figura 7 – Agências financiadoras

Fonte: Própria (2023). Extraída de *Web of Science*.

As áreas que mais publicaram (Figura 8) sobre o tema foram a Educação (49,36%), seguida pela Ciência da Computação, Economia e a Psicologia Multidisciplinar (somam 23,78%).

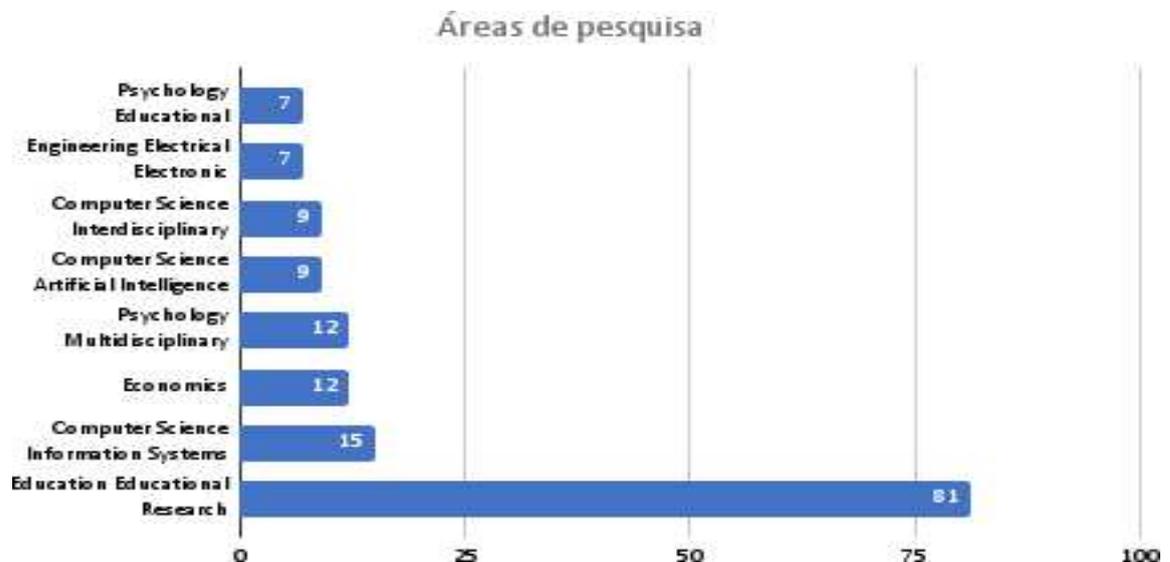


Figura 8 – Áreas de pesquisa

Fonte.: Própria (2023). Extraída de *Web of Science*.

Apesar da análise quantitativa de dados, utilização de modelagens e técnicas estatísticas, os estudos concentram-se na área educacional, prevalecendo a temática da evasão.

Os 164 trabalhos catalogados estão distribuídos em 147 revistas diferentes (Figura 9). Dentre estas, a que mais publicou foi a *Frontiers in Education*, considerada uma das principais revistas científicas neste campo e fonte de referência para pesquisadores, educadores e profissionais da educação, com oito trabalhos.

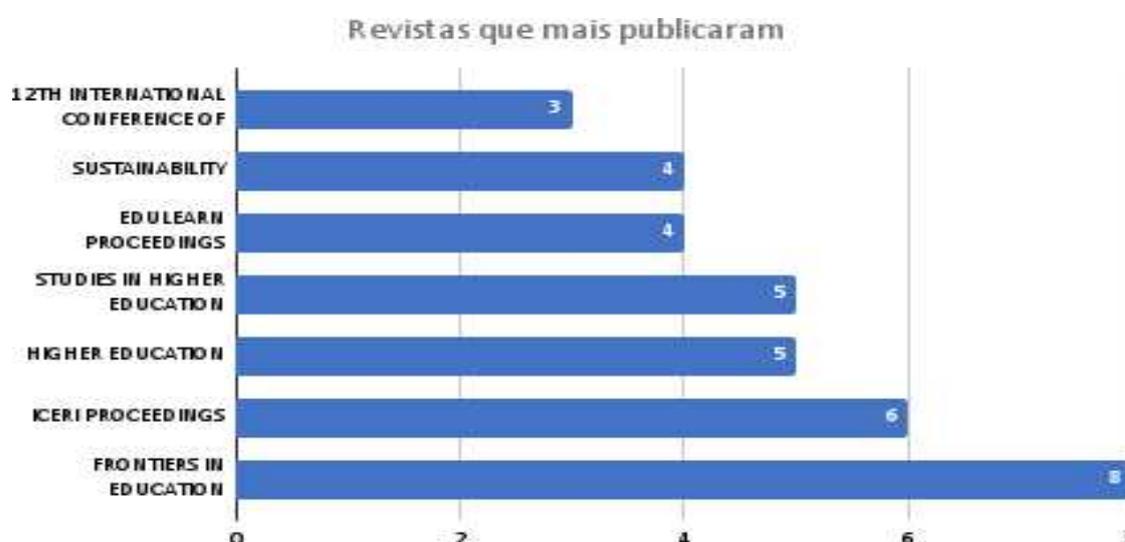


Figura 9 – Revistas que mais publicam

Fonte: Própria (2023). Extraída de *Web of Science*.

A *Iceri Proceedings* publica os trabalhos apresentados na Conferência Internacional sobre Educação, Pesquisa e Inovação (*International Conference of Education, Research and Innovation – ICERI*), sendo um importante fórum de compartilhamento de pesquisas.

Este levantamento apontou que as universidades que mais publicam são espanholas, seguidas de Alemanha e Itália. A Corporación Universitaria de Asturias, figura neste ranking com três publicações (Figura 10).

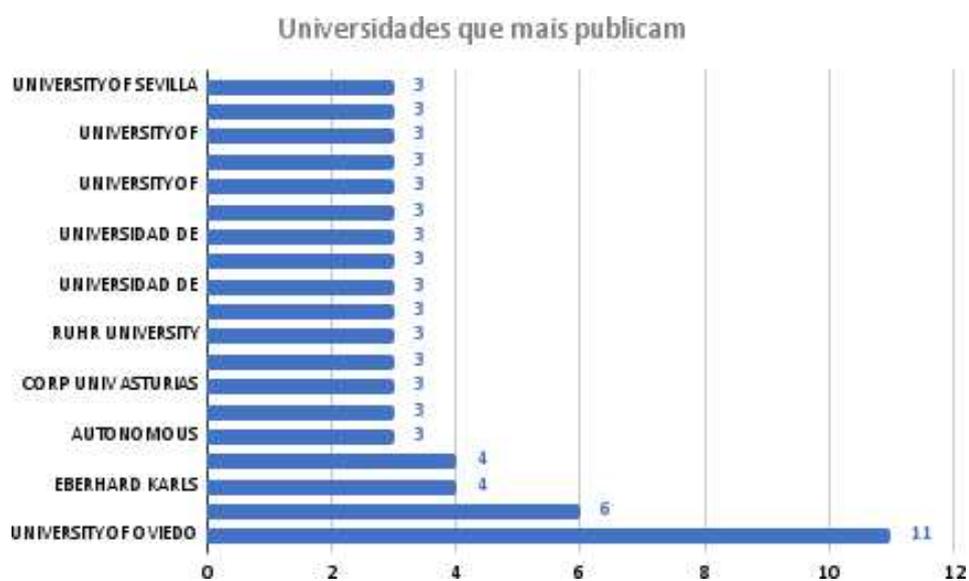


Figura 10 – Universidades que mais publicam

Fonte: Própria (2023). Extraída de *Web of Science*.

Entre as instituições brasileiras apresentam-se a Universidade Federal de Ouro Preto com duas publicações, e o Centro Universitário SENAC e a Fundação Getúlio Vargas com uma publicação.

Consolidam-se as palavras mais encontradas nos títulos das publicações (Figura 11) e é possível perceber núcleos estruturados e intensos em torno de termos como *university*, *university dropout*, *analysis*, *determinant* e *academic performance*.

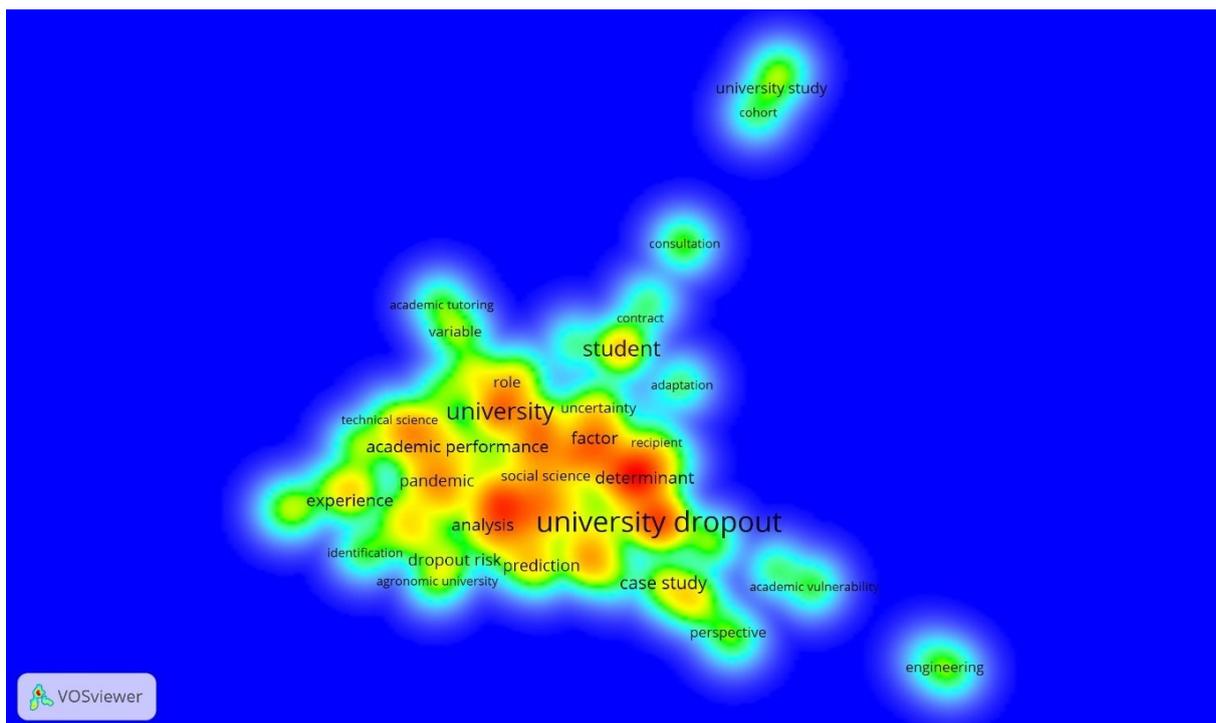


Figura 11 – Palavras mais frequentes nos títulos dos artigos

Fonte: Própria (2023). Extraído de *Web of Science*.

Além dos núcleos já desenvolvidos, pode-se identificar novos polos emergentes, indicando novas perspectivas sendo buscadas dentro da temática, como *academic vulnerability*, *perspective*, *academic tutoring* e *engineering*.

3.3 Detalhamento, Modelo Integrador e Validação por Evidências

Nesta etapa é apresentado o detalhamento da pesquisa, utilizando o conceito de *cocitation* para identificar os documentos que normalmente são citados conjuntamente em outros trabalhos acadêmicos, demonstrando uma proximidade conceitual e grupos de pesquisa, e *coupling*, utilizada para explorar as fronteiras do campo e novas tendências que compartilham a mesma área de interesse.

A análise de *cocitation* usa um algoritmo de clusterização identificando núcleos de autores cocitados nos trabalhos encontrados sobre o tema, já o *coupling*, usa os três últimos anos para localizar os trabalhos que citam conjuntamente os mesmos trabalhos (Mariano & Rocha, 2017).

O mapa de cocitação (Figura 12) indica as abordagens mais representativas e foram selecionados autores com o mínimo de cinco cocitações.

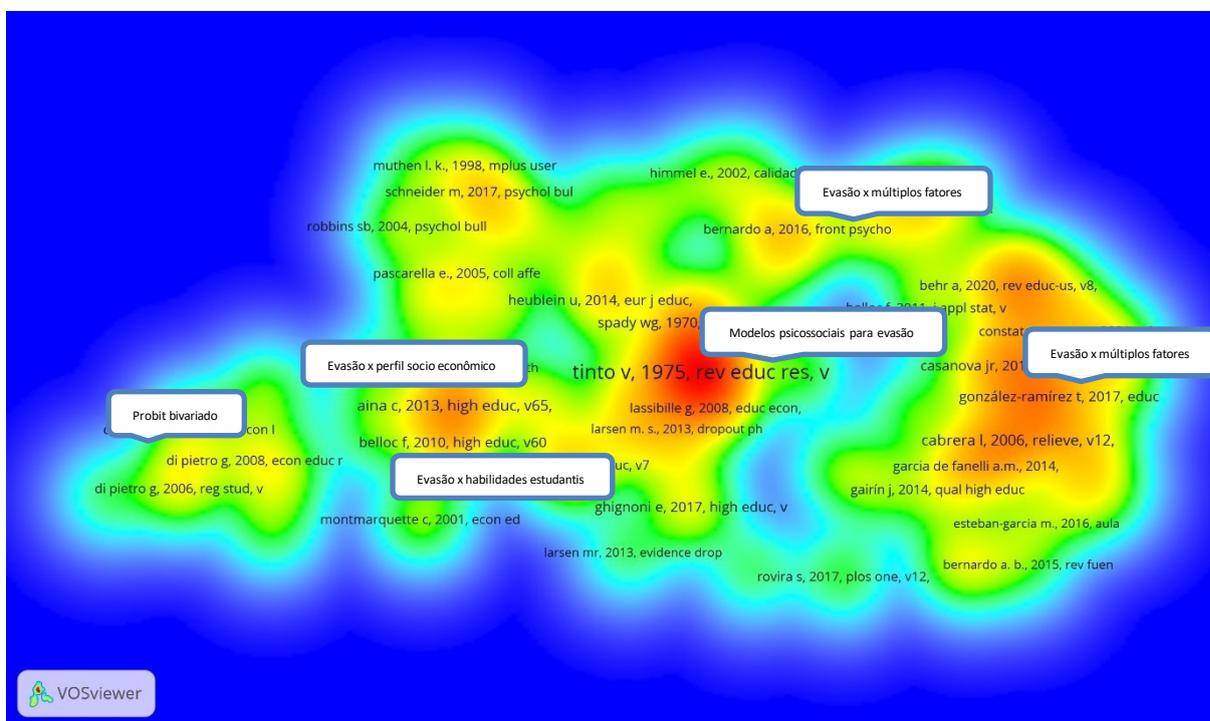


Figura 12 – Mapa de Cocitation

Fonte: Própria (2023). Extraída do VOSviewer.

Observa-se um núcleo mais intenso em Tinto (1975), indicando uma fonte de referência que trata de modelos psicossociais utilizados para explicar a evasão, Aina (2013) com enfoque na análise do perfil socioeconômico como variáveis para a evasão e Gonzalez-Ramirez (2017) e Bernardo (2016), com abordagens similares, ressaltando que a evasão se deve a múltiplos fatores interrelacionados. O núcleo de Di Pietro (2006) traz maior ênfase no Probit bivariado.

Na análise de *coupling* (Figura 13) observam-se dois núcleos bastante fortalecidos, com Palacios (2021) e Aina (2022), outros dois núcleos em crescimento com Constate-Amores (2021) e Castro-Lopez (2022) e outros autores ainda emergindo nesta temática. Bernardo (2022) permanece sendo relevante no tema da evasão.

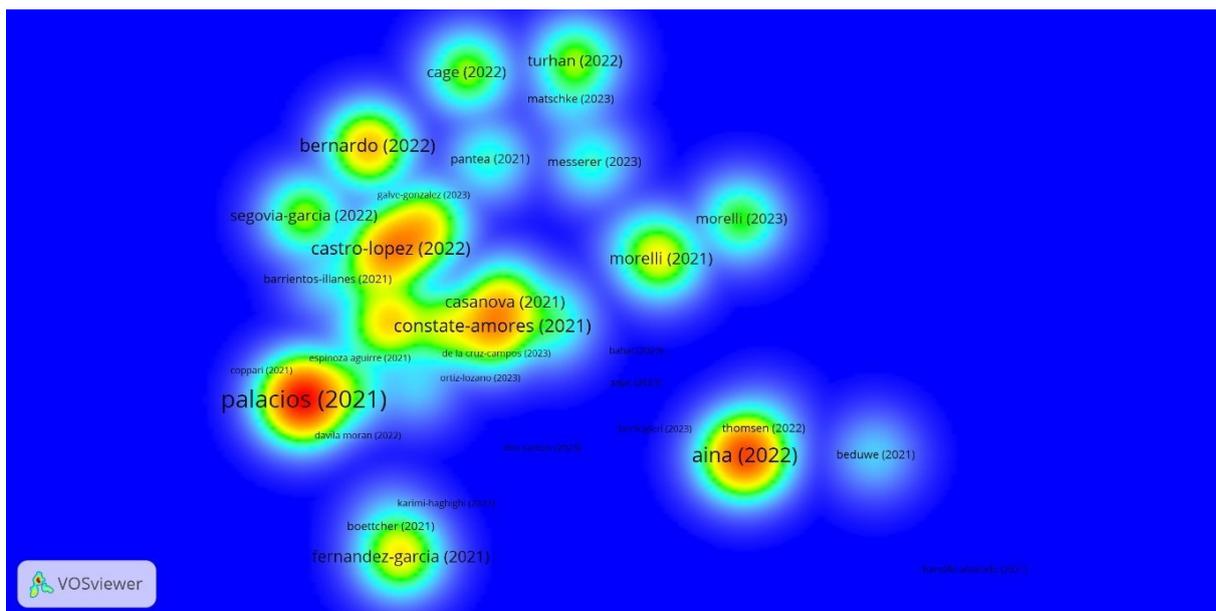


Figura 13 – Mapa de *Coupling*

Fonte: Própria (2023). Extraída de VOSviewer.

Os dois núcleos mais aquecidos tratam da relação entre evasão e habilidades individuais (Palacios, 2021) e evasão e fatores socioeconômicos (Aina, 2022), enquanto os núcleos emergentes abordam questões motivacionais e acadêmicas (Constata-Amores, 2021) e conferem especial atenção às abordagens de mineração empregadas (Castro-Lopez, 2022), indicando a ampliação do espectro das abordagens aplicadas ao campo de estudo.

Em acordo com a leitura realizada durante as três etapas deste processo, foram selecionados os documentos mais relevantes para a temática a serem validados por sua adesão ao tema.

O coeficiente Kappa (κ) é uma medida estatística que avalia a concordância entre dois avaliadores ou métodos de medição em um estudo. É frequentemente utilizado em estudos de confiabilidade ou validação, onde é importante avaliar o grau de concordância além do que seria esperado ao acaso. O coeficiente Kappa leva em consideração a proporção de concordância observada e a concordância esperada por acaso (Landis et. al, 2014).

A fórmula geral para o coeficiente Kappa é a seguinte:

$$k = \frac{Co - Ca}{1 - Ca}$$

Onde:

- C_o é a proporção de concordância observada, encontrada na soma entre o percentual de estudos incluídos em concordância por ambos os pesquisadores, adicionado ao percentual de estudos excluídos em concordância por ambos.
- C_a é a proporção de concordância esperada por acaso.

O coeficiente Kappa varia de -1 a 1, onde:

- $K = 1$: Concordância perfeita.
- $K = 0$: Concordância esperada ao acaso.
- $K = -1$: Discordância perfeita.

Foi realizado o cálculo da confiabilidade seguindo as instruções de Felizardo et al. (2017) (Tabela 1):

Tabela 1. Cálculo de Kappa

		Pesquisador 1		Total marginal da linha
		I	E	
Pesquisador 2	I	16,5%	11,6%	28,1%
	E	0,6%	71,3%	71,9%
Total marginal da linha		17,1%	82,9%	

$$Kappa = \frac{0,878 - 0,645}{1 - 0,645} * (100\%) \quad (4)$$

$$Kappa = 65,69\% \quad (5)$$

Sendo assim, aplicando a fórmula do Kappa, obtém-se o valor de 65,69%, ou 0,656 de confiabilidade. A literatura (Landis, 1997) estabelece que o valor de 0,656 é considerada uma concordância substancial.

Assim, os artigos selecionados estão listados no Quadro 5.

Quadro 5. Artigos selecionados

Autores/Ano	Título	Justificativa	Resultados
Agrusti, F; Bonavolontà, G; Mezzini, M (2019)	University dropout prediction through Educational Data Mining techniques: A systematic review.	Revisão sistemática de técnicas de EDM; Mezzini está entre os autores que mais publicam.	Apresenta 73 artigos, 6 técnicas de classificação, 53 algoritmos e 14 ferramentas.
Mezzini, M; Bonavolontà, G; Agrusti, F (2019)	Predicting university dropout by using convolutional neural networks.	Artigo apresenta análise baseada em <i>Convolutional neural networks</i> .	Redes Neurais, amostra 6000 alunos da Universidade Roma Tre, acurácia entre 67,1% e 88,7%.
Agrusti, F; Mezzini, M; Bonavolontà, G (2020)	Deep learning approach for predicting university dropout: a case study at Roma Tre University.	Estudo de caso, mesmo grupo de autores (Mezzini), Bayesian network.	Redes Bayesianas, amostra 6000 alunos da Universidade Roma Tre, acurácia entre 67,1% e 94,3%.
Santos, KJD; Menezes, AG; de Carvalho, AB; Montesco, CAE (2019).	Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout.	Artigo brasileiro, UFSE, trata especificamente de Educational Data Mining e aponta várias técnicas.	Uso de algoritmos como ferramentas de apoio a decisões de gestão da evasão.
Behr, A; Giese, M; Teguim, KHD; Theune, K (2020)	Early Prediction of University Dropouts - A Random Forest Approach.	Artigo alemão, está entre os 03 países que mais publicam.	Fornecimento de informações para implementação de sistema de alerta precoce e serviço de aconselhamento.
Palacios, CA; Reyes-Suárez, JA; Bearzotti, LA; Leiva, V; Marchant, C (2021)	Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile.	Palacios está entre os autores mais citados.	Escore educacional secundário e o índice de pobreza da comunidade são variáveis preditivas importantes.
Montmarquette, C; Mahseredjian, S; Houle, R (2001)	The determinants of university dropouts: a bivariate probability model with sample selection.	Montmarquette centro da nuvem de autores mais citados.	Dados longitudinais da Universidade de Montreal, variáveis críticas de interesse e habilidades.
Ferreira, F; Santos, BS; Marques, B; Dias, P (2020)	FICAVIS: Data Visualization to Prevent University Dropout.	Proposta de abordagem que permite explorar visual e interativamente os dados; autores que mais publicam.	Ferramentas para Identificação e Combate ao Abandono Escolar com metodologia de design participativo.
Castro-Lopez, A; Cervero, A; Galve-González, C; Puente, J; Bernardo, AB (2022)	Evaluating critical success factors in the permanence in Higher Education using multi-criteria decision-making.	Artigo engloba diferentes tipos de análises de dados numa amostra de 1912 linhas.	Múltiplas variáveis críticas que interagem para a intenção de permanência.
Vignoles, AF; Powdthavee, N (2009)	The Socioeconomic Gap in University Dropouts.	Artigos mais citados.	Existência de uma lacuna socioeconômica na evasão universitária.
Contini, D; Cugnata, F; Scagni, A (2018)	Social selection in higher education. Enrolment, dropout and timely degree attainment in Italy.	Artigos mais citados, proveniente da Itália, países que mais publicam.	Desfavorecimento econômico e baixas taxas de desemprego juvenil contribuem para a evasão.
Massa, S; Puliafito, PP (1999)	An application of data mining to the problem of the University students' dropout using Markov chains.	Referência mais antiga.	Apresenta a aplicação de nova técnica de modelagem baseada na cadeia de Markov.

