

ESTUDO DO FENÔMENO DA EVASÃO ESCOLAR: REVISÃO DE LITERATURA E AVALIAÇÃO DO CENÁRIO DA MINERAÇÃO DE DADOS EDUCACIONAIS ENQUANTO CAMPO DE PESQUISA

Roberto Peleteiro Galhardi
Celso Barreto da Silva
Fabio Fonseca Barbosa Gomes
José Vicente Cardoso Santos

RESUMO

Com o recente avanço de novas tecnologias, em especial, o aprendizado de máquina, a Mineração de Dados Educacionais (MDE) emergiu como um novo campo de estudo e método de análise de evasão escolar. Agora, alguns anos depois da fundação deste novo ramo de pesquisa, as condições para uma investigação geral a respeito de seu progresso estão estabelecidas. Assim, este artigo tem como objetivo geral construir uma revisão de literatura para a MDE. Os objetivos específicos envolvem: a) a avaliação do cenário geral da MDE e b) a criação de gráficos de interesse que elucidem mais a respeito de tópicos específicos. Para tal, foram investigados mais de 30 estudos publicados ainda neste século, de origem nacional e internacional.

Palavras-chave: Revisão de Literatura. Mineração de Dados Educacionais. Evasão Escolar.

1 INTRODUÇÃO

O início do estudo da evasão escolar foi na década de 30, com as primeiras pesquisas voltadas para o entendimento da saída de estudantes de universidades, apesar do campo ter se consolidado enquanto objeto de estudo somente em 1960. Antes desse ponto, graduação era um evento muito exclusivo e prestigiado para que qualquer estudo a respeito de evasão fosse considerado. (Demetriou e Sciborski, 2011)

Assim, nessa época, em virtude das ferramentas de análise disponíveis, os modelos preditivos envolviam aspectos sociológicos, psicológicos e educacionais, destacando-se, entre todos, os exemplos de Spady (1970), Spady (1971) e Tinto (1975).

As pesquisas que seguiram após esta era expandiram sobre estes modelos, incorporando novos fatores sociais daquele tempo ou reflexos de novos estudos. (Demetriou e Sciborski, 2011)

De todo modo, conforme é revisitado em Tinto (2006), a evasão escolar ainda é um fenômeno bastante perceptível mais de 3 décadas depois. Neste trabalho, o autor, uma das referências no assunto, clama por ações práticas a respeito do assunto, crescendo para além do campo teórico de pesquisa.

Paralelo a este desenvolvimento, a tecnologia evoluía e despontava enquanto ferramenta para a solução de problemas. Em 1959, Arthur Samuel definiu o aprendizado de máquina como “o campo de estudo que possibilita aos computadores a capacidade de aprender sem explicitamente programá-los” (Gérion, 2021).

As técnicas desse campo fundamentaram o surgimento de um novo campo de pesquisa: a Mineração de Dados Educacionais (MDE), que busca utilizar modelos preditivos de aprendizado de máquina para um melhor entendimento e tratamento do fenômeno da evasão escolar. (Baker, 2011)

Agora, com alguns anos desde o início dessa abordagem, é hora de avaliar-se o cenário atual do campo de pesquisa, avaliar seu desempenho e apontar suas falhas, visando guiar a comunidade para uma direção melhor.

2 ANALISANDO O CENÁRIO

Para a revisão, foram estudados mais de 30 artigos, de origem nacional e internacional, conforme mostra a Tabela 1.

Tabela 01 - Artigos Analisados para a Revisão

Identificador	Autor	Identificador	Autor
A01	(Lottering, 2020)	A20	(Kantorski et al, 2016)
A02	(Santana et al, 2015)	A21	(Melo et al, 2022)
A03	(Soares et al, 2020)	A22	(Lanes e Alcântara, 2018)
A04	(Teodoro et al, 2020)	A23	(Jesus et al, 2021)
A05	(Pereira e Zambrano, 2017)	A24	(Filho et al, 2020)
A06	(Perez, 2018)	A25	(Bittencourt e Ferrero, 2019)
A07	(Viloria et al, 2020)	A26	(Yuksel Turk et al, 2014)
A08	(Muniz, 2019)	A27	(Marquez-Vera et al, 2013)
A09	(Dharmawan et al, 2018)	A28	(Dekker et al, 2009)
A10	(Lim Sathit Wong, 2018)	A29	(Meedech, 2016)
A11	(Silva et al, 2020)	A30	(Rocha et al, 2017)
A12	(Paz e Cazella, 2017)	A31	(Suharjito, 2019)
A13	(Utari, 2020)	A32	(Maria et al, 2016)
A14	(Gonçalves e Beltrame, 2019)	A33	(Pradeep, 2015)

