A identificação de alunos em risco de reprovação acadêmica através da Previsão Precoce: Um Mapeamento Sistemático

Rodrigo Miranda Feitosa, DCCMAPI, UFMA, rodrigo.feitosa@ifma.edu.br,
https://orcid.org/0009-0000-9120-1841
Luiz Aurélio Batista Neto, DCCMAPI, UFMA, luiz.neto@ifma.edu.br,
https://orcid.org/0009-0002-4403-2517
Vinicius Ponte Machado, UFPI, vinicius@ufpi.edu.br,
https://orcid.org/0000-0003-3391-8443
André Luis Silva Santos, IFMA, andresantos@ifma.edu.br,
https://orcid.org/0000-0002-9590-6686
André Macedo Santana, UFPI, andremacedo@ufpi.edu.br,
https://orcid.org/0000-0002-0062-1806

Resumo: Com base nas técnicas de Mineração de Dados Educacionais, os modelos de Previsão Precoce na área educacional são reconhecidos por sua capacidade de antecipar resultados e comportamentos do desempenho acadêmico. Este mapeamento sistemático tem como objetivo investigar a aplicação desses modelos de predição no desempenho acadêmico, voltados para a detecção dos alunos em risco de reprovação em cursos. Além disso, esta pesquisa descreve as estratégias adotadas na aplicação desse tema e quais são os elementos determinantes para a identificação dos fatores de insucesso acadêmico entre os discentes. Neste contexto, este mapeamento sistemático reúne informações que esclarecem o cenário da Previsão Precoce de alunos em risco de reprovação, tais como: a classificação das abordagens utilizadas, as técnicas aplicadas, os conjuntos de variáveis determinantes na identificação de insucesso acadêmico e quais são as fontes de dados utilizadas neste contexto.

Palavras-chave: Previsão Precoce, Mineração de Dados Educacionais, Risco, Inteligência Artificial.

Identifying students at risk of academic failure through Early Prediction: A Systematic Mapping

Abstract: Based on Educational Data Mining techniques, Early Prediction models in the educational field are recognized for their ability to anticipate results and behaviors of academic performance. This systematic mapping aims to investigate the application of these prediction models in academic performance, aimed at detecting students at risk of failing courses. In addition, this research describes the strategies adopted in the application of this theme and which are the determining elements for identifying the factors of academic failure among students. In this context, this systematic mapping gathers information that clarifies the scenario of Early Prediction of students at risk of failing, such as: the classification of the approaches used, the techniques applied, the sets of variables determining the identification of academic failure and which are the data sources used in this context.

Keywords: Early Prediction, Educational Data Mining, Risk, Artificial Intelligence.

1. Introdução

A aplicabilidade de modelos de predição na área educacional alcançou avanços para identificar precocemente alunos em risco de reprovação acadêmica. A Previsão Precoce (PP) pode ser definida como a aplicação de modelos preditivos que usam variáveis-chave para antecipar o fracasso ou abandono do aluno em um curso

antecipadamente (Berens *et al.*, 2019). A Previsão Precoce é uma tarefa desafiadora no campo da Mineração de Dados Educacionais (MDE), devido aos muitos fatores que podem influenciar o desempenho final de um aluno. Trata-se de uma questão que afeta muitos estudantes em todas as fases do ensino em instituições educacionais e universidades ao redor do mundo (López-Zambrano; Torralbo e Romero, 2021). O desempenho dos alunos é um dos principais critérios utilizados para avaliar a eficácia das instituições de ensino superior em seus processos de ensino-aprendizagem. O desenvolvimento de pesquisas nesta área torna-se essencial para identificar identificar os aspectos que contribuem para o baixo desempenho do aluno e, com base nas informações extraídas, identificar as dificuldades de aprendizagem antecipando uma possível reprovação acadêmica.

Os pesquisadores (López-Zambrano; Torralbo e Romero, 2021) desenvolveram uma revisão sistemática de modelos de Previsão Precoce do desempenho acadêmico listando os algoritmos que mais se destacaram nos resultados das previsões e nos níveis de precisão. No entanto, não se aprofundaram nas variáveis determinantes na identificação dos índices de desempenho dos alunos nos sistemas educacionais citados, nem na categorização de modelos de predição na área. Por outro lado, os autores Batista e Fagundes (2023) investigaram a previsão do desempenho de alunos em concursos educacionais avaliando as variáveis de entrada e saída delimitadas nos dados educacionais, demográficos, comportamentais, ou ainda a combinação de todos esses. Entretanto, a pesquisa anterior não contempla uma descrição das fontes de dados adotadas para extração das variáveis dos estudantes e quais foram as abordagens de previsão do desempenho encontradas. Os autores Pelima, Sukmana e Rosmansyah (2024) investigaram as abordagens dos modelos de predição, as técnicas de MDE que foram implementadas, as fontes de dados e as variáveis selecionadas dos alunos. No entanto, a revisão investigou apenas as pesquisas escritas em inglês, focando no contexto da previsão de término de cursos de graduação e nas tomadas de decisão pelos alunos.