Autores/Ano	Título	Justificativa	Resultados
Ortiz-Lozano, JM; Aparicio-Chueca, P; Triadó-Ivern, XM; Arroyo-Barrigüete, JL (2023)	Early dropout predictors in social sciences and management degree student.	Referência mais recente.	Logit e Redes Neurais/ Floresta Randômica com acurácia de 77% e 69%, respectivamente, em amostra de 3.583 da Universidade de Barcelona.
Da Re, L; Zago, G (2011)	Academic Tutoring and Dispersion: A Project of Research of the University of Padua (Italy).	Artigo utiliza programa de tutorias (dois anos de estudo) para minimizar evasão.	Compreensão de diferentes tipos de tutorias oferecidas na Itália.
Sommer, M; Dumont, K (2011)	Psychosocial factors predicting academic performance of students at a historically disadvantaged university.	Artigo da África do Sul, traz a perspectiva dos modelos aplicados em contexto de desvantagem histórica.	Identificação de fatores psicossociais e especificidades do modelo estudado no desempenho acadêmico.
Rodríguez-Gómez, D; Feixas, M; Gairín, J; Muñoz, JL (2012)	Understanding Catalan University dropout from a comparative approach.	Oferece extenso panorama de como a evasão tem sido estudada na América, Europa e Espanha.	Panorama da evasão universitária, em especial da Espanha e Catalunha e oferta de planos de ação para elevar a retenção.
Aina, C (2013)	Parental background and university dropout in Italy.	Autores mais citados.	Baixa renda como fator negativo de ingresso no ensino superior, sugerindo intervenção política.
Aina, C; Baici, E; Casalone, G; Pastore, F (2022)	The determinants of university dropout: A review of the socio-economic literature.	Revisão da literatura socioeconômica sobre os determinantes da evasão; Aina como autores mais citados	Persistência/desgaste universitário depende de combinação de fatores individuais, institucionais e econômicos, mediados pela integração ao sistema acadêmico.
Vallejos, CA; Steel, MFJ (2017)	Bayesian survival modelling of university outcomes.	Descrição de modelos aplicados.	Identificação das principais covariáveis para o risco de evasão.
Viloria, A; Padilla, JG; Vargas-Mercado, C; Hernández-Palma, H; Llinas, NO; David, MA (2019)	Integration of Data Technology for Analyzing University Dropout.	Árvore de Decisão, Rede Bayesiana, Redes Neurais, big data.	Índice de retenção de 78,3% por Árvore de Decisão. Acurácia de 76%, 75% e 83% para rede bayesiana, Árvore de decisão e rede neural, respectivamente.
Rodríguez-Muñiz, LJ; Bernardo, AB; Esteban, M; Díaz, I (2019)	Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques?	Método de aprendizado de máquina com abordagem holística, compatível com big data sets	Influência de variáveis pessoais e contextuais e a importância do desempenho acadêmico no primeiro ano.
Medina, EC; Chunga, CB; Armas-Aguirre, J; Grandón, EE (2020)	Predictive model to reduce the dropout rate of university students in Peru: Bayesian Networks vs. Decision Trees.	Redes Bayesianas x Árvore de Decisão	Redes bayesianas se comportam melhor que Árvores de Decisão em métricas de precisão, exatidão, especificidade e taxa de erro.
Santos, JD; Magalhoes, AMD; Real, GCM (2020)	Dropout management in brazilian educational policies: from	Artigo brasileiro, usa fonte do INEP e Capes;	Ações governamentais de controle à evasão tem caminhos distintos na

Autores/Ano	Título	Justificativa	Resultados
	undergraduate to <i>stricto sensu</i> postgraduate.	Universidade Federal Grande Dourados.	graduação e pós-graduação <i>stricto sensu</i> . Implementação de políticas de assistência estudantil e medidas avaliativas.
Pedraza-Navarro, I; González-Ramírez, T (2021)	Educational quality and dropout risk: a causal analysis of the university dropout phenomenon.	Análise causal da evasão; Universidade de Sevilla.	Variáveis independentes "idade", "estado civil" e "nº irmãos" influenciam na conclusão do ensino superior.
Flores, V; Heras, S; Julian, V (2022)	Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education.	Estudo comparativo entre 08 modelos preditivos baseados em métodos de data mining.	Modelo preditivo baseado em Floresta Randômica é o que apresenta maior precisão e robustez, de oito avaliados.
Flores, V; Heras, S; Julian, V (2022)	A New Methodological Framework for Project Design to Analyse and Prevent Students from Dropping Out of Higher Education.	Metodologia CRISP-DM; Mineração de Dados Educacionais; Metodologia de quadro lógico; Design de projeto.	Proposta de framework que mescla as seis fases do CRISP-DM com a primeira etapa do Quadro Lógico (LFM) para aumentar a retenção universitária.
Segura, M; Mello, J; Hernández, A (2022)	Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?	Sobre aprendizado de máquina para predição da evasão.	Múltiplas variáveis críticas, relevância para desempenho acadêmico. Aprendizado de máquina tem melhores resultados que Regressão Logística.
Gutierrez-Pachas, DA; Garcia-Zanabria, G; Cuadros-Vargas, E; Camara-Chavez, G; Gomez-Nieto, E (2023)	Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic, Socioeconomic, and Equity Factors through Machine Learning and Survival Analysis Methods in the Latin American Context.	Sobre processo de tomada de decisão sobre evasão com dataset longitudinal, diferentes métodos e dados robustos.	Ampliação do risco de evasão associado às variáveis: departamento acadêmico, sexo, e grupo socioeconômico.
Melucci, D; Marcuccio, M; Zappoli, S; Boga, C; Silva, L; Strocchi, E (2019).	The Results of QPCS as Part of an Entry Guidance with University Students. The Experience of the Scientific Degrees Plan (PLS) - Chemistry of the University of Bologna.	Utilização de Questionário de Percepção de Competências Estratégicas (QPCS) e PLS	Tutoria metacognitiva estratégica como abordagem didática inovadora para contrastar a evasão universitária.
Urbina-Nájera, AB; Camino-Hampshire, JC; Barbosa, RC (2020)	University dropout: Prevention patterns through the application of educational data mining.	Mineração de Dados Educacionais com aplicação de Árvore de Decisão e 27 fatores relevantes.	Aponta 27 fatores relevantes e 3 fatores principais e 7 padrões de comportamento encontrados por Árvore de Decisão.
Viloria, A; Sierra, DM; Samper, MG; Basto, WOC; Pichón, AR; Hernández-Palma, H; Orozco, VD; Kamatkar, SJ (2020).	Dropout-Permanence Analysis of University Students Using Data Mining.	Estudo da Universidade de Mumbai, Índia, ampliando o espectro de abrangência da pesquisa. Utilização de	Variáveis críticas: cursos aprovados, quantidade e resultados dos cursos frequentados, origem e idade de entrada do aluno.

Autores/Ano	Título	Justificativa	Resultados
		Árvore de Decisão, Redes Bayesianas.	
Velasco, JCP; Velasco, IMP; Irala, IAE (2020).	Analysis of student dropout at a public university in Bolivia.	Estudo boliviano, trazendo cenário socioeconômico de país em desenvolvimento, como o Brasil. Amostra de 46.347 estudantes.	Análise por meio de equações estruturais: fatores "Econômicos" e "Famíliares" são críticos na evasão estudantil, além de "Motivação", "Saúde", "Social" e "Acadêmico".

Fonte: Própria (2023).

Os artigos selecionados, além de apresentarem alta aderência ao tema da pesquisa conforme comprovado pelo cálculo de Kappa, também proporciona um amplo espectro de nuances da temática da evasão, permitindo que sua utilização agregue a pesquisa as perspectivas de diferentes metodologias, ferramentas e algoritmos variados, técnicas de classificação diversas, diferentes variáveis críticas, sugestões de programas de apoio acadêmico para gestão da evasão, além de englobar estudos nacionais e de diferentes continentes, autores e artigos mais citados, compondo um panorama bastante diverso.

Assim, conforme listado no Quadro 5, foram aprofundados os estudos tomando por base os 32 artigos destacados, que possibilitaram a elaboração de um modelo integrador para este trabalho (Figura 14).

Um modelo integrador é uma abordagem teórica e metodológica que busca reunir diferentes conceitos, teorias ou variáveis em uma estrutura coesa, permitindo uma análise mais abrangente e multidimensional de um fenômeno, sintetizando informações de diversas fontes ou áreas do conhecimento, proporcionando uma visão mais completa e detalhada do objeto de estudo ao integrar diferentes perspectivas. Um modelo integrador pode revelar interações complexas entre variáveis, ampliar a compreensão do problema e oferecer soluções mais eficazes e contextualizadas, além de facilitar a aplicação prática dos resultados.



Figura 14 - Modelo Integrador.

Fonte: Própria (2023).

O processo educativo é uma jornada complexa e progressiva, constituída por uma sequência de etapas fundamentais que moldam o desenvolvimento acadêmico e intelectual dos indivíduos. Inicialmente, este percurso se desdobra na educação fundamental, onde os alicerces do conhecimento são estabelecidos, proporcionando aos estudantes as bases essenciais para o aprendizado futuro.

Na sequência, o processo educativo avança para o Ensino Médio, uma fase que busca aprofundar e diversificar os saberes, preparando os estudantes para os desafios acadêmicos e profissionais que os aguardam. É neste período que se delineiam interesses específicos e se delineiam trajetórias educacionais mais especializadas.

O ápice deste processo é atingido nos aprendizados de nível superior, compreendendo tanto os cursos de graduação quanto os de pós-graduação. Os cursos de graduação representam uma imersão mais aprofundada em áreas específicas do conhecimento, proporcionando uma formação mais especializada e crítica. Paralelamente, a pós-graduação oferece oportunidades para a pesquisa avançada, a especialização profissional e a contribuição ativa para o avanço do saber em diversas áreas.

Assim, o encadeamento destas etapas contribui para a construção de uma base sólida de conhecimento ao mesmo tempo em que fomenta o desenvolvimento

intelectual, a capacidade crítica e a especialização nas áreas de interesse. O processo educativo, ao abranger desde as bases fundamentais até os patamares mais elevados do ensino, desempenha um papel crucial na formação de indivíduos preparados para enfrentar os desafios de uma sociedade cada vez mais complexa e dinâmica.

Este processo encadeado faz com que tanto o aprendizado quanto o desempenho de uma etapa se reflitam nos estágios seguintes, assim como peças que se encaixam uma formam uma figura única de formação e aprendizagem.

Neste conceito de trajetória, é possível identificar marcos críticos na incidência da evasão, ocorrendo de forma mais acentuada no primeiro e no segundo anos de formação superior. Múltiplos são os fatores apontados para esta ocorrência e também as abordagens que podem ser utilizadas para seu estudo e para a proposição de ações que visam reduzir estes índices de abandono.

4 MÉTODOLOGIA DA PESQUISA

Este capítulo se dedica a descrever os procedimentos desenvolvidos para atender os objetivos, geral e específicos, propostos neste estudo. Na etimologia, método é um processo de pesquisa organizado de forma sistemática e lógica, um conjunto de meios que permitem alcançar um objetivo (Ferreira, 1999).

Assim, é apresentado o detalhamento do caminho percorrido neste processo, seus formatos e etapas, contemplando o tipo de pesquisa e seu desenho, o local e o objeto da pesquisa, a coleta de dados e as opções para a análise realizada

4.1 Tipo de Pesquisa

Para delimitar a taxonomia da pesquisa aqui proposta, são utilizadas as classificações propostas por Vergara (1990), utilizando-se como guia dois critérios básicos: quanto aos fins e quanto aos meios da pesquisa.

O método de pesquisa adotado para este trabalho é uma abordagem quantitativa, quanto aos fins, é de natureza aplicada, realizando o levantamento teórico para posterior aplicação prática com a utilização de análises por Regressão Logística binominal e da modelagem de Árvore de Decisão nos dados de evasão da UF SB, segundo as variáveis críticas estudadas, visando contribuir com as práticas da instituição a que se destina.

Quanto aos objetivos gerais este trabalho se propõe a realizar uma pesquisa Exploratória Explicativa para analisar a evasão acadêmica nos cursos de graduação da Universidade Federal do Sul da Bahia (UF SB) a partir da utilização da abordagem de Mineração de Dados Educacionais.

A pesquisa exploratória é aquela inédita, que busca aprofundar o conhecimento de determinada realidade investigada, dentro de uma perspectiva específica. Este tipo de pesquisa se propõe a levantar indagações sobre o problema estudado. A pesquisa explicativa busca no histórico dos dados a elucidação, as causas do problema investigado (Vergara, 1990).

Quanto aos procedimentos, ou meios, de acordo com a classificação de Vergara (1990) este trabalho realizou inicialmente um estudo bibliométrico, composto

pelo referencial teórico e revisão bibliográfica com aplicação do TEMAC (Mariano e Rocha, 2017). A segunda parte será documental, investigando dados secundários, quantitativos, relativos aos números da evasão nos cursos de graduação da UFSB, fornecidos pela própria instituição para estudar as variáveis críticas relacionadas ao problema.

Deste modo, a terceira parte adotará como estratégia de pesquisa o estudo de caso (Vergara, 1990) aplicado à UFSB, utilizando dados internos da Diretoria de Percursos Acadêmicos (DPA), com análise por Regressão Logística e modelagem por Árvore de Decisão, para apontar possíveis intervenções para a gestão da evasão (Figura 15).



Figura 15 – Taxonomia da Pesquisa

Fonte: Própria (2023)

Com base na taxonomia adotada e na estratégia metodológica definida, esta pesquisa visa contribuir para uma compreensão mais aprofundada do fenômeno da evasão acadêmica nos cursos de graduação da UFSB. Por meio da abordagem

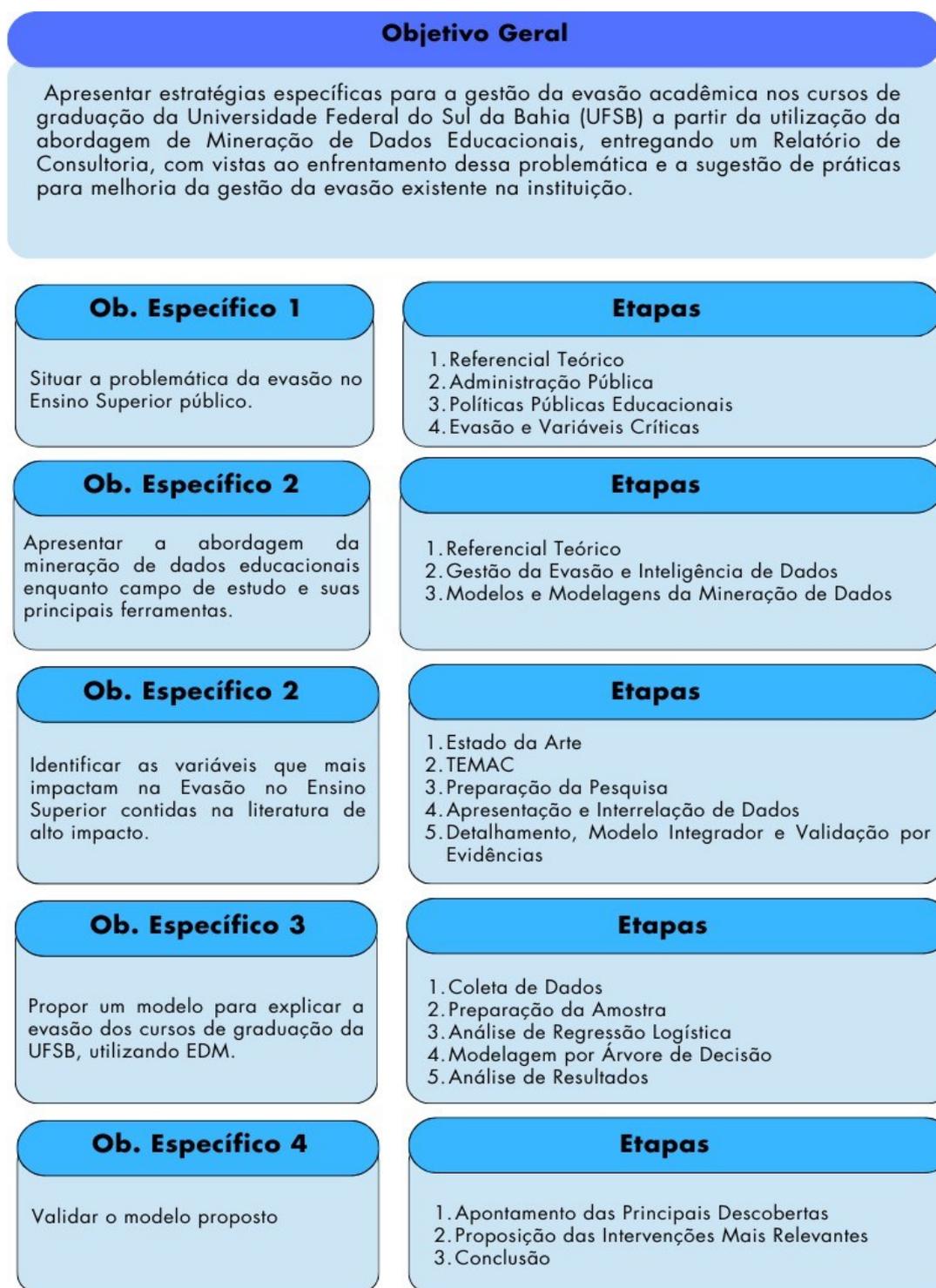
quantitativa, Regressão Logística e da modelagem de Árvore de Decisão, espera-se identificar padrões e fatores críticos que influenciam a evasão, permitindo que a instituição desenvolva práticas mais eficazes de gestão e intervenção.

O estudo de caso específico da UFSB, aliado à aplicação de técnicas de Mineração de Dados Educacionais, busca fornecer um panorama explicativo, ao mesmo tempo em que apresenta subsídios práticos para a implementação de soluções voltadas à melhoria das taxas de retenção.

4.2 Desenho da Pesquisa

O desenho da pesquisa é uma representação gráfica, que fornece uma estrutura metodológica clara que orientará a condução do estudo. Este recurso permite uma visualização rápida e global das etapas a serem executadas, conferindo validade e confiabilidade ao método escolhido (Quadro 6).

Quadro 6 – Desenho da Pesquisa



Fonte: Própria (2024).

Todos os objetivos específicos e etapas estabelecidas foram cumpridas e compõem o presente trabalho.

A Universidade Federal do Sul da Bahia foi fundada em 2013, com sede em Itabuna, multi *campi*, conta com 5.560 estudantes distribuídos entre 52 cursos de primeiro e segundo ciclos, segundo dados de 2021 da própria instituição, distribuídos equitativamente nos *campi* (Figura 17).

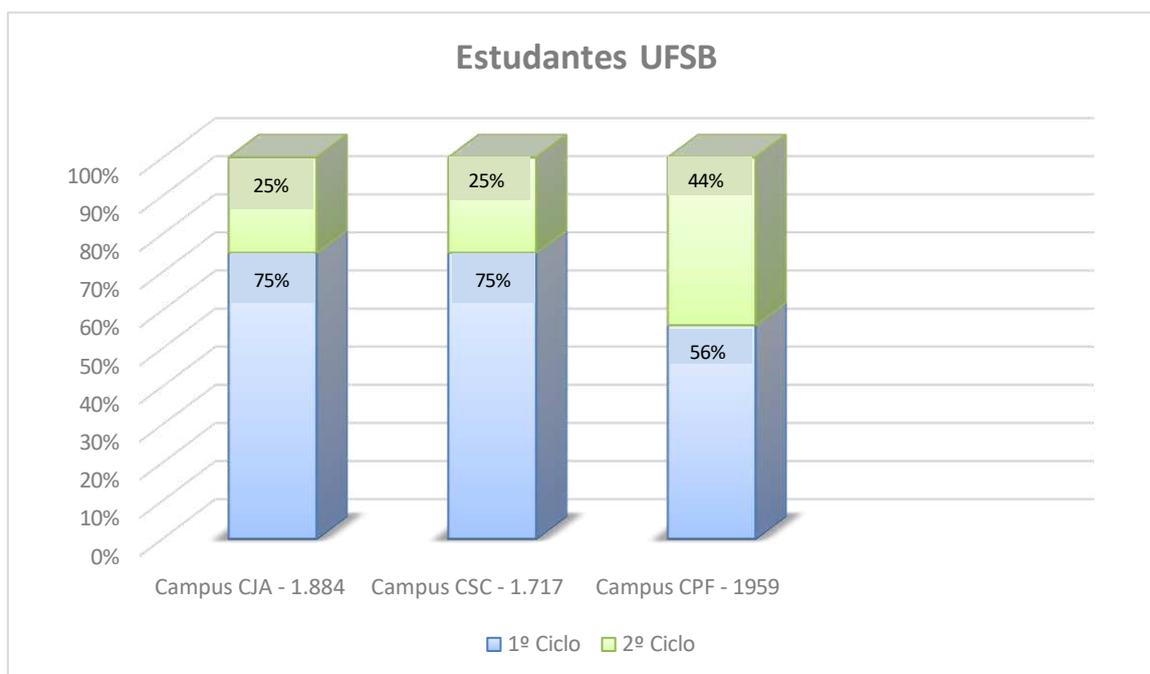


Figura 17 – Estudantes UFSB.
Fonte: Relatório de Gestão UFSB (2022).

A Bahia, segundo dados do IBGE (2021) registra um Índice de Desenvolvimento Humano de 0,69 (IBGE, 2021). O IDH é uma ferramenta de estatística utilizada pelo Programa das Nações Unidas para medir o desenvolvimento de um local, considerando aspectos como saúde, renda e educação. Considerando que não é possível avaliar uma região por um único índice, esta medida pode sugerir que alguns progressos e confortos já foram adquiridos localmente, mas que ainda existe margem para significantes melhorias na região.

Como reflexo da situação econômica da região, a UFSB assume o papel social de proporcionar ensino público superior, atingindo marcas singulares de popularização da educação com índices de estudantes cotistas que chegam aos 85%, o que torna seu público acadêmico bastante singular (UFSB, 2023).

4.4 Objeto da Pesquisa

O objeto da pesquisa está delimitado em torno dos dados da evasão nos cursos de graduação da Universidade Federal do Sul da Bahia (UFSB) utilizando a modelagem de Árvore de Decisão na busca da identificação de decisões que possam impactar a gestão desta problemática, oferecendo informações imparciais para auxiliar nos processos de gestão da educação.

4.5 Coleta e Tratamento de Dados

A pesquisa teórica realizada, englobando a fundamentação teórica e o Estado da Arte, foram primordiais para, inicialmente, compreender o universo maior da Administração Pública e da Educação e, num segundo momento, compreender as dificuldades enfrentadas especificamente na Educação. Dentre os diversos desafios identificados, a problemática da evasão se destacou por suas implicações sociais, financeiras e de governança. Para esta pesquisa foi utilizada exclusivamente a base de dados da *Web of Science*, em sua coleção completa.

Para o estudo acerca dos índices de evasão dos cursos de graduação da Universidade Federal do Sul da Bahia foram solicitados dados institucionais, por e-mail, à Diretoria de Percursos Acadêmicos (DPA), instância responsável pela coleta primária destas informações e responsável pela gerência de dados acadêmicos da Universidade e complementados pela Superintendência de Tecnologia e Informação (STI).

A amostra inicial é composta por uma base de dados secundários com 11.626 linhas relacionadas aos cursos de graduação e 502.403 relacionadas as disciplinas, com informações demográficas, sociais e acadêmicas sobre todos os estudantes de graduação da Universidade Federal do Sul da Bahia, incluindo todo o seu percurso acadêmico. Selecionamos 23 variáveis que serão nomeadas, definidas e tipificadas na Tabela 2.

Tabela 2. Descrição da amostra

Nome da Variável	Definição da Variável	Tipo da Variável
Data de Nascimento	Data de nascimento do estudante.	Quantitativa Contínua
Naturalidade	Local de nascimento do estudante.	Catagórica Nominal
Curso	Curso de graduação de vínculo do estudante.	Catagórica Nominal
Status da Matrícula	Se Ativo, Trancado, Cancelado, Formando ou Formado.	Catagórica Nominal
Período Letivo Inicial	Marco do início do vínculo institucional.	Catagórica Ordinal
Período Letivo Atual	Período letivo no momento da coleta do dado.	Catagórica Ordinal
Período Letivo de Saída	Período letivo em que o estudante deixa a instituição, como egresso ou evadido.	Catagórica Ordinal
Suspensões	Solicitações de suspensão de matrícula.	Catagórica Nominal
Prorrogações	Solicitações de prorrogação de prazo para integralização do curso.	Catagórica Ordinal
Tipo de Saída	Formado ou Evadido.	Catagórica Nominal
Data de Saída	Data de encerramento do vínculo com a instituição.	Quantitativa Contínua
CCs Cursados	Componentes Curriculares já realizados no percurso acadêmico.	Catagórica Nominal
CCs Cursando	Componentes Curriculares em curso no momento da coleta dos dados.	Catagórica Nominal
Status do CC Cursado	Aprovado por média, Aprovado por nota, Cancelado, Reprovado por média, Reprovado por nota, Reprovado por falta, Cumpriu.	Catagórica Nominal
Frequência	Frequência acadêmica nos CC.	Quantitativa Contínua
Atividade Complementar	Atividades completares cursadas.	Catagórica Nominal
CH Integralizada	Carga Horária cumprida do curso.	Quantitativa Discreta
CH Pendente	Carga Horária restante para integralizar o curso.	Quantitativa Discreta
Índice Acadêmico (CR)	Índice de desempenho Acadêmico global.	Quantitativa Contínua
Semestres Cursados	Quantidade de semestres cursados.	Quantitativa Discreta
Se Evadiu	Status final do discente.	Catagórica Nominal
Nota Média CCs Básicos	Média das notas dos CCs comuns aos cursos de graduação.	Quantitativa Contínua
Nota Média CC Específicos	Média das notas dos CCs específicos do curso do estudante.	Quantitativa Contínua

Fonte: Própria (2023).

A amostra foi preparada sob três critérios iniciais, observando a curva de normalidade, identificando dados faltantes e os outliers (dados que se diferenciam drasticamente do restante da amostra, podendo induzir o resultado a erro).

Inicialmente foi realizado um estudo por Regressão Logística sob duas perspectivas: por curso e por disciplinas. Como os estudos apontaram achados que

destoavam da literatura científica estudada sobre o tema, foi decidido ampliar a pesquisa, complementando a análise por Árvore de Decisão.

Para a modelagem por Árvore de Decisão a amostra foi dividida em dois grupos: um grupo de treino e um grupo de teste. Os programas usados foram o *Rapid Miner*, e *SmartPLS*.

Como parte dos resultados desta pesquisa, foi desenvolvido um *dashboard* interativo que organiza e apresenta os dados coletados de maneira visual e dinâmica. Este dashboard permite a visualização clara das informações relacionadas à evasão acadêmica na UFSB, facilitando a análise de padrões, tendências e fatores críticos que influenciam o problema. Com ele, gestores e pesquisadores poderão explorar os dados de forma intuitiva, auxiliando na tomada de decisões e na criação de estratégias mais eficazes para a gestão da evasão.

O tratamento de dados seguiu o fluxo CRISP-DM para organizar os dados, que estão expostos no próximo capítulo para melhor linearidade.

A análise dos dados provenientes deste estudo proporcionará informações para o desenvolvimento de práticas e estratégias específicas, alinhadas às características da UFSB, visando aprimorar a gestão das iniciativas de combate à evasão, o que envolve a criação de um Produto Técnico-Tecnológico (PTT) na forma de um Relatório de Consultoria. Este relatório apresenta uma proposta de ações mais impactantes para gerenciar a evasão nos cursos de graduação da UFSB, além de sugestões de ajustes nas políticas institucionais.

5 ANÁLISE SITUACIONAL

Neste capítulo são apresentados os resultados apurados da análise dos dados secundários, fornecidos pela Universidade Federal do Sul da Bahia (UFSB), para análise da evasão nos cursos de graduação da instituição, sua correlação com a literatura levantada nos capítulos de referencial teórico e Estado da Arte (pelo método TEMAC) e um painel de visualização dos resultados.

Para o processo de análise dos dados, foi adotado o modelo do CRISP-DM e seu respectivo conjunto de etapas, conforme descrito no referencial teórico. Estas etapas serão relatadas a seguir, constituindo o roteiro de análise dos dados, resultados e avaliação, finalizando este capítulo com a discussão proposta.

5.1 Compreensão do Negócio

Trata-se da compreensão do problema investigado na pesquisa, neste caso, de forma mais abrangente a evasão no Ensino Superior, e de forma mais específica a evasão nos cursos de graduação da Universidade Federal do Sul da Bahia.

Muito deste tema já foi reportado nos capítulos do Referencial Teórico e do Estado da Arte. Assim, cabe aqui um breve relato dos principais pontos identificados nestes capítulos.

A evasão acadêmica é um fenômeno complexo e multifatorial, com variáveis acadêmicas, demográficas e financeiras frequentemente apontadas como críticas. No entanto, não há consenso sobre quais são as mais relevantes. Estudos internacionais destacam a importância de uma abordagem holística para lidar com a evasão, considerando fatores como habilidades e interesses dos estudantes, integração ao sistema acadêmico e contexto socioeconômico, o que sugeriria uma abordagem qualitativa associada aos métodos quantitativos.

Assim, a evasão acadêmica apresenta-se como um fenômeno global multifacetado, afetando instituições de ensino de diferentes países de maneira variada. Estudos realizados na Europa, Canadá e Chile demonstram que fatores como

desigualdades econômicas, preparo escolar insuficiente durante o ensino médio, desconexão entre o ensino superior e o mercado de trabalho, além da falta de suporte emocional e financeiro, são causas recorrentes.

No Canadá, por exemplo, o número de alunos em cursos obrigatórios do primeiro ano e as escolhas econômicas dos estudantes influenciam a evasão (Montmarquette, 2001). Na Itália e Espanha, estudos reforçam a interdependência de fatores individuais, institucionais e econômicos, destacando a necessidade de ações de integração acadêmica e social para reduzir o abandono (Aina, 2022; Castro-Lopez, 2022).

No Reino Unido, embora o acesso ao ensino superior esteja desvinculado de fatores econômicos, a permanência e conclusão dos cursos ainda dependem fortemente de variáveis socioeconômicas (Vignoles, 2009). Em Portugal, a Universidade de Aveiro identificou sete variáveis críticas, como taxa de sucesso acadêmico e satisfação com a vida, para monitorar o risco de evasão (Ferreira, 2020). No Chile, variáveis como notas do ensino médio e quintil econômico desempenham papel importante no abandono dos cursos, especialmente nas áreas de engenharia, onde as taxas de evasão são mais elevadas (Palácios, 2021).

No Brasil, a situação é amplificada por questões estruturais, socioeconômicas e regionais, impactando diretamente a democratização do acesso à educação e a qualificação profissional. Embora políticas públicas tenham contribuído para a redução da evasão, a relação direta entre estas políticas e a diminuição dos índices ainda não é plenamente comprovada, dada a complexidade do fenômeno (Santos, 2020).

A Universidade Federal do Sul da Bahia (UFSB), por exemplo, enfrenta desafios relacionados à diversidade de perfis estudantis, carências infra estruturais e dificuldades de adaptação acadêmica, o que demanda ações diversas e assertivas para mitigar a evasão. A evasão na UFSB tem repercussões significativas tanto no percurso dos discentes quanto nos objetivos institucionais da universidade, que busca oferecer uma educação inclusiva e de qualidade. A instituição, sendo nova (10 anos) e com alto índice de adesão a políticas de cotas (até 85%), enfrenta o desafio de manter os alunos vinculados, especialmente aqueles em situação de vulnerabilidade socioeconômica.

5.2 Compreensão dos Dados

Os dados foram coletados junto à Universidade Federal do Sul da Bahia, cedidos pela Diretoria de Processos Acadêmicos (DPA) e pela Superintendência de Tecnologia e Informação, em planilhas de extensão *.xls* para programa da *Microsoft Excel®*. Foram recebidos dados de 2014 a 2023, divididos em duas perspectivas de análise, por cursos de graduação e por componentes curriculares (disciplinas), conforme os Quadros 7 e 8 apresentados a seguir:

Quadro 07 – Dados dos alunos por Curso

Variável	Descrição	Tipo
Planilha 01 – Alunos vs. Cursos		
Matrícula	Número de identificação do estudante junto à instituição de ensino.	Categórica nominal
Data de Nascimento	Título de identificação do componente curricular (disciplina).	Categórica nominal
Naturalidade	Cidade de nascimento do estudante	Categórica nominal
UF de Naturalidade	Unidade federativa de nascimento do estudante	Categórica nominal
Curso de Ingresso	Curso de ingresso do estudante para a matrícula analisada nos dados.	Categórica nominal
Local de Ingresso	Ano de referência do período letivo em que o estudante foi matriculado no componente curricular.	Categórica ordinal
Curso Atual	Período de referência e que o estudante cursou o componente curricular.	Categórica ordinal
Campus/CUNI	Quantidade de faltas do estudante registradas na disciplina.	Quantitativa discreta
Status de Matrícula	Percentual de frequência do estudante durante o período letivo.	Quantitativa contínua
Ano/Período de Ingresso	Carga horária total do componente curricular.	Quantitativa discreta
Período Letivo Atual	Período letivo no momento da coleta dos dados.	Categórica ordinal
Data de Saída	Data de encerramento do vínculo com a instituição.	Quantitativa contínua
Data de Cancelamento	Data de cancelamento da matrícula do estudante.	Quantitativa contínua
Quantidade de Trancamentos	Número de ocorrência de solicitação de trancamento do curso atual.	Quantitativa discreta
Prorrogações	Solicitações de prorrogação de prazo para integralização do curso.	Categórica ordinal
Carga Horária (CH) Total	Carga horária cumprida do curso de graduação.	Quantitativa discreta
Coefficiente de Rendimento (CR)	Índice de desempenho acadêmico global.	Quantitativa contínua

Fonte: Própria (2024). Dados fornecidos pela UF SB.

Inicialmente foi realizada uma Regressão Logística para saber os fatores que mais estavam associados a evasão e quais não influenciavam significativamente o resultado. Achadas as variáveis críticas, foi percebido alguns resultados diferentes da literatura científica sobre o tema.

Assim, em seguida, foi feita a análise destes dados por meio de uma Árvore de Decisão, que se propõe a traçar um panorama da evasão em cada curso de graduação listado, identificando não apenas os possíveis fatores que mais contribuem para este cenário, mas entregando uma ordem de como estes fatores se apresentam, organizando-os e permitindo um processo decisório.

O Quadro 08 traz os dados dos estudantes em função dos componentes curriculares, o que amplia a análise anterior, considerando que componentes curriculares são comuns a vários cursos, possibilitando uma análise que extrapola o universo mais uniforme proporcionado pela mesma graduação.

Quadro 08 – Dados dos alunos por Disciplina

Variável	Descrição	Tipo
Planilha 02 – Alunos vs. Disciplinas		
Matrícula	Número de identificação do estudante junto à instituição de ensino.	Categórica nominal
Nome do Componente	Título de identificação do componente curricular (disciplina).	Categórica nominal
Tipo de Integralização		
Média Final	Média final do estudante no componente curricular analisado.	Quantitativa contínua
Situação da Matrícula		Categórica nominal
Ano	Ano de referência do período letivo em que o estudante foi matriculado no componente curricular.	Categórica ordinal
Período	Período de referência em que o estudante cursou o componente curricular.	Categórica ordinal
Número de Faltas	Quantidade de faltas do estudante registradas na disciplina.	Quantitativa discreta
Porcentagem de Frequência	Percentual de frequência do estudante durante o período letivo.	Quantitativa contínua
Carga Horária (CH) Total	Carga horária total do componente curricular.	Quantitativa discreta

Fonte: Própria (2024). Dados fornecidos pela UFSB.

A análise destes dados permite identificar se existe um comportamento de evasão que independe do curso escolhido ou outras variáveis que possam ser determinantes e comuns a mais de um curso de graduação. Estes dados também oportunizam a

elaboração de uma Árvore de Decisão, com o mesmo propósito de permitir o mapeamento ordenado dos fatores e possibilitar as tomadas de decisão.

Não foram utilizadas variáveis que identificassem os estudantes, utilizando como referência sempre o número de matrícula, a anonimização garante que não haja riscos à privacidade dos participantes. Desta forma, esta pesquisa não é classificada como procedimento de investigação que envolve seres humanos, o que dispensa a necessidade de submissão ao Comitê de Ética em Pesquisa (CEP).

A apresentação dos dados foi feita em plataforma digital (Figura 18), no formato de *dashboard*, no endereço virtual <https://www.unbdatalab.org/dashevasao>, que permite que ocorram atualizações periódicas com nova entrada de dados, possibilitando, assim, que a UFSB continue reformulando as análises, mesmo após encerrada desta pesquisa. Este é o Produto Técnico-Tecnológico desenvolvido e será aprofundado no tópico 5.6.1.1.

PROFIA P
MESTRADO PROFISSIONAL EM
ADMINISTRAÇÃO PÚBLICA

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

UFSB
UNIVERSIDADE FEDERAL
DO SUL DE BAHIA

DataLab
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Evasão Analytics - Retention Radar

Cursos

Evasão

Disciplinas

Mapa

Programa de Pós-Graduação do Mestrado Profissional em Administração Pública
Universidade Tecnológica Federal do Paraná.
Discente: Anna Carla Freire Luna Campêlo Bastos
Orientador: Prof. Dr. Oséias Santos de Oliveira (UTFPR)
Coorientador: Prof. Dr. Ari Melo Mariano (UNB)

UNIVERSITY DROPOUT

Figura 18 – Plataforma digital *dashboard*.

Fonte: Própria (2024).

Esta opção foi adotada para que fosse possível conhecer os dados por inteiro, permitindo identificar discrepâncias que poderiam ser tratadas na fase de preparação dos dados minimizando a possibilidade de resultados distorcidos.

5.3 Preparação dos Dados

A preparação dos dados é uma das etapas mais críticas no processo de Mineração de Dados, pois envolve a transformação dos dados brutos em um formato adequado para análise. No contexto da metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*), esta fase antecede a modelagem e visa garantir que os dados estejam limpos, integrados, formatados e estruturados de forma a maximizar a eficiência e a precisão dos modelos preditivos subsequentes.

Aqui estão os principais processos de organização dos dados, que incluem a identificação e tratamento de elementos críticos como dados repetidos, valores ausentes (*missing data*), e outliers.

Os dados repetidos são registros duplicados que precisam ser removidos para evitar distorções nas análises. Os dados faltantes (*missing*) exigem estratégias de preenchimento ou exclusão, garantindo que a ausência de informações não comprometa os resultados. Os outliers são dados extremos, muito distantes da média de comportamento ou valor mediano e que requerem um olhar sobre sua permanência ou exclusão.

A primeira etapa da preparação foi aplicada na planilha de cursos (Figura 19) com a limpeza dos dados, retirada de um número de matrícula que constava repetido na base de dados e em seguida, de linhas com dados em branco ou incompletos:

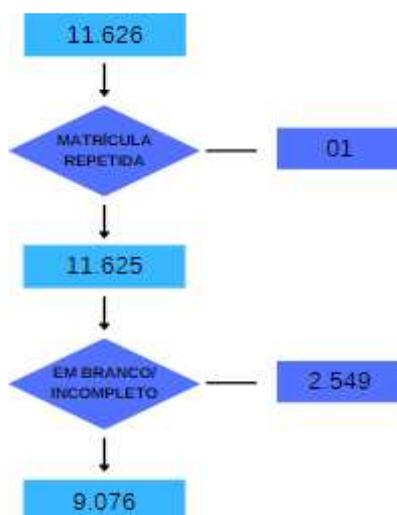


Figura 19 – Limpeza de dados da planilha de Cursos.

Fonte: Própria (2024).

A segunda etapa realizada na preparação do banco de dados foi a atribuição de novas etiquetas às variáveis que categorizavam a situação do estudante no momento da coleta dos dados, reduzindo à dois novos rótulos, num processo de enriquecimento de dados (Quadro 9).

Quadro 9 – Atribuição de novas etiquetas

Situação	Nova Etiqueta
Concluído	Não evadiu
Cancelado	Evadiu
Ativo	Não evadiu
Formado	Não evadiu
Formando	Não evadiu
Excluído	Evadiu

Fonte: Própria (2024).

Para isto, adotou-se como conceito de “não evadiu” todas as situações em que o estudante concluiu o curso de graduação em que estava matriculado, estava em processo de conclusão ou cursando. Todas as hipóteses em que o estudante suspendeu as atividades, cancelou a matrícula ou encerrou seu vínculo com a instituição de ensino foram etiquetadas como “evadiu” neste estudo.

Foram localizados os outliers da amostra, ou seja, os valores ou observações no conjunto de dados que se diferenciavam significativamente dos demais pontos, sendo considerados extremos ou atípicos, pois estavam muito distantes da maioria dos outros valores. Isto pode ocorrer tanto em variáveis numéricas quanto categóricas, sendo que em análises estatísticas e de mineração de dados, os outliers podem afetar a média, a variância e outras métricas.

Assim, a Figura 20 apresenta a distribuição dos coeficientes de rendimento (CR) dos estudantes da UFSB, destacando possíveis padrões relacionados à evasão. Utilizou-se um gráfico tipo boxplot, por ser uma ferramenta estatística eficaz para visualizar a dispersão e identificar outliers nos dados.

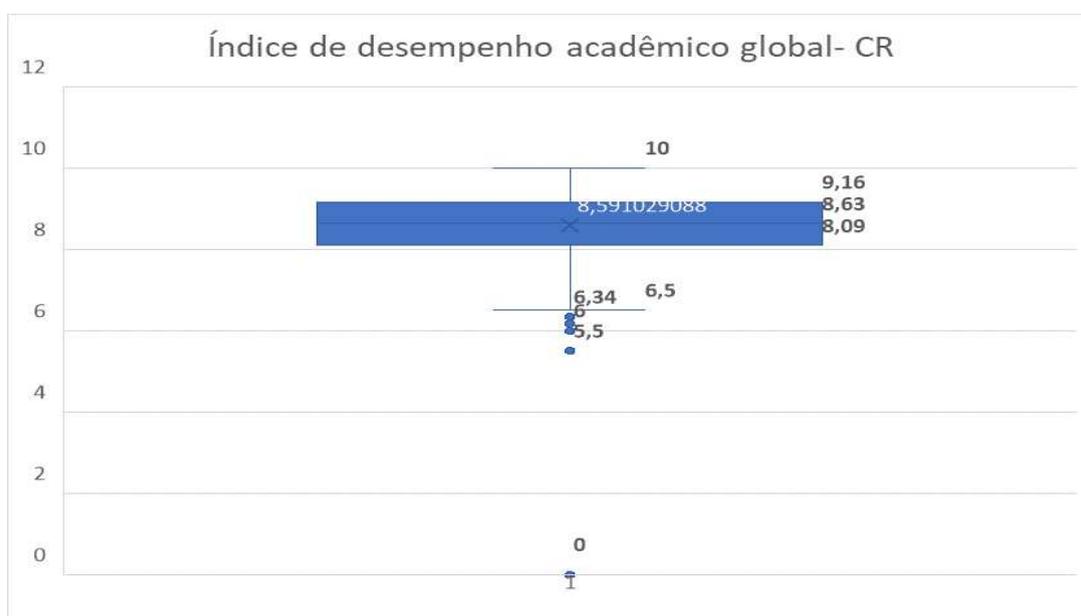


Figura 20 – Índice de desempenho acadêmico global - CR

Fonte: Própria (2024).

No boxplot, a "caixa" representa o intervalo interquartil (IQR), ou seja, os 50% centrais dos dados, enquanto a linha interna indica a mediana. As "extensões" (*whiskers*) mostram a amplitude dos dados, exceto possíveis outliers, que são indicados por pontos fora dos limites das extensões.

Em seguida foi realizada a limpeza dos dados da planilha por disciplinas (Figura 21), retirando um conjunto informações conforme indicado (matrículas excluídas e indeferidas), sendo também retirado da amostra as notas entre 0 e 0,4 pela inconsistência que apresentavam ao serem incorporadas nas análises:

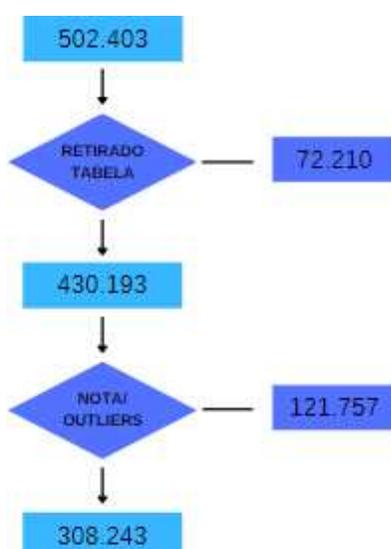


Figura 21 – Limpeza de dados da planilha de Disciplinas.

Fonte: Própria (2024).

Após os procedimentos de limpeza, com a retirada dos rótulos apresentados no Quadro 7 e as notas de CR outliers, a amostra passou a ser constituída por 308.243 linhas a serem trabalhadas.

Na identificação dos outliers no boxplot de médias, foram percebidos pontos muito afastados da maioria dos valores e que poderiam influenciar métricas como a média e a variância, impactando a interpretação dos resultados, sendo especialmente relevante em análises estatísticas, pois eles podem indicar variabilidade natural, erros

de medição ou características específicas do fenômeno estudado. A Figura 22 permite a visualização dos outliers pelo gráfico de boxplot das médias dos estudantes.