Identificador	Autor	Identificador	Autor
A15	(Pal, 2012)	A34	(Amorim et al, 2008)
A16	(Rodriguez-Maya et al, 2017)	A35	(Gonçalves et al, 2018)
A17	(Solis et al, 2018)	A36	(Ramos et al, 2018)
A18	(Beltran et al, 2019)	A37	(Silva et al, 2015)
A19	(Bittencourt et al, 2022)	A38	(Saraiva et al, 2019)

Fonte: Autores (2023)

Fazendo uso dos identificadores providos pela Tabela 1, é possível identificar as principais características de cada estudo, entregues pela Tabela 2.

Tabela 2 - Características dos Artigos

Identificador	Alunos Utilizados na Pesquisa	Número de Variáveis Usadas	Variáveis de Maior Relevância
A01	4417	17	- Idade - Número de Créditos - Auxílio Financeiro - Língua Nativa - Módulos Completos - Módulos Repetidos - Módulos ainda a ser Feitos - Persistência - Esteve ativo no último ano? - Anos no Sistema
A02	1800	11	Não é informado pelo estudo.
A03	2319	10	Não é informado pelo estudo.

A04	376.746	160	1. Se o aluno participa de Atividade Extracurricular 2. Se o aluno entrou pelo Vestibular 3. Se o aluno participa de Pesquisa 4. Quantidade Total de Concluintes do Curso 5. Se o aluno recebe suporte financeiro.
A05	6870	21	Foram avaliadas a relevância de um conjunto de atributos em união, ao invés de variáveis individuais. Dentre as avaliações, entretanto, algumas variáveis se destacaram: - A Nota Média é a variável de principal relevância em todas as análises - “Se perdeu o semestre” também é uma variável frequente - O número de matérias reprovadas aparece frequentemente também
A06	802	43	O estudo não informa explicitamente as variáveis mais relevantes. Mas, em uma análise de cluster, é possível notar um conjunto de variáveis que teve correlação relevante com a saída do curso. São elas: - Média de Pontos no Programa de Administração - Média de Pontos no Programa de Biologia - Número de Semestres - Desvio-Médio da Nota Acumulada
A07	1268	89	Não é informado pelo estudo.
A08	1055	17	Esse aspecto foi avaliado por uma árvore de decisão; podemos afirmar as seguintes variáveis como mais relevantes: 1. Número Total de Créditos Inscritos 2. Número Total de Créditos Passados 3. Forma de Ingresso 4. Razão para a Escolha dos Estudos 5. Idade
A09	103	17	- Número de Membros na Família - Relacionamento com os Docentes

A10	28801	6	Não é informado pelo estudo.
A11	2.991	18	O nível de relevância a das variáveis foi calculado individualmente para cada curso avaliado. Em todos os casos, a IAA (o Índice de Aproveitamento Acumulado) foi a variável de maior relevância, ao ponto que todas as outras variáveis se mostraram praticamente inutilizáveis. O aumento, em 1, desse índice de notas, levava a uma redução do risco de evasão de 75% até 91%, a depender do caso.
A12	4601	5	Incentivo
A13	2406	14	As performances do aluno no 5º, 6º, 4º e 3º semestres.
A14	Não é Informado	9	1. Frequência de Presença em Sala 2. Cidade 3. Escola de Origem
A15	1650	10	1. Notas no Senior Secondary 2. Notas no Ensino Médio 3. Casta dos Estudantes 4. Renda Familiar 5. Ocupação do Pai
A16	671	Não é informado pela pesquisa	1. Média de Qualificação no Ensino Médio 2. Percentil que o aluno ficou no Exame de Seleção (Vestibular) 3. Quantos por cento acima do CNE no Exame de Seleção 4. Percentual de qualificação nas questões de raciocínio lógico matemático 5. Posição do estudante no exame
A17	15.720	17	Não é informado pelo estudo.
A18	6.089	36	Não é informado pelo estudo.

A19	Não é informado pela pesquisa.	18	A avaliação da relevância das variáveis foi feita individualmente para cada período (semestre). - Durante o processo de matrícula, os principais variáveis de evasão foram: 1. Idade 2. Renda Familiar 3. Número de Pessoas da Família - No Primeiro e Segundo Períodos, as principais variáveis foram: 1. Nota Média do Período 2. Coeficiente de Rendimento do Período 3. Frequência Relativa do Período - No Terceiro, Quarto e Quinto Períodos, as principais variáveis foram: 1. Renda Familiar 2. Coeficiente de Rendimento do Período 3. Nota Média do Período
A20	791	24	Não é informado pelo estudo.
A21	1852	11	1. Percentual de Reprovação no 1º Semestre 2. Quantidade de Aprovações e Reprovações 3. Carga Horária 4. O estudante recebe auxílio? 5. Tempo em Ingresso
A22	916	9	O resultado foi feito em formato de árvore de decisão. As variáveis mais relevantes para a árvore foram: 1. Último Coeficiente 2. É bolsista? 3. Faixa de Idade
A23	733	7	O estudo não informa.
A24	1942	8	Não é informado pelo estudo.
A25	764	8	Não é informado pelo estudo.

A26	189	9	Foi avaliado a importância das variáveis para diversos algoritmos. Em média, os fatores mais relevantes foram: 1. Auto Eficácia 2. Preparação 3. Se teve Experiência Online Anterior (EAD) 4. Gênero 5. Idade
A27	670	77	1. Nota Média 2. Nota em Disciplinas 3. Nível de Motivação
A28	648	11	- Nota Média em Matérias de Ciência no VWO - Nota Média em Matérias de Matemática no VWO - Número de Cursos de Matemática ou Ciência no VWO
A29	509	9	Não é informado pelo estudo.
A30	1154	21	Não é informado pelo estudo.
A31	17.432	13	1. Frequência de Atendimento às Aulas 2. Pontuação no Dever de Casa 3. Pontuação no Teste Final 4. Pontuação no Teste Intermediário 5. Número Total de Créditos
A32	666	11	Não é informado pelo estudo.
A33	670	38	1. Notas nas Matérias (Inglês, Matemática, Química, Física, Ciência da Computação e Leitura e Escrita) 2. Nível de Motivação 3. Percentual de Frequência no Ensino Médio (10th Grade) 4. Fuma? 5. Nota média do ADMSN-EXM 6. Idade 7. Possui Deficiência Física?
A34	8073	8	Não é informado pelo estudo.

A35	574	21	- Tempo de Curso - Percentual de Presença - Média de Faltas Semestre - Média de Notas - Período Atual - Forma de Ingresso - Número de Disciplinas no Semestre - Em que tipo de escola completou o 1º ou o 2º grau.
A36	11440	7	Não é informado pelo estudo.
A37	353	6	O resultado foi feito em formato de árvore de decisão. As variáveis mais relevantes para a árvore foram: 1. Nota Média em Fóruns 2. Quantidade de Fóruns de o Aluno Participou 3. Quantidade de Posts em Fóruns
A38	6345	16	1. Total de Matérias Aprovadas 2. Total de Matérias Aprovadas com Dependências 3. Total de Matérias Trancadas 4. Total de Matérias Aprovadas Parcialmente 5. Frequência Total as aulas

Fonte: Autores (2023)

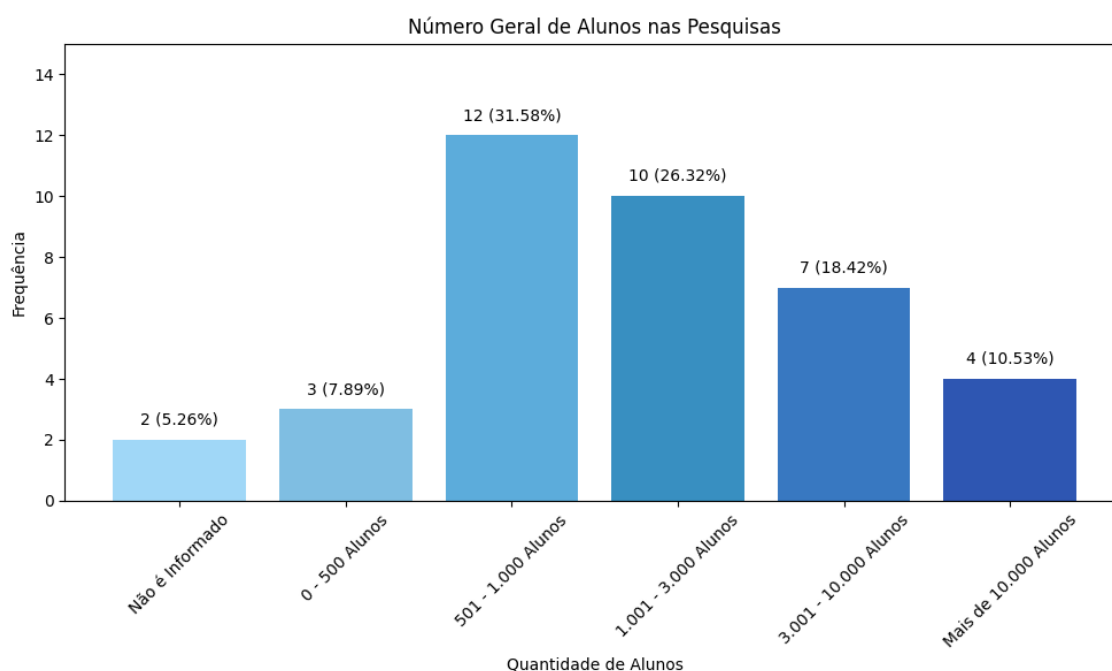
Com estes dados, é possível construir visualizações à respeito do cenário atual do campo de estudo.

2.1 - Número de Alunos em Pesquisas

Focalizando no número de alunos utilizados em pesquisas, temos um resultado bastante esparsos nas frequências, com trabalhos coletando dados de mais de trezentos mil alunos e outros de menos de duzentos.

Apesar de tais extremos, contudo, mais de 70% dos trabalhos (71,05%) utilizaram menos de 3000 alunos na construção do estudo, como demonstrado na Figura 1. Esta é uma tendência que parece se manter em outras revisões, a exemplo de Santos et al (2020).

Figura 1 - Número Geral de Alunos nas Pesquisas



Fonte: Autores (2023)

2.2 - Número de Variáveis Utilizadas

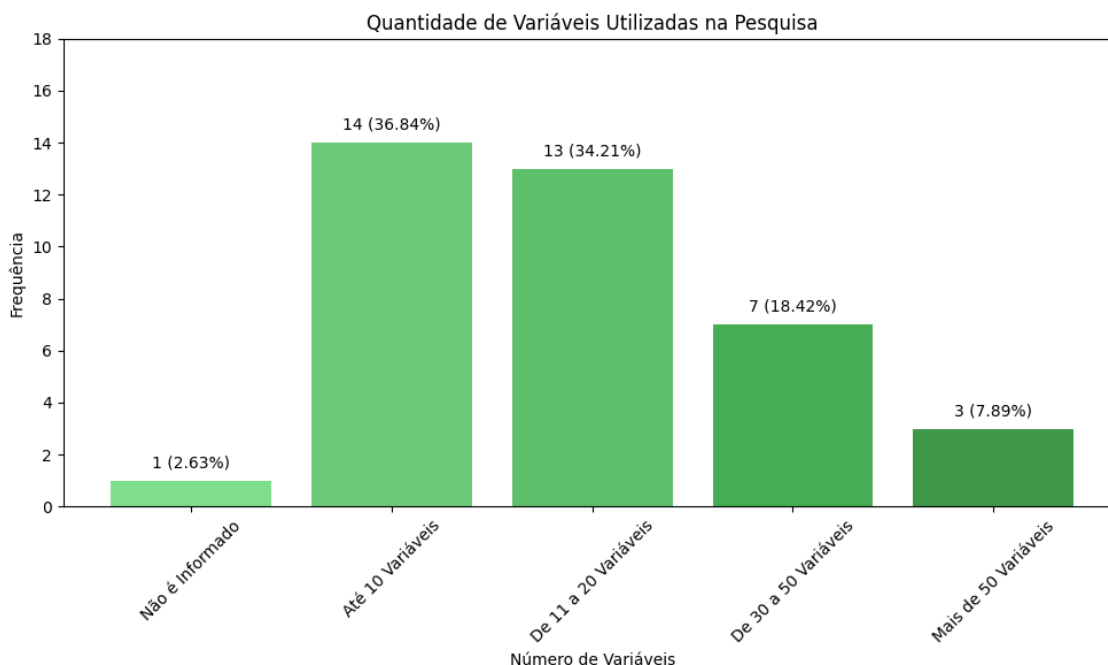
Passando a uma análise dos atributos considerados para a criação de modelos, temos a Figura 2.

Analisando o que foi encontrado, temos que a maioria dos artigos (73,68%) fazem uso de menos de 20 variáveis para a construção de classificadores preditivos.

Outros estudos que optam por uma ampla quantidade de atributos para a análise, também acabam, em algum momento de sua metodologia, aplicando técnicas para a seleção de categorias, a fim de evitar a chamada “Maldição da

Dimensionalidade”, um problema bem documentado em caso de aplicação de aprendizado de máquina (Gérion, 2021).

Figura 2 - Gráfico de Variáveis em Pesquisa



Fonte: Autores (2023)

2.3 - Variáveis Frequentes

É possível verificar essa seleção de variáveis analisando, em específico, os atributos que mais se repetem em pesquisas, demonstrado na Figura 3.

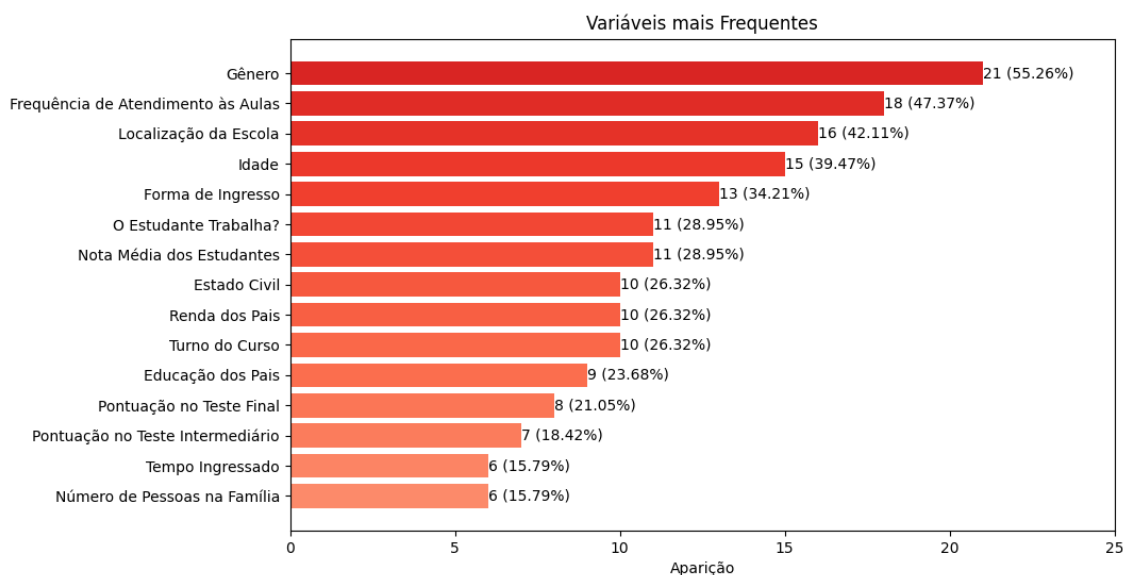
É válido lembrar que as variáveis abaixo não são as que mais impactam no resultado do modelo, mas sim aquelas que mais aparecem como constituintes do mesmo.

Com isso em mente, é encontrado, como é de se esperar, uma boa posição para variáveis pessoais, como “Gênero” e “Idade”, já que tais atributos costumam ser dados obtidos pelas instituições na própria matrícula.

Ademais, variáveis acadêmicas, a exemplo de “Frequência de Atendimento às Aulas” e “Nota Média dos Estudantes” também estão entre as mais frequentes, demonstrando que a performance acadêmica é um dos aspectos relevantes para a construção de modelos.

Outro ponto notório são características que indagam a respeito da situação familiar do aluno, como “Renda dos pais”, “Educação dos Pais” e “Número de Pessoas na Família”.

Figura 3 - Gráfico de Variáveis mais Frequentes



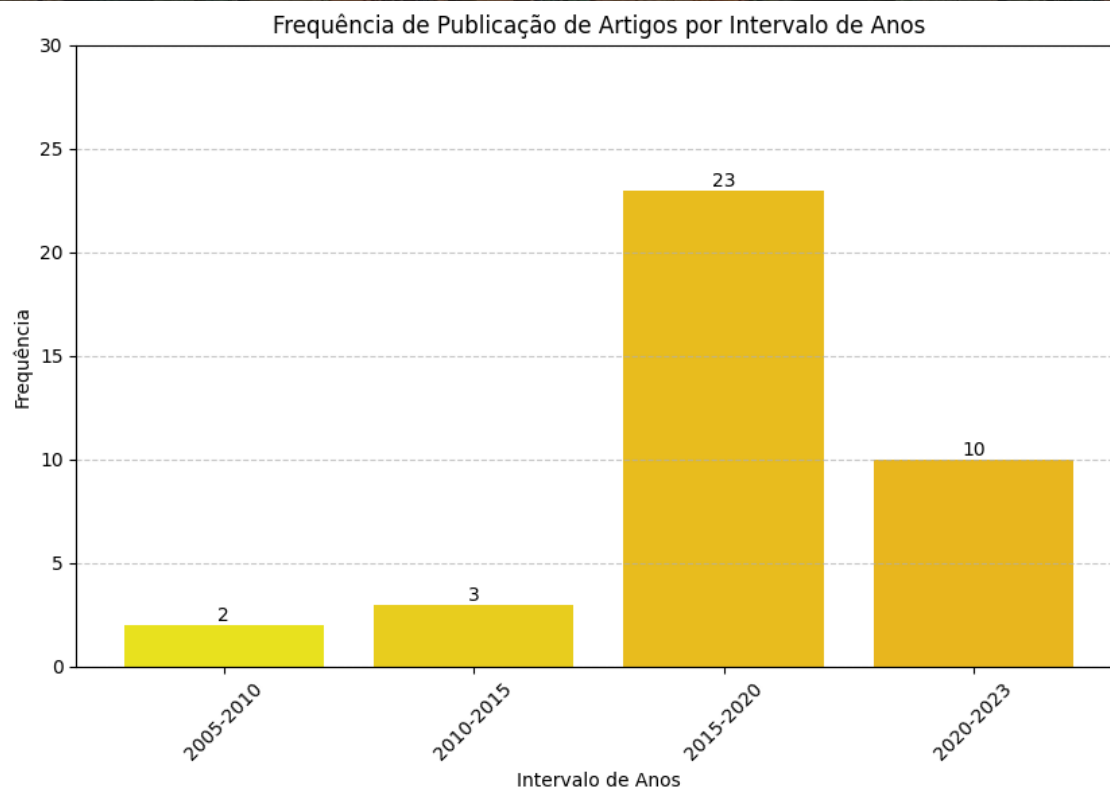
Fonte: Autores (2023)

2.4 - Anos de Publicação

Em seguida, temos o enfoque sobre os anos de publicações dos estudos, demonstrado pela Figura 4.

Aqui, vemos a MDE como de fato um campo emergente, com a maioria das pesquisas sendo de data recente.

Figura 4 - Gráfico de Publicações por Ano



Fonte: Autores(2023)

2.5 - Relevância em Variáveis

Em retrospecto, a análise da relevância individual de cada variável para a conclusão de um estudo de evasão escolar é o ponto mais controverso para análise, como consequência dos vastos métodos de abordagem ao problema.

Assim, em quase 40% dos artigos analisados (39.47%), este dado foi completamente desconsiderado e não informado, e, mesmo tratando-se de casos que tenham tal informação, os resultados variam, sejam devido ao meio utilizado para chegar a conclusão ou por natureza do estudo em si.

Variáveis diferentes são apontadas em casos diferentes possuindo graus de relevância diferentes na comparação de estudos de evasão escolar. Não existe um padrão certo ou fórmula específica para traçar definitivos.

O mais próximo deste ponto seria a questão da performance acadêmica, que se mostra dentre as variáveis mais relevantes em aproximadamente 74% (73.91%) dos estudos que informam tal fator, sob a forma de vários possíveis aspectos, como *Nota Média nas Disciplinas*, *Número de Créditos* ou *Número de Matérias Reprovadas*.

3 CONCLUSÃO

Em Baker (2011), a Mineração de Dados Educacionais é introduzida como um campo emergente promissor, com potencial para o desenvolvimento e delineamento de soluções concretas para o meio educacional.

Analisando os resultados dessa revisão, essa imagem se mantém. A alta incidência de produções, particularmente se considerado o histórico recente, denota o interesse acadêmico e científico por detrás deste novo ramo de pesquisa.

Infelizmente, uma metodologia clara de investigação, além dos métodos tradicionais de solução de problemas de aprendizado de máquina, ainda não parece ter sido delineada. A grande variância em características chave aos estudos, como o número de variáveis e de alunos analisados, demonstra isso.

Assim, a MDE, com algumas exceções, se resume, no contexto atual, a uma série de estudos de casos e análises individuais. Não existe um esforço centrado em uma direção específica, mas sim o teste de uma nova metodologia em um problema já conhecido.

Para que o campo chegue ao próximo estágio, alguns fatores serão essenciais. E o principal deles é um esforço na construção de um modelo de variáveis base, com o conjunto de características mais relevantes para a evasão.

Esse foi um ponto desconsiderado por quase 40% dos estudos, e, ainda que esta investigação aponte para variáveis relacionadas ao desempenho educacional como sendo um ponto de partida para a criação de tal modelo, a evidência ainda não é concreta.

Logo, conclui-se que, assim como mencionado em Baker (2011), a criação de um ambiente integrado para o compartilhamento de dados entre instituições educacionais, para análise científica posterior de uma visão abrangente é essencial.

REFERÊNCIAS

BAKER, R.S.J.d., CARVALHO, A. M. J. B. d., ISOTANI, Seiji., **Mineração de Dados Educacionais: Oportunidades para o Brasil.** *Revista Brasileira de Informática na Educação*, Volume 19, Número 2, 2011. DOI: 10.5753/RBIE.2011.19.02.03

DEMETRIOU, Cynthia., SCHMITZ-SCIBORSKI, Amy., **Integration, Motivation, Strengths and Optimism: Retention Theories Past, Present and Future.** Em R. Hayes (Org.), *Proceedings of the 7th National Symposium on student retention* (pp. 300-312). Norman, OK: The University of Oklahoma.

GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras & Tensorflow: conceitos, ferramentas e técnicas para a construção de sistemas inteligentes.** 2021, 2ª Edição.

MARQUES, Leonardo Torres, CASTRO, Angélica Félix de, MARQUES, Bruno Torres, SILVA, Jesaias Carvalho Pereira, QUEIROZ, Paulo Gabriel Gadelha., **Mineração de dados auxiliando na descoberta das causas da evasão escolar: Um Mapeamento Sistemático da Literatura.** *Revista Novas Tecnologias na Educação*, Volume 17, Número 3, 2019.

SANTOS, Vitor Hugo Barbosa dos, SARAIVA, Daniel Victor, OLIVEIRA, Catarina Teixeira de., **Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar.** *X Congresso Brasileiro de Informática na Educação*, 2021. DOI: 10.5753/sbie.2021.218167

SPADY, W. G. (1970). **Dropouts from higher education: An interdisciplinary review and synthesis.** *Interchange*, 1(1), 64–85. doi:10.1007/bf02214313

SPADY, W. G. (1971). **Dropouts from higher education: Toward an empirical model.** *Interchange*, 2(3), 38–62. doi:10.1007/bf02282469

TINTO, V. (1975). **Dropout from Higher Education: A Theoretical Synthesis of Recent Research.** *Review of Educational Research*, 45(1), 89.
doi:10.2307/1170024

TINTO, V. (2006). **Research and Practice of Student Retention: What Next?** *Journal of College Student Retention: Research, Theory & Practice*, 8(1), 1–19.
doi:10.2190/4ynu-4tmb-22dj-an4w