

Integrando Dados Textuais e Estruturados para Compreender Fatores Associados à Infrequência Escolar

Samuel Dario da Silva^{1,2}, Karin Becker¹

¹ Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil
samuellds@procempa.com.br, karin.becker@inf.ufrgs.br

² Companhia de Processamento de Dados do Município de Porto Alegre (PROCEMPA), Porto Alegre, RS, Brazil

Abstract. School dropout is influenced by both academic and socio-demographic factors, yet Educational Data Mining (EDM) studies predominantly rely on structured quantitative data. As a result, contextual information contained in textual records remains underexplored. This paper proposes a Knowledge Discovery in Databases (KDD) process that integrates unstructured and structured data from FICAI, a school absenteeism monitoring system used in southern Brazil. The approach applies topic modeling to reports produced by child protection counselors to identify latent topics associated with school absenteeism. These topics are then combined with demographic and institutional attributes and analyzed through Association rules. The study analyzes records from approximately 89,800 students. The results show that text mining uncovers factors overlooked by traditional institutional diagnoses. The discovered associations indicate that the pedagogical triad reflects symptoms rather than causes of disengagement, reveal distinct risk profiles across educational stages, and provide evidence to support targeted dropout prevention policies.

CCS Concepts: • **Information systems** → **Data mining**; • **Computing methodologies** → **Natural language processing**; • **Applied computing** → *Education*.

Keywords: association rules, educational data mining, natural language processing, school dropout, topic modeling

1. INTRODUÇÃO

A evasão escolar constitui um importante desafio social e econômico no Brasil. No Rio Grande do Sul, seu monitoramento é realizado por meio da Ficha de Comunicação de Aluno Infrequente (FICAI), instrumento intersetorial utilizado há quase 30 anos. Sua versão mais recente, o FICAI 4.0, integra o fluxo de atuação entre escolas, Conselhos Tutelares e o Ministério Público do Rio Grande do Sul (MPRS)¹. Do ponto de vista pedagógico e de políticas públicas, a compreensão da evasão escolar exige uma visão ampla do fenômeno. No Brasil, a evasão na Educação Básica resulta de uma complexa interação entre fatores intraescolares (e.g., dificuldades de aprendizagem e histórico de reprovação) e extraescolares (e.g., vulnerabilidade socioeconômica, inserção precoce no trabalho e falta de infraestrutura) [Silva Filho and de Lima Araújo 2017].

O fluxo de informações no FICAI começa quando a escola identifica a infrequência contínua de um dado aluno e abre uma ficha na qual constam diagnósticos preliminares, selecionados dentre categorias pré-definidas (e.g., *histórico de reprovação*, *defasagem de idade*), possivelmente acompanhada de um relato textual dos professores. Na sequência, o Conselho Tutelar é acionado para investigar a situação específica junto à família ou responsáveis, elaborando relatos que descrevem os motivos reais do distanciamento escolar. Em casos extremos, o MPRS é acionado para as devidas providências. Embora a ficha de infrequência contenha relatos envolvendo determinantes sociais da evasão, seu uso permanece majoritariamente restrito a análises dos diagnósticos pré-definidos por meio de *dashboards*.

¹O funcionamento do FICAI e a interação das escolas, dos conselhos tutelares e do Ministério Público são detalhados em <https://www.mprs.mp.br/hotsite/ficai/> e <https://prefeitura.poa.br/procempa/projetos/ficai-40>.

Pesquisa parcialmente apoiada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, CNPq (processo 308246/2026-8), e INCT TILD-IAR (Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação) (processo CNPq/SECTICS/CAPES/FAPs 408490/2024-1). Os autores agradecem a PROCEMPA e o MPRS pelos dados que viabilizaram esta pesquisa.

Os diagnósticos iniciais registrados pelas escolas no FICAI costumam concentrar-se em uma “tríade pedagógica” que estabelece um ciclo vicioso entre defasagem idade-série, histórico de reprovação e dificuldades de aprendizagem [Silva Filho and de Lima Araújo 2017]. Embora relevantes, esses indicadores refletem principalmente manifestações intraescolares do problema, sem necessariamente captar suas causas extraescolares subjacentes. Os relatos textuais dos conselheiros documentam esses fatores extraescolares e contextos sociais, motivando a aplicação de técnicas de processamento de linguagem natural (PLN) e processos de descoberta de conhecimento.

Na literatura de *Educational Data Mining* (EDM), predominam estudos voltados à predição de evasão com base em dados quantitativos, especialmente no Ensino Superior [de Jesus and de Gusmão 2024; Dol and Jawandhiya 2024]. Alguns trabalhos também exploram Regras de Associação para investigar padrões de absenteísmo na Educação Básica [Haque et al. 2024]. Técnicas de Modelagem de Tópicos focam tipicamente a análise de interações em fóruns de ensino a distância e *feedback* de alunos [Chen et al. 2020; Kastrati et al. 2020]. A modelagem de tópicos em dados não estruturados para compreender os fatores associados à evasão escolar na Educação Básica permanece pouco explorada.

Este trabalho propõe um processo de descoberta de conhecimento para integrar informações não estruturadas e estruturadas em sistemas de monitoramento da evasão escolar. Diferentemente das abordagens tradicionais, o processo explora relatos produzidos durante a investigação dos casos para identificar fatores contextuais da infrequência e relacioná-los aos perfis dos estudantes. Dessa forma, torna-se possível analisar conjuntamente fatores intra e extraescolares, ampliando a compreensão dos mecanismos associados ao afastamento escolar. Para implementar esse processo, utilizou-se o algoritmo LDA (Latent Dirichlet Allocation) [Blei et al. 2003] para extrair tópicos latentes dos relatos dos conselheiros tutelares, representando padrões recorrentes nos contextos sociais associados à infrequência. Em seguida, esses tópicos são incorporados ao conjunto de atributos dos estudantes e analisados juntamente com informações demográficas e diagnósticos institucionais por meio de Regras de Associação [Han et al. 2011]. Essa integração permite identificar combinações de fatores associadas a diferentes perfis de risco para a infrequência escolar.

As seguintes questões de pesquisa são investigadas: (QP1) Quais são os principais fatores associados à evasão escolar identificados nos relatos textuais da FICAI e como eles complementam as categorias estruturadas existentes? e (QP2) Como os tópicos extraídos dos relatos se relacionam com diagnósticos institucionais e etapas de ensino, revelando padrões de risco para a infrequência escolar?

O estudo demonstra que a incorporação sistemática de informações textuais amplia significativamente a compreensão dos fatores associados à infrequência escolar. Os resultados revelam quatro dimensões subjacentes à evasão, complementando diagnósticos limitados ao desempenho acadêmico. Tomados em conjunto, os dados evidenciam um deslocamento progressivo de risco: barreiras de infraestrutura predominam nos anos iniciais, enquanto na adolescência destacam-se fatores agudos como violência, trabalho precoce, gravidez e saúde mental. Essa integração textual permite mapear perfis de risco, viabilizando intervenções intersetoriais direcionadas.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve os materiais e métodos. A Seção 4 discute os resultados. Por fim, a Seção 5 apresenta as conclusões e perspectivas de trabalhos futuros.

2. TRABALHOS RELACIONADOS

A literatura em EDM sobre evasão escolar concentra-se predominantemente na aplicação de Aprendizagem de Máquina supervisionada sobre dados quantitativos e estruturados. Revisões sistemáticas indicam que a maior parte dos estudos está voltada ao Ensino Superior e à construção de modelos preditivos de desempenho e abandono estudantil [de Jesus and de Gusmão 2024; Dol and Jawandhiya 2024]. Embora eficazes para previsão, essas abordagens frequentemente dependem de variáveis previamente definidas e oferecem limitada compreensão dos fatores contextuais associados à evasão.

Nesse contexto, técnicas de PLN vêm sendo utilizadas para explorar fontes textuais. A Modelagem de Tópicos tem sido aplicada principalmente à análise de interações em ambientes virtuais de apren-

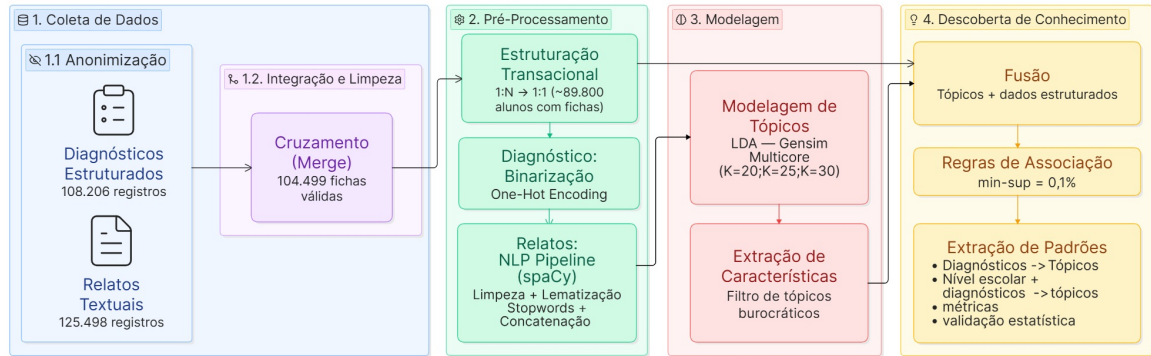


Fig. 1. Processo de KDD sobre dados do sistema FICAI.

dizagem e *feedback* de estudantes [Chen et al. 2020; Kastrati et al. 2020]. Entretanto, sua utilização para analisar relatos institucionais sobre a infrequência escolar permanece pouco explorada.