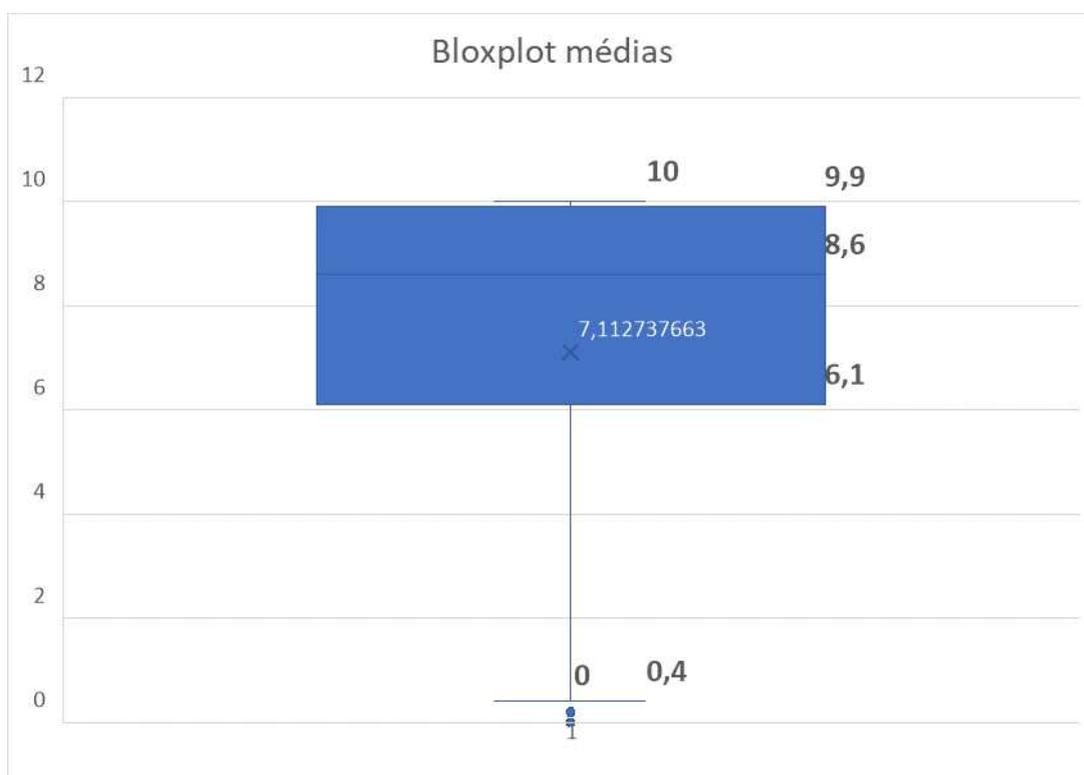


Figura 22 – Boxplot de Médias.

Fonte: Própria (2024).

Ao concluir a análise dos outliers na amostra das médias em disciplinas, é importante reconhecer que, embora estes pontos atípicos existam e revelem casos pontuais, sua incorporação nos resultados gerais pode distorcer o comportamento padrão dos dados. Estes outliers representam exceções que, se incluídas sem a devida cautela, poderiam desviar a interpretação e prejudicar a análise mais ampla das tendências de evasão.

No entanto, eles indicam situações específicas que podem merecer uma investigação separada e devem ser tratados como exceções que podem enriquecer o estudo em termos qualitativos, mas sua influência sobre o comportamento geral dos dados precisa ser minimizada para garantir que as conclusões sejam representativas da maioria dos casos analisados.

Os dados também passaram por um processo de enriquecimento, ou seja, informações adicionais e relevantes foram incorporadas a um conjunto de dados existente para aumentar seu valor, qualidade e precisão. Este processo visa aprimorar as análises, oferecendo um panorama mais completo e contextualizado sobre os dados originais.

Desta forma, na preparação dos dados da planilha por componente curricular, foram criadas novas etiquetas para os dados coletados, agrupando as variáveis e identificando dados inconsistentes que foram excluídos, sob a etiqueta “Retirado” (Quadro 10):

Quadro 10 – Enriquecimento dos dados

Situação	Nova etiqueta
Aguardando Deferimento	Aguardando
Aprovado por Média	Aprovado
Aprovado por Nota	Aprovado
Cancelado	Evasão
Cumpriu	Aprovado
Desistência	Evasão
Dispensado	Aprovado
Em Espera	Aguardando
Excluída	Retirado
Indeferido	Retirado
Matriculado	Aguardando
Null	Retirado
Reprovado por Faltas	Reprovado
Reprovado por Média	Reprovado
Reprovado por Média e por Faltas	Reprovado
Reprovado por Nota	Reprovado
Reprovado por Nota e Falta	Reprovado
Trancado	Aguardando
Transferido	Retirado

Fonte: Própria (2024).

A adição de novos atributos permitiu uma compreensão mais completa dos padrões e comportamentos, levando a decisões mais informadas, segmentar grupos de maneira mais detalhada, tornando os modelos preditivos mais robustos, aumentando os níveis de acurácia.

5.4 Modelagem

A modelagem de dados desempenha um papel crucial no entendimento de fenômenos complexos, como a evasão no ensino superior. No contexto desta pesquisa, foram utilizadas técnicas de Mineração de Dados para explorar os padrões e relações subjacentes que contribuem para este fenômeno, a partir de dados coletados de estudantes da UFSB e suas trajetórias acadêmicas. O foco central da análise está em determinar como diferentes variáveis influenciam o status dos estudantes, sejam os componentes curriculares ou no curso como um todo.

Existem três abordagens de análise possíveis: a supervisionada, a semi supervisionada e a não supervisionada. A análise supervisionada, que é o foco principal desta seção, foi escolhida por possibilitar o uso de dados de entrada já conhecidos (como características dos alunos, histórico acadêmico, etc.) e dados de saída previamente estabelecidos (status de aprovação ou evasão). O objetivo desta abordagem é identificar como os dados de entrada se relacionam com os resultados acadêmicos, ou seja, de que forma os fatores analisados contribuem para o desfecho final do estudante.

Os resultados da pesquisa foram obtidos em duas etapas principais de tratamento de dados. Primeiramente, utilizou-se a Regressão Logística para uma análise inicial das variáveis mais relevantes para a evasão via o software SmartPLS. Em seguida, a plataforma *RapidMiner* foi utilizada para construir uma Árvore de Decisão, permitindo uma visualização clara dos principais fatores que influenciam o fenômeno.

Estas análises não só fornecem uma compreensão detalhada dos fatores que afetam a evasão no ensino superior, mas também possibilitam a identificação de padrões previsíveis, que podem ser úteis para a tomada de decisões institucionais e a criação de estratégias de intervenção.

5.4.1 Análise por Regressão Logística

A análise por Regressão Logística é uma técnica estatística amplamente utilizada para modelar e prever a probabilidade de ocorrência de um evento binário (Fávero, 2017), ou seja, um resultado que pode ter dois valores possíveis, como "sucesso/falha", "aprovado/reprovado" ou "evadiu/não evadiu", sendo especialmente útil em contextos em que o desfecho que se deseja prever é categórico e dicotômico. Segundo Han (2022), a Regressão Logística não apenas tem a capacidade preditiva de um resultado binário, mas também informa o quão confiável é esta predição.

Os coeficientes β na Regressão Logística variam de $-\infty$ a $+\infty$, mas a probabilidade resultante da regressão estará sempre entre 0 e 1. Esta probabilidade de um evento acontecer, com o resultado expresso em termos de *odds* (razão de chances) ou de uma probabilidade entre 0 e 1. O *intercept* pode variar de $-\infty$ a $+\infty$, assim como os outros coeficientes da regressão. O intercepto na Regressão Logística representa a *log-odds* do evento ocorrer quando todas as variáveis explicativas são zero. Ele afeta diretamente a probabilidade prevista para o evento base (quando $X_1, X_2, \dots, X_k=0$) e pode ser interpretado em termos de *odds* (razão de chances) ou convertido em probabilidade usando a função logística.

A Regressão Logística é, especificamente, interessante por permitir interpretar diretamente as relações entre variáveis preditoras e o resultado alcançado e como uma variável específica afeta a probabilidade de ocorrência de um evento (*odds ratio*). A Regressão Logística tem a versatilidade de poder ser usada com variáveis preditoras de tipos diferentes e é menos suscetível a outliers que outros métodos de regressão.

Neste estudo, foram utilizadas múltiplas variáveis independentes (variáveis preditoras) quantitativas contínuas e discretas, para prever o desfecho binário "Evadiu" ou "Não Evadiu". Cada variável preditora tem um coeficiente estimado, que representa a força e a direção da associação (diretamente ou inversamente proporcional) entre a variável e o desfecho.

A primeira Regressão Logística tratou das variáveis CR, Distância, Idade e Permanência em função do curso do estudante e apresentou os seguintes resultados (Tabela 3):

Cursos:

Total de dados - cursos: **9.076**

Análise de Regressão Logística:

	Coeficientes	SE	Estatística z	Wald	Valor P	Exp(Coeficientes)	Suportada?
CR	-0,362	0,030	-12,176	148,266	0,000	0,697	Sim
Distância	-0,104	0,056	-1,870	3,499	0,061	0,901	Não
Idade	-0,005	0,003	-1,677	2,813	0,093	0,995	Não
Permanência	0,139	0,004	34,189	1168,882	0,000	1,150	Sim
Intercept	1,852	0,271	6,847	46,881	0,000	6,374	Sim

Fonte: Própria (2024).

Quando:

- **Coeficientes:** Indicam o impacto de cada variável sobre o logaritmo das chances (*log odds*). Se o coeficiente é negativo, a variável reduz as chances do evento acontecer (inversamente proporcional), e se for positivo, aumenta as chances (diretamente proporcional).
- **SE (Erro padrão):** Mede a precisão da estimativa do coeficiente. Um erro padrão menor indica uma estimativa mais precisa, logo, são desejáveis SEs que tendam a zero.
- **Estatística z:** A razão entre o coeficiente e o erro padrão. Valores maiores em magnitude indicam maior confiança de que a variável é significativa.
- **Wald:** Um teste que mede a importância de cada coeficiente. Quanto maior o valor, mais significativa é a variável.
- **Valor P:** Indica se o efeito da variável é estatisticamente significativo. Valores abaixo de 0,05 (ou um limite predefinido) sugerem que a variável tem um efeito significativo.

- **Exp(Coeficientes):** Representa as "odds ratio". Um valor acima de 1 indica que a variável aumenta as chances do evento ocorrer; valores abaixo de 1, diminuem estas chances.

A ilustração da Regressão Logística realizada com os dados da UFSB de 2014 a 2023, utilizando como variáveis independentes CR (coeficiente de rendimento), Distância (entre o ponto de naturalidade e o ponto onde estuda), Idade e Permanência - número de períodos letivos que o estudante esteve vinculado à Universidade está apresentada abaixo (Figura 23).

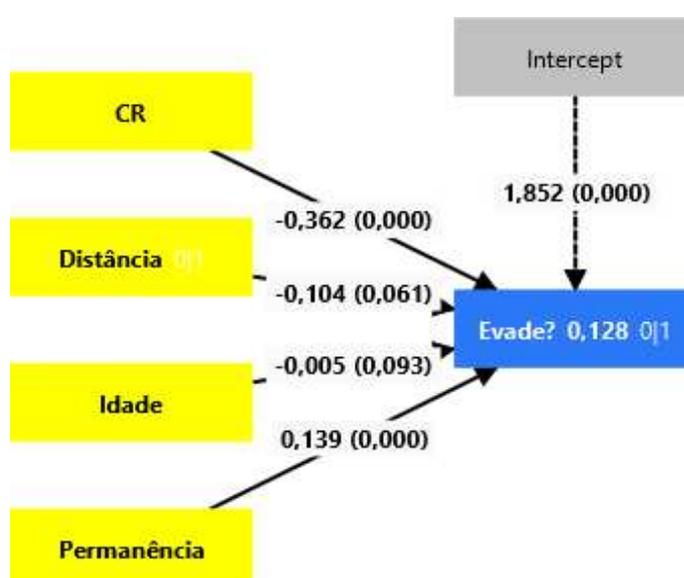


Figura 23 – Regressão Logística por Cursos.

Fonte: Própria (2024).

É oportuno destacar que o *Intercept* (ou intercepto) numa Regressão Logística é o valor da constante que representa o ponto onde a curva da função logística intercepta o eixo y (ou, mais precisamente, o logaritmo das probabilidades odds ratio) quando todas as variáveis independentes são iguais a zero. Em termos simples, ele é o valor base da função logística antes da influência das variáveis explicativas (ou preditoras) serem consideradas.

Os resultados da análise de Regressão Logística apontam para alguns fatores-chave que influenciam a evasão no ensino superior na UFSB. Entre as variáveis analisadas, o Coeficiente de Rendimento (CR) e a Permanência (número de

quadrimestres cursados) destacaram-se como os fatores mais relevantes e estatisticamente significativos para prever a probabilidade de evasão dos estudantes.

O Coeficiente de Rendimento (CR) apresentou um efeito protetor contra a evasão. Ou seja, quanto maior o desempenho acadêmico do estudante, menor a probabilidade de ele abandonar o curso. Este achado é consistente com a literatura, que frequentemente associa um bom desempenho acadêmico a uma maior permanência no ensino superior.

Por outro lado, o tempo de permanência mostrou-se um fator que aumenta o risco de evasão. Conforme os estudantes permanecem na instituição de ensino, a probabilidade de evasão aumenta. Este fenômeno pode estar relacionado a fatores como o não avanço no percurso acadêmico, apesar da permanência ao longo dos períodos letivos, à trancamentos e suspensões de atividades ou à matrícula no período sem o comparecimento às aulas, por exemplo.

Neste ponto indica-se a necessidade de ampliação das variáveis analisadas com dados demográficos e um estudo qualitativo que possa identificar possíveis fatores psicológicos, pessoais, motivacionais, entre outros, que possam ser determinantes ou necessários para o desfecho apontado.

Em relação às demais variáveis, a distância entre a residência do estudante e a instituição apresentou um efeito marginal sobre a evasão, indicando que, embora não seja um fator tão forte quanto o CR e o tempo de permanência, ainda pode influenciar levemente a decisão de evasão. Fatores como a concessão de auxílios permanência podem ter impacto neste resultado, favorecendo a permanência. Tal achado também pode ser aprofundado com a adição de outras variáveis, permitindo uma análise de eficácia dessa política educacional.

No entanto, a idade não pareceu ser relevante para prever a evasão neste contexto específico, sugerindo que a faixa etária dos alunos não tem um impacto significativo no desfecho de permanência ou abandono nesta ocasião.

Apesar destes achados, o modelo apresentou um ajuste considerado tímido (12,8%), indicando que ele capta apenas uma parte dos fatores que influenciam a evasão, mas ainda há espaço para melhorias. Reitera-se que a inclusão de mais variáveis, como fatores socioeconômicos, apoio familiar ou aspectos psicológicos,

podem aprimorar o modelo, fornecendo uma compreensão mais abrangente do fenômeno da evasão e possibilitando a criação de intervenções mais eficazes.

A segunda Regressão Logística tratou das variáveis por componente curricular cursado pelo estudante e apresentou os seguintes resultados (Tabela 4):

Tabela 4: Disciplinas: Total de dados - disciplinas: 284.455 - Análise de Regressão Logística

	Coeficiente	SE	Estatística z	Wald	Valor P	Exp(Coeficiente Suportada ?)
ch_total	-0,007	0,004	-2,033	4,132	0,042	0,993
media_final	-1,759	0,061	-28,993	840,577	0,000	0,172
numero_faltas	-0,057	0,030	-1,888	3,564	0,059	0,944
período	-0,047	0,103	-0,451	0,204	0,652	0,954
porcentagem_frequencia	-0,227	1,135	-0,200	0,040	0,842	0,797
Intercept	6,111	1,168	5,232	27,376	0,000	450,913

Fonte: Própria (2024).

Sendo:

- **Coefficientes:** Indicam o impacto de cada variável sobre o logaritmo das chances (*log odds*). Se o coeficiente é negativo, a variável reduz as chances do evento acontecer (inversamente proporcional), e se for positivo, aumenta as chances (diretamente proporcional).
- **SE (Erro padrão):** Mede a precisão da estimativa do coeficiente. Um erro padrão menor indica uma estimativa mais precisa, logo, são desejáveis SEs que tendam a zero.
- **Estatística z:** A razão entre o coeficiente e o erro padrão. Valores maiores em magnitude indicam maior confiança de que a variável é significativa.
- **Wald:** Um teste que mede a importância de cada coeficiente. Quanto maior o valor, mais significativa é a variável.
- **Valor P:** Indica se o efeito da variável é estatisticamente significativo. Valores abaixo de 0,05 (ou um limite predefinido) sugerem que a variável tem um efeito significativo.

- **Exp(Coeficientes):** Representa as "odds ratio". Um valor acima de 1 indica que a variável aumenta as chances do evento ocorrer; valores abaixo de 1, diminuem estas chances.

Na Figura 24 tem-se a ilustração da Regressão Logística realizada com os dados da UFSB de 2014 a 2023, utilizando como variáveis independentes a Média Final, CH (carga horária do componente curricular), Número de Faltas e Período (número de períodos letivos que o estudante esteve vinculado à Universidade).

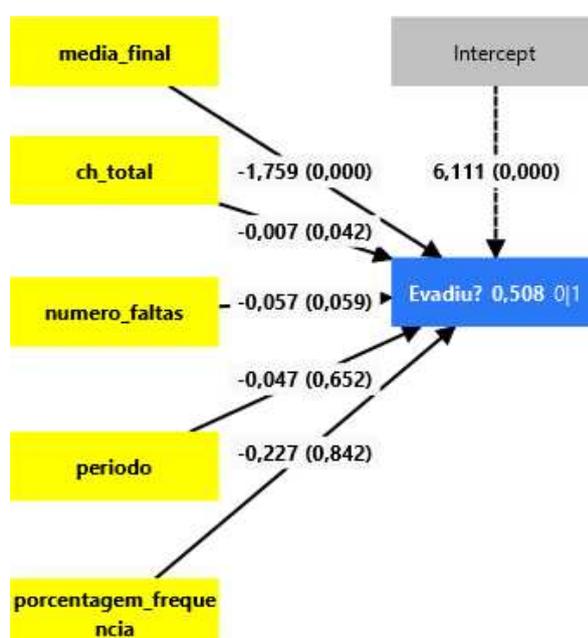


Figura 24 – Regressão Logística por Disciplinas.

Fonte: Própria (2024).

Os resultados da análise de Regressão Logística evidenciam que a Média Final é a variável mais importante na previsão da evasão dos estudantes, sendo altamente significativa com um Valor-p inferior a 0,001. Este dado indica que a média final exerce um forte impacto na probabilidade de evasão: à medida que a média final aumenta, as chances de evadir caem substancialmente. Este efeito protetor da média final reflete o fato de que estudantes com melhor desempenho acadêmico ao longo das disciplinas têm muito menos probabilidade de abandonar.

Além disto, a carga horária total também se mostrou uma variável estatisticamente significativa na previsão da evasão, embora a relação inversamente proporcional não seja usual e o seu efeito seja relativamente pequeno. Isto sugere que o volume de horas cursadas pode influenciar a probabilidade de evadir, mas de forma menos pronunciada em comparação com a média final.

Por outro lado, outras variáveis analisadas, como o número de faltas e a frequência, não apresentaram significância estatística no modelo. Isto significa que, de acordo com os dados analisados, estas variáveis não têm um impacto relevante na previsão de abandono dos estudantes nesta ocasião e a partir desta amostra de dados. Embora estes fatores possam ser importantes em outros contextos ou para determinados grupos de discentes, nesta análise eles não se destacaram como preditores robustos.

De forma geral, os resultados indicam que o desempenho acadêmico medido pela média final é o fator decisivo para prever a evasão, enquanto outros aspectos, como carga horária e frequência, desempenham um papel menor ou insignificante neste processo. Estes achados sugerem que intervenções voltadas para melhorar o desempenho acadêmico dos estudantes podem ser mais eficazes na redução das taxas de abandono.

O resultado corrobora com estudos europeus destacam como relevantes para a gestão da evasão as estratégias de orientação estudantil antes da entrada na universidade, além de ações de integração acadêmica e social, e intervenções de aconselhamento, mentoria e tutorias após o ingresso (Aina, 2022).

Este tipo de análise tem algumas restrições, pois, apesar de ser uma técnica não linear, a Regressão Logística pressupõe uma relação linear entre as variáveis independentes e o logaritmo das *odds*, o que pode limitar sua aplicabilidade em casos com relações mais complexas entre as variáveis.

Pode-se perceber que quando se realizou a análise por disciplina o poder de predição aumentou para 50,8%, significativamente maior que a análise por curso. Observando estas considerações iniciais, decidiu-se realizar a Árvore de Decisão apenas para os cursos. A Árvore de Decisão trabalha a partir de uma perspectiva acumulativa de fatores, considerando que existem fatores prioritários sobre os outros,

oferecendo combinações úteis para tomada de decisão em um contexto muitas vezes não linear, capturando relações mais complexas e sobrepostas no processo de evasão.

A Regressão Logística, por outro lado, é essencialmente linear nos parâmetros (mesmo que o resultado seja probabilístico). Se a relação entre as variáveis preditoras e a evasão não for linear, a regressão logística pode não capturar adequadamente estas interações, enquanto a Árvore de Decisão pode segmentar o espaço de dados em regiões apropriadas.

Deste modo, como existem resultados esperados pela literatura que não se confirmaram, foi decidido ampliar a análise para a Árvore de Decisão.

5.4.2 Análise por Árvore de Decisão

A Árvore de Decisão é uma técnica de aprendizado de máquina usada para classificação e regressão, que divide os dados em subconjuntos com base em uma série de decisões sequenciais, criando uma estrutura de árvore.

Segundo Han (2023), cada nó interno da árvore representa uma decisão baseada em uma variável preditora, um teste de atributo, e cada ramo corresponde a uma das possíveis respostas ou valores desta variável, continuando o processo até que um nó final, chamado de nó folha (ou nó terminal), seja alcançado, representando o resultado final (rótulo de classe ou valor predito).

5.4.2.1 Análise dos Nós e Folhas da Árvore de Decisão

Neste tópico apresenta-se os resultados da pesquisa com os dados recebidos da Universidade Federal do Sul da Bahia, englobando o período de 2014 a 2023 e suas análises a partir do tratamento de dados pelo *software RapidMiner 9.10.013* a partir da estrutura detalhada (Figura 25).

Durante a construção da Árvore de Decisão, medidas de seleção de atributos são usadas para selecionar o atributo que melhor particiona as sequências de dados em classes (Hans, 2023).

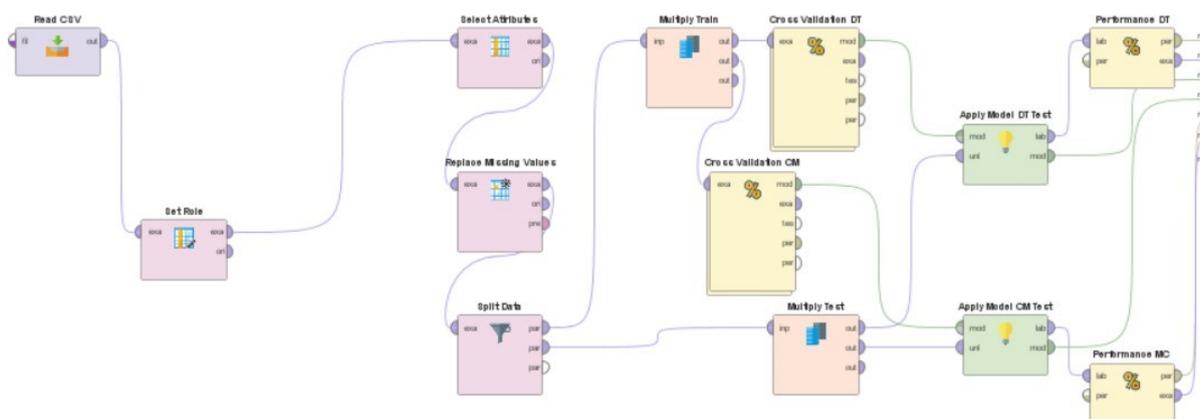


Figura 25 – Tratamento de dados.
 Fonte: Própria (2024). Extraída de *RapidMiner*.

Iniciou-se a sequência com a Leitura dos Dados (*Read CSV*), onde o bloco 1 do diagrama representa a importação dos dados de um arquivo CSV, que é lido e os dados passam para o próximo estágio do processo.

A etapa seguinte é a Definição dos Papéis, ou *Set Role*, onde é definido o papel de cada atributo (coluna) no universo dos dados. É identificado o atributo que se deseja prever, denominado atributo-alvo, e as demais colunas são tratadas como variáveis independentes, ou *features*. Esta etapa está representada pelo bloco 2. A variável “Evadiu” foi usada como atributo alvo.

O bloco 3 corresponde à Seleção dos Atributos (*Select Attributes*), onde as variáveis mais relevantes para este modelo são selecionadas. Esta etapa é fundamental para eliminar variáveis irrelevantes ou redundantes, que poderiam vir a baixar o desempenho do modelo apresentado. Foram selecionadas as variáveis Permanência, CR, *Campus*, Idade e Distância como atributos para predição.

Em seguida, é realizado o Tratamento de Valores Faltantes (*Replace Missing Values*), representado no bloco 4, quando os valores ausentes nos dados são substituídos ou imputados, garantindo que o modelo não sofra erros durante as fases de treinamento e teste por dados faltantes.

No bloco 5 tem-se a Divisão dos Dados (*Split Data*). Neste ponto os dados são divididos em duas partes, a saber, treino e teste. Este procedimento é feito para avaliar o desempenho do modelo em dados que não foram vistos durante o treinamento. Os cinco blocos podem ser vistos em detalhes na Figura 26.

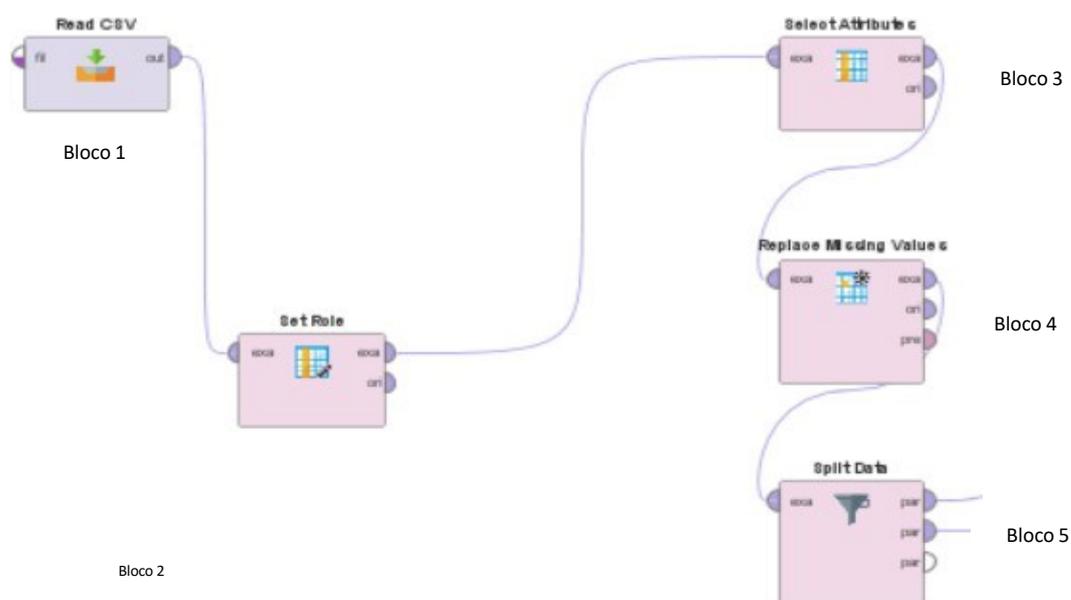


Figura 26 – Tratamento de dados Blocos 01 a 05

Fonte: Própria (2024). Extraída de *RapidMiner*.

No bloco 6, referente à Multiplicação dos Dados (*Multiply Train*), o operador cria cópias dos dados de treinamento, permitindo que diferentes métodos de validação e modelagem possam ser aplicados separadamente, testando um maior número de hipóteses.

No bloco 7 temos a Validação Cruzada da (*Decision tree DT* e *Confusion Matrix CM*), uma técnica essencial para avaliar a performance de um modelo de aprendizado de máquina, verificando sua capacidade de generalização, ou seja, quão bem ele se comporta em novos dados, reduzindo o risco de *overfitting*. Este processo envolve dividir os dados em partes e testar o modelo em diferentes subconjuntos, garantindo uma avaliação mais robusta.

Na Validação Cruzada com Divisão Treinamento/Teste (*Decision tree -DT*), os dados são divididos em um conjunto de treino e um conjunto de teste, o modelo é testado no conjunto de treino e avaliado no conjunto de teste, sendo uma validação de fácil implementação que avalia o modelo de decisão (Árvore de Decisão).

Na Validação Cruzada com Matrizes de Confusão (*Cross Validation CM*), a matriz de confusão é gerada a cada interação da validação cruzada, analisando mais profundamente a performance do modelo, fornecendo uma visão mais detalhada das verdadeiras classificações positivas e negativas. No entanto, requer um processo mais complexo de avaliação pela necessidade de interpretar várias métricas, avaliando outro algoritmo alternativo.

O bloco 8 consiste na Aplicação do Modelo aos Dados de Teste (*Apply Model DT e Apply Model CM*). Após a validação cruzada, os modelos treinados são aplicados ao conjunto de dados de teste para verificar seu desempenho real. Assim, dois modelos são aplicados: um modelo baseado na Árvore de Decisão (DT) e outro modelo alternativo (CM).

Por fim, no bloco 9 é realizada a Avaliação de Desempenho (*Performance DT e Performance CM*), onde o funcionamento de cada modelo é avaliado individualmente, usando métricas como precisão, recall, acurácia, entre outras, com o objetivo de comparação e identificação de qual tem melhor desempenho. Os blocos de 6 a 9 podem ser vistos em detalhes na Figura 27.

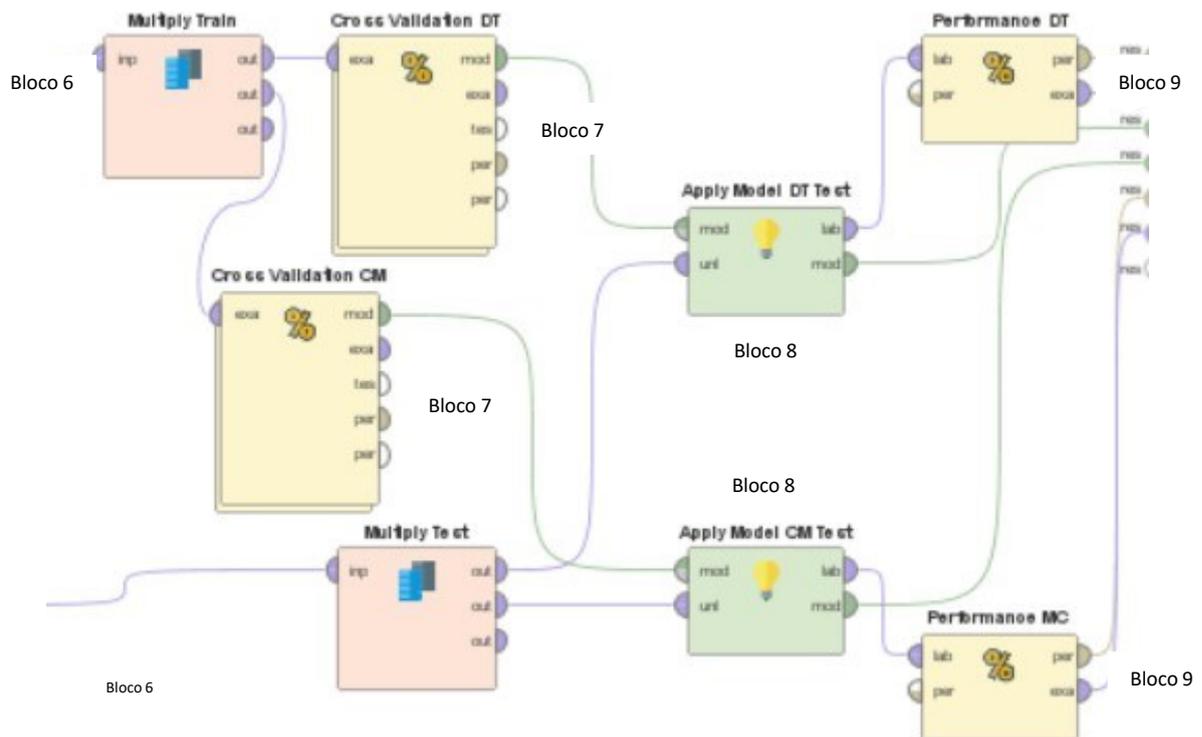


Figura 27 – Tratamento de dados Blocos 06 a 09.

Fonte: Própria (2024). Extraída de *RapidMiner*.

Com o modelo de predição elaborado, a Matriz de Confusão demonstrou acurácia superior a 86,5%, podendo-se então considerar a Árvore de Decisão classificatória (Grus, 2016) proposta como válida para o universo estudado. A partir desta confirmação, a análise da estrutura da árvore foi elaborada.

Diversas Árvores de Decisão foram modeladas segundo o processo descrito até se alcançar este nível de acurácia e a Árvore de Decisão final (Figura 28), com as tendências de comportamento mais assertivas, devido as correções realizadas é exibida abaixo.

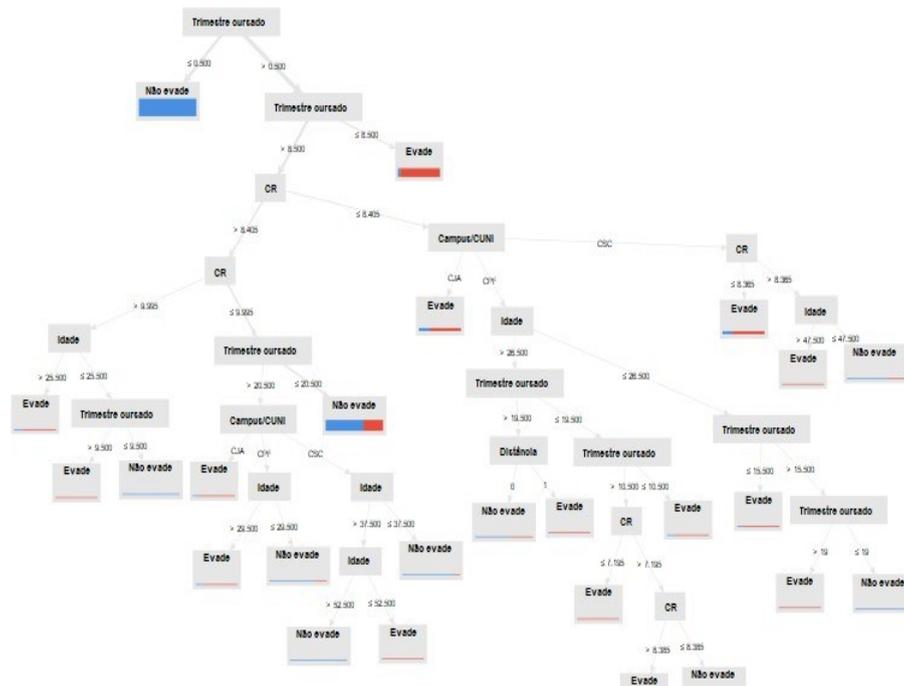


Figura 28 – Árvore de Decisão.

Fonte: Própria (2024).

Apresenta-se aqui a estrutura geral de uma Árvore de Decisão, necessária para uma leitura apropriada da árvore elaborada. A Árvore de Decisão começa com o nó raiz, localizado no topo da árvore, que utiliza a variável Permanência para fazer a primeira divisão. Dependendo do valor desta variável, a árvore segue para as próximas etapas, onde estão os nós internos.

Cada um destes nós representa uma bifurcação baseada em características específicas, como Idade, Coeficiente de Rendimento (CR), ou o *campus* do estudante. Estes nós conduzem, por fim, às folhas, que são as classificações finais da árvore. Nas folhas, a decisão sobre se o estudante evade ou não evade é feita com base nas características analisadas.

Para facilitar a leitura da Árvore de Decisão alguns recursos visuais foram adotados (Figura 29): o primeiro caminho, ou caminho da esquerda, foi grifado na ilustração principal e nas figuras ampliadas em verde e representa o universo de estudantes com CR maior que 8,405 e seus percursos; o segundo caminho, ou caminho da direita, foi grifado na ilustração principal e nas figuras ampliadas em roxo e representa o universo dos estudantes com CR menor ou igual a 8,405; para os

resultados de “evadiu” (e seus percentuais) foi adotada a cor vermelha e resultados de “não evadiu” (e seus percentuais) a cor azul; números romanos foram adicionados à ilustração principal e figuras ampliadas como marcadores do caminho e não se confundem com outros valores expressos.

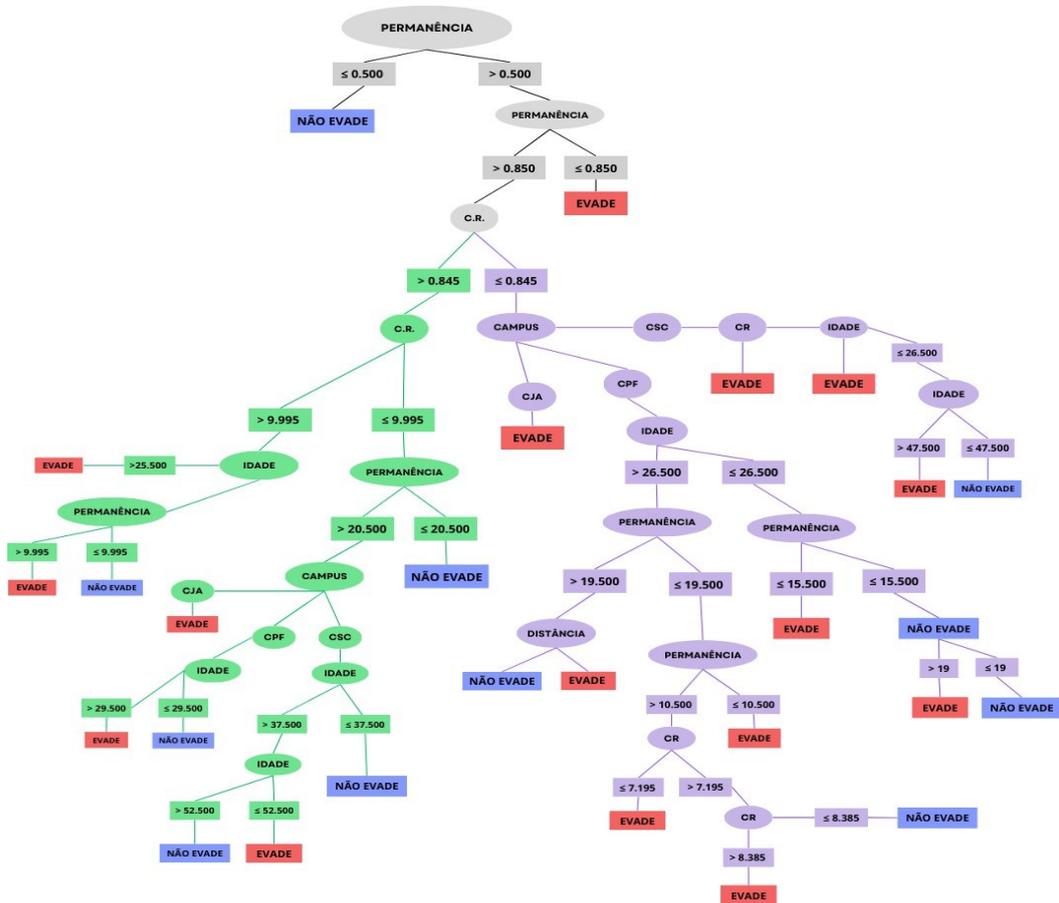


Figura 29 – Árvore de Decisão identificada.

Fonte: Própria (2024). Elaborada com *Canva*.

A árvore é composta por 1 nó raiz, sendo esse a variável de tempo de Permanência do estudante da UFSB, 19 nós e 27 folhas. São 5 fatores não repetidos, sendo 28 fatores no total. A Permanência é o fator crítico e aparece sete vezes, seguido da variável CR com quatro, *Campus* com duas, Idade com sete e Distância com uma ocorrência.

A árvore de decisão apresentada aponta como raiz, ou elemento principal, para a evasão ou não evasão dos estudantes da UFSB, o tempo de Permanência que este discente esteve na instituição. Se a Permanência do estudante for menor ou igual a

cinco períodos letivos (05 quadrimestres, aproximadamente um ano e meio) ele não evade (I), se for maior que cinco períodos letivos (II), é necessário verificar novamente a Permanência. Se o estudante passou mais de cinco e menos de nove períodos (III) na instituição ele evade, se ele passou mais de nove períodos (IV) temos a primeira ramificação da árvore de decisão, no caminho da esquerda. Assim, identifica-se um intervalo de potencial risco para evasão – entre cinco e nove períodos, ou, entre 01 ano e meio e três anos (Figura 30). Na árvore de decisão apresentada, dois caminhos são possíveis, o caminho da esquerda e o caminho da direita, e passaremos à descrição e análise de cada um deles.

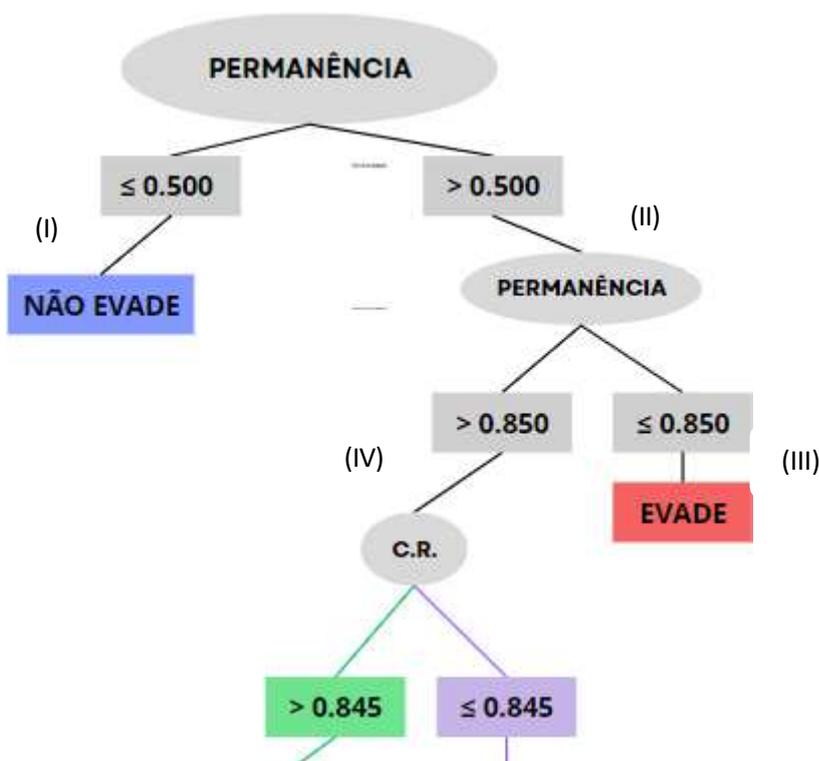


Figura 30 – Raiz da Árvore de Decisão, percurso I a IV.

Fonte: Própria (2024).

A árvore nos informa que se o tempo de Permanência do estudante na UFSB é maior que 9 períodos letivos (quadrimestres, à época), a árvore desce e verifica o valor do CR desse estudante. Nesse momento temos o ponto onde os caminhos se diferenciam. O fato do estudante ter um CR maior ou menor/igual a 8.405 vai definir se ele segue o caminho da esquerda ou da direita na árvore de decisão.

O Primeiro Caminho (esquerda superior), mostra que se o CR desse estudante for maior que 8,405 (V), a árvore analisa novamente o CR. Se o estudante tem CR maior que 9,995 (VI) é relevante adicionar o fator Idade. Se o estudante tem CR maior que 9,995 e Idade maior que 25,5 anos (VII), ele evade. Se ele tem CR maior que 9,995 e Idade menor ou igual a 25,5, (VIII) dependerá da Permanência. Se a Permanência for maior que 10 períodos letivos (IX) ele evade, se a Permanência for menor ou igual a 10 períodos letivos (X) ele não evade (Figura 31).

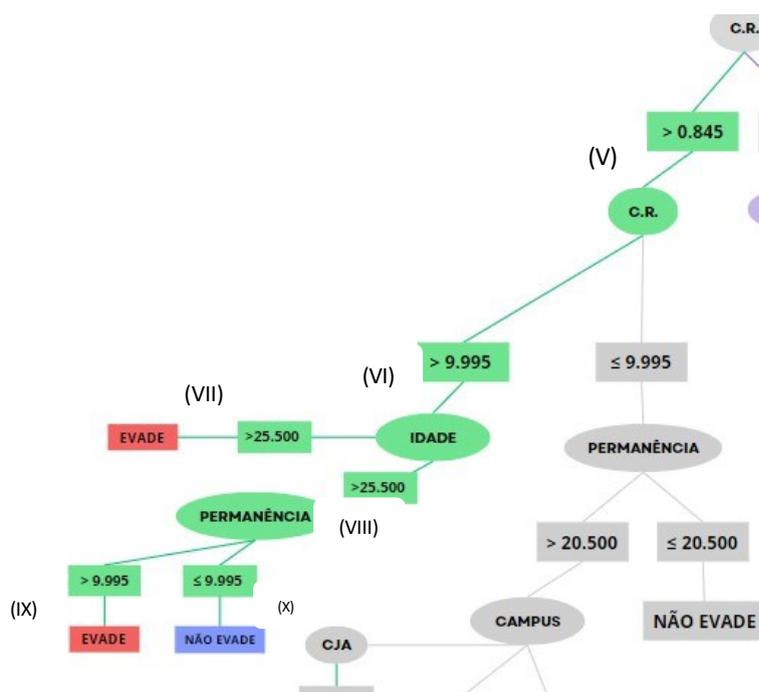


Figura 31 – Primeiro Caminho, percurso V a X.

Fonte: Própria (2024).

Ainda no Primeiro Caminho, se o CR do estudante for menor ou igual a 9,995 (XI) é preciso considerar a Permanência. Se a Permanência for menor ou igual a 21 períodos letivos (XII), sete anos, ele não evade. Se a Permanência for maior que 21 períodos letivos (XIII) é necessário considerar o *campus* ao qual o aluno está vinculado. Se a Permanência for maior que 21 períodos e o *campus* for o CJA (XIV), ele evade. Se o *campus* for o CPF é preciso considerar a Idade, se maior que 29,5 anos (XV) ele evade, se menor ou igual a 29,5 (XVI) ele não evade. Se o *campus* for o CSC também é preciso considerar a Idade, se menor ou igual a 37,5 (XVII) ele

evade, se maior que 37,5 (XVIII) e menor ou igual a 52,5 (XIX) ele evade, se maior que 52,5 (XX) ele não evade (Figura 32).

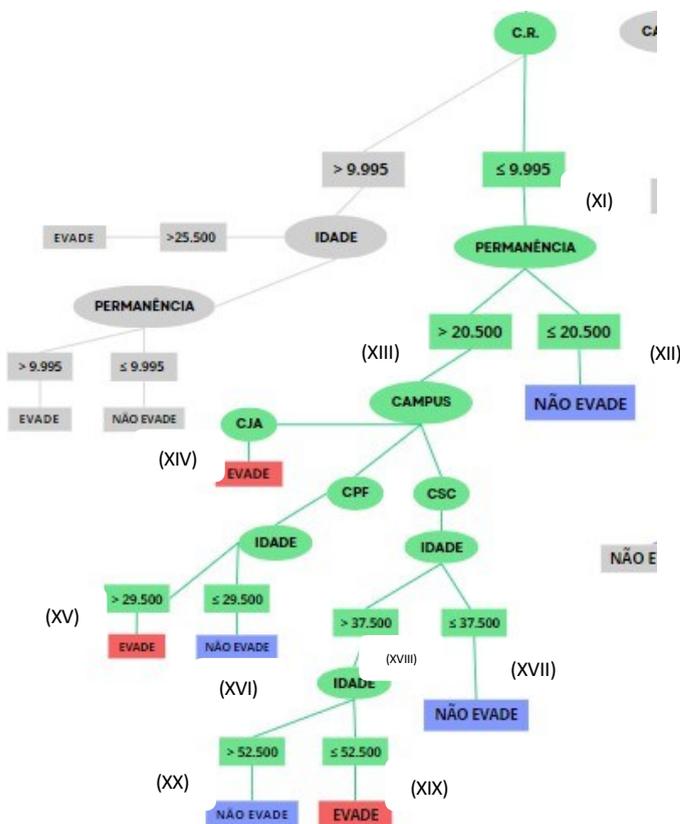


Figura 32 – Primeiro Caminho, percurso XI a XX

Fonte: Própria (2024).

Este primeiro caminho nos permite dizer que estudantes com menos de três anos cursados, com maiores notas e mais novos, tendem a não permanecer na instituição, tendem a evadir.

No Segundo Caminho (direita superior), se o CR desse estudante for menor ou igual a 8,405 (XXI) é necessário considerar o *campus* ao qual o estudante está vinculado – *campus* Jorge Amado (CJA), *campus* Paulo Freire (CPF) ou *campus* Sosígenes Costa (CSC):

Se o estudante pertence ao *campus* CJA (XXII), ele evade (Figura 33).

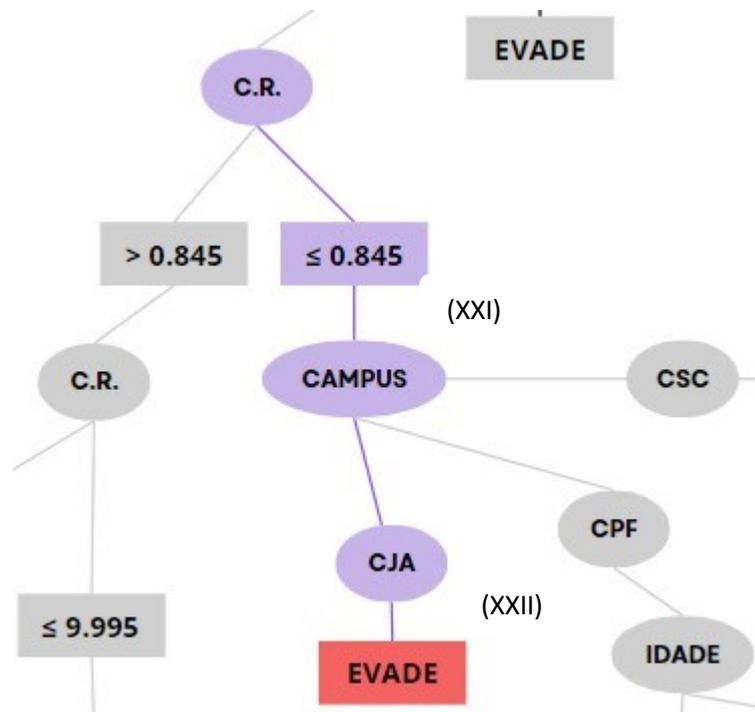


Figura 33 – Segundo Caminho, percurso XXI a XXII.

Fonte: Própria (2024).

Se ele pertence ao *campus* CPF é necessário considerar a Idade (XXIII), se for menor ou igual a 26,5 anos (XXIV) é necessário adicionar o tempo de Permanência. Se o estudante permaneceu 16 períodos ou menos (XXV) ele evade. Se o estudante permaneceu mais de 16 períodos (XXVI) e menor ou igual a 19 períodos (XXVII), ele não evade. Se o estudante estiver na universidade há mais de 19 períodos (XXVIII) ele evade (Figura 34).

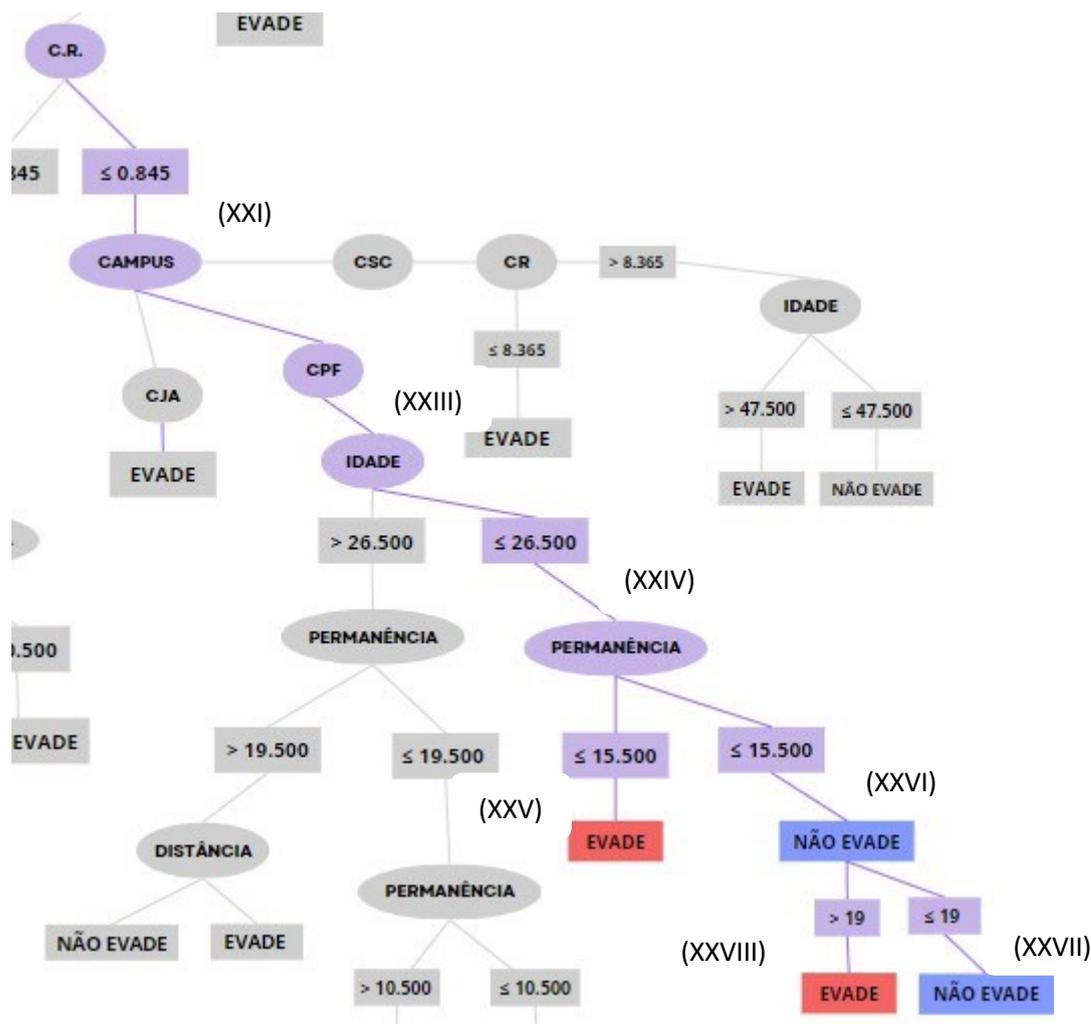


Figura 34 – Segundo Caminho, percurso XXIII a XXVIII.

Fonte: Própria (2024).

Se ele pertence ao *campus* CPF e tem Idade maior que 26,5 anos (XXIX) é necessário considerar o tempo de Permanência. Se o estudante está a mais de 20 períodos letivos (XXX), considera-se o fator Distância, se ele é oriundo da cidade do *campus* (XXXI) ele não evade, se ele é de outra localidade (XXXII) ele evade. Se a Permanência é menor ou igual a 20 períodos (XXXIII) e menor ou igual a 11 períodos (XXXIV), ele evade. Se a Permanência é menor ou igual a 20 períodos (XXXIII) e maior que 11 períodos (XXXV), considera-se o CR. Se o CR for menor ou igual a 7,195 (XXXVI), ele evade. Se o CR for maior que 7,195 (XXXVII) e menor ou igual a 8,385 (XXXVIII) ele não evade, se o CR for maior que 8,385 (XXXIX), ele evade (Figura 35).

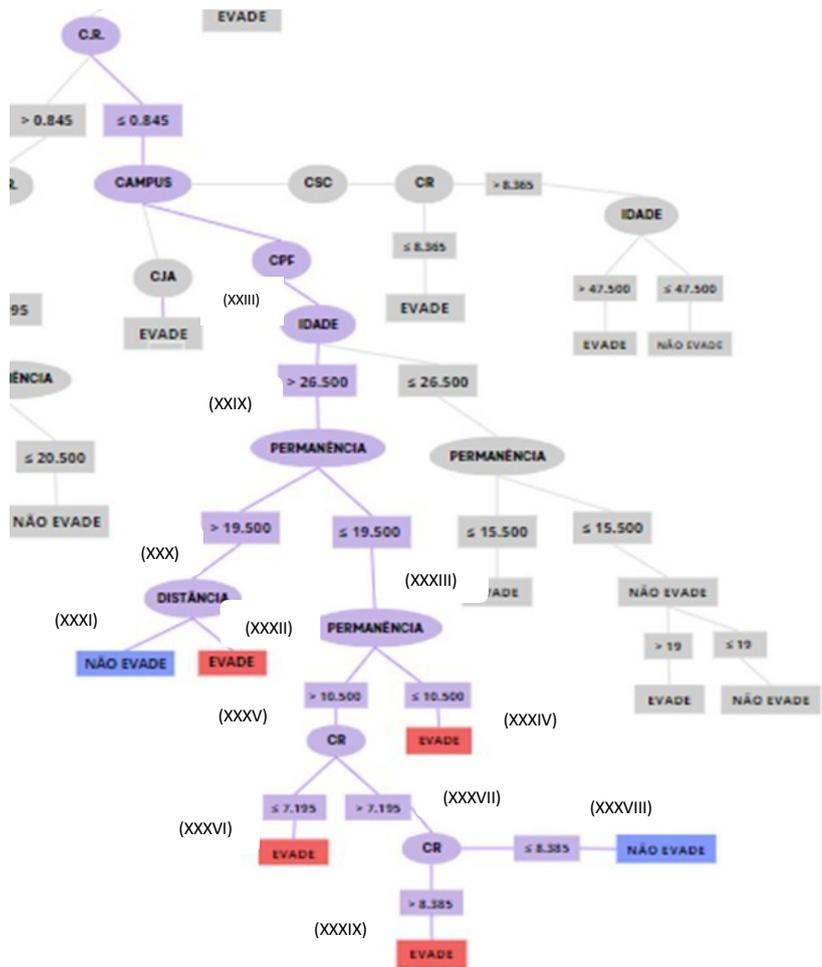


Figura 35 – Segundo Caminho, percurso XXIX a XXXIX.

Fonte: Própria (2024).

Se o estudante pertence ao *campus* CSC (XL), é necessário considerar o CR. Se o estudante tem CR menor ou igual a 8,365 (XLI) ele evade. Se o CR é maior que 8,365 (XLII) considera-se a Idade, se menor ou igual a 47,5 (XLIII), o estudante não evade, se maior que 47,5 (XLIV) ele evade (Figura 36).

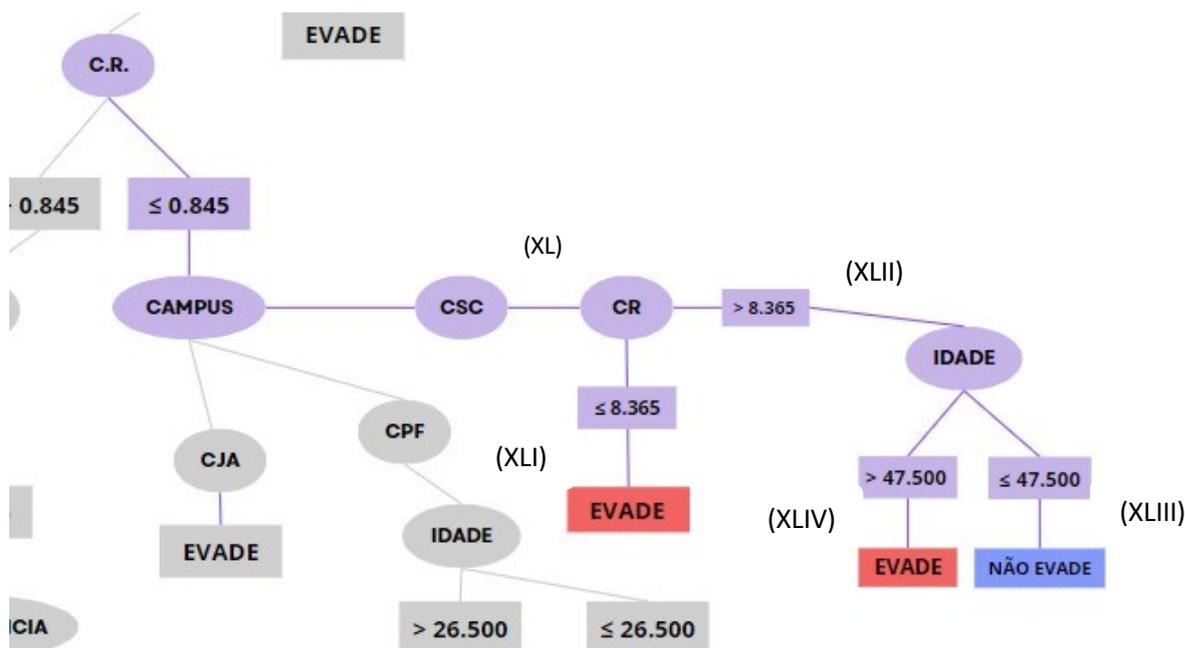


Figura 36 – Segundo Caminho, percurso XL a XLIV.

Fonte: Própria (2024).

A análise da Árvore de Decisão demonstra que o fator mais relevante para prever a evasão ou permanência dos estudantes na UFSB é o tempo de permanência na instituição. Estudantes que permanecem menos de cinco períodos letivos tendem a não evadir, enquanto aqueles que permanecem entre cinco e nove períodos estão em uma zona de risco. Após nove períodos, a decisão sobre evasão passa a depender de outras variáveis como o Coeficiente de Rendimento (CR), Idade, *Campus*, e, em alguns casos, a Distância da cidade de origem em relação à universidade.

A análise sugere que, para reduzir a evasão, estratégias diferenciadas devem ser adotadas, levando em conta estas variáveis críticas. Identificar os estudantes em risco com base na permanência, CR e idade pode permitir intervenções mais eficazes, como programas de apoio acadêmico e políticas voltadas para a retenção, especialmente nos *campi* e grupos mais vulneráveis.

Dentro da ciência de dados e da mineração de dados, as árvores de decisão programadas são uma evolução das árvores tradicionais, otimizadas automaticamente para encontrar o melhor modelo de predição ou classificação. Sua automatização e capacidade de gerar modelos mais complexos as tornam úteis em

problemas de grande escala e alta dimensionalidade. É possível ver em anexo (Anexo 1) a árvore programada referente à Árvore de Decisão apresentada neste estudo.

Após a descrição e análise detalhada da Árvore de Decisão, é fundamental avaliar a performance do modelo gerado, verificando o quanto suas previsões se aproximam da realidade. Para isto, utiliza-se a matriz de confusão, permitindo uma análise mais aprofundada da qualidade do modelo de Árvore de Decisão, identificando as taxas de acertos e erros em relação à evasão acadêmica, e oferecendo *insights* sobre o desempenho global do algoritmo.

5.5 Avaliação

Após a construção do modelo, a fase de avaliação envolve verificar se o modelo atende aos requisitos definidos na fase de compreensão do negócio. Isto inclui a avaliação da precisão, revocação, e outras métricas de desempenho com base no conjunto de teste.

5.5.1 Matriz de Confusão

A Matriz de Confusão (Tabela 5) é uma ferramenta fundamental para avaliar o desempenho de classificadores, como uma árvore de decisão, especialmente em problemas de mineração de dados aplicados à evasão no ensino superior. Ao lidar com a predição de evasão estudantil, o uso de uma Matriz de Confusão permite uma análise detalhada dos acertos e erros do modelo, além de fornecer métricas essenciais para mensurar sua eficácia, uma vez que seus resultados têm implicações diretas na formulação das políticas públicas que serão desenhadas e nas ações preventivas.

Tabela 5: Accuracy: 86,61%

	true Não Evade	true Evade	class precision
pred. Não Evade	1075	168	86,48%
pred. Evade	75	497	86,89%
class recall	93,48%	74,74%	

Fonte: Própria (2024).

A Matriz de Confusão consiste em uma tabela que resume os resultados da classificação, organizando-os em quatro categorias principais:

- Verdadeiros Negativos (VN): Casos em que o modelo previu corretamente que o estudante não evadiu. O modelo apresentado previu corretamente 1075 ocorrências onde o resultado verdadeiro era “não evade”. Houve casos em que o modelo previu que o estudante não evadiria, mas ele acabou evadindo. O modelo previu "Não Evade", mas o resultado verdadeiro era "Evade" em 168 casos (erro de classificação).
- Verdadeiros Positivos (VP): Casos em que o modelo previu corretamente que o estudante evadiu. O modelo previu corretamente 497 instâncias onde o resultado verdadeiro era "Evade". Ocorreram casos em que o modelo previu evasão, mas o estudante não evadiu. O modelo previu "Evade", mas o resultado verdadeiro era "Não Evade" em 75 casos (erro de classificação).

A Matriz de Confusão também agrega métricas importantes para o modelo preditivo, sendo as principais a acurácia, a precisão e a revocação (*recall*), essenciais para medir a qualidade das predições.

A acurácia (*accuracy*) é uma métrica usada para medir o desempenho de um modelo de classificação, indicando a porcentagem de previsões corretas feitas pelo modelo em relação ao total de previsões. A fórmula da acurácia é:

$$a. \text{ Acurácia} = \frac{\text{Previsões corretas}}{\text{Total de Previsões}}$$

No modelo apresentado a Acurácia foi de 86,61%, ou seja, isto significa que, de todas as previsões feitas, 86,61% estavam corretas.

A precisão é indicada pela proporção de predições corretas, tanto para os que evadiram quanto para os que não evadiram. Para a classe “não evade”, das previsões feitas 86,48% estavam corretas e para a classe “evade”, das previsões feitas 86,89% estavam corretas no modelo apresentado. Verifica-se que as precisões para ambas

as classes são equilibradas, 86,61% em média, indicado que o modelo elaborado tem desempenho semelhante para prever a evasão e a não evasão.

A revocação ou *recall* é a razão entre o número de verdadeiros positivos e o número total de casos positivos. O modelo desenvolvido identificou corretamente 93,48% dos alunos na classe “não evade” e corretamente 74,74% dos alunos na classe “evade”. O *recall* para a classe "evade" (74,74%) é menor que o *recall* para "não evade" (93,48%), o que indica que o modelo tem mais dificuldade em prever corretamente os casos de evasão devido falta de mais dados.

Com as análises e métricas adotadas, podemos dizer que a acurácia geral do modelo é alta, registrando 86,61% de acertos, o que significa que a maioria das previsões foram corretas.

Assim, pode-se observar que o modelo consegue prever melhor o aluno que não evade, seguido do estudante que evade. Isto confirma (estudos de permanência), onde compreender os motivos que ficam passam a ser mais importantes, por serem mais controláveis.

5.6 Implementação

5.6.1 Implicações do Modelo Proposto e Apresentação do *Dashboard*

Uma vez identificadas as variáveis críticas para o estudo da evasão na UFSB, é possível sinalizar comportamentos, situações ou períodos de maior risco da evasão. Isto permite propor ações e monitorar acontecimentos, com menor esforço e recurso e maior resultado, como preconizado no Princípio de Pareto (Korom, 2015).

Foram encontrados neste estudo evidências de que a Permanência é o fator mais relevante para o comportamento de evasão/não evasão e pode estar associado a uma política de retenção de estudantes. O estudo apontou algumas situações diferentes: no primeiro cenário, um determinado intervalo de permanência com risco de evasão, entre cinco e nove períodos letivos (entre um ano e meio e três anos), consistente

com os estudos Castro-Lopez (2022) e Aina (2022), considerando que existem múltiplos fatores associados a esta permanência.

No segundo cenário claro que se define, é a divisão de percursos no CR 8,405 apontando diferentes comportamentos para evasão e diferentes fatores envolvidos. Estudantes com notas mais altas, idade inferior a 25,5 anos e menos de 3 anos de cursos tendem a evadir da instituição, o que nos mostra um cenário de não dificuldade acadêmica.

Este tipo de achado pode levar a hipótese de insatisfação com o curso, caso o estudante evada para outro curso de graduação dentro da própria instituição, ou a insatisfação com a instituição de ensino, caso ele evada para outra IE. Seria recomendado uma pesquisa qualitativa para apurar se este estudante, de fato, deixa o ensino superior ou se ele deixa a IE. Outros estudos que também consideram a variável Idade como relevante ao processo de evasão são Viloría (2020) e Pedraza-Navarro (2021).

Outro perfil extremo também encontrado na pesquisa são estudantes que permanecem por períodos maiores que o padrão para uma graduação (mais de 20 períodos letivos, ou sete anos) e evadem da instituição, levantando duas hipóteses: estudantes com baixo CR que permanecem matriculados na instituição e não avançam nos seus respectivos cursos; e estudantes com CR alto, longa permanência e evasão sem conclusão do curso.

O primeiro cenário pode indicar que, embora os alunos estejam permanecendo matriculados no curso, eles estão enfrentando dificuldades significativas para progredir em suas trajetórias acadêmicas. Isto pode ser um sintoma de vários problemas sistêmicos, acadêmicos e pessoais, com impactos profundos na qualidade da formação dos estudantes e na eficiência do sistema de ensino. Comportamentos distintos em função do *campus* ou da unidade acadêmica ao qual o estudante está vinculado também aparecem nos estudos de Gutierrez-Pachas (2023).

O segundo cenário também adentra o tema da retenção, que se refere à capacidade de uma instituição de manter os estudantes matriculados ao longo de seu curso, garantindo que eles progridam nos estudos e eventualmente concluem sua graduação. A retenção é o oposto da evasão e representa um dos principais

indicadores de sucesso institucional, tanto para o estudante quanto para a própria universidade.

Porém, quando a retenção ou o longo período de permanência em um curso de graduação não é acompanhado por índices de aprovação, mas sim por reprovação em disciplinas ou forma de cumprir créditos para outras graduações que não a de matrícula, surgem desafios significativos tanto para os estudantes quanto para as instituições.

Este tipo de retenção pode ser um sintoma de vários problemas sistêmicos, acadêmicos e pessoais, com impactos profundos na qualidade da formação dos estudantes e na eficiência do sistema de ensino, tais como baixa qualidade do progresso acadêmico, reprovações repetidas, abandono emocional do curso apesar das sucessivas matrículas nos períodos letivos, prolongamento do tempo de conclusão, impacto na saúde mental e na motivação, lacunas de aprendizagem do ensino médio, entre outras.

Isto também acarreta consequências institucionais, impactando negativamente o desempenho global da IE, refletindo uma baixa qualidade de progressão acadêmica apesar da taxa de retenção estável, custo operacional elevado incluindo possíveis auxílios estudantis por todo o período, além dos custos sociais pela mobilização de recurso público e não formação do profissional pretendido.

Desta forma, para munir a gestão acadêmica de uma compreensão panorâmica e dinâmica da problemática da evasão na UFSB, foi desenvolvido como Produto Técnico-Tecnológico (PTT) um *dashboard*, intitulado Observatório da Evasão, que será entregue à instituição, permitindo que novos dados sejam inseridos periodicamente, atualizando as decisões e estratégias aos cenários futuros.

A apresentação dos dados feita em plataforma digital (Figura 37), no endereço virtual <https://www.unbdatalab.org/dashevasao>, também contribui para a publicidade de dados, fortalecendo uma gestão transparente e fornecendo parâmetros para que outras instituições de ensino possam adotar estratégias semelhantes, otimizando a gestão pública.

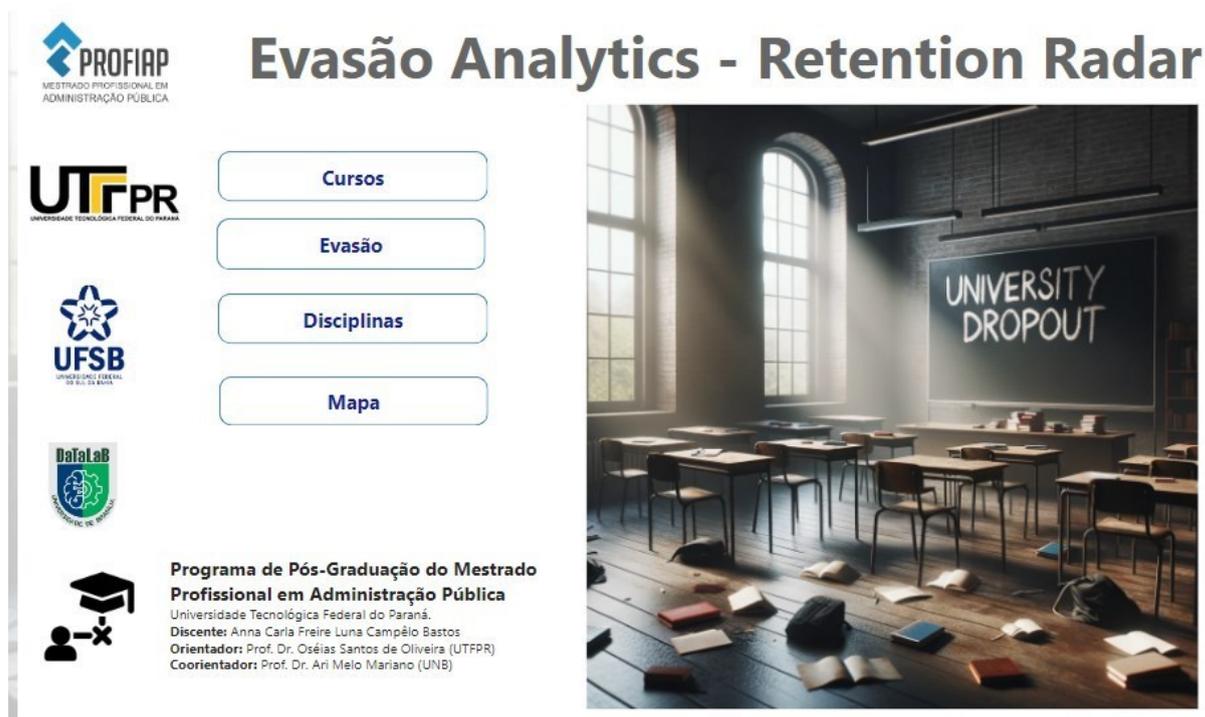


Figura 37 – *Dashboard* da Evasão.

Fonte: Própria (2024).

Este modelo de *Business Intelligence* (BI) no formato de dashboard é uma ferramenta essencial para a visualização de dados, permitindo que as informações sejam apresentadas de forma clara e concisa para facilitar a tomada de decisões. Um dashboard é uma interface visual que reúne dados selecionados por sua importância em um único local, permitindo que gestores e usuários monitorem métricas de desempenho e identifiquem tendências rapidamente, mas também permite perceber a necessidade de modificações antes da modelagem dos dados.

Neste produto são apresentados quatro grandes blocos de análises, cada um com diversas leituras possíveis:

- **Cursos:** oferece o percentual de evasão por curso de graduação da UFSB, podendo ser filtrado por ano dentro do período de 2015 a 2023, e por *campus*, gráfico de percentual de evasão por idade, números absolutos e outras leituras.
- **Evasão:** oferece gráficos de percentual de evasão em função do curso ao longo dos anos e de percentual de evasão em função da disciplina cursada

também ao longo do tempo, oferece a possibilidade de filtrar dados por ano, por *campus* e por *status* (evadiu e não evadiu);

- **Disciplina:** oferece o percentual de evasão por disciplina cursada, podendo ser filtrado por ano dentro do período de 2015 a 2023, e por campus, gráfico de percentual de estudante por status na disciplina, tais como aprovado, reprovado, aguardando e evadido, números absolutos e outras leituras.
- **Mapa:** oferece um mapa da evasão usando a cidade de origem do discente como referência para o mapeamento, podendo apontar necessidades de ações de incentivo a permanência e outros programas de auxílio, além da distribuição de evadidos por cursos e seus percentuais de retenção.

Entre as limitações percebidas pode-se reportar a falta de dados demográfico como sexo e renda. Acredita-se que quanto mais variáveis possam ser introduzidas no modelo de BI entregue, maior será a capacidade do sistema de monitorar a evasão, estabelecendo relações de causa e efeito que poderiam não ser percebidas numa análise fora da abordagem da Mineração de Dados.

O modelo foi desenvolvido para oferecer uma visão dinâmica e sistematizada da evasão na UFSB (Figura 38), e utiliza como base os dados fornecidos pela instituição e que foram utilizados posteriormente nas análises de Regressão Logística e de Árvore de Decisão.

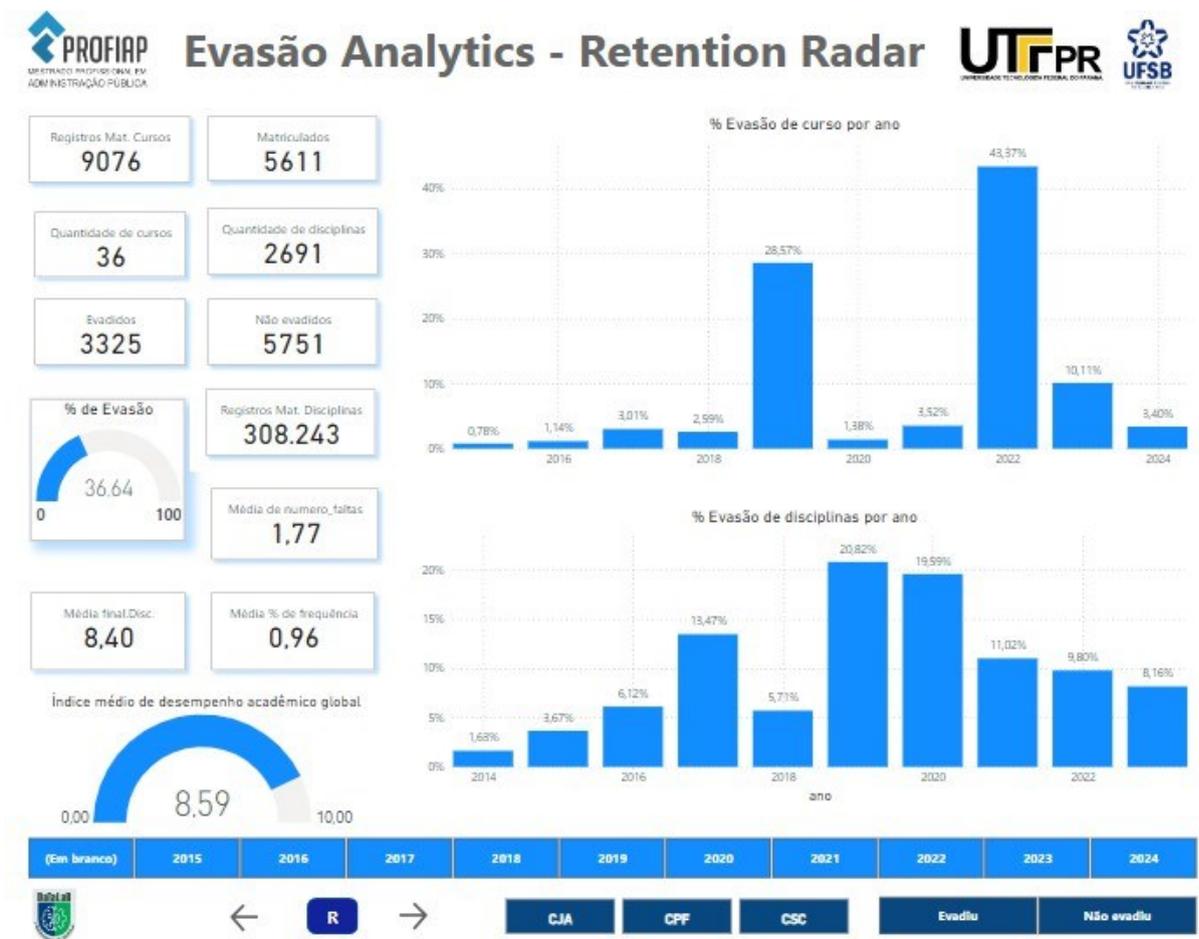


Figura 38 – Evasão na UFSB

Fonte: Própria (2024).

O *dashboard* apresentado traz funcionalidades como gráficos de coluna que informam os percentuais de evasão por curso, percentuais de evasão por disciplina, análise temporal, permitindo que o usuário filtre as informações por *campus*, ano de ingresso, cursos específicos e outras variáveis críticas, entre outras análises, de forma dinâmica e de fácil leitura.

Também se considera importante que as soluções que sejam propostas para enfrentar os problemas identificados e compreendidos, cheguem não apenas à gestão, mas diretamente aos estudantes impactados, contribuindo para a diminuição da evasão e melhoria da experiência acadêmica.

Outra proposta sugerida é o desenvolvimento de um aplicativo para *smartphone* de acompanhamento acadêmico (Figura 39).



Figura 39 – Aplicativo de Acompanhamento Acadêmico.

Fonte: Própria (2024).

Uma das possíveis estratégias que recomendamos é o desenvolvimento futuro de um aplicativo bastante interativo e intuitivo, que possa entregar os resultados destas análises em tempo real não apenas aos gestores institucionais, mas aos estudantes, em risco ou não de evasão, para uma melhor performance acadêmica em geral.

5.6.2 Produto Técnico Tecnológico: Recomendações para Uso do *Dashboard* Observatório da Evasão

Para otimizar o uso de um Produto Técnico Tecnológico (PTT), algumas recomendações essenciais devem ser consideradas para garantir seu máximo potencial e relevância para a gestão acadêmica. Primeiramente, é importante que todos os usuários, especialmente gestores e técnicos responsáveis, tenham

capacitação para navegar e interpretar os dados e funcionalidades da ferramenta. Esta capacitação pode envolver treinamento inicial e uma documentação de suporte, como manuais e FAQs, que detalhem as principais funções.

Seguem as recomendações para o uso e aprimoramento do Dashboard "Observatório da Evasão", desenvolvido como parte de Produto Técnico-Tecnológico (PTT) no âmbito do PROFIAP. Este Dashboard foi criado com o objetivo de auxiliar a gestão no monitoramento e na compreensão da evasão nos cursos de graduação da Universidade Federal do Sul da Bahia (UFSB).

As sugestões elencadas no Quadro 11 visam otimizar seu uso, aumentar sua eficácia e propor possíveis desenvolvimentos futuros.

Quadro 11 – Recomendações de uso do Produto Técnico Tecnológico

RECOMENDAÇÕES	DETALHAMENTO DAS AÇÕES
1) Integração com outras fontes de dados	<ul style="list-style-type: none"> a) Aprimorar a base de dados: aumentar a abrangência dos dados disponíveis no <i>Dashboard</i>, incorporando informações demográficas como sexo, renda, e histórico escolar prévio, de modo que a ferramenta tenha uma visão mais completa do perfil dos estudantes e dos fatores de risco associados à evasão. b) Atualização periódica: garantir que os dados sejam atualizados periodicamente, de preferência semestralmente, para refletir a situação real e permitir a tomada de decisões baseadas em dados atuais. A integração direta com os sistemas acadêmicos da UFSB, como o sistema de gestão acadêmica, pode automatizar este processo. c) Pesquisa qualitativa: implementar um módulo de coleta de feedback qualitativo com os estudantes que evadiram ou estão em risco de evasão. Esta funcionalidade permitirá uma análise mais profunda das razões subjacentes à evasão, como insatisfação com o curso ou desafios pessoais, enriquecendo as interpretações dos dados quantitativos.
2) Monitoramento de padrões de risco e alertas proativos	<ul style="list-style-type: none"> a) Configuração de alertas automáticos: incorporar alertas automáticos com vistas à identificação de estudantes com alta probabilidade de evasão, baseado em critérios como baixo Coeficiente de Rendimento (CR), tempo de permanência prolongado, reprovações sucessivas, e outros fatores de risco identificados, sendo que estes alertas poderão ser configurados para notificar automaticamente os tutores ou coordenadores de curso. b) Mapeamento de períodos críticos: criar uma funcionalidade que permita monitorar períodos críticos de evasão identificados no estudo, como o intervalo entre cinco e nove períodos letivos e após 20 períodos, sendo que isto poderá orientar a aplicação de medidas específicas de apoio acadêmico nesses intervalos.

3) Intervenção personalizada e apoio direto ao estudante	<p>a) Desenvolvimento de um aplicativo de acompanhamento acadêmico: criar um aplicativo para <i>smartphones</i> que funcione integrado ao <i>Dashboard</i> pode ser uma ferramenta poderosa. Este aplicativo permitirá aos estudantes visualizarem seu progresso acadêmico e receberem alertas personalizados sobre possíveis fatores de risco para evasão, assim como orientações e recursos de apoio.</p> <p>b) Feedback e comprometimento do estudante: permitir que os estudantes forneçam feedback diretamente no Dashboard ou aplicativo a ser criado, avaliando suas experiências e dificuldades acadêmicas de modo que isto possibilite a identificação dos fatores de insatisfação que podem não ser capturados por dados quantitativos.</p>
4) Expansão e personalização dos relatórios de evasão	<p>a) Relatórios personalizáveis para gestores: incluir uma funcionalidade para que coordenadores e gestores personalizem relatórios de acordo com as necessidades específicas de suas unidades acadêmicas sendo que estes relatórios poderão destacar dados de evasão por curso, por campus e por outros filtros específicos, permitindo um monitoramento mais focado.</p> <p>b) Análises de causalidade e tendências de longo prazo: incorporar recursos de análise preditiva, permitindo que o Dashboard faça projeções de tendências de evasão com base em dados históricos sendo que esta funcionalidade auxiliará na antecipação de cenários e no planejar ações preventivas de longo prazo.</p>
5) Treinamento e capacitação dos usuários	<p>a) Capacitação de gestores e coordenadores: realizar treinamentos periódicos para que gestores, coordenadores de curso e membros da equipe administrativa estejam capacitados a utilizar o Dashboard de forma eficaz, extraindo o máximo das funcionalidades disponíveis.</p> <p>b) Manuais e suporte técnico: disponibilizar manuais detalhados e um canal de suporte técnico para auxiliar no uso do Dashboard e solucionar possíveis dúvidas e problemas técnicos, o que facilitará a utilização desta ferramenta de modo pleno, possibilitando, ainda, que os erros operacionais sejam minimizados.</p>
6) Aprimoramento da interface e usabilidade	<p>a) Interface intuitiva e acessível: revisar a interface do Dashboard periodicamente com vistas a garantir que esta seja intuitiva e acessível, inclusive para usuários com baixa familiaridade em ferramentas de Business Intelligence (BI). Adicionalmente, poderão ser adotadas práticas de design responsivo também para facilitar o uso em dispositivos móveis.</p> <p>b) Feedback do usuário: implementar uma ferramenta para coleta de feedback sobre a interface do Dashboard, permitindo que os usuários façam sugestões de melhorias e indiquem eventuais dificuldades, visto que isto poderá servir como orientação para futuras atualizações e aprimoramentos.</p>
7) Parcerias e comparação com outras instituições	<p>a) Benchmarking interinstitucional: realizar parcerias com outras universidades que tenham problemas similares de evasão para comparar dados e estratégias de retenção sendo que, estas trocas de experiências poderão gerar perspectivas de análises valiosas e soluções inovadoras,</p>

	<p>além de permitir que a UFSB avalie sua eficácia em relação a instituições semelhantes.</p> <p>b) Publicação de relatórios anuais: divulgar relatórios anuais públicos com as principais estatísticas de evasão, as ações adotadas e os resultados obtidos. As publicações e produções resultantes poderão oportunizar o fortalecimento da transparência e da responsabilidade social, além de servir como um recurso para outras instituições interessadas em adotar práticas semelhantes.</p>
8) Pesquisa e desenvolvimento contínuo	<p>a) Expansão do modelo de Mineração de Dados Educacionais: investir na ampliação do modelo de mineração de dados educacionais para incluir novos algoritmos de aprendizado de máquina, permitindo análises mais complexas e precisas dos dados de evasão.</p> <p>b) Parcerias de pesquisa com outras Instituições de Ensino Superior: estabelecer colaborações com outras universidades e mesmo centros de pesquisa para aprimorar as funcionalidades do Dashboard e explorar novas metodologias de análise o que, potencialmente, favorecerá a ampliação de uma perspectiva inovadora e possibilitará que a UFSB esteja alinhada com as melhores práticas e avanços tecnológicos.</p>

Fonte: Própria (2024).

Estas recomendações visam tornar o "Observatório da Evasão" uma ferramenta de referência para a gestão educacional na UFSB, promovendo uma análise mais aprofundada, responsiva e eficaz sobre a evasão nos cursos de graduação. É relevante considerar que a adoção destas práticas poderão ser uma grande contribuição para que a universidade possa mitigar o impacto da evasão e implementar políticas de retenção mais eficientes e orientadas por dados.

6 CONSIDERAÇÕES FINAIS

Muito tem-se discutido sobre a evasão na educação, em especial no Ensino Superior, com o propósito de predizer as causas deste fenômeno. Inúmeras pesquisas, distribuídas por todos os continentes, têm buscado as variáveis críticas que motivam o estudante ao abandono desta etapa da formação educativa, desde a década de 1970.

Mesmo com estudos extensivos e inúmeras abordagens metodológicas, ainda não se chegou a um consenso sobre os fatores determinantes da evasão ou mesmo uma definição padronizada do que ela representa, indicativos da complexidade e multiplicidade de fatores impactantes nesta problemática.

A evasão acadêmica no Brasil se apresenta como um desafio complexo, ampliado pelas desigualdades regionais e pela diversidade característica de um país com dimensões continentais. Os índices de evasão variam amplamente entre os cursos e instituições, dificultando a definição de um padrão nacional, enquanto as políticas públicas, apesar de focadas em fatores socioeconômicos, não têm alcançado os resultados esperados num fenômeno multifatorial que exige soluções integradas, demonstrando sua insuficiência diante da complexidade deste fenômeno e evidenciando a dificuldade de estabelecer diretrizes nacionais únicas.

Consiste num desafio imenso estabelecer uma política pública nacional que consiga atender satisfatoriamente diferentes regiões, instituições, cursos de graduação e estudantes sobre um problema que tem resultados diferentes para cada um desses aspectos. Acredita-se que, determinando os fatores críticos para a evasão, avança-se também na direção que norteará essa política pública.

Assim, o problema de pesquisa elaborado foi determinar: quais os fatores relevantes para o enfrentamento da evasão nos cursos de graduação da Universidade Federal do Sul da Bahia (UFSB), potencializando os resultados destas ações para a qualificação da gestão pública educacional? Buscou-se aprofundar a compreensão da evasão na Universidade, identificando fatores críticos e estabelecendo relações entre eles.

Para compreender os dados coletados, que mostravam o panorama dos percursos acadêmicos de todos os estudantes, de todos os cursos de graduação, no

período de 2014 a 2023, foram realizados dois tipos de investigação: por Regressão Logística, subdivida em análises por Curso e análises por Disciplinas, e por Árvore de Decisão.

Após as análises de Regressão Logística por Cursos foram identificadas como variáveis críticas a Permanência e o Coeficiente de Rendimento, indicando esses fatores como os mais impactantes no desfecho da evasão. Na variável Permanência foi encontrada uma relação diretamente proporcional, ou seja, alunos com muito tempo de permanência (mais de 20 períodos letivos) tendem a evadir sem concluir a graduação, o que nos sugere um não avanço acadêmico.

Esse achado traz a forte recomendação de análise aprofundada dos índices de retenção da instituição, que podem apontar um problema oculto, visto que, em geral, altas taxas de retenção são percebidas como indicativos de sucesso institucional, mas as análises dos dados indicam uma outra conclusão: a estagnação no percurso acadêmico.

Nas análises de Regressão Logística por Disciplinas os fatores mais relevantes foram a Média Final e a Carga Horária total. Esses achados não descartam a existência de outros fatores vinculados ao processo de evasão, mas oferecem assertividade no planejamento de ações de contenção destes índices, pretendendo resultados mais eficientes e eficazes. A Média Final pode servir de referência para a sugestão de matrícula dos estudantes em programas acadêmicos de Monitoria ou Tutorias, tanto como monitores e tutores para os discentes de melhor desempenho acadêmico, quanto como tutorandos para os alunos com Médias Finais que indiquem possibilidade de evasão.

A variável de carga horária pode servir como indicativo para os formatos de componentes curriculares que têm oferecido os melhores resultados ao perfil da instituição, sendo indicado verificar se existem fatores associados à Carga Horária que possam modificar os resultados encontrados nesta primeira investigação.

Quando realizada a modelagem por Árvore de Decisão, este estudo encontrou como variáveis críticas para a evasão nos cursos de graduação da UFSB a Permanência, o Coeficiente de Rendimento, a Idade, o *Campus* onde o estudante está matriculado e a Distância entre a cidade de origem e de estudo. O achado de cinco

variáveis críticas corrobora com a literatura científica da área que indica múltiplos fatores associados à evasão.

Quanto à Permanência, ressalta-se novamente a importância de aprofundamento do estudo da evasão em paralelo com o entendimento da retenção na instituição. Compreender o percurso do aluno que se mantém vinculados à universidade durante anos e evade sem concluir a graduação passa por entender o quanto a retenção é, de fato, um índice de sucesso, uma vez que esse entendimento faz com que as instituições de ensino não criem estratégias para geri-la ou mesmo não que percebam esses dados como um alerta de problema.

O Coeficiente de Rendimento provavelmente é a variável mais facilmente percebida como indicativo de sucesso acadêmico, mas o presente estudo aponta também outras perspectivas que essa variável pode oferecer, como evasão de estudantes com altas notas e alta permanência na universidade. Esses achados, apesar de não indicarem dificuldades acadêmicas, podem indicar uma insatisfação com a universidade e a perda de estudantes com alto desempenho para outras instituições. É sugerida uma investigação qualitativa para identificar a causa deste tipo de abandono.

A variável Idade pode nortear abordagens distintas para faixas etárias específicas, necessidade de oferecimento de cursos de capacitação dos estudantes para uso de ferramentas acadêmicas digitais, ampliação de programas de suporte financeiro para estudantes que já são chefes de família, apoio psicológico, entre outras estratégias que considerem os diferentes perfis etários, suas habilidades e dificuldades.

Quanto ao *Campus*, percebe-se na *Árvore de Decisão*, que existem diferentes tomadas de decisão em função do local onde o estudante está vinculado. No caso estudado, esses diferentes comportamentos podem estar associados à área do curso de graduação, visto que cada *campus* concentra mais cursos voltados para o campo das Exatas, Humanas ou Saúde. Também pode estar associado a fatores socioeconômicos, como as cidades em que estão localizados, seus entornos, suas realidades econômicas, quantidade de instituições de ensino na região, entre outros fatores não abordados nesta análise.

Quanto à Distância, tanto a Árvore de Decisão quanto o *dashboard* do Observatório da Evasão, oferecem uma leitura rica e interessante que pode ser amplamente utilizada pela Pró-Reitoria de Ações Afirmativas (PROAF), podendo indicar, por exemplo, a necessidade de ampliação da oferta do Programa de Bolsa Permanência. A leitura desta variável permite identificar tanto a evasão quanto a permanência dos estudantes em função da cidade de origem.

Em relação ao modelo preditivo desenvolvido, foi aferida uma acurácia de 86,61% para o modelo preditivo apresentado, indicando boa confiabilidade, o trabalho mostrou-se promissor na identificação de padrões e na sugestão de intervenções eficazes. Estes resultados abordam tópicos que não foram encontrados em outras literaturas científicas sobre o tema da evasão.

Desta forma, o objetivo geral deste trabalho consistiu em apresentar estratégias específicas para a gestão da evasão acadêmica nos cursos de graduação da Universidade Federal do Sul da Bahia (UFSB) a partir da utilização da abordagem de Mineração de Dados Educacionais, entregando um Relatório de Consultoria, com vistas ao enfrentamento desta problemática e a sugestão de práticas para melhoria da gestão da evasão existente na instituição, foi completamente alcançado.

São limitações desta pesquisa a ausência de dados demográficos como sexo, etnia e renda, que contribuiriam para diversificar os fatores pesquisados, ampliando o número de possibilidades de fatores significativos para o desfecho da evasão. A limitação da ausência de dados demográficos ressalta a importância de incluir variáveis adicionais para expandir a robustez da análise.

Como futuros aprimoramentos desta pesquisa poderão realizadas a consolidação do Observatório da Evasão, com ampliação dos resultados exibidos neste *dashboard*, incluindo novas variáveis, revisando a confiabilidade e a qualidade dos dados inseridos. Também podem ser realizadas novas análises por meio de equações estruturais, com maior robustez das predições e o desenvolvimento de aplicativo para *smartphone* para monitoramento e prevenção da evasão, aproximando os estudantes das estratégias de gestão da evasão.

REFERÊNCIAS

BRASIL. Constituição (1988). Constituição da República Federativa do Brasil. Brasília, DF: Senado Federal, 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 18 out. 2023.

_____. Lei nº 9.394, de 20 de dezembro de 1996. Estabelece as Diretrizes e Bases da Educação Nacional. Brasília, DF, 1996. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9394.htm. Acesso em: 18 out. 2023.

_____. Lei nº 11.096, de 13 de janeiro de 2005. Institui o Programa Universidade para Todos – PROUNI, regula a atuação de entidades beneficentes de assistência social no ensino superior; altera a Lei nº 10.891, de 9 de julho de 2004, e dá outras providências. Brasília, DF, 2005. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/lei/l11096.htm. Acesso em: 23 out. 2023.

_____. Decreto nº 7.234, de 19 de julho de 2010. Dispõe sobre o Programa Nacional de Assistência Estudantil – PNAES. Brasília, DF, 2010. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2010/decreto/d7234.htm. Acesso em: 30 dez. 2020.

_____. IBGE – Instituto Brasileiro de Geografia Estatística. Cidades e Estados: Bahia. Disponível em: <https://www.ibge.gov.br/cidades-e-estados/ba/>. Acesso em: 19 dez. 2023.

_____. Ministério da Educação. Portaria Normativa nº 21, de 5 de novembro de 2012. Dispõe sobre o Sistema de Seleção Unificada – SISU. Disponível em: <https://sisugestao.mec.gov.br/docs/portaria-2012-21.pdf>. Acesso em: 25 out. 2023.

_____. Ministério da Educação. Portaria Normativa nº 389, de 9 de maio de 2013. Dispõe sobre o Programa de Bolsa Permanência. Disponível em: http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=68911-portaria-389-09052013&category_slug=programa-bolsa-permanencia&Itemid=30192. Acesso em: 25 out. 2023.

_____. Ministério da Educação; INEP – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Indicadores de Fluxo da Educação Superior (2018-2022). Publicado em 21 out. 2020 e atualizado em 10 out. 2023. Disponível em: <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/indicadores->

[educacionais/indicadores-de-fluxo-da-educacao-superior](#). Acesso em: 9 nov. 2023.

_____. Ministério da Educação; INEP – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Sinopse Estatística da Educação Superior 2022: Indicadores de Trajetória de Cursos de Graduação – Brasil – 2022. Publicado em 17 nov. 2020 e atualizado em 10 out. 2023. Disponível em: <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/sinopses-estatisticas/educacao-superior-graduacao>. Acesso em: 12 nov. 2023.

_____. SEI – Superintendência de Estudos Econômicos e Sociais da Bahia. Mapa Sul Baiano (2015). Disponível em: https://sei.ba.gov.br/site/geoambientais/mapas/pdf/regional/mesorregiao_geografica/mapa_sul_baiano_1.pdf.

_____. UFSB – Universidade Federal do Sul da Bahia. Dados e Indicadores da Graduação: Evasão (2023). Publicado: Segunda, 16 de dezembro de 2019, 15h25; última atualização em: Quarta, 27 de setembro de 2023, 14h55. Disponível em: <https://ufsb.edu.br/progeac/a-progeac/dados-e-indicadores-da-graduacao>.

AGRUSTI, F.; BONAVOLONTÀ, G.; MEZZINI, M. University dropout prediction through educational data mining techniques: A systematic review. **Journal of e-learning and knowledge society**, v. 15, n. 3, p. 161-182, 2019.

AINA, C. Parental background and university dropout in Italy. **Higher Education**, v. 65, p. 437-456, 2013.

AINA, C.; BAICI, E.; CASALONE, G.; PASTORE, F. The determinants of university dropout: A review of the socio-economic literature. **Socio-Economic Planning Sciences**, v. 79, p. 101102, 2022.

ALVES, M. A.; SEGATTO, C. I.; PINEDA, A. M. Changes in Brazilian education policy and the rise of right-wing populism. **British Educational Research Journal**, v. 47, n. 2, p. 332-354, 2021.

ANSELL, C.; GASH, A. Collaborative governance in theory and practice. **Journal of public administration research and theory**, v. 18, n. 4, p. 543-571, 2008.

ASSIS, V. D.; NEVES, V. N. S.; ARAGÃO, W. H. Um balanço da produção científica brasileira sobre trabalho docente (2008-2023). **Revista Educação & Formação**, v. 8, 2023.

BEHR, A.; GIESE, M.; TEGUIM K, H. D.; THEUNE, K. Early prediction of university dropouts—a random forest approach. **Jahrbücher für Nationalökonomie und Statistik**, v. 240, n. 6, p. 743-789, 2020.

CASTRO-LOPEZ, A.; CERVERO, A.; GALVE-GONZÁLEZ, C.; PUENTE, J.; BERNARDO, A. B. Evaluating critical success factors in the permanence in Higher Education using multi-criteria decision-making. **Higher Education Research & Development**, v. 41, n. 3, p. 628-646, 2022.

CHIAVENATO, I. **Introdução à teoria geral da administração**. Brasil: Elsevier, 2003.

CONSTATE-AMORES, A.; FLORENCIANO MARTÍNEZ, E.; NAVARRO ASENCIO, E.; FERNÁNDEZ-MELLIZO, M. Factors associated with university dropout. **Educación XXI**, v. 24, p. 17-44, 2021.

CONTINI, D.; CUGNATA, F.; SCAGNI, A. Social selection in higher education. Enrolment, dropout and timely degree attainment in Italy. **Higher Education**, v. 75, n. 5, p. 785-808, 2018.

DA RE, L.; ZAGO, G. Academic tutoring and dispersion: a project of research of the university of Padua (Italy). In: **ICERI2011 Proceedings**, p. 1457-1467. IATED, 2011.

DI PIETRO, M. S. Z. **Direito administrativo**. São Paulo: Atlas, 2010. v. 8.

DUNLEAVY, P.; MARGETTS, H.; BASTOW, S.; TINKLER, J. New public management is dead—long live digital-era governance. **Journal of public administration research and theory**, v. 16, n. 3, p. 467-494, 2006.

FARAH, M. F. S. Administração pública e políticas públicas. **Revista de Administração Pública**, v. 45, p. 813-836, 2011.

FÁVERO, L. P.; BELFIORE, P. **Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®**. Elsevier Brasil, 2017.

FERREIRA, A. B. D. H. **Novo dicionário Aurélio-século XXI**. Rio de Janeiro: Nova Fronteira, 1999. v. 1.

FERREIRA, F.; SANTOS, B. S.; MARQUES, B.; DIAS, P. FICAVIS: data visualization to prevent university dropout. In: *2020 24th International Conference Information Visualisation (IV)*, p. 57-62. IEEE, 2020.

FLORES, V.; HERAS, S.; JULIAN, V. Comparison of predictive models with balanced classes using the SMOTE method for the forecast of student dropout in higher education. **Electronics**, v. 11, n. 3, p. 457, 2022.

GRUS, J. **Data science do zero**. Rio de Janeiro: Alta Books, 2016. v. 1.

GUIMARÃES, T. D. A. **A nova administração pública e a abordagem da competência**. 2000.

GUSMÃO, G. M.; de HOLANDA, J. T.; FIALHO, H. V. M. L.; MOREIRA, F. D. O. P.; de ALENCAR GOMES, D.; CEOLIN, A. C.; CABRAL, R. M. Monitoramento e controle no gerenciamento de projetos do setor público brasileiro: uma revisão sistemática da literatura. **Revista de Gestão e Secretariado (Management and Administrative Professional Review)**, v. 14, n. 4, p. 6494-6505, 2023.

GUTIERREZ-PACHAS, D. A.; GARCIA-ZANABRIA, G.; CUADROS-VARGAS, E.; CAMARA-CHAVEZ, G.; GOMEZ-NIETO, E. A comparative study of WHO and WHEN prediction approaches for early identification of university students at dropout risk. In: **2021 XLVII Latin American Computing Conference (CLEI 2021)**, DOI: 10.1109/CLEI53233.2021.9640119.

GUTIERREZ-PACHAS, D. A.; GARCIA-ZANABRIA, G.; CUADROS-VARGAS, E.; CAMARA-CHAVEZ, G.; GOMEZ-NIETO, E. Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic, Socioeconomic, and Equity Factors through Machine Learning and Survival Analysis Methods in the Latin American Context. **Education Sciences**, v. 13, n. 2, p. 154, 2023. Disponível em: <https://doi.org/10.3390/educsci13020154>.

HAN, J.; PEI, J.; TONG, H. **Data mining: concepts and techniques**. Morgan Kaufmann, 2022.

KOROM, P.; PLANK, M. Elites: History of the concept. **International Encyclopedia of the Social & Behavioral Sciences**, v. 7, p. 390-395, 2015.

LANDIS, J. R.; KOCH, G. G. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. **Biometrics**, p. 363-374,

1977.

LIBÂNEO, J. C. O dualismo perverso da escola pública brasileira: escola do conhecimento para os ricos, escola do acolhimento social para os pobres. **Educação e Pesquisa**, v. 38, p. 13-28, 2012.

MASSA, S.; PULIAFITO, P. P. An application of data mining to the problem of the university students' dropout using markov chains. In: **European conference on principles of data mining and knowledge discovery**. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 51-60, set. 1999.

MARIANO, A. M.; ROCHA, M. S. **Revisão da Literatura**: Apresentação de uma Abordagem Integradora. AEDM International Conference – Economy, Business and Uncertainty: Ideas for a European and Mediterranean industrial policy. Reggio Calabria (Italia), 2017.

MEZZINI, M.; BONAVOLONTÀ, G.; AGRUSTI, F. Predicting university dropout by using convolutional neural networks. In: **INTED2019 Proceedings**. IATED, p. 9155-9163, 2019.

MISRA, D.; STENSAKER, B. Balancing control and legitimacy in higher education policy implementation. **Studies in Higher Education**, p. 1-14, 2024.

MONTMARQUETTE, C.; MAHSEREDIJAN, S.; HOULE, R. The determinants of university dropouts: a bivariate probability model with sample selection. **Economics of education review**, v. 20, n. 5, p. 475-484, 2001.

MOTTA, F. C. P.; ALCADIPANI, R. Teoria Crítica. *Revista de Administração de Empresas* (FGV), São Paulo, v. 41, n. 4, p. 108, 2001.

MOTTA, F. C. P. Administração e participação: reflexões para a educação. *Educação e Pesquisa*, São Paulo, v. 29, n. 2, p. 369-373, jul.-dez. 2003.

NIEROTKA, R. L.; SALATA, A.; KLITZKE MARTINS, M. Fatores associados à evasão no ensino superior: um estudo longitudinal. **Cadernos de Pesquisa**, v. 53, e09961, 2023.

ORTIZ-LOZANO, J. M.; APARICIO-CHUECA, P.; TRIADÓ-IVERN, X. M.; ARROYO-

BARRIGÜETE, J. L. Early dropout predictors in social sciences and management degree students. **Studies in Higher Education**, p. 1-14, 2023.

OSBORNE, D. Reinventing government. **Public productivity & management Review**, p. 349-356, 1993.

OSBORNE, S. P. From public service-dominant logic to public service logic: are public service organizations capable of co-production and value co-creation? **Public Management Review**, v. 20, n. 2, p. 225-231, 2018.

PALACIOS, C. A.; REYES-SUÁREZ, J. A.; BEARZOTTI, L. A.; LEIVA, V.; MARCHANT, C. Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. **Entropy**, v. 23, n. 4, p. 485, 2021.

PEDRAZA-NAVARRO, I.; GONZÁLEZ-RAMÍREZ, T. Educational quality and dropout risk: a causal analysis of the university dropout phenomenon. In: **7th International Conference on Higher Education Advances (HEAD'21)**. Editorial Universitat Politècnica de València, p. 881-888, jul. 2021.

ROBINSON, L. et al. Digital Inclusion Across the Americas and Caribbean. **Social Inclusion**, v. 8, n. 2, p. 244-259, 2020. DOI: <https://doi.org/10.17645/si.v8i2.2632>.

RODRIGUEZ-GÓMEZ, D.; FEIXAS, M.; GAIRÍN, J.; MUÑOZ, J. L. Understanding Catalan University Dropout from a Comparative Approach. **Social and Behavioral Sciences**, v. 46, p. 1424-1429, 2012. DOI: <https://doi.org/10.1016/j.sbspro.2012.05.314>.

SALTZ, J. S. CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. **IEEE International Conference on Big Data (Big Data)**, 2021.

SANTOS, C. O.; PILATTI, L. A.; BONDARIK, R. Evasão no ensino superior brasileiro: conceito, mensuração, causas e consequências. **Debates em Educação**, v. 14, n. 35, p. 294-314, 2022.

SANTOS, K. J. D. O.; MENEZES, A. G.; CARVALHO, A. B.; MONTESCO, C. A. Supervised learning in the context of educational data mining to avoid university students dropout. In: **2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)**. IEEE, v. 2161, p. 207-208, jul. 2019.

SANTOS JÚNIOR, J. D. S.; MAGALHÃES, A. M. D. S.; REAL, G. C. M. Dropout management in Brazilian educational policies: from undergraduate to stricto sensu postgraduate. *ETD Educação Temática Digital*, v. 22, n. 2, p. 460-478, 2020.

DA SILVA, L. M.; DIAS, L. P. S.; RIGO, S.; BARBOSA, J. L. V.; LEITHARDT, D. R. F.; LEITHARDT, V. R. Q. A literature review on intelligent services applied to distance learning. *Education Sciences*, v. 11, n. 11, p. 666, 2021. Disponível em: <https://doi.org/10.3390/educsci11110666>. Acesso em: [data de acesso].

SILVA, M. D. S. T.; CORREIA, S.; ARAÚJO MACHADO, P.; OLIVEIRA, V. M. Adoption of information technology in public administration: a focus on the organizational factors of a Brazilian federal university. *Teoria e Prática em Administração*, v. 10, n. 2, p. 138-153, 2020.

SOMMER, M.; DUMONT, K. Psychosocial factors predicting academic performance of students at a historically disadvantaged university. *South African Journal of Psychology*, v. 41, n. 3, p. 386-395, 2011.

TAMADA, R. C. P.; CUNHA, I. C. K. O. Gestão por competências na administração pública brasileira: uma revisão integrativa da literatura. *Revista do Serviço Público (RSP)*, v.73, n. 3, p. 426-450, 2022.

TINTO, V. Dropout from higher education: a theoretical synthesis of recent research. *Review of Educational Research*, v. 45, n. 1, p. 89-125, 1975.

VALENTE, J. A.; ALMEIDA, M. E. B. Brazilian technology policies in education: history and lessons learned. *Education Policy Analysis Archives*, v. 28, n. 94, 2020.

VALLEJOS, C. A.; STEEL, M. F. Bayesian survival modelling of university outcomes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, v. 180, n. 2, p. 613-631, 2017.

VERGARA, S. C. *Tipos de pesquisa em administração*. São Paulo: Atlas. 1990.

VIGNOLES, A. F.; POWDTHAVEE, N. The socioeconomic gap in university dropouts. *The BE Journal of Economic Analysis & Policy*, v. 9, n. 1, 2009.

VILORIA, A. et al. Integration of data technology for analyzing university dropout. **Procedia Computer Science**, v. 155, p. 569-574, 2019.

VILORIA, A. et al. Dropout-permanence analysis of university students using data mining. In: *Intelligent Computing, Information and Control Systems: ICICCS 2019*. p. 374-383. **Springer International Publishing**, 2020.

VYGOTSKY, L. S. **Pensamento e linguagem**. 4. ed. São Paulo: Editora, 2008.

ANEXO 1

Árvore programada:

```

Trimestre cursado > 0.500
|   Trimestre cursado > 8.500
|   |   CR > 8.405
|   |   |   CR > 9.995
|   |   |   |   Idade > 25.500: Evade {Não evade=6, Evade=24}
|   |   |   |   Idade ≤ 25.500
|   |   |   |   |   Trimestre cursado > 9.500: Evade {Não evade=0,
Evade=2}
|   |   |   |   |   Trimestre cursado ≤ 9.500: Não evade {Não evade=4,
Evade=0}
|   |   |   |   CR ≤ 9.995
|   |   |   |   |   Trimestre cursado > 20.500
|   |   |   |   |   |   Campus/CUNI = CJA: Evade {Não evade=3, Evade=15}
|   |   |   |   |   |   Campus/CUNI = CPF
|   |   |   |   |   |   |   Idade > 29.500: Evade {Não evade=2, Evade=8}
|   |   |   |   |   |   |   Idade ≤ 29.500: Não evade {Não evade=4,
Evade=1}
|   |   |   |   |   |   |   Campus/CUNI = CSC
|   |   |   |   |   |   |   |   Idade > 37.500
|   |   |   |   |   |   |   |   |   Idade > 52.500: Não evade {Não evade=2,
Evade=0}
|   |   |   |   |   |   |   |   |   Idade ≤ 52.500: Evade {Não evade=0,
Evade=3}
|   |   |   |   |   |   |   |   |   Idade ≤ 37.500: Não evade {Não evade=7,
Evade=1}
|   |   |   |   |   |   |   Trimestre cursado ≤ 20.500: Não evade {Não evade=1185,
Evade=609}
|   |   |   |   |   CR ≤ 8.405
|   |   |   |   |   |   Campus/CUNI = CJA: Evade {Não evade=94, Evade=273}
|   |   |   |   |   |   Campus/CUNI = CPF
|   |   |   |   |   |   |   Idade > 26.500
|   |   |   |   |   |   |   |   Trimestre cursado > 19.500
|   |   |   |   |   |   |   |   |   Distância = 0: Não evade {Não evade=3,
Evade=2}
|   |   |   |   |   |   |   |   |   Distância = 1: Evade {Não evade=1, Evade=13}
|   |   |   |   |   |   |   |   |   Trimestre cursado ≤ 19.500
|   |   |   |   |   |   |   |   |   Trimestre cursado > 10.500
|   |   |   |   |   |   |   |   |   |   CR > 7.195
|   |   |   |   |   |   |   |   |   |   |   CR > 8.385: Evade {Não evade=2,
Evade=6}
|   |   |   |   |   |   |   |   |   |   |   |   CR ≤ 8.385: Não evade {Não evade=133,
Evade=47}
|   |   |   |   |   |   |   |   |   |   |   |   |   CR ≤ 7.195: Evade {Não evade=0, Evade=7}
|   |   |   |   |   |   |   |   |   |   |   |   |   Trimestre cursado ≤ 10.500: Evade {Não
evade=3, Evade=13}
|   |   |   |   |   |   |   |   |   |   |   |   |   Idade ≤ 26.500
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Trimestre cursado > 15.500
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Trimestre cursado > 19: Evade {Não evade=0,
Evade=2}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Trimestre cursado ≤ 19: Não evade {Não
evade=3, Evade=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Trimestre cursado ≤ 15.500: Evade {Não evade=4,
Evade=27}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Campus/CUNI = CSC
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   CR > 8.365
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Idade > 47.500: Evade {Não evade=0, Evade=2}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Idade ≤ 47.500: Não evade {Não evade=15, Evade=6}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   CR ≤ 8.365: Evade {Não evade=88, Evade=290}
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Trimestre cursado ≤ 8.500: Evade {Não evade=70, Evade=1309}

```