Este trabalho desenvolveu um Mapeamento Sistemático de Literatura (MSL) sobre o tema de Previsão Precoce para a descoberta de alunos em risco de reprovação acadêmica. Os resultados deste MSL apresentam informações essenciais sobre o tema, tais como: a classificação das estratégias de PP, a atualização das técnicas utilizadas neste campo, as fontes de dados de onde foram extraídas as informações sobre os alunos e quais foram as variáveis-chave selecionadas para identificação do fracasso acadêmico. O mapeamento sistemático ainda proporcionou uma análise das variáveis-chave e quais abordagens de Previsão Precoce foram relacionadas. Esta contribuição é importante na construção de estratégias de intervenção e personalização do ensino para o aluno. Esta pesquisa está dividida da seguinte forma: a seção 2 descreve a metodologia deste mapeamento sistemático; a seção 3 apresenta os resultados e as discussões; e a seção 4 trata das considerações finais.

2. Metodologia

Este mapeamento sistemático teve como referência a metodologia conduzida pelos pesquisadores Kitchenham *et al.* (2010) para a realização de estudos secundários. Os autores apresentaram a estrutura de uma revisão sistemática que consistia nas fases de definição das questões de pesquisa, definição das estratégias de busca e definição dos critérios de inclusão e exclusão. Após a execução dessas fases, foi acrescentada uma última etapa para avaliar a qualidade dos artigos. Este MSL incluiu os trabalhos disponíveis até o primeiro trimestre de 2024.

2.1. Questões de Pesquisa

Com o propósito de atingir os objetivos traçados, a pesquisa de Silva, Pimentel e Botelho (2022) foi utilizada como referência para o desenvolvimento das seguintes questões de pesquisa: **(Q1)**: Como são utilizados os modelos de Previsão Precoce para identificar os alunos em risco de reprovação acadêmica? **(Q2)**: Quais são as técnicas adotadas nos modelos de Previsão Precoce neste contexto? **(Q3)**: Quais são as fontes de dados de onde as variáveis foram extraídas para os modelos de predição? **(Q4)**: Quais variáveis investigadas contribuem para a identificação dos alunos em risco de reprovação acadêmica através da Previsão Precoce?

2.2. *String* de Busca e Veículos de publicação

Com o propósito de elucidar as questões de pesquisa definidas anteriormente, a estratégia de busca foi organizada em dois momentos: criação da *string* de busca e definição dos veículos de publicação para coleta dos trabalhos. A definição da *string* de busca utilizou a junção de alguns termos e seus sinônimos, que pudessem selecionar todas as variações que preenchessem o objetivo desta pesquisa. O trabalho dos autores Silva, Pimentel e Botelho (2022) foi utilizado como referência para a criação desta *string*. Abaixo, segue a *string* de busca com termos definidos em inglês para abranger o maior número possível de trabalhos:

(”early prediction”OR ”early detection”) AND (”academic perfomance”OR ”machine learning”OR ”data mining”) AND (”educational data”) AND (”students”OR ”at-risk students”)

As bases de dados selecionadas para a escolha dos trabalhos publicados foram definidas pela grande importância na difusão de pesquisa nesta temática. Os veículos de publicação internacional selecionados foram IEEE, ACM *Digital Library* e Scopus. Quanto aos eventos e periódicos de Informática na Educação no Brasil, foram listados os seguintes veículos para uma busca manual: Simpósio Brasileiro de Educação em Computação (EduComp), Simpósio Brasileiro de Informática na Educação (SBIE), Workshop sobre Educação em Computação (WEI), Workshop de Informática na Escola (WIE), Revista Brasileira de Informática na Educação (RBIE), Revista Informática na Educação: Teoria & Prática (IETEP), Workshop de Desafios da Computação Aplicada à Educação (DesafIE), Revista de Informática Teórica e Aplicada (RITA) e Revista Novas Tecnologias na Educação (RENOTE). A próxima subseção trata dos critérios de inclusão e exclusão aplicados aos artigos coletados que atenderam ao objetivo deste mapeamento sistemático.