Regras de Associação também têm sido empregadas para identificar fatores associados ao absentismo e à evasão escolar. [Haque et al. 2024] utilizou o algoritmo *Apriori* para investigar padrões de evasão na Educação Básica australiana, enquanto que [Hossain and Mondal 2025] combinaram seleção de atributos e Regras de Associação para identificar fatores relacionados à evasão. Esses estudos limitam-se à análise de atributos estruturados.

O presente trabalho integra informações textuais e estruturadas em um mesmo processo de descoberta de conhecimento, explorando modelagem de tópicos para transformar relatos produzidos durante a investigação dos casos em variáveis analíticas e Regras de Associação para analisar sua relação com as variáveis estruturadas disponíveis no FICAI.

3. MATERIAIS E MÉTODOS

A Figura 1 resume o processo de KDD, cujo fluxo contempla coleta, integração e o pré-processamento dos dados, a modelagem de tópicos dos relatos textuais e a mineração de Regras de Associação. As análises foram implementadas em Python utilizando as bibliotecas *pandas*, *spaCy* e *Gensim*, com execução em infraestrutura computacional de alto desempenho gerenciada pelo SLURM². Código e dados não podem ser disponibilizados por questões de privacidade de menores. O repositório público³ inclui experimentações e justificativas para definição de parâmetros, e os resultados detalhados que não são mostrados por completo nesse artigo por restrições de espaço.

3.1 Dados e Pré-processamento

Os dados utilizados neste estudo foram extraídos do sistema FICAI 4.0. Cada ficha registra informações sociodemográficas do estudante, diagnósticos pré-definidos selecionados pela escola e, quando o caso for encaminhado ao Conselho Tutelar, relatos textuais com diagnóstico e encaminhamentos relativos ao afastamento escolar. Um mesmo aluno pode estar associado a muitas fichas. Foram extraídos dados referentes aos anos de 2024 e 2025, correspondendo a 125.498 fichas com relatos textuais e 108.206 registros de diagnósticos estruturados. Após a integração das bases, foram mantidos apenas os casos que possuíam simultaneamente diagnósticos escolares e relatos do Conselho Tutelar, totalizando 104.499 fichas completas. Todos os dados foram anonimizados por meio da remoção de identificadores de estudantes e instituições.

Como um mesmo estudante poderia possuir múltiplas fichas, os registros foram agregados por aluno, produzindo uma base final com aproximadamente 89.800 estudantes únicos. Nessa etapa, os

²<http://pcad.inf.ufrgs.br>

³https://github.com/samuelsilva/artigo_kdmile_dados

atributos demográficos (e.g., gênero, raça) foram consolidados, enquanto as categorias de diagnósticos assinaladas pela escola foram transformadas por meio de one-hot encoding. Existem 20 possíveis diagnósticos institucionais, apontando questões pedagógicas e determinantes sociais (e.g., falta de transporte escolar, suspeita de violência, gestação). Com a binarização dessas variáveis, aplicou-se uma regra de ocorrência máxima: se o aluno teve um problema específico marcado em qualquer uma de suas fichas históricas, a respectiva coluna dummy recebeu o valor verdadeiro no seu perfil consolidado.

Os relatos textuais foram pré-processados com a biblioteca *spaCy*, incluindo conversão para minúsculas, remoção de números e pontuação, lematização e eliminação de *stopwords*. Adicionalmente, foram removidos termos institucionais recorrentes (e.g., “encaminhar”, “contato” e “FICAI”) para reduzir ruídos e enfatizar conteúdos relacionados aos fatores associados à infrequência escolar.

3.2 Modelagem de Tópicos (QP1)

Para identificar fatores associados à infrequência escolar (QP1), utilizou-se o algoritmo LDA [Blei et al. 2003]. A escolha deve-se ao tamanho dos documentos (relatos concatenados) e a sua natureza multitemática. Os relatos pré-processados foram representados por meio de uma matriz *Bag-of-Words* construída com a biblioteca *Gensim*. O vocabulário foi filtrado removendo termos presentes em mais de 50% dos documentos e termos com frequência inferior a cinco documentos (e.g., infrequentes, erros de digitação). Os modelos foram treinados utilizando a implementação *LdaMulticore* com 16 *workers*, 30 *passes* e *random_state*=42. Para definir o número de tópicos (K), foram avaliadas configurações com 20, 25 e 30 tópicos. Após inspeção qualitativa dos tópicos gerados, adotou-se $K = 30$, por apresentar maior granularidade semântica e menor sobreposição entre temas.

Por fim, adicionou-se aos dados estruturados de cada estudante os tópicos extraídos de seus respectivos relatos concatenados. Considerando o caráter multifatorial da infrequência escolar, foram mantidos até cinco tópicos por relato concatenado de estudante, desde que apresentassem probabilidade igual ou superior a 10%. Justificativas mais detalhadas para seleção de K e experimentos para seleção de número de tópicos por estudantes são fornecidas no material suplementar.

3.3 Mineração de Regras de Associação (QP2)

Para responder à questão QP2, utilizou-se o algoritmo *FP-Growth* [Han et al. 2011] para identificar regras de associação entre os tópicos extraídos dos relatos e os atributos estruturados dos estudantes. Uma regra $X \rightarrow Y$, onde X e Y são itemsets, tem suporte e confiança. $\text{Suporte}(X \rightarrow Y) = |\{t \in D : X \cup Y \subseteq t\}|/|D| = P(X \cap Y)$, mede a frequência relativa da ocorrência conjunta de X e Y na base de dados D . $\text{Confiança}(X \rightarrow Y) = P(X \cap Y)/P(X)$ mede a probabilidade de Y ocorrer dado X . $\text{Lift}(X \rightarrow Y) = P(X \cap Y)/P(X)P(Y)$ compara a frequência observada de coocorrência entre os itemsets X e Y à esperada sob a hipótese de independência entre os dois conjuntos. Valores de *lift* superiores a 1 indicam associação positiva entre X e Y , enquanto valores inferiores a 1 indicam associação negativa em relação à independência dos dois conjuntos.

A análise foi realizada sobre a matriz binária consolidada, composta por características sociodemográficas, diagnósticos institucionais e tópicos obtidos pelo LDA a partir dos relatos textuais. Considerando o interesse em identificar padrões associados a grupos minoritários, adotou-se um suporte mínimo de 0.1%. Para identificação de regras interessantes usou-se filtragem estrutural e a métrica *lift*. Filtram-se apenas regras $X \rightarrow Y$ cujos antecedentes X correspondem a um ou mais atributos estruturados extraídos do FICAI (e.g., diagnóstico, nível escolar) e consequentes Y correspondem a um tópico extraído dos relatos textuais. Para a interpretação dos resultados, filtraram-se regras com $\text{Lift} > 1.1$ (associação positiva) ou $\text{Lift} < 0.8$ (associação negativa).

4. RESULTADOS E DISCUSSÕES

4.1 Extração e Categorização dos Tópicos Latentes (QP1)

A Tabela I apresenta os 30 tópicos identificados sobre os relatos concatenados dos professores e conselheiros tutelares, em resposta à QP1. Cada tópico é representado por um nome sucinto e pelas 4

Table I. Tópicos identificados e suas palavras mais representativas, ordenados por dimensão (1 a 4).

Tópico	Top 4 palavras	Tópico	Top 4 palavras
(1) Desastre Climático	retornar, informar, enchente, data	(3) Faltas Matutinas	vir, dia, pouco, ficar
(1) Vulnerabilidade Socioeconômica	transporte, condição, recurso, saúde	(4) Pedidos de Transferência	dizer, ir, falar, entrar
(1) Mudança de Cidade	cidade, mudar, município, residir	(4) Faltas Intercaladas	falta, muito, intercalar, motivo
(1) Trabalho e EJA	trabalhar, relatar, trabalho, noite	(4) Encaminhamentos	caso, encaminhar, haver, informação
(1) Violência e Vulnerabilidade	relatar, casa, relato, sofrer	(4) Abandono Pós-Matrícula	frequentar, ficha, matrícula, matricular
(2) Gravidez Adolescente	trazer, menina, conseguir, ficar	(4) Maioridade no Ensino Médio	ensino, ano, adolescente, genitora
(2) Afastamento Médico	atestado, justificar, médico, falta	(4) Busca Ativa	busca, ativo, alegre, porto
(2) Negligência Familiar	familiar, negligência, problema, suspeita	(4) Promessas de Retorno	faltar, continuar, conversar, voltar
(2) Saúde Mental	atendimento, saúde, psicológico, acompanhamento	(4) Mensagens Ignoradas	comparecer, mensagem, enviar, responder
(2) Mudança de Guarda	morar, casa, avó, irmã	(4) Contato Telefônico	tentativa, ligação, atender, várias
(3) Desinteresse Escolar	estudo, demonstrar, desinteresse, conversa	(4) Transferência Escolar	vaga, transferir, transferência, conseguir
(3) Conflitos Escolares	colega, professor, interação, turma	(4) Baixa Frequência	frequência, falta, baixo, apresentar
(3) Indisciplina Escolar	colega, sala, comportamento, professor	(4) Direito à Educação	educação, direito, realizar, medida
(3) Resistência às Atividades	atividade, realizar, participar, conversa	(4) Histórico Escolar	histórico, anterior, ano, cra
(3) Dificuldades de Aprendizagem	dificuldade, aprendizagem, apresentar, frequência	(4) Atendimento Domiciliar	ser, realizar, atividade, domiciliar

palavras mais representativas. O material complementar apresenta esses tópicos com mais detalhes.

De forma geral, os tópicos identificados podem ser organizados em quatro grandes dimensões: (1) **social e ambiental**, contemplando fatores como vulnerabilidade socioeconômica, trabalho precoce, migração territorial e mesmo impactos de desastres climáticos; (2) **familiar e de saúde**, incluindo gravidez na adolescência, negligência familiar, mudanças de guarda, problemas de saúde mental e afastamentos médicos; (3) **escolar e pedagógica**, composta por tópicos relacionados a conflitos escolares, indisciplina, desinteresse e dificuldades de aprendizagem; e (4) **processo de monitoramento institucional**, englobando tentativas de contato, busca ativa, encaminhamentos e outros procedimentos realizados pela rede de proteção. Destaca-se que essa última dimensão é não negligenciável (mais de 30% dos relatos). Os achados corroboram que a infrequência e a evasão escolar são resultado de uma complexa sobreposição de fatores intra/extra-escolares [Silva Filho and de Lima Araújo 2017].

Um resultado importante é que os tópicos complementam significativamente os diagnósticos estruturados disponíveis no FICAI. Enquanto os diagnósticos pré-definidos tendem a registrar categorias amplas, como dificuldades de aprendizagem, reprovação ou defasagem idade-ano, a modelagem de tópicos permite explicitar os contextos sociais associados a esses registros. Dessa forma, a modelagem de tópicos atua como um mecanismo de aprofundamento e complementação, capturando nuances que os diagnósticos pré-definidos de múltipla escolha são incapazes de registrar.

4.2 O Comportamento da Tríade Pedagógica

Para responder à QP2, analisamos, primeiramente, a relação entre os diagnósticos relativos à “Tríade Pedagógica” (*Defasagem Idade-Ano*, *Reprovação* e *Dificuldade de Aprendizagem*) e os tópicos identificados por meio de regras de associação. O conjunto completo de regras está disponível no material complementar.

As regras de associação envolvendo diagnósticos da tríade pedagógica e tópicos confirmam alinhamentos diretos e esperados, todas com $lift > 1.8$. Por exemplo, o diagnóstico *DIFICULDADES DE APRENDIZAGEM* relaciona-se com o tópico *Dificuldades de Aprendizagem*; o diagnóstico *REPROVAÇÃO* associa-se ao tópico *Desinteresse Escolar*; e o diagnóstico *DISTORÇÃO IDADE-ANO* é o principal antecedente do tópico *Maioridade no Ensino Médio*.

Embora algumas associações confirmem relações esperadas entre diagnósticos e tópicos, os resultados sugerem que a tríade pedagógica captura predominantemente manifestações observáveis da infrequência escolar, mas não necessariamente fatores subjacentes. A Tabela II apresenta alguns exemplos interessantes. Por exemplo, *DISTORÇÃO IDADE-ANO* apresentou forte associação com o atingimento da *Maioridade no Ensino Médio*, quando os mecanismos de proteção voltados à frequência escolar não podem ser ativados. Também se observa associação positiva com tópicos relacionados à *Violência e Vulnerabilidade social* e *Trabalho e EJA*. Finalmente, há fatores de desmotivação, considerando-se a associação com os tópicos de *Desinteresse Escolar* e *Promessas de Retorno*. De forma semelhante, *DIFICULDADES DE APRENDIZAGEM* e histórico de *REPROVAÇÃO* mostraram associações relevantes com *Saúde Mental*, *Desinteresse Escolar* e *Conflitos Escolares*. Em conjunto, esses achados sugerem que indicadores tradicionalmente utilizados para caracterizar difi-

Table II. Exemplos ilustrativos de regras interessantes entre diagnósticos predefinidos e tópicos textuais.

Regra	Lift
DISTORCAO_IDADE_ANO → Maioridade no Ensino Médio	1.88
DISTORCAO_IDADE_ANO → Violência e Vulnerabilidade	1.50
DISTORCAO_IDADE_ANO → Trabalho e EJA	1.41
DISTORCAO_IDADE_ANO → Desinteresse Escolar	1.68
DISTORCAO_IDADE_ANO → Promessas de Retorno	1.41
REPROVAÇÃO → Desinteresse Escolar	1.78
DISTORCAO_IDADE_ANO → Desinteresse Escolar	1.68
DIFICULDADE_ENSINO-APRENDIZAGEM → Saúde Mental	1.51
DIFICULDADE_ENSINO-APRENDIZAGEM → Gravidez Adolescente	0.78
DISTORÇÃO_IDADE-ANO → Gravidez Adolescente	0.57
REPROVAÇÃO → Gravidez Adolescente	0.59
DISTORÇÃO_IDADE-ANO → Afastamento Médico	0.67
REPROVAÇÃO → Afastamento Médico	0.70

culdades escolares frequentemente estão associados a contextos sociais, emocionais e familiares mais amplos, reforçando a importância de complementar a análise desses indicadores com as informações extraídas dos relatos textuais.

As associações com *lift* abaixo de 0,8 indicam que determinadas combinações de fatores ocorrem com menor frequência do que o esperado sob a hipótese de independência. Nossos achados evidenciam diferentes trajetórias associadas à infrequência escolar. Por exemplo, observamos uma associação negativa entre *Gravidez Adolescente* e diagnósticos como *REPROVAÇÃO*, *DISTORÇÃO IDADE-ANO* e *DIFICULDADES DE APRENDIZAGEM*. Esse padrão é compatível com a hipótese de que a maternidade precoce também possa estar associada à interrupção da trajetória escolar de estudantes sem histórico registrado de fracasso acadêmico. Outro resultado refere-se aos *Afastamentos Médicos* temporários, que apresentaram associação negativa com *REPROVAÇÃO* e *DISTORÇÃO IDADE-ANO*. Esse padrão pode indicar que esses afastamentos estão associados a episódios pontuais de saúde, sem necessariamente estarem acompanhados de indicadores prévios de dificuldades escolares. Em conjunto, as associações negativas sugerem diferentes padrões de registro da infrequência escolar, sugerindo que quando a instituição dispõe de uma causa conhecida para a infrequência, ela tende a registrar essa causa, enquanto que em situações nas quais não há uma explicação extraescolar identificada (potencialmente incluindo casos em que o paradeiro do estudante ser desconhecido), o registro tende a se concentrar nos indicadores pedagógicos disponíveis.

Em conjunto, esses resultados reforçam a ideia de que a infrequência escolar não decorre de um único fator causal [Silva Filho and de Lima Araújo 2017]. Enquanto alguns perfis estão associados a processos graduais de vulnerabilização e desconexão com a escola, outros decorrem de eventos externos que interrompem trajetórias acadêmicas anteriormente regulares. Essa distinção evidencia o potencial das regras de associação para identificar perfis heterogêneos de infrequência que não seriam observados apenas por meio dos diagnósticos pré-definidos.

4.3 Cruzamentos Estruturais e Interseccionalidade (QP2)

Finalmente, para responder à QP2, os tópicos extraídos dos relatos foram combinados aos diagnósticos estruturados e analisados por meio de regras de associação, em conjunto com os dados do nível escolar. A adoção de suporte mínimo de 0,1% permitiu identificar padrões envolvendo grupos minoritários, frequentemente negligenciados em análises baseadas apenas em frequência. Todas as regras discutidas foram corroboradas pelo teste do Qui-Quadrado ($p < 0.01$), fornecendo evidências de que as associações observadas não decorrem de variações aleatórias da amostra. Alguns exemplos relevantes são apresentados na Tabela III. A lista completa de regras encontra-se no material complementar.

A análise das regras revelou padrões distintos de infrequência ao longo das etapas de ensino. Os elevados valores de *lift* indicam associações positivas entre diagnósticos estruturados e fatores contextuais associados aos tópicos. Eles evidenciam que perfis de infrequência emergem em momentos distintos da trajetória escolar. Agrupamos os padrões encontrados nas quatro dimensões: familiar e saúde, social e ambiental, escolar e pedagógica, e monitoramento institucional.

Table III. Padrões por Etapa de Ensino (Diagnósticos Estruturados vs. Tópicos Textuais).

Etapa Escolar	Regra de Associação ($X \rightarrow Y$)	Suporte	Confiança	Lift
Educação Infantil	PRE_ESCOLA $\rightarrow$ Transferência Escolar	4.9%	78.7%	6.08
1º Ano (Fund.)	NAO_OFERTA_DE_EDUCACAO_ESPECIAL $\rightarrow$ Saúde Mental	0.3%	85.0%	9.76
7º Ano (Fund.)	FILHO_PEQUENO_SEM_CRECHE $\rightarrow$ Gravidez Adolescente	0.1%	39.1%	8.55
8º Ano (Fund.)	GESTACAO $\rightarrow$ Gravidez Adolescente	0.8%	53.4%	11.59
Ensino Médio - 1º Ano	GESTACAO $\rightarrow$ Gravidez Adolescente	0.8%	48.4%	10.97
Ensino Médio - 1º Ano	FILHO_PEQUENO_SEM_CRECHE $\rightarrow$ Gravidez Adolescente	0.3%	43.2%	9.80
Ensino Médio - 1º Ano	SUSPEITA_DE_VIOLENCIA_DOMESTICA $\rightarrow$ Violência e Vulnerabilidade	0.2%	36.6%	9.04
Ensino Médio - 2º Ano	ENSINO_MEDIO $\rightarrow$ Transferência Escolar	0.2%	100.0%	18.89
Ensino Médio - 2º Ano	VIOLENCIA_ENTORNO_ESCOLA $\rightarrow$ Violência e Vulnerabilidade	0.2%	44.0%	11.49
Ensino Médio - 2º Ano	RELACIONAMENTO_NA_ESCOLA $\rightarrow$ Conflitos Escolares	0.4%	43.3%	6.51
Ensino Médio - 2º Ano	FALTA_DE_TRANSPORTE_ESCOLAR $\rightarrow$ Vulnerabilidade Socioeconômica	0.9%	30.7%	7.41
Ensino Médio - 2º Ano	FALTA_DE_ACESSO_A_TRANSPORTE_PUBLICO $\rightarrow$ Vulnerabilidade Socioeconômica	0.3%	30.7%	7.39
Ensino Médio - 3º Ano	NAO_OFERTA_DE_EDUCACAO_ESPECIAL $\rightarrow$ Saúde Mental	0.1%	50.0%	10.26
EJA	NAO_OFERTA_DE_EDUCACAO_ESPECIAL $\rightarrow$ Saúde Mental	0.1%	55.6%	15.56

A **dimensão social e ambiental** torna-se particularmente relevante no Ensino Médio. Diagnósticos relacionados à violência doméstica e à violência no entorno escolar desde o início do ensino médio apresentaram algumas das associações mais fortes observadas com tópicos de *Violência e Vulnerabilidade* (*lift* até 11.49), evidenciando a influência de fatores territoriais e familiares sobre a permanência escolar. Os adolescentes enfrentam contextos de ameaça direta à integridade física, tanto fora quanto dentro do núcleo familiar, além de questões de infraestrutura. Observa-se a limitação socioeconômica no Ensino Médio, em que diagnósticos de transporte público e escolar estão fortemente associados ao tópico de *Vulnerabilidade Socioeconômica* (*lifts* de 7.39 e 7.41).

Na **dimensão familiar e de saúde**, os impactos mais severos giram em torno da gravidez na adolescência. O diagnóstico de *GESTACAO* apresenta forte associação com o tópico *Gravidez Adolescente* ao longo de todo o Ensino Médio (*lift* entre 9.82 e 11.01). De forma semelhante, o diagnóstico *FILHO PEQUENO SEM CRECHE* também se associa fortemente a esse tópico nos anos finais do Ensino Fundamental e no Ensino Médio (*lift* entre 8.55 e 11.59), indicando que a maternidade precoce e as responsabilidades de cuidado constituem fatores relevantes de afastamento escolar. Destaca-se ainda a associação entre *NÃO OFERTA DE EDUCAÇÃO ESPECIAL* e o tópico *Saúde Mental* no final do Ensino Médio (*lift*=10.26). Esse padrão também foi observado no início do Ensino Fundamental e no EJA (*lifts* de 9.76 e 15.56), sugerindo uma relação consistente entre exclusão educacional e sofrimento psíquico/emocional ao longo da trajetória escolar.

Nas dimensões **escolar/pedagógica** e de **monitoramento institucional**, os dados revelam os pontos de ruptura relacional e a saturação do fluxo assistencial. No 2º ano do Ensino Médio, o diagnóstico de *RELACIONAMENTO NA ESCOLA* associa-se ao tópico de *Conflitos Escolares* (*lift*=6.51), demarcando o esgotamento dos vínculos internos entre discentes e corpo docente. Quando esses laços se rompem, o sistema registra padrões de organização territorial forçada: as regras que vinculam atributos estruturais ao tópico de *Transferência Escolar* operam com confiança absoluta de 100.0% no 2º ano e no 3º ano do Ensino Médio (*lifts* de 18.89 e 15.95). Esse padrão sinaliza que a evasão de estudantes veteranos muitas vezes não se caracteriza como um abandono passivo, mas como uma busca ativa e/ou burocrática por realocação institucional ou de bairro, gerando um rastro procedimental que sobrecarrega a rede de proteção.

No material complementar são listadas as regras com *lift* inferior a 0.8. Aquelas observadas no Ensino Médio, revelam coocorrências que ajudam a desconstruir ideias pré-concebidas sobre infrequência. No 1º ano, os diagnósticos de *TRABALHO* e *NECESSIDADE DE CUIDADOS NA FAMÍLIA* apresentam baixa coocorrência com o tópico *Desinteresse Escolar* (*lifts* de 0.43 e 0.40). Padrão semelhante ocorre no 2º ano, em que *TRABALHO* também apresenta baixa associação com os tópicos *Saúde Mental* (0.38) e *Desinteresse Escolar* (0.59). Esses resultados sugerem que estudantes que precisam trabalhar ou assumir responsabilidades familiares não se enquadram no perfil típico de desengajamento escolar identificado nos relatos. Sua infrequência parece estar mais associada a restrições econômicas e responsabilidades de cuidado do que à perda de interesse pelos estudos.

5. CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresentou um processo de KDD para integrar informações estruturadas e não estruturadas do sistema FICAI na análise da infrequência escolar. Por meio da combinação de modelagem de tópicos e regras de associação, foi possível transformar relatos textuais dos professores e conselheiros tutelares em evidências analíticas complementares aos diagnósticos institucionais.

Os resultados mostram que os diagnósticos institucionais, em particular aqueles relacionados à tríade pedagógica, descrevem principalmente manifestações do problema, enquanto os relatos textuais revelam fatores contextuais associados à vulnerabilidade social, saúde mental, violência, maternidade precoce, trabalho e barreiras de infraestrutura. Além disso, a análise conjunta com dados escolares permitiu identificar diferentes perfis de risco ao longo das etapas de ensino, evidenciando a natureza multifatorial da evasão escolar. Conclui-se que a evasão escolar deve ser compreendida como resultado de uma cadeia de falhas de proteção social, demandando políticas públicas integradas e intersetoriais, para além de intervenções exclusivamente pedagógicas.

Como principal contribuição, o estudo demonstra que a incorporação sistemática de informações textuais amplia significativamente a compreensão dos fatores associados à infrequência escolar, revelando padrões que permanecem ocultos em análises baseadas exclusivamente em dados estruturados. Esses resultados podem subsidiar políticas públicas mais direcionadas e intervenções intersetoriais voltadas à permanência dos estudantes na escola.

Como trabalhos futuros, propõe-se (i) a análise longitudinal dos padrões de evasão ao longo do calendário letivo, permitindo compreender a dinâmica temporal dos fatores associados; (ii) a incorporação de modelos mais avançados de representação textual, como embeddings contextuais, para refinar a identificação de perfis de risco; (iii) o desenvolvimento de técnicas de clusterização de estudantes, com o objetivo de identificar padrões latentes de evasão em diferentes contextos demográficos, econômicos e geográficos; e (iv) a criação de um painel de monitoramento em tempo real integrado aos sistemas das secretarias de educação, possibilitando a geração automática de evidências para apoio à tomada de decisão e intervenção precoce.

Declaração do uso de ferramentas de IA

Ferramentas de IA generativa foram utilizadas apenas para a melhoria do texto e formatação, bem como apoio à geração de imagens. Os autores declaram ser responsáveis pelo texto apresentado.

REFERENCES

- BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (Jan): 993–1022, 2003.
- CHEN, X., ZOU, D., CHENG, G., AND XIE, H. Detecting Latent Topics and Trends in Educational Technologies Over Four Decades Using Structural Topic Modeling: A Retrospective of all Volumes of Computers & Education. *Computers & Education* vol. 151, pp. 103855, 2020.
- DE JESUS, J. A. AND DE GUSMÃO, R. P. Investigação da Evasão Estudantil por meio da Mineração de Dados e Aprendizagem de Máquina: Um Mapeamento Sistemático. *Revista Brasileira de Informática na Educação* vol. 32, pp. 807–841, mar., 2024.
- DOL, S. M. AND JAWANDHIYA, P. M. Systematic Review and Analysis of EDM for Predicting the Academic Performance of Students. *Journal of The Institution of Engineers (India): Series B* 105 (4): 1021–1071, 2024.
- HAN, J., KAMBER, M., AND PEI, J. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Waltham, MA, 2011.
- HAQUE, U. M., KABIR, E., AND KHANAM, R. Investigating school absenteeism and refusal among Australian children and adolescents using Apriori association rule mining. *Scientific Reports* vol. 14, pp. 1443, 2024.
- HOSSAIN, M. AND MONDAL, M. N. I. Mining Hidden Patterns in Student Dropout: Using Random Forest-Based Feature Selection and Association Rule Mining. In *Proceedings of the 2nd International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*. IEEE, New York, NY, USA, pp. 1–6, 2025.
- KASTRATI, Z., IMRAN, A. S., AND KURTI, A. Weakly Supervised Framework for Aspect-Based Sentiment Analysis on Students' Reviews of MOOCs. *IEEE Access* vol. 8, pp. 106598–106610, 2020.
- SILVA FILHO, R. B. AND DE LIMA ARAÚJO, R. M. Evasão e abandono escolar na educação básica no Brasil: fatores, causas e possíveis consequências. *Educação por Escrito* 8 (1): 35–48, 2017.