INTEGRAÇÃO ENTRE CIÊNCIA DA INFORMAÇÃO E CIÊNCIA DE DADOS: análise de conteúdo aplicada à seleção de variáveis para modelagem preditiva

INTEGRATION BETWEEN INFORMATION SCIENCE AND DATA SCIENCE: content analysis applied to variable selection for predictive modeling

Alexsandra Schmidt Fabrin[1]
https://orcid.org/0009-0004-2727-7218
Jordan Paulesky Juliani[2]
https://orcid.org/0000-0001-7823-6644

RESUMO
As variáveis utilizadas em modelos de predição constituem elementos fundamentais para a qualidade dos resultados obtidos, tornando sua identificação e seleção etapas relevantes no processo de modelagem. O presente artigo tem como objetivo apresentar a aplicação da análise de conteúdo, técnica amplamente utilizada na Ciência da Informação, na exploração e seleção de variáveis para predição do abandono escolar dos alunos no ensino médio da rede estadual de ensino de Santa Catarina. A pesquisa caracteriza-se como aplicada, de objetivos exploratórios, abordagem predominantemente qualitativa e procedimentos técnicos documental e estudo de caso. O corpus foi constituído por 16 estudos selecionados por meio de levantamento bibliográfico, dos quais foram identificadas 269 variáveis, posteriormente agrupadas em 114 agrupamentos por similaridade semântica e conceitual e classificadas segundo as características Disponibilidade, Relevância e Dimensão. Os resultados demonstram que a análise de conteúdo possibilitou organizar e sistematizar as variáveis de

Artigo submetido em 14/05/2026 e aceito para publicação em 31/07/2026.

[1] Doutoranda em Gestão de Unidades de Informação na Universidade do Estado de Santa Catarina (Udesc). Mestre em Gestão de Unidades de Informação pela Universidade do Estado de Santa Catarina (Udesc, 2025). Especialista em Educação Matemática pela Universidade do Sul de Santa Catarina (Unisul, 2006). Licenciada em Matemática pela Universidade Federal de Santa Catarina (UFSC, 2002). Técnica da Secretaria de Estado da Educação de Santa Catarina (SED/SC). E-mail: alexsandras@sed.sc.gov.br.
[2] Doutor em Engenharia e Gestão do Conhecimento pela Universidade Federal de Santa Catarina (UFSC, 2008). Mestre em Engenharia de Produção pela Universidade Federal de Santa Catarina (UFSC, 2002). Especialista em Ciência de Dados e Inteligência Artificial pela Pontifícia Universidade Católica do Rio Grande do Sul (PUC-RS, 2024). Graduado em Administração pela Universidade Federal de Santa Catarina (UFSC, 2005) e em Ciência da Computação pela Universidade do Vale do Itajaí (Univali, 1999). Professor da Universidade do Estado de Santa Catarina (Udesc) e da University of Wisconsin – Sevens Point. E-mail: jordan.juliani@udesc.br e jordan.juliani@uwsp.edu.

forma contextualizada e estruturada, subsidiando sua seleção para modelagem preditiva. Conclui-se que a aplicação da análise de conteúdo favorece a integração entre princípios da Ciência da Informação e métodos da Ciência de Dados, oferecendo uma sistematização inicial das variáveis identificadas na literatura que poderá subsidiar pesquisas futuras voltadas à seleção de variáveis e à modelagem preditiva do abandono escolar.

Palavras-Chave: ciência da informação; ciência de dados; predição do abandono escolar; variáveis; análise de conteúdo.

ABSTRACT
The variables used in predictive models are fundamental to the quality of the results obtained, making their identification and selection relevant steps in the modeling process. This article aims to present the application of content analysis, a technique widely used in Information Science, in the exploration and selection of variables for predicting high school dropout among students in the Santa Catarina State public school system. The research is characterized as applied, with exploratory objectives, a predominantly qualitative approach, and documentary research and a case study as its technical procedures. The corpus consisted of 16 studies selected through a bibliographic survey, from which 269 variables were identified, subsequently grouped into 114 groupings based on semantic and conceptual similarity and classified according to the characteristics of Availability, Relevance, and Dimension. The results demonstrate that content analysis made it possible to organize and systematize the variables in a contextualized and structured manner, supporting their selection for predictive modeling. It is concluded that the application of content analysis fosters the integration of Information Science principles and Data Science methods, providing an initial systematization of the variables identified in the literature that may support future research on variable selection and predictive modeling of school dropout.

Keywords: *information science; data science; school dropout prediction; variables; content analysis.*

1 INTRODUÇÃO

A Ciência da Informação e a Ciência de Dados, disciplinas interdisciplinares e interligadas, compartilham preocupações com a necessidade, o uso e o aproveitamento estratégico da informação. Conforme discutido por Moutinho *et al*. (2024), essa convergência evidencia o potencial da primeira em contribuir diretamente para o desenvolvimento de modelos baseados em dados, especialmente no que se refere à organização e curadoria, aspectos fundamentais para garantir a qualidade, a relevância e a compreensão das variáveis envolvidas.

Neste sentido, a análise de conteúdo, tradicionalmente aplicada à categorização e interpretação de informações textuais, é aqui proposta como

metodologia que fortalece essa interface. Tal articulação torna-se ainda mais evidente quando se considera a definição da Ciência de Dados como “[...] processos, modelos e tecnologias que estudam os dados durante todo o seu ciclo de vida: da produção ao descarte [...]” (Amaral, 2016, p. 4).

As variáveis, nesse contexto, configuram-se como unidades de informação que descrevem entidades e permitem que modelos[3] preditivos da Ciência de Dados sejam aplicados, correlacionando as informações e gerando *insights* valiosos para o contexto.

Estudos da Ciência de Dados, como os de Barros *et al*. (2019), Barros (2020), Chinnasamy e Balasubramanian (2022a, 2022b), Colpani (2018), Krüger, Britto Jr. e Barddal (2023), Lopes Filho (2021), Maheshwari *et al*. (2020), Masood e Begum (2022), Mduma (2023), Queiroga *et al*. (2022), Queiroga *et al*. (2024), Silva (2021), Souza (2020), Teles (2023) e Wongvorachan, He e Bulut (2023), se utilizam das variáveis para a aplicação de algoritmos[4] que geram modelos para predição/detecção/previsão dos alunos propensos ao abandono escolar.

O abandono escolar, no qual o aluno deixa de frequentar a escola durante o ano letivo, é uma triste realidade. Em 2025, no ensino médio, as taxas de abandono no Brasil foram de 2,2% e em Santa Catarina de 3,3%, sendo que para a rede estadual de ensino de Santa Catarina a taxa foi ainda maior, de 4,0% (Brasil, 2026b).

Ainda em 2025, segundo dados do Censo Escolar da Educação Básica, o Brasil atendeu 7.370.879 matrículas no ensino médio e Santa Catarina foi responsável por atender 269.965 dessas matrículas, sendo 216.109 (80,05%) atendidas pela rede estadual de ensino (Brasil, 2026a).

Identificar fatores que possam contribuir para que o abandono escolar aconteça, com o auxílio de Ciência de Dados, requer a utilização de variáveis disponíveis e significativas para o contexto, o que leva a seguinte proposição: quais variáveis devem ser consideradas na aplicação de modelos de predição do abandono

[3] Modelo: produto resultante da aplicação de um algoritmo sobre uma base de dados, “[...] representa o que foi aprendido pelo algoritmo ao utilizar a base de dados” (Sicsú; Samartini; Barth, 2023, p. 17).

[4] Algoritmo: procedimentos a serem executados, contendo instruções e regras, que transformam um conjunto de dados de entrada em um ou mais valores de saída (Sicsú; Samartini; Barth, 2023).

escolar dos alunos no ensino médio da rede estadual de ensino de Santa Catarina?

Diante do exposto, o objetivo geral desta pesquisa é apresentar um estudo de caso da aplicação da análise de conteúdo na exploração e seleção de variáveis voltadas à modelagem preditiva do abandono escolar dos alunos no ensino médio da rede estadual de ensino de Santa Catarina.

2 DESENVOLVIMENTO

Apresenta-se a seguir breve contextualização dos trabalhos relacionados, de Ciência de Dados, e suas variáveis, que apresentam enfoque na predição/detecção/previsão dos alunos propensos ao abandono escolar no ensino médio.

Aborda-se também a análise de conteúdo, que será utilizada na seleção das variáveis para aplicação em modelos de predição do abandono escolar dos alunos no ensino médio da rede estadual de ensino de Santa Catarina.

2.1 TRABALHOS RELACIONADOS E SUAS VARIÁVEIS

Os trabalhos de Barros *et al*. (2019) e Barros (2020) aplicaram modelos de Ciência de Dados, de predição, para identificar alunos com risco de abandono no Instituto Federal do Rio Grande do Norte e o de Silva (2021) no Instituto Federal de Mato Grosso.

Silva (2021, p. 16) enfatizou que “é necessário escavar montanhas de dados à procura de preciosidades, associações, recorrências e padrões imperceptíveis de início e que passam a ser revelados”.

Krüger, Britto Jr. e Barddal (2023), salientaram que utilizar técnicas para a previsão do abandono escolar não é ideia nova, sendo o campo desafiador e com obstáculos, principalmente aqueles ligados à obtenção de dados sobre os alunos, abordaram a previsão do abandono para instituições da rede privada de ensino brasileira e Souza (2020) para uma instituição privada de Belo Horizonte.

Colpani (2018) trabalhou com o contexto do Pará, Lopes Filho (2021) abordou a problemática para a rede estadual de ensino de São Paulo, Queiroga *et al*. (2024) para o estado do Espírito Santo e Teles (2023) para o Ceará. Teles (2023) enfatiza a importância de serem realizadas análises qualitativas e quantitativas dos elementos envolvidos no processo de predição do abandono escolar de maneira a facilitar a leitura e interpretação, auxiliando a tomada de decisão.

Os trabalhos de Chinnasamy e Balasubramanian (2022a, 2022b), Maheshwari *et al*. (2020) e Masood e Begum (2022) dizem respeito ao contexto da Índia e o de Mduma (2023), além do contexto indiano trabalhou também com o da Tanzânia. Maheshwari *et al*. (2020) também apontaram para a importância de ajudar os educadores a se concentrarem nos fatores associados e assim orientar os alunos na tentativa de evitar o abandono escolar.

Já a pesquisa de Queiroga *et al*. (2022) foi focada para alunos do Uruguai e a de Wongvorachan, He e Bulut (2023), para os dos Estados Unidos da América.

Levando em consideração as variáveis listadas como selecionadas nos trabalhos relacionados, segue o Quadro 1:

Quadro 1 - Variáveis listadas como selecionadas nos trabalhos relacionados

Autor(es)	Título do trabalho	Variáveis
Barros *et al*. (2019)	*Predictive models for imbalanced data: a school dropout perspective*	Nota português, Nota matemática, Número de dependências português, Número de dependências matemática, Porcentagem de presença português, Porcentagem de presença matemática, se cursou ensino fundamental somente em escola pública, área residência do aluno, Companhia doméstica, Estado civil, Curso, Situação do imóvel, Escolaridade mãe, Escolaridade pai, Escolaridade responsável legal, Raça, Quem é o responsável legal, Descrição trabalho do aluno, Sexo, Necessidade especial, Número de pessoas no domicílio, Campus estudantil, Quantidade computador ou notebook, Renda familiar bruta, Tempo entre conclusão do ensino fundamental e ingresso no IFRN.
Barros (2020)	Um processo orientado a dados para geração de modelo de predição de evasão escolar	Média aritmética das disciplinas no primeiro ano, Desvio padrão da média aritmética das disciplinas, Se o aluno parou de estudar, Residência informada, Campus João Câmara, Campus Mossoró, Campus Natal Central, Curso técnico em têxtil, Responsável financeiro cônjuge, Curso técnico em controle ambiental, Curso técnico em geologia.

Chinnasamy e Balasubramanian (2022a)	*Rank biserial stochastic feature embed bivariate kernelized regressive bootstrap aggregative classifier for school student dropout prediction*	Abandono, Identificação (ID) do aluno, Gênero, Casta, Notas de matemática, Notas de inglês, Notas de ciências, Professor de ciências, Professor de línguas, Guardião, Internet, Identificação (ID) da escola, Total de alunos, Total de banheiros e Ano do estabelecimento.
Chinnasamy e Balasubramanian(2022b)	*Tucker's congruence regressive feature projected tversky discriminant multiple instance learning boost data classification for school student dropout prediction*	Abandono, Identificação (ID) do aluno, Gênero, Casta, Notas de matemática, Notas de inglês, Notas de ciências, Professor de ciências, Professor de línguas, Guardião, Internet, Identificação (ID) da escola, Total de alunos, Total de banheiros e Ano do estabelecimento.
Colpani (2018)	Mineração de dados educacionais: um estudo da evasão no ensino médio com base nos indicadores do Censo Escolar	Taxa de distorção idade-série.
Krüger, Britto Jr. e Barddal (2023)	*An explainable machine learning approach for student dropout prediction*	Escola, Série, Ano escolar, Classe, Estágio educacional, Idade esperada na série, Idade, Sexo, Faltas, Anos nessa escola, Envolvido em aulas religiosas, Envolvido em aulas extracurriculares, Estudo em tempo integral, Taxa de Matrícula, Aumento da taxa de matrícula, Descontos na matrícula, Profissão dos pais, Estado civil dos pais, Nível de escolaridade dos pais, Se mora com os pais, Localização da residência, Produto Interno Bruto (PIB), Índice de Desenvolvimento Humano Municipal (IDHM), Expectativa de vida, Expectativa de estudo, Notas em cada trimestre, Notas acumuladas dos trimestres, Diferença entre a nota de um trimestre e outro, Abandono.

Lopes Filho (2021)	Detecção de estudantes em risco de evasão escolar usando aprendizagem de máquina	Sexo, Cor/raça, Idade, Idade sugerida para a série, Distorção idade/série, Tipo ensino se fundamental ou médio, Ano ou série, Turno, Nota IDESP por escola, Repasse verbas por escola, Valor repasse verbas, Valor Indicador Nível Socioeconômico (Inse) escola, Total alunos escola, Total alunos evadidos escola, Porcentagem alunos evadidos escola, Nota no bimestre português, geografia, história, matemática, Frequência português, geografia, história, matemática, Situação setor censitário, Setor censitário considerado aglomerado subnormal, Grupo do IPVS escola, Número domicílios particulares e coletivos setor escola, Média moradores domicílios particulares permanentes setor escola, Proporção crianças 0 a 5 anos setor escola, Rendimento médio domicílios setor escola, Renda per capta domicílios, Proporção domicílios sem renda per capta, Idade média pessoas responsáveis, Rendimento médio responsáveis, Proporção pessoas alfabetizadas, Proporção responsável 10 a 29 anos, Participação renda responsável nos domicílios, Idade média mulheres responsáveis, Proporção mulheres responsáveis alfabetizadas, Proporção mulheres responsáveis menos 30 anos, Rendimento médio mulheres responsáveis, Proporção abastecimento água, Proporção banheiros domicílios, Proporção domicílios coleta de lixo, Proporção domicílios acesso energia elétrica, Proporção total munícipes situados no setor da escola.
Maheshwari *et al*. (2020)	Prediction of factors associated with the dropout rates of primary to high school students in India using data mining tools	Percentual de escolas com banheiro para meninos e meninas e Proporção aluno/professor.
Masood e Begum (2022)	*Comparison of resampling techniques for imbalanced datasets in student dropout prediction*	Gênero, Casta, Notas de matemática, Notas de inglês, Notas de ciências, Professor de ciências, Professor de línguas, Guardião, Internet, Abandono.
Mduma (2023)	*Data balancing techniques for predicting student dropout using machine learning*	Para dados da Tanzânia: Principal fonte de renda familiar, Proporção de banheiros para meninos, Se a escola possui sala privada para meninas, Região, Distrito, Aldeia, Gênero, Se os pais olham o caderno de exercícios da criança uma vez por semana, Refeições domésticas por dia, Se o aluno leu um livro com seus pais na semana passada, Se os pais discutiram sobre o progresso da criança no último semestre com o professor, Idade, Tipo de área, Tamanho da família, Razão de banheiros femininos, Razão de reuniões de pais e professores, Razão de alunos em sala de aula, Razão de professor por aluno e Abandono. Para dados da Índia: Abandono, Identificação (ID) do aluno, Gênero, Casta, Notas de matemática, Notas de inglês, Notas de ciências, Professor de ciências, Professor de línguas,

		Guardião, Internet, Identificação (ID) da escola, Total de alunos, Total de banheiros e Ano do estabelecimento.
Queiroga *et al*. (2022)	*Early prediction of at-risk students in secondary education: a countrywide K-12 learning analytics initiative in Uruguay*	Utilizou 9 sistemas de gestão educacional para composição dos dados, como banco de dados de trajetória e desempenho dos alunos, programas de assistência social, notas e faltas dos alunos, informações sobre as escolas, entre outros, gerando uma série temporal do ensino primário e secundário básico.
Queiroga *et al*. (2024)	*Anticipating student abandonment and failure: predictive models in high school settings*	Idade, Gênero, Turno, Assistência governamental, Etnia, Deficiência, Necessidade de Assistência especial, Nota em Matemática, Nota em Sociologia, Nota em português, Nota em Filosofia, Nota em Inglês, Nota em Biologia, Nota em Física, Nota em Educação Física, Nota em Arte, Nota em Química, Faltas em Matemática, Faltas em Sociologia, Faltas em Português, Faltas em Filosofia, Faltas em Inglês, Faltas em Biologia, Faltas em Física, Faltas em Educação Física, Faltas em Arte, Faltas em Química.
Silva (2021)	Mineração de dados aplicada à predição da evasão escolar no ensino médio	Se trabalha, Dependência administrativa da escola onde realizou a maior parte dos estudos, Reserva de vagas por cotas, Conceito obtido nas provas do processo seletivo, Conceito acadêmico obtido durante o curso, Bolsa de apoio financeiro.
Souza (2020)	*Machine learning* e a evasão escolar - análise preditiva no suporte à tomada de decisão	Necessidade especial, Idade, Sexo, Cidade do aluno, Estado civil, Titulação, Cor/Raça, Ocupação, Faixa de renda mensal, Escola fundamental, Escola médio, Turno, Código do curso, Carga horária, Unidade de negócio, Cidade da escola, Região da escola, Situação do aluno, Frequência, Disciplina, Aluno, Situação Financeira, Tipo Baixa Renda, Modalidade de ensino.
Teles (2023)	Aprendizagem de máquina para predição do abandono e evasão dos estudantes do estado do Ceará	Desempenho em português, Desempenho em matemática, Cor/Raça, Etapa de ensino, Presença da mãe em bases de violência.
Wongvorachan, He e Bulut (2023)	*A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in*	Status socioeconômico, Autoeficácia matemática, Interesse curso matemática outono, Percepção sobre utilidade científica, Autoeficácia científica, Sentimento pertencimento escolar, Engajamento escola, Avaliação clima escolar, Percepção conselheiro expectativas do professor, do conselheiro, do diretor, Média das notas (*Grade Point Average* (*GPA*)), Horas gastas lição casa/estudo dia, Nível escolaridade mãe/ responsável, Nível escolaridade pai/ responsável, quão longe na escola aluno pensa que irá chegar, quão longe na escola

	educational data mining	pais acham aluno irá chegar, Frequência não faz dever de casa, vai sem lápis ou papel, vai sem livros, vai atrasado, Acha que estudar compensa com bom emprego, Acha que mesmo estudando não vai entrar faculdade, Acha que mesmo estudando família não vai conseguir pagar faculdade, Acha trabalhar mais importante que faculdade, Amigo mais próximo tira boas notas, Amigo mais próximo tem interesse na escola, Amigo mais próximo frequenta aulas regularmente, Amigo mais próximo planeja ir para faculdade, Horas gastas lição de casa/estudo de matemática, ciências, Está certo que se formará no ensino médio, Pais são contactados sobre comportamento problemático, sobre frequência ruim, sobre desempenho ruim, Abandonou ensino médio 2016 e situação matrícula ensino superior fevereiro 2016.

Fonte: Elaborado pelos autores (2025).

O levantamento dessas variáveis, indicadas como utilizadas pelos autores, aponta para possíveis variáveis a serem utilizadas em modelos de predição do abandono escolar dos alunos no ensino médio da rede estadual de ensino de Santa Catarina.

Visto esse contexto, apresenta-se a análise de conteúdo.

2.2 ANÁLISE DE CONTEÚDO

Análise de conteúdo é uma técnica que auxilia na realização do tratamento de dados predominantemente qualitativos, interpretação e categorização de conteúdos textuais, entrevistas e outros tipos de documentos. Consiste na pré-análise do conteúdo, na exploração do material, agrupando e classificando em categorias, sendo essas, essenciais para interpretação, síntese e conclusão dos achados, com a interpretação dos resultados (Bardin, 2009).

Nesse sentido, ao utilizar a análise de conteúdo para estruturar e organizar o conjunto de variáveis que compõem os modelos preditivos, estabelece-se uma aplicação prática de uma técnica utilizada na Ciência da Informação em benefício da Ciência de Dados. Essa abordagem valoriza a curadoria informacional e o tratamento sistemático da informação, contribuindo para a qualidade e a contextualização do modelo gerado.

A análise de conteúdo, segundo Bardin (2009), pode ser utilizada na criação de categorias para um grupo de informações. Por meio da análise quantitativa pode-se observar a frequência com que as variáveis são citadas, criando agrupamentos, e pela análise qualitativa, a presença ou ausência de certas características, que auxiliam na classificação, criando categorias para esses agrupamentos.

Como regras básicas para a análise de conteúdo ser válida, Bardin (2009) ressalta que as categorias devem ser adequadas ao contexto (segundo ao qual o estudo se aplica), objetivas (de modo a terem clareza de entendimento), exclusivas (com classificação única), homogêneas (com pontos em comum) e exaustivas (de forma que compreendam a totalidade do universo analisado).

O pesquisador exerce papel primordial na análise de conteúdo, que pode ser tanto quantitativa quanto qualitativa, com sua competência e sensibilidade, num exercício de leitura, imersão e reflexão perante os dados, as informações. Segundo Bardin (2009, p. 41): “o analista é como um arqueólogo. Trabalha com vestígios: os documentos que pode descobrir ou suscitar”.

De forma a garantir qualidade e comunicação clara dos resultados obtidos, a atuação do pesquisador na análise de conteúdo envolve a organização do material, a interpretação crítica, a codificação e categorização dos dados, e a interpretação dos mesmos e de seus resultados (Bardin, 2009).

Dito isso, apresenta-se a seguir os procedimentos metodológicos.

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa caracteriza-se, quanto à natureza, como aplicada, tratando e buscando solução para um problema específico, concreto e prático. Possui objetivos exploratórios, busca familiaridade com o problema e a identificação de elementos que possam auxiliar na predição do abandono escolar.

A abordagem é predominantemente qualitativa, e utiliza a análise de conteúdo, conforme proposta por Bardin (2009), como método para identificação, organização e

seleção de variáveis para modelagem preditiva do abandono escolar dos alunos no ensino médio da rede estadual de ensino de Santa Catarina.

Quanto aos procedimentos técnicos, a pesquisa classifica-se como documental e estudo de caso, uma vez que utiliza documentos científicos como corpus de análise e tem como contexto específico a rede estadual de ensino de Santa Catarina. O corpus da pesquisa foi constituído a partir de levantamento bibliográfico realizado em 2024, em dissertações, teses e artigos científicos das bases BDTD, Oasisbr, Catálogo de Teses e Dissertações da CAPES, Portal de Periódicos CAPES, SciELO, Web of Science, Scopus e EBSCOhost. As estratégias de busca foram elaboradas em português e inglês, combinando termos relacionados ao abandono escolar no ensino médio ("ensino médio", "abandono escolar", "evasão escolar", "*high school*", "*school dropout*", "*truancy*") e à Ciência de Dados ("predição", "ciência de dados", "mineração de dados", "análise preditiva", "*data science*", "*data mining*" e "*prediction*").

Foram recuperados 1.747 documentos. Após a remoção de duplicidades e a aplicação do recorte temporal (publicações a partir de 2018), 888 documentos seguiram para leitura dos títulos e verificação de aderência à pesquisa. Dessa etapa, 116 documentos foram selecionados para leitura dos resumos, com o objetivo de identificar estudos sobre a predição do abandono escolar no ensino médio que apresentassem as variáveis utilizadas nos modelos preditivos. Ao final desse processo, foram selecionados 16 estudos, que constituíram o corpus submetido à análise de conteúdo.

3.1 APLICAÇÃO DA ANÁLISE DE CONTEÚDO NA SELEÇÃO DE VARIÁVEIS

A leitura dos 16 estudos que compuseram o corpus da pesquisa permitiu identificar as variáveis selecionadas pelos autores em seus modelos de predição do abandono escolar, apresentadas no Quadro 1. Em seguida, cada variável foi registrada individualmente em planilha do Microsoft Excel, totalizando 269 variáveis, que constituíram o material submetido à pré-análise e à exploração, conforme proposto por Bardin (2009).

Ainda dentro da análise de conteúdo, seguindo os passos descritos por Bardin (2009), foi realizado o agrupamento das variáveis por similaridade. Para esse agrupamento, foram considerados a equivalência semântica e conceitual entre as variáveis identificadas na literatura, reunindo em um mesmo grupo aquelas que apresentavam o mesmo significado ou representavam um mesmo atributo, ainda que descritas por terminologias distintas pelos diferentes autores.

O conhecimento da pesquisadora, decorrente de sua atuação na extração de dados e disseminação de informações educacionais, auxiliou na interpretação dessas equivalências e na definição dos agrupamentos, considerando o contexto investigado. Conforme Bardin (2009), o pesquisador desempenha papel fundamental nesse processo por interpretar o material à luz do contexto da pesquisa.

Cada grupo assim formado recebeu uma denominação representativa do conceito comum compartilhado pelas variáveis que o compunham, buscando facilitar sua identificação e utilização em etapas subsequentes da análise, resultando em 114 agrupamentos.

Para cada agrupamento foram definidas três características (Disponibilidade, Relevância e Dimensão), às quais foram atribuídas condições específicas, permitindo avaliar seu potencial de aplicação em modelos preditivos do abandono escolar para o ensino médio da rede estadual de ensino de Santa Catarina.

• **Disponibilidade**: indicando se a variável está disponível ou não nos bancos de dados para as escolas e alunos da rede estadual de ensino de Santa Catarina (**Sim** ou **Não**);

• **Relevância**: indicando se a variável é relevante para o contexto em questão (**Sim** ou **Não**), ou seja, para os alunos e escolas da rede estadual de ensino de Santa Catarina;

• **Dimensão**: indicando a quem o dado se refere, se à **Escola** ou ao **Aluno**.

Parte-se ainda do pressuposto de que:

> Na identificação das variáveis é melhor pecar pelo excesso. É muito comum que certas variáveis não sejam consideradas em determinado estudo, pelo simples fato que o analista da área de atividade, ainda que muito experiente, não acredita serem importantes. Ou pior, assegura que não tem utilidade.

> Quando utilizadas posteriormente, podem eventualmente mostrar-se poderosas informações. Outrossim, certas variáveis, historicamente consideradas importantes pelos envolvidos na empresa, mostram-se pouco ou nada relevantes para a melhoria dos resultados do modelo (Sicsú; Samartini; Barth, 2023, p. 26).

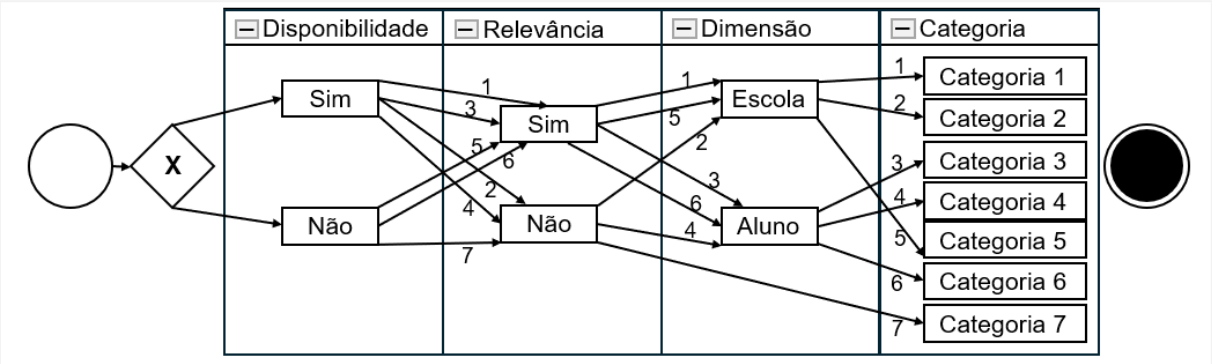
Desse modo, a característica considerada primordial é a Disponibilidade, pois constitui requisito indispensável para a utilização da variável em modelos preditivos.

Para verificação da disponibilidade foram considerados dados e indicadores disponibilizados publicamente, que em sua maioria possuem desagregação para Brasil, Regiões, Estados e municípios, alguns chegando até a escola, e dados provenientes do Sistema de Gestão Educacional de Santa Catarina (SISGESC), sendo essa uma base de dados própria e restrita à Secretaria de Estado da Educação de Santa Catarina (SED/SC).

Por meio da análise e associação das características/condições imputadas a cada agrupamento, seguindo o fluxo apresentado na Figura 1, foi possível caracterizá-los da seguinte forma:

- Categoria 1: Disponibilidade **Sim** – Relevância **Sim** – Dimensão **Escola**.
- Categoria 2: Disponibilidade **Sim** – Relevância **Não** – Dimensão **Escola**.
- Categoria 3: Disponibilidade **Sim** – Relevância **Sim** – Dimensão **Aluno**.
- Categoria 4: Disponibilidade **Sim** – Relevância **Não** – Dimensão **Aluno**.
- Categoria 5: Disponibilidade **Não** – Relevância **Sim** – Dimensão **Escola**.
- Categoria 6: Disponibilidade **Não** – Relevância **Sim** – Dimensão **Aluno**.
- Categoria 7: Disponibilidade **Não** – Relevância **Não**.

Figura 1 – Fluxo de categorização dos agrupamentos



Fonte: Elaborado pelos autores (2026).

A seguir são apresentadas e interpretadas as sete categorias associadas aos 114 agrupamentos, passos importantes dentro da análise de conteúdo.

3.1.1 Categoria 1: Disponibilidade Sim – Relevância Sim – Dimensão Escola

A Categoria 1 é formada pelos agrupamentos disponíveis nos bancos de dados, que contêm informações referentes às escolas que ofertam ensino médio para a rede estadual de ensino de Santa Catarina, e foram considerados relevantes para o contexto da pesquisa. Os agrupamentos são apresentados no Quadro 2:

Quadro 2 – Agrupamentos da Categoria 1: Disponibilidade **Sim** – Relevância **Sim** – Dimensão **Escola**

Agrupamentos	**Agrupamentos**
Abandono/Escola	Município/Localização
Ano letivo	Socioeconômico - IDHM
Identificador/Escola	Socioeconômico/Escola - Indicador de Nível Socioeconômico (Inse)
Infraestrutura tecnológica - Computador	

Fonte: Elaborado pelos autores (2026).

Os agrupamentos selecionados foram os de: ano letivo, identificador da escola, município/localização, Índice do Desenvolvimento Humano Municipal (IDHM)[5],

[5] O Índice do Desenvolvimento Humano Municipal é composto por três indicadores que abrangem as dimensões de longevidade, educação e renda. (PNUD; Ipea; FJP, 2017).

Indicador de Nível Socioeconômico (Inse)[6], infraestrutura tecnológica - computador e abandono da escola.

O IDHM possui em sua composição dimensões importantes, que representam fatores de vida longa e saudável, com o indicador de longevidade, de acesso ao conhecimento, com o indicador de educação e do padrão de vida de um município, com o indicador de renda. Já o Inse, que possui divulgação bianual, agrega informações socioeconômicas específicas dos alunos, combina elementos de escolaridade dos pais e a posse de bens e serviços, e é divulgado para a escola.

Variáveis na dimensão da escola possibilitam que os modelos identifiquem o diferencial entre uma escola e outra, porém não são suficientes para determinar o diferencial entre o aluno que abandona e aquele que não.

3.1.2 Categoria 2: Disponibilidade Sim – Relevância Não – Dimensão Escola

A Categoria 2 apresenta os agrupamentos da dimensão Escola, que possuem disponibilidade Sim e relevância considerada para o contexto da pesquisa como Não, conforme Quadro 3:

Quadro 3 – Agrupamentos da Categoria 2: Disponibilidade **Sim** – Relevância **Não** – Dimensão **Escola**

Agrupamentos	**Agrupamentos**
Financeiro - Entorno da escola - abastecimento de água	Financeiro - Entorno da escola - renda per capta
Financeiro - Entorno da escola - banheiros por domicílio	Financeiro - Entorno da escola - rendimento médio domicílios
Financeiro - Entorno da escola - coleta de lixo	Financeiro - Entorno da escola - rendimento médio mulheres responsáveis
Financeiro - Entorno da escola - crianças de 0 a 5 anos	Financeiro - Entorno da escola - rendimento médio responsáveis
Financeiro - Entorno da escola - domicílios particulares e coletivos	Financeiro - Entorno da escola - responsável 10 a 29 anos

[6] Construído a partir do questionário do aluno, que é aplicado no Sistema Nacional de Avaliação da Educação Básica (Saeb), o Indicador de Nível Socioeconômico (Inse) é composto por oito níveis socioeconômicos e tem como objetivo contextualizar resultados obtidos em avaliações e exames aplicados pelo Inep, possibilitar ações de planejamento, monitoramento e avaliação de políticas públicas, objetivando o aumento da qualidade e a equidade educacional (Brasil, 2023).

Financeiro - Entorno da escola - energia elétrica	Financeiro - Entorno da escola - sem renda per capta
Financeiro - Entorno da escola - idade média mulheres responsáveis	Financeiro - Entorno da escola - setor censitário aglomerado subnormal
Financeiro - Entorno da escola - idade média pessoas responsáveis	Financeiro - Entorno da escola - situação setor censitário
Financeiro - Entorno da escola - média de moradores domicílios particulares permanentes	Financeiro - Entorno da escola - total munícipes
Financeiro - Entorno da escola - mulheres responsáveis alfabetizadas	Financeiro - PIB
Financeiro - Entorno da escola - mulheres responsáveis menos 30 anos	Infraestrutura - Banheiro
Financeiro - Entorno da escola - participação renda responsável nos domicílios	Infraestrutura tecnológica - Internet
Financeiro - Entorno da escola - pessoas alfabetizadas	

Fonte: Elaborado pelos autores (2026).

Os agrupamentos relacionados ao entorno da escola, provenientes do Censo Demográfico[7] e divulgados pelo Instituto Brasileiro de Geografia e Estatística (IBGE), foram considerados como não relevantes para o contexto da pesquisa por possuírem periodicidade de divulgação muito espaçada, dez anos, e por considerar o Inse (Categoria 1) mais relevante, por estar diretamente relacionado à escola.

O Produto Interno Bruto (PIB)[8] foi considerado não relevante pois mesmo que um município apresente um PIB alto, a distribuição de renda entre seus munícipes pode ser desigual, sendo o fator financeiro da escola melhor associado ao Inse (Categoria 1).

Já os agrupamentos relacionados a infraestrutura, banheiro e internet, para o contexto da pesquisa não são relevantes, pois todas as escolas de ensino médio da

[7] Com periodicidade decenal, o Censo Demográfico constitui a principal fonte de referências para o conhecimento da população brasileira, sendo que a coleta de dados abrange todas as pessoas residentes no Território Nacional, com resultados divulgados para Brasil, Grandes Regiões, Unidades da Federação, Mesorregiões, Microrregiões, Regiões Metropolitanas, Municípios, Distritos, Subdistritos e Setores Censitários, permitindo a desagregação das informações para o entorno de um determinado local (IBGE, 2023).

[8] O Produto Interno Bruno (PIB) é um indicador de fluxo de novos bens e serviços finais, medidos de acordo com o preço final ao consumidor, representando a economia de um país, ou de um município, durante um período específico (IBGE, 2024).

rede estadual de ensino de Santa Catarina possuem banheiro, tanto feminino quanto masculino, e possuem acesso à internet.

O contexto deve ser levado em consideração, assim como visto nos trabalhos realizados por Chinnasamy e Balasubramanian (2022a, 2022b) (Índia), Maheshwari *et al*. (2020) (Índia) e Mduma (2023) (Índia e Tanzânia), por exemplo, que consideraram importante o fator banheiro, inclusive na desagregação banheiro feminino.

3.1.3 Categoria 3: Disponibilidade Sim – Relevância Sim – Dimensão Aluno

As variáveis mais desejadas para modelos de predição são aquelas relativas a menor dimensão, ou seja, que possibilitam diferenciar um aluno do outro, o propenso ao abandono daquele que não. Sendo assim, ilustra-se no Quadro 4, a Categoria 3, dos agrupamentos da dimensão Aluno, que possuem disponibilidade Sim e relevância Sim.

Quadro 4 – Agrupamentos da Categoria 3: Disponibilidade **Sim** – Relevância **Sim** – Dimensão **Aluno**

Agrupamentos	**Agrupamentos**
Abandono	Financeiro - Bolsa de estudo
Carga horária	Frequência
Cor/Raça	Gênero/Sexo
Desempenho	Idade/Distorção idade-série
Docente	Identificador/Aluno
Educação especial	Município/Localização
Engajamento aluno - Anos nessa escola	Residência - Tipo de residência
Engajamento aluno - Aulas extracurriculares	Tipo Ensino Médio
Estado civil	Total de alunos/Sala/Escola
Etapa/Ano/Série	Turno

Fonte: Elaborado pelos autores (2026).

3.1.4 Categoria 4: Disponibilidade Sim – Relevância Não – Dimensão Aluno

O Quadro 5 ilustra a Categoria 4, da dimensão Aluno com disponibilidade Sim e relevância Não.

Quadro 5 – Agrupamentos da Categoria 4: Disponibilidade **Sim** – Relevância **Não** – Dimensão **Aluno**

Agrupamentos
Professor por aluno

Fonte: Elaborado pelos autores (2026).

Para o contexto da pesquisa foi considerado não relevante o agrupamento professor por aluno, visto que o campo total de alunos (Categoria 3), da turma e da escola, já representa uma característica relacionada à complexidade do contexto escolar.

3.1.5 Categoria 5: Disponibilidade Não – Relevância Sim – Dimensão Escola

A Categoria 5, ilustrada no Quadro 6, apresenta os agrupamentos da dimensão Escola, que possuem disponibilidade Não e relevância Sim.

Quadro 6 – Agrupamentos da Categoria 5: Disponibilidade **Não** – Relevância **Sim** – Dimensão **Escola**

Agrupamentos	**Agrupamentos**
Percepções variadas/Comunidade escolar - Clima escolar	Percepções variadas/Responsável escola - Sobre expectativas

Fonte: Elaborado pelos autores (2026).

Ambos os agrupamentos foram listados no trabalho de Wongvorachan, He e Bulut (2023) e são relativos a um banco de dados de escolas dos Estados Unidos da América (EUA).

As percepções do conselheiro sobre as expectativas do professor, do conselheiro e do diretor, apresentando as percepções dentre os profissionais de educação atuantes nas escolas, assim como também a percepção da comunidade escolar sobre o clima escolar, poderiam apresentar relevância dentro do modelo de predição do abandono escolar, porém, não se encontram disponíveis para a rede estadual de ensino catarinense.

3.1.6 Categoria 6: Disponibilidade Não – Relevância Sim – Dimensão Aluno

A Categoria 6, ilustrada no Quadro 7, apresenta os agrupamentos da dimensão Aluno, que possuem disponibilidade Não e relevância Sim.

Quadro 7 – Agrupamentos da Categoria 6: Disponibilidade **Não** – Relevância **Sim** – Dimensão **Aluno**

Agrupamentos	**Agrupamentos**
Engajamento/Aluno - Atrasado	Percepções variadas/Aluno - Sobre emprego
Engajamento/Aluno - Em relação à escola	Percepções variadas/Aluno - Sobre entrar na faculdade
Engajamento/Aluno - Frequência que não faz tarefa	Percepções variadas/Aluno - Sobre família conseguir pagar faculdade
Engajamento/Aluno - Horas gastas estudo por dia	Percepções variadas/Aluno - Sobre trabalhar ser mais importante
Engajamento/Aluno - Sem material	Percepções variadas/Aluno - Sobre utilidade científica
Engajamento/Aluno e familiar - Leitura com pais	Percepções variadas/Família - Quão longe pensam que o aluno chegará
Engajamento/Família - Se olham o material	Refeições domésticas por dia
Engajamento/Família - Se procuram a escola	Relacionamento/Aluno - Amigo querer ir para faculdade
Engajamento/Família - Se são contactados sobre comportamento	Relacionamento/Aluno - Amigo ser frequente
Financeiro - Assistência governamental	Relacionamento/Aluno - Amigo tem interesse na escola
Financeiro - Classificação baixa renda	Relacionamento/Aluno - Amigo tira boas notas
Financeiro - Renda familiar - Principal fonte	Residência - Se mora com os pais
Financeiro - Renda familiar bruta	Responsável - Base de violência (mãe)
Financeiro - Renda mensal familiar	Responsável - Escolaridade
Financeiro - Responsável financeiro	Responsável - Estado civil
Financeiro - Situação financeira	Responsável - Profissão
Ocupação/Aluno	Reuniões de pais e professores
Percepções variadas/Aluno - Está certo que se formará no EM	Socioeconômico
Percepções variadas/Aluno - Pertencimento escolar	Tamanho da família
Percepções variadas/Aluno - Quão longe pensa chegar	

Fonte: Elaborado pelos autores (2026).

Variáveis na dimensão aluno apresentam características individualizadas e se mostram potencialmente relevantes, como uma forma de amplificar a eficiência dos modelos. Porém, variáveis da dimensão aluno não estão disponíveis publicamente,

além disso, nem sempre estão disponíveis nos bancos de dados das instituições ou mantenedoras. Como é o caso das aqui listadas, que não se encontram no banco de dados da SED/SC.

Agrupamentos de engajamento do aluno: se o aluno chega atrasado, em relação à escola, a frequência que não faz tarefa, horas gastas com estudo por dia e se vai para a aula sem material, assim como suas percepções: se considera que irá se formar no ensino médio, de pertencimento escolar, quão longe pensa chegar em relação aos estudos, sobre entrar na faculdade, sobre conseguir pagar a faculdade, sobre a utilidade científica, sobre emprego e sobre trabalhar ser mais importante do que estudar, foram indicados no trabalho de Wongvorachan, He e Bulut (2023).

Assim como agrupamentos listados no mesmo trabalho de Wongvorachan, He e Bulut (2023), que envolvem o relacionamento do aluno com o amigo mais próximo: se quer ir para a faculdade, se é frequente, se tem interesse na escola e se tira boas notas.

Mduma (2023) apresenta o engajamento da família: se procuram a escola, se participam de reuniões de pais e professores, se olham o material escolar, se leem com o aluno, e o trabalho de Wongvorachan, He e Bulut (2023): se os pais são contactados sobre comportamento e quão longe acham que o aluno chegará.

O nível de escolaridade dos pais também foi apontado nos trabalhos de Wongvorachan, He e Bulut (2023) e Krüger, Britto Jr. e Barddal (2023), e nesse último ainda aparecem variáveis ligadas aos responsáveis, como estado civil, profissão e se o aluno mora com eles.

Demais variáveis ligadas ao financeiro, socioeconômicas, ocupação do aluno, tamanho da família, refeições por dia e se a mãe está presente em bases de violência também foram elencadas nos trabalhos relacionados, porém não se encontram disponíveis nas bases de dados da rede estadual de ensino catarinense.

A Categoria 6 apresenta informações qualitativas, sobre o aluno e sua família, e apresenta potencial para contribuir significativamente na construção de modelos da Ciência de Dados eficazes no seu propósito, predizer o abandono.

3.1.7 Categoria 7: Disponibilidade Não – Relevância Não

A Categoria 7, ilustrada no Quadro 8, apresenta os agrupamentos das dimensões Escola e/ou Aluno, que possuem disponibilidade Não e relevância Não.

Quadro 8 – Agrupamentos da Categoria 7: Disponibilidade **Não** – Relevância **Não**

Agrupamentos	**Agrupamentos**
Dependência administrativa - estudo prévio	Financeiro - Taxa de matrícula
Desempenho - GPA	Financeiro/Escola - Repasse de verbas (escola)
Desempenho - IDESP	Financeiro/Escola - Valor repasse verbas
Desempenho prévio - Processo seletivo	Infraestrutura - Sala privada meninas
Desempenho/Dependência	Matrícula curso superior posterior
Engajamento aluno - Aulas religiosas	Reserva de vagas por cotas
Engajamento/Aluno - Curso matemática extra	Responsável (guardião)
Financeiro - Aumento taxa de matrícula	Tempo entre conclusão do ensino fundamental e ingresso no IFRN
Financeiro - Desconto na matrícula	Titulação
Financeiro - Grupo social (casta)	Vulnerabilidade (IPVS)

Fonte: Elaborado pelos autores (2026).

Os agrupamentos apresentados na Categoria 7 são todos muito específicos ao contexto em que os modelos propostos pelos autores relacionados estavam atuando, e não condizem ou não são significativos para o contexto da rede estadual de ensino de Santa Catarina.

Agrupamentos de casta e guardião, por exemplo, são listadas nos trabalhos de Chinnasamy e Balasubramanian (2022a, 2022b), Masood e Begum (2022) e Mduma (2023), são ligadas à situação apresentada na Índia e/ou na Tanzânia e não condizem com o contexto catarinense, assim como a escola possuir sala privada para meninas.

Variáveis ligadas ao financeiro da escola não foram consideradas relevantes, pelo fato de que todas as escolas da rede estadual de ensino possuem a própria SED/SC como mantenedora.

Outros agrupamentos também são específicos aos trabalhos em que foram utilizados, como por exemplo a situação da matrícula em curso de ensino superior em

fevereiro 2016, interesse em curso extra de matemática e o desempenho no Grade Point Average (GPA)[9], apresentadas no trabalho de Wongvorachan, Hee Bulut (2023).

Também são específicos do contexto utilizado no trabalho de Lopes Filho (2021), o Índice Paulista de Vulnerabilidade Social (IPVS)[10] da escola e o Índice de Desenvolvimento da Educação do Estado de São Paulo (IDESP)[11], assim como a titulação, no trabalho de Souza (2020).

O contexto educacional dos institutos federais, nos quais o ingresso no ensino médio é realizado via prova/processo de seleção traz agrupamentos relativos ao conceito obtido nas provas do processo seletivo e reserva de vagas por cotas, do trabalho de Silva (2021), o tempo de conclusão do ensino fundamental e ingresso no Instituto Federal do Rio Grande do Norte (IFRN) e o número de dependências em português e matemática, do trabalho de Barros *et al*. (2019).

Agrupamentos relativos à dependência administrativa do estudo prévio do aluno, utilizados nos modelos relativos aos institutos federais e de instituição que oferta somente o ensino médio integrado à educação profissional, como visto nos trabalhos de Barros *et al*. (2019), Silva (2021) e Souza (2020), não foram considerados como relevantes, visto a grande maioria dos alunos do ensino médio da rede estadual de ensino de Santa Catarina serem provenientes de escolas públicas.

Já agrupamentos relativos à taxa, aumento e desconto de matrícula, assim como se os alunos são envolvidos com aulas religiosas, listadas no trabalho de Krüger, Britto Jr. e Barddal (2023), condizem ao contexto de um trabalho aplicado à rede privada de ensino.

Posto isso, e tendo todos os passos descritos por Bardin (2009), como essenciais para a análise de conteúdo, sido executados, sendo eles: exploração do

[9] O Grade Point Average (GPA), média de pontos das notas, utilizado por muitas instituições no mundo, com uma escala de 1 a 5, visa aferir um valor para o desempenho acadêmico do aluno ao longo de uma etapa dos seus estudos (Sumares, 2024).

[10] O Índice Paulista de Vulnerabilidade Social (IPVS) é uma combinação das dimensões socioeconômicas e demográficas do Estado de São Paulo, que classifica os setores censitários em seis grupos de vulnerabilidade social (Alesp, 2004).

[11] O Índice de Desenvolvimento da Educação de São Paulo (IDESP) é um indicador da qualidade de ensino da rede estadual de ensino paulista (São Paulo, 2008).

material, agrupamento e classificação em categorias, interpretação, segue a síntese e conclusão dos achados:

Os agrupamentos da Categoria 1 (Disponibilidade **Sim** – Relevância **Sim** – Dimensão **Escola**) e da Categoria 3 (Disponibilidade **Sim** – Relevância **Sim** – Dimensão **Aluno**) podem ser considerados na aplicação de modelos de predição do abandono escolar dos alunos no ensino médio da rede estadual de ensino de Santa Catarina.

4 CONSIDERAÇÕES FINAIS

A análise de conteúdo, por meio de suas técnicas para tratamento dos dados, neste contexto predominantemente qualitativo, possibilitou que o processo de escolha/seleção das variáveis a serem consideradas na aplicação de modelos de predição do abandono escolar dos alunos no ensino médio da rede estadual de ensino de Santa Catarina seguisse um fluxo coerente e estruturado.

Ao integrar princípios da organização do conhecimento, da Ciência da Informação, à lógica estatística, da Ciência de Dados, esta pesquisa evidencia o potencial da abordagem interdisciplinar na construção de modelos preditivos**.**

A exploração do material bibliográfico permitiu listar variáveis citadas pelos autores dos trabalhos relacionados com o abandono escolar no ensino médio. O agrupamento das variáveis, juntamente com a atribuição de características/condições (Disponibilidade, Relevância e Dimensão), possibilitou sua classificação em categorias, resultando em uma sistematização inicial das variáveis identificadas na literatura para o contexto investigado.

Por meio da interpretação de cada uma das categorias, tendo os pesquisadores conhecimento e domínio do contexto, primordial na análise de conteúdo, chegou-se à síntese e conclusão dos achados, indicando que os agrupamentos da Categoria 1 (Disponibilidade **Sim** – Relevância **Sim** – Dimensão **Escola**) e da Categoria 3 (Disponibilidade **Sim** – Relevância **Sim** – Dimensão **Aluno**) reúnem variáveis com

potencial para subsidiar etapas de modelagem preditiva do abandono escolar dos alunos no ensino médio da rede estadual de ensino de Santa Catarina.

Observa-se ainda que os agrupamentos da Categoria 5 (Disponibilidade **Não** – Relevância **Sim** – Dimensão **Escola**) e da Categoria 6 (Disponibilidade **Não** – Relevância **Sim** – Dimensão **Aluno**) representam um conjunto de variáveis que poderão subsidiar pesquisas futuras voltadas à incorporação de novas informações em modelos preditivos, mediante etapas posteriores de coleta e modelagem.

Os resultados obtidos permitem responder à questão que orientou esta pesquisa ao demonstrar que a seleção de variáveis para modelos preditivos pode ser conduzida por meio de um processo sistemático de análise de conteúdo, fundamentado na literatura, contextualizado à realidade investigada e apoiado na interpretação do pesquisador. Dessa forma, o estudo alcança seu objetivo ao apresentar a aplicação da análise de conteúdo como estratégia para exploração, organização e seleção de variáveis voltadas à modelagem preditiva do abandono escolar.

Como contribuição para a Ciência da Informação, esta pesquisa evidencia que a análise de conteúdo pode apoiar processos de organização, sistematização e seleção de informações para utilização em modelos de Ciência de Dados, ampliando as possibilidades de integração entre esses dois campos. Ao apresentar uma sistematização inicial das variáveis identificadas na literatura e sua contextualização para a rede estadual de ensino de Santa Catarina, o estudo oferece um referencial metodológico que poderá subsidiar pesquisas futuras voltadas à seleção de variáveis e à construção de modelos preditivos em diferentes contextos educacionais.

REFERÊNCIAS

ALESP. Assembleia Legislativa do Estado de São Paulo. **Assembleia lança o Índice Paulista de Vulnerabilidade Social**. São Paulo: 2004. Disponível em: https://www.al.sp.gov.br/noticia/?id=264220. Acesso em: 14 out. 2024.

AMARAL, Fernando. **Introdução a ciência de dados**: mineração de dados e Big Data. 1. ed. Rio de Janeiro: Alta Books, 2016.

BARDIN, Laurence. **Análise de conteúdo**. 4. ed. Lisboa: Edições 70, 2009.

BARROS, Thiago Medeiros; SOUZA NETO, Plácido A.; SILVA, Ivanovitch; GUEDES, Luiz Affonso. Predictive models for imbalanced data: a school dropout perspective. **Education Sciences**, v. 9, n. 275, 2019. Disponível em: https://doi.org/10.3390/educsci9040275. Acesso em: 22 jan. 2024.

BARROS, Thiago Medeiros. **Um processo orientado a dados para geração de modelo de predição de evasão escolar**. 2020. Tese (Doutorado em Engenharia Elétrica e de Computação) - Universidade Federal do Rio Grande do Norte (UFRN), Natal RN, 2020. Disponível em: https://repositorio.ufrn.br/handle/123456789/31933. Acesso em: 08 jan. 2024.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **Saeb 2021**: Indicador de Nível Socioeconômico do Saeb 2021: nota técnica. Brasília, DF: Inep, 2023. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/nivel-socioeconomico. Acesso em: 14 out. 2024.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **Sinopse estatística da educação básica 2025**. Brasília: Inep, 2026a. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/sinopses-estatisticas/educacao-basica. Acesso em: 03 jul. 2026.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **Taxas de rendimento escolar**. Brasília: Inep, 2026b. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/taxas-de-rendimento-escolar. Acesso em: 03 jul. 2026.

CHINNASAMY, Rajagopal; BALASUBRAMANIAN, Thangavel. Rank biserial stochastic feature embed bivariate kernelized regressive bootstrap aggregative classifier for school student dropout prediction. **Concurrency and Computation**: Practice and Experience, v. 34, n. 21, 2022a. Disponível em: https://doi.org/10.1002/cpe.7133. Acesso em: 22 jan. 2024.

CHINNASAMY, Rajagopal; BALASUBRAMANIAN, Thangavel. Tucker's congruence regressive feature projected tversky discriminant multiple instance learning boost data classification for school student dropout prediction. **Concurrency and Computation**: Practice and Experience, v. 34, n. 18, 2022b. Disponível em: https://doi.org/10.1002/cpe.7021. Acesso em: 22 jan. 2024.

COLPANI, Rogério. Mineração de dados educacionais: um estudo da evasão no ensino médio com base nos indicadores do Censo Escolar. **Informática na Educação**: teoria & prática, Porto Alegre, v. 21, n. 3, p. 109-123, 2018. Disponível em: https://seer.ufrgs.br/index.php/InfEducTeoriaPratica/article/view/87880. Acesso em: 08 jan. 2024.

IBGE. **Censo Demográfico**: O que é. [S.l.]: IBGE, 2023. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/trabalho/22827-censo-demografico-2022.html?=&t=o-que-e. Acesso em: 10 dez. 2024.

IBGE. **Produto Interno Bruno – PIB**: O que é o PIB. [S.l.]: IBGE, 2024. Disponível em: https://www.ibge.gov.br/explica/pib.php. Acesso em: 04 out. 2024.

KRÜGER, João Gabriel Corrêa; BRITTO JR., Alceu de Souza; BARDDAL, Jean Paul. An explainable machine learning approach for student dropout prediction. **Expert Systems With Applications**. V. 233, 2023. Disponível em: https://doi.org/10.1016/j.eswa.2023.120933. Acesso em: 20 set. 2024.

LOPES FILHO, José Ahirton Batista. **Detecção de estudantes em risco de evasão escolar usando aprendizagem de máquina**. Tese (Doutorado em Engenharia Elétrica e Computação) - Universidade Presbiteriana Mackenzie, São Paulo SP, 2021. Disponível em: https://adelpha-api.mackenzie.br/server/api/core/bitstreams/178b32ac-cd0a-44e3-9804-87566f733e15/content. Acesso em: 22 jan. 2024.

MAHESHWARI, Ekansh; ROY, Chandrima; PANDEY, Manjusha; RAUTRAY, Siddharth. Prediction of factors associated with the dropout rates of primary to high school students in India using data mining tools. **Frontiers in Intelligent Computing**: Theory and Applications, v. 1.013, p. 242-251, 2020. Disponível em: https://doi.org/10.1007/978-981-32-9186-7_26. Acesso em: 22 jan. 2024.

MASOOD, Sheikh Wakie; BEGUM, Shahin Ara. Comparison of resampling techniques for imbalanced datasets in student dropout prediction. **IEEE Silchar Subsection Conference** (SILCON), Silchar, India, p. 1, 2022. Disponível em: https://ieeexplore.ieee.org/document/10028915. Acesso em: 22 jan. 2024.

MDUMA, Neema. Data balancing techniques for predicting student dropout using machine learning. **Data**, v. 8, n. 49, 2023. Disponível em: https://doi.org/10.3390/data8030049. Acesso em: 22 jan. 2024.

MOUTINHO, Sônia Oliveira Matos; MARTINS, Paulo George Miranda; ALENCAR, Danila Fernandes; CONEGLIAN, Caio Saraiva. Ciência da Informação e Ciência de Dados: convergências interdisciplinares. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis/SC, Brasil, v. 29, p. 1–26,

2024. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/99127. Acesso em: 14 maio. 2025.

PNUD; IPEA; FJP. Atlas do desenvolvimento humano nas regiões metropolitanas brasileiras: Florianópolis, Sorocaba, Ride Grande Teresina, Ride Petrolina-Juazeiro. Brasília: PNUD: IPEA: FJP, 2017. Disponível em: http://www.atlasbrasil.org.br/acervo/biblioteca. Acesso em: 20 jan. 2025.

QUEIROGA, Emanuel Marques; MACHADO, Matheus Francisco Batista; PARAGARINO, Virgínia Rodés; PRIMO, Tiago Thompsen; CECHINEL, Cristian. Early prediction of at-risk students in secondary education: a countrywide K-12 learning analytics initiative in Uruguay. **Information**, v. 13, n. 9, 2022. Disponível em: https://doi.org/10.3390/info13090401. Acesso em: 22 jan. 2024.

QUEIROGA, Emanuel Marques; SANTANA, Daniel; SILVA, Marcelo da; AGUIAR, Martim de; SANTOS, Vinícius dos; MELLO, Rafael Ferreira; BITTENCOURT, Ig Ibert; CECHINEL, Cristian. Anticipating student abandonment and failure: predictive models in high school settings. In: International Conference on Artificial Intelligence in Education (AIED) 2024. **Lecture Notes in Computer Science** (LNCS), v. 14829, 2024. Disponível em: https://doi.org/10.1007/978-3-031-64302-6_25. Acesso em: 22 set. 2024.

SÃO PAULO (Estado). Resolução SE - 74 de 6 de nov. de 2008. Institui sobre o Programa de Qualidade da Educação (PQE) e o Índice de Desenvolvimento da Educação do Estado de São Paulo (IDESP). **Diário Oficial do Estado de São Paulo**, São Paulo, v.118, n.211, p.19. 07 nov. 2008. Disponível em: https://www.imprensaoficial.com.br/DO/BuscaDO2001Documento_11_4.aspx?link=%2f2008%2fexecutivo%2520secao%2520i%2fnovembro%2f07%2fpag_0019_129M23AVIEGM5e2IK32DK4U06FF.pdf&pagina=19&data=07/11/2008&caderno=Executivo%20I&paginaordenacao=100019. Acesso em: 12 out. 2024.

SICSÚ, Abraham L.; SAMARTINI, André; BARTH, Nelson L. **Técnicas de machine learning**. Editora Blucher, 2023.

SILVA, Alex Sandro Siqueira da. **Mineração de dados aplicada à predição da evasão escolar no ensino médio**. 2021. Tese (Doutorado em Engenharia Elétrica) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Ilha Solteira, SP, 2021. Disponível em: http://hdl.handle.net/11449/213799. Acesso em: 08 jan. 2024.

SOUZA, Alex Marques de. **Machine learning e a evasão escolar - análise preditiva no suporte à tomada de decisão**. 2020. Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento) - Universidade FUMEC, Belo Horizonte, MG, 2020. Disponível em: https://repositorio.fumec.br/xmlui/bitstream/handle/123456789/420/alex_souza_mes_sigc_2020.pdf. Acesso em: 08 jan. 2024.

SUMARES, Gustavo. **O que é GPA e como calcular o seu**. [S.l.]: Fundação Estudar, 2024. Disponível em: https://www.estudarfora.org.br/gpa-o-que-e-e-como-calcular-o-seu/. Acesso em: 12 out. 2024.

TELES, Igor Antônio Gomes. **Aprendizagem de máquina para predição do abandono e evasão dos estudantes do estado do Ceará**. 2023. Dissertação (Mestrado em Engenharia Elétrica e Computação) - Universidade Federal do Ceará, Sobral, CE, 2023. Disponível em: http://repositorio.ufc.br/handle/riufc/73132. Acesso em: 08 jan. 2024.

WONGVORACHAN, Tarid; HE, Surina; BULUT, Okan. A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. **Information**, v. 14, n. 54, 2023. Disponível em: https://doi.org/10.3390/info14010054. Acesso em: 22 jan. 2024.

Copyright (c) Este trabalho autores mantêm os direitos autorais e concedem à revista o direito de primeira publicação, com o trabalho licenciado sob a licença Creative Commons Attribuição 4.0 International, que permite o compartilhamento do trabalho com reconhecimento da autoria. Os artigos são de acesso aberto e uso gratuito.