2.3. Critérios de Inclusão e Exclusão

A aplicação dos critérios de inclusão e exclusão foi realizada em duas etapas. Na primeira etapa, os artigos foram avaliados com base na leitura do título e resumo para eliminação dos trabalhos que não atenderam ao propósito deste MSL. A segunda etapa foi realizada uma análise detalhada do artigo para verificar se atendia ao propósito deste mapeamento sistemático, respondendo às questões de pesquisa levantadas. Os critérios de inclusão (CI) e exclusão (CE) que foram aplicados são: **(CI1)**: o trabalho contempla aplicação de Previsão Precoce para medir o desempenho acadêmico; **(CI2)**: o trabalho utiliza abordagens de Previsão Precoce para identificar alunos em risco de reprovação; **(CE1)**: o trabalho não contempla aplicação de Previsão Precoce para medir o desempenho acadêmico; **(CE2)**: o trabalho não utiliza abordagens de Previsão Precoce para identificar alunos em risco de reprovação; **(CE3)**: O artigo está inacessível; e **(CE4)**: o artigo está duplicado.

2.4. Avaliação de Qualidade

Esta última etapa teve como objetivo garantir a confiabilidade e precisão dos trabalhos encontrados após a aplicação dos critérios de inclusão e exclusão, incluindo a avaliação da qualidade dos artigos. Os critérios de qualidade foram desenvolvidos com base nos trabalhos de (Dybå; Dingsøyr e Hanssen, 2007) e (Pelima; Sukmana e Rosmansyah, 2024). A utilização dos critérios de qualidade foi realizada após uma análise aprofundada dos artigos, com aplicação de uma pontuação para cada critério de qualidade, cuja divisão é: ”0”quando o artigo não for apto; ou ”1”caso o artigo seja apto. Os tipos de critérios de qualidade foram classificados em três grupos: **Rigor**, se a abordagem contém os principais métodos de pesquisa no estudo; **Credibilidade**, se a pesquisa é explicada adequadamente; e **Relevância**, se as descobertas têm utilidade para a indústria de software e a comunidade científica. A pontuação máxima a ser contabilizada é 8, e os artigos que apresentaram pontuação a partir de 6 foram filtrados nesta etapa. Abaixo, a Tabela 1 descreve os oito critérios de qualidade utilizados para avaliar os artigos.

Tabela 1. Critérios de qualidade. Fonte: Os Autores

ID	Descrição	Tipo
CQ1	O conteúdo do trabalho está escrito de maneira clara e compreensível ?	Rigor
CQ2	O artigo responde as questões de pesquisa?	Rigor
CQ3	O artigo relata as técnicas desenvolvidas para atingir o objetivo?	Credibilidade
CQ4	O artigo descreve o contexto de aplicação do modelo de predição?	Credibilidade
CQ5	O artigo descreve a motivação da pesquisa?	Credibilidade
CQ6	O artigo descreve as dificuldades encontradas na aplicação do modelo de predição?	Relevância
CQ7	Os resultados são descritos de forma clara?	Relevância
CQ8	O estudo evidencia sua contribuição para a área de pesquisa?	Relevância

3. Resultados

Nesta seção, são apresentadas as respostas para as questões de pesquisa deste MSL com a análise e discussão dos elementos investigados. No total, 27 artigos foram selecionados. A Tabela 2 apresenta a listagem total, com 5 artigos no veículo IEEE, 3 artigos no veículo ACM, 10 artigos no veículo *Scopus*, 4 artigos no SBIE, 3 artigos no RBIE e 2 artigos na RENOTE. A listagem completa dos artigos filtrados está disponível no *link*: https://docs.google.com/spreadsheets/d/1lZeznITiSAMiyNsiU6X26gcuDFs-xiJOYd5f-xnEIUY/edit?usp=sharing. A identificação de cada artigo é estruturada através de um ID único.

3.1. Análise das Questões de Pesquisa

Q1: Como são utilizados os modelos de Previsão Precoce para identificar os alunos em risco de reprovação acadêmica? Este Mapeamento Sistemático de Literatura analisou os trabalhos filtrados e constatou uma padronização dos objetivos das pesquisas já realizadas. Por meio do trabalho de Pelima, Sukmana e Rosmansyah (2024), usado como referência neste MSL, os trabalhos foram classificados em tópicos indicando as principais estratégias de Previsão Precoce relacionadas a este tema:

- **Previsão nas Fase Iniciais**: identificação de alunos em risco nas primeiras semanas de curso;
- **Previsão por Progresso do Aluno**: acompanhamento frequente do aluno, traçando seu percurso para evitar ou minimizar problemas de aprendizagem;
- **Sistemas de Alerta**: sinalização de alunos com risco potencial de reprovação;
- **Previsão Comportamental**: caracterização do comportamento (saúde mental, dados fisiológicos, relações interpessoais) do aluno, que pode indicar problemas.

Os trabalhos referentes à **Previsão nas Fases Iniciais** têm a maior aplicabilidade neste campo de PP, com total de 14 trabalhos, enquanto os trabalhos de **Previsão por**

Tabela 2. Filtragem dos trabalhos selecionados. Fonte: Os Autores

Veículo de Publicação	Artigos Extraídos	Etapa 1	Etapa 2	Etapa 3
IEEE	25	12	7	5
ACM	18	9	4	3
Scopus	86	51	12	10
EduComp	0	0	0	0
SBIE	12	9	4	4
WEI	1	0	0	0
WIE	0	0	0	0
RBIE	10	3	3	3
IETEP	1	0	0	0
DesafIE	0	0	0	0
RITA	0	0	0	0
RENOTE	21	5	2	2
Total	175	90	32	27

Progresso do Aluno ocupam o segundo lugar, com um total de 8 trabalhos listados. Os tópicos de **Sistemas de Alerta** e textbfPrevisão Comportamental apresentam um total de 2 e 3 trabalhos, respectivamente. Abaixo, a Tabela 3 apresenta a descrição de quais são os trabalhos (com o código identificador) que correspondem aos tópicos definidos.

Tabela 3. Levantamento das estratégias de Previsão Precoce. Fonte: Os Autores

Identificação das Estratégias	Identificação dos Estudos Primários	Total
Previsão nas Fases Iniciais	[A1], [A2], [A3], [A8], [A13], [A16], [A17], [A18], [A19], [A21], [A22], [A23], [A24], [A25]	14
Previsão por Progresso do Aluno	[A5], [A9], [A10], [A11], [A12], [A14], [A15], [A20]	8
Previsão Comportamental	[A4], [A26], [A27]	3
Sistemas de Alerta	[A6], [A7]	2

Este MSL revelou trabalhos na estratégia de **Previsão nas Fases Iniciais** que utilizaram os dados acadêmicos e demográficos dos alunos para detectar aqueles que apresentam desempenho insatisfatório ou que já estão em fase de reprovação em disciplinas iniciais. Os trabalhos apontaram a oportunidade de intervenções pontuais, com melhorias no desempenho do aluno e na gestão da aprendizagem [A1],[A2],[A18],[A19],[A21],[A25].

Outras alternativas desenvolvidas foram: prever as dificuldades de aprendizagem para melhorias do sistema educacional por meio de fatores internos ou externos (problemas sociais dos alunos) [A3]; prever os resultados dos exames dos alunos com base nos testes de admissão combinados com as primeiras avaliações do curso [A8]; prever o desempenho do aluno com base em dados demográficos no contexto do ensino presencial [A13]; prever a reprovação com a utilização de diferentes fontes de dados em um ensino semipresencial, visando à geração dos resultados antes do término da disciplina [A16]; construir um repositório de trajetórias acadêmicas para caracterizar as reprovações [A17]; prever a reprovação apenas com as notas dos exames [A22]; coletar informações dos alunos e das opiniões dos educadores para proposição de personalização do ensino com base nos resultados [A23]; e gerar um modelo de predição genérico para qualquer tipo de aluno e curso com a seleção apenas de características acadêmicas e demográficas [A24].

Em relação à estratégia de **Previsão por Progresso do Aluno**, outros trabalhos buscaram alcançar os resultados no uso das plataformas de Ambientes Virtuais de Aprendizagem, por meio dos registros de interações dos alunos e análise de tarefas semanais desenvolvidas [A5],[A9],[A12],[A14],[A15]. A finalidade é identificar os alunos em risco de reprovação bem antes do término de uma disciplina, na qual o aluno pode acompanhar seu progresso e obter *feedback* das suas atividades entregues. Outra opção para o registro de interações do aluno é a utilização de sistemas de avaliação

online, contemplando um ambiente de programação em um curso da área computacional com o intuito de combinar as variáveis extraídas dos alunos às atividades desenvolvidas para alcançar os resultados esperados em um modelo de predição contínua [A10],[A11]. Outros autores [A20] tomaram como base a análise comportamental do aluno no ensino presencial, com base em avaliações periódicas ao longo do semestre letivo para que o aluno passe por uma supervisão de seu desempenho e tome a decisão de continuar ou mudar de curso.

Na estratégia de **Previsão Comportamental**, o aspecto mais importante é a utilização de características psicológicas e relações sociais (algumas dessas informações extraídas das redes sociais dos alunos) as quais são analisadas para a seleção de variáveis que serão aplicadas às técnicas de PP [A4],[A26],[A27], para que, juntamente com as variáveis acadêmicas, obtenham maior precisão nos resultados das aplicações. As abordagens de **Sistemas de Alerta** [A6],[A7] apresentam em suas pesquisas não apenas a previsão antecipada do insucesso acadêmico do aluno, mas propõem ferramentas de alerta de reprovação do aluno com resultados ágeis, como por exemplo, obtidos em menos de um mês.

Em relação às revisões sistemáticas descritas na seção de Introdução, este MSL acrescenta informações à temática de Previsão Precoce com a elaboração de um levantamento quantitativo e suas respectivas classificações. Esta estrutura ajuda a compreender quais tópicos da Previsão Precoce, voltados para a identificação de alunos em risco carecem de mais investigações, tais como: o desenvolvimento de **Sistemas de Alerta**, e uma análise mais detalhada do comportamento do aluno nos ambientes de ensino (**Previsão Comportamental**). O estudo aponta a preocupação das pesquisas em identificar os alunos com risco de reprovação nos primeiros períodos, utilizando apenas os dados prévios do histórico e/ou informações do aluno obtidas nas primeiras semanas. É importante o desenvolvimento de outras estratégias de Previsão Precoce no que diz respeito ao percurso acadêmico do aluno, para a construção de ferramentas de alerta e diagnóstico do desempenho acadêmico.

Q2: Quais são as técnicas adotadas nos modelos de Previsão Precoce neste contexto? Com base no levantamento das técnicas de aprendizagem de máquina adotadas no tema de aplicação da PP, foram listados os respectivos algoritmos: *Long Short Term Memory* (LSTM), *Stochastic Gradient Descent*, *Gradient Boost Regression Tree*, *Gradient Boost*, *Stacking Classifier* (STC), *CatBoost*, *Association Rules*, *PART Classifier*, *Bagging* (BAG), *MultiBoost* (MB), *Voting Classifier*, *Ensemble*, *Rule Learners*, *LightGBM*, *AdaBoost*, *Deep Learning*, *Neural Network*, *XGBoost*, *Multilayer Perceptron*, *K-Nearest Neighbors* (KNN), *Logistic Regression*, *Support Vector Machine*, *Naives Bayes*, *Decision Tree* e *Random Forest*. Abaixo, a Figura 1 apresenta a quantidade do uso destas técnicas no trabalhos investigados.

Este mapeamento sistemático revelou que a técnica de *Random Forest* é amplamente utilizada neste tema de previsão de alunos em risco, com um total de 18 abordagens de sua aplicação. As outras técnicas com grande utilização são: *Naives Bayes* e *Support Vector Machine*, com 10 abordagens. Por sua vez, a técnica *Decision Tree* apresenta 9 utilizações. Em contrapartida aos trabalhos de revisão sistemática citados na seção 1, outras técnicas cresceram neste âmbito: *LightGBM*, *AdaBoost* e *Deep Learning*. A pesquisa identificou que as técnicas referentes aos algoritmos de Classificação são de grande relevância e atraíram um maior uso, pois contribuem para as abordagens de PP que utilizam dados extraídos previamente dos discentes no início do curso. As pesquisas na abordagem de **Previsão nas Fases Iniciais** atuam não apenas utilizando técnicas de Classificação binária, mas se desdobram para Classificação multiclasse com a combinação

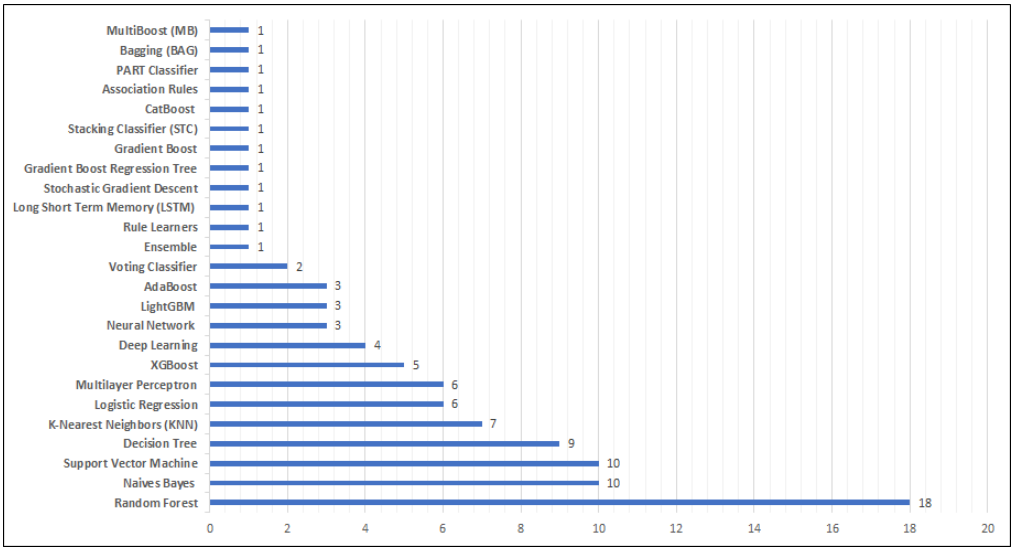

Figura 1. Quantitativo de técnicas de Previsão Precoce nos artigos listados. Fonte: Os Autores

de variáveis de diferentes tipos para que os resultados sejam mais eficazes [A8][A10]. Um outro aspecto é o incremento de estratégias junto às técnicas de aprendizagem de máquina, tais como a utilização da Inteligência Artificial Explicável em etapas finais, mesclagem de algoritmos de diferentes famílias, fusão de classificadores, a adoção de conceitos de *Learning Analytics* [A11], a obtenção dos dados dos alunos por meio de Metodologias Ativas (Sala de Aula Invertida) [A12] e o acompanhamento do percurso acadêmico através da Aprendizagem Autorregulada [A20]. Foi realizada uma breve análise comparativa das principais técnicas de PP encontradas nos trabalhos investigados para ajudar os docentes no entendimento de suas aplicações e resultados:

- As técnicas de *Decision Tree* e *Random Forest* apresentam uma precisão maior, sobretudo para prever atributos-alvos relacionados as notas dos discentes com conjunto de dados de informações escassas nos sistemas acadêmicos. Enquanto, o *SVM* já apresenta um resultado melhor com uso de dados acadêmicos e demográficos dos alunos. A técnica *Ensemble* atua com a junção de algoritmos de árvore junto ao *SVM* para ter maior precisão e sem erros inesperados nos resultados para os docentes;
- O algoritmo *Naive Bayes* é o mais simples em comparação as técnicas de árvore e produção de resultados similares. No entanto, pode apresentar alterações inesperadas nos resultados finais;
- A técnica *KNN* tem alta precisão, porém, se o modelo tiver uma grande conjunto de dados com características dos alunos pode dificultar o agrupamento destes que possuem pontuação similar;
- A técnica de *Logistic Regression* responde com maior precisão pela influência de características demográficas dos alunos em detrimento às informações acadêmicas;
- As técnicas de *Multilayer Perceptron*, *Neural Network*, *Deep Learning* apresentam o diferencial em relação as demais técnicas por apresentar modelos de maior precisão quando trabalham com várias informações comportamentais dos alunos incluindo o desenvolvimento de tarefas acadêmicas;
- As técnicas *XGBoost*, *CatBoost*, *LightBM*, *AdaBoost* e *GradientBoost* são aprimoramentos que apresentam maior eficiência com a melhoria contínua de modelos considerados fracos (por exemplo, atributos ausentes e maior número

de atributos categóricos) para construção de modelos preditivos fortes com base em um treinamento frequente;
- As outras técnicas citadas no início da seção precisam de mais experimentos que configurem o grau de eficiência em comparação as demais.

Q3: Quais são as fontes de dados de onde as variáveis foram extraídas para os modelos de predição? Esta pesquisa revelou quais são as fontes de dados utilizadas para a seleção das variáveis que foram aplicadas nas técnicas de Previsão Precoce. Abaixo, a Figura 2 apresenta o levantamento das fontes de dados que foram utilizadas.

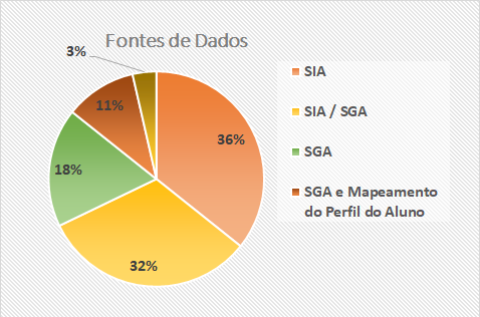

Figura 2. Levantamento das fontes de dados utilizadas nas pesquisas. Fonte: Os Autores

Os **Sistemas de Informação do Aluno (SIA)** apresentam 10 utilizações, correspondendo a 37% dos trabalhos encontrados neste MSL. Este registro é composto de características específicas do aluno, tais como: nome, sexo, gênero, idade, endereço, curso, notas, frequência, entre outras informações que são pertinentes para as fases iniciais do curso. A junção das fontes **SIA** e os **Sistemas de Gestão de Aprendizagem (SGA)** aparecem em segundo lugar, com 8 utilizações encontradas, correspondendo a 30% dos trabalhos encontrados. Essas fontes de dados nas pesquisas mapeadas apontam para informações pertinentes extraídas das interações dos alunos em Ambientes Virtuais de Aprendizagem (AVA), com a identificação dos dados acadêmicos dos alunos e índices de comportamento e dificuldades na aprendizagem. Sobre o uso apenas da fonte **SGA**, foram registradas 5 utilizações nas pesquisas, correspondendo a 18% dos estudos primários de Previsão Precoce extraídos, cujo foco é a utilização de registros de notas, histórico do aluno e outras informações dos cursos, proporcionando ao aluno o acompanhamento do seu desenvolvimento e progresso. A junção da fonte de dados **SGA** e a coleta de informações como o **Mapeamento do Perfil do Aluno** apresentam 3 utilizações, correspondendo a 11% das pesquisas do MSL. Esta fonte de dados mescla as informações acadêmicas do SGA com os questionários aplicados aos alunos para compreensão dos seus hábitos e dados fisiológicos. E a junção da fonte de dados **SGA** e as informações das **Mídias Sociais dos Alunos** apresentam 1 utilização, correspondendo a 4% das pesquisas do MSL. Esta fonte de dados mescla os registros acadêmicos do SGA com as informações obtidas pelas redes sociais dos alunos para entendimento das relações sociais estabelecidas entre todos os alunos.

Ao analisar esta terceira questão de pesquisa em comparação com os trabalhos de revisão citados na seção 1, este MSL apresenta dados atuais do tema com as fontes de dados mais utilizadas e aquelas que têm aplicações práticas insuficientes. Entre as oportunidades de pesquisa, o mapeamento constatou que a captura de características dos alunos por meios de questionários e entrevistas pode acrescentar informações novas, as quais os SIA e SGA não conseguem extrair, além de possibilitar a obtenção de informações pertinentes para a personalização do ensino do discente.

Q4: Quais variáveis investigadas contribuem para a identificação dos alunos em risco de reprovação acadêmica através da Previsão Precoce? Este Mapeamento Sistemático de Literatura identificou um padrão de variáveis-chave (características) dos alunos que contribuíram para determinar qual deve ser a abordagem de Previsão Precoce a ser adotada para detectar os alunos em risco de reprovação acadêmica. Abaixo, a Tabela 4 descreve a correlação entre variáveis-chave, os trabalhos extraídos neste mapeamento sistemático e as abordagens de Previsão Precoce que correspondem a cada trabalho.

Tabela 4. Levantamento dos tipos de variáveis selecionados. Fonte: Os autores

Nº	Tipos de variáveis-chaves	Artigo/Trabalhos	Abordagens
1	Acadêmicos	[A6], [A11], [A17], [A19], [A20], [A22]	Previsão nas Fases Iniciais; Previsão por Progresso do Aluno e Sistemas de Alerta
2	Acadêmicos e Demográficos	[A13]	Previsão nas Fases Iniciais
3	Acadêmicos, Demográficos e Comportamento	[A2]	Previsão nas Fases Iniciais
4	Acadêmicos e Redes Sociais	[A26]	Previsão Comportamental
5	Demográfico e Notas	[A3], [A5], [A23], [A24]	Previsão nas Fases Iniciais e Previsão por Progresso do Aluno
6	Acadêmicos e Atividades Programadas	[A12]	Previsão por Progresso do Aluno
7	Acadêmicos, Demográficos, Notas e Interações em Ambientes Virtuais de Aprendizagem	[A1], [A7]	Previsão nas Fases Iniciais e Sistemas de Alerta
8	Acadêmicos e Interações em Ambiente Virtual de Aprendizagem	[A9], [A10], [A14], [A15], [A18], [A27]	Previsão nas Fases Iniciais; Previsão por Progresso do Aluno e Previsão Comportamental
9	Acadêmicos e Atividades Programadas	[A25]	Previsão nas Fases Iniciais
10	Acadêmicos e Perfil do Aluno	[A4], [A16], [A21]	Previsão nas Fases Iniciais e Previsão Comportamental
11	Demográficos e Notas de Admissão	[A8]	Previsão nas Fases Iniciais

O conjunto destas características dos números 1 e 8 apresenta a maior utilização nos artigos filtrados deste mapeamento sistemático, refletindo a preocupação dos autores em obter respostas sobre o desempenho do aluno em risco de fracasso no ambiente acadêmico. Portanto, não precisaram de informações socioculturais e utilizaram informações cadastrais do aluno e notas que correspondiam aos períodos iniciais do curso como alternativa para acompanhar o progresso inicial do discente. A utilização dos Sistemas de Gestão de Aprendizagem contribuiu para o entendimento do comportamento do aluno e seu grau de interação em um Ambiente Virtual de Aprendizagem, permitindo a construção de ferramentas de alerta dos riscos de reprovação, ou ainda a análise de mudanças no comportamento do aluno que ocasionariam uma futura reprovação.

As variáveis-chave de número 5 são relacionadas a trabalhos que buscam características de dados populacionais (como, por exemplo, idade, gênero, número de familiares no mesmo domicílio, histórico escolar de pais, classe econômica e uso de que tipo de transporte) e que ajudam antecipadamente a mapear perfis de alunos em risco de reprovação em um período letivo. O conjunto de variáveis-chave 10 também remete ao mapeamento de um perfil do aluno através de entrevistas e questionários sobre interesses do aluno e que, somadas às notas avaliativas, contribuem na execução das abordagens de PP desejadas. As variáveis-chave de número 7 têm uma preocupação maior com a combinação das informações acadêmicas, demográficas e interações do aluno em Ambientes Virtuais de Aprendizagem, não apenas para obter previsões no início do curso, mas também emitir alertas precoces sobre o desempenho acadêmico do aluno. O aluno fica ciente do risco de não ter sucesso na aprovação em uma componente curricular. Outro ponto que chama a atenção são as variáveis-chave **4** que apresenta um trabalho que coleta dados acadêmicos e informações das redes sociais dos alunos para obtenção das relações sociais de amizade e as preferências de pesquisa (para, no futuro, personalizar o ensino). As variáveis dos números **2** e **3** visam identificar, nas primeiras semanas de curso, os

alunos que se enquadram em risco de reprovação; as variáveis-chaves de número **6** são direcionadas à coleta das informações de atividades semanais entregues pelos alunos para traçar o percurso estudantil em uma componente curricular. E as variáveis-chave de número **11** têm a proposta de coletar dados demográficos junto às respectivas notas da admissão de cada aluno no curso, com o intuito de mapear o perfil inicial dos alunos.

Em relação aos trabalhos de revisão sistemática citados na seção 1, este MSL contribuiu para traçar uma correlação entre os tipos de variáveis-chave apresentam a melhor indicação para o uso de cada estratégia de Previsão Precoce, que identifica os alunos em risco de reprovação acadêmica. Ou seja, esta pesquisa estruturou a classificação das abordagens de Previsão Precoce, dos trabalhos filtrados e das variáveis-chave selecionadas para a previsão do risco de reprovação. O MSL ainda oportuniza a descrição de correlações entre tipos de variáveis-chave e abordagens de PP apresentam poucos trabalhos na área.

4. Considerações Finais

Este MSL investigou os trabalhos referentes à Previsão Precoce para identificação de alunos em risco de reprovação em cursos de ensino médio e superior com as estratégias desenvolvidas. Em contraste com os trabalhos das revisões sistemáticas citadas na Introdução, este mapeamento apresentou uma estruturação das variáveis-chaves selecionadas, dos métodos adotados e das fontes de dados utilizadas nas aplicações deste tema. Esta investigação contribuiu para esclarecimentos dos campos de atuação da Previsão Precoce com a classificação dos modelos de predição em tópicos que delineiam as abordagens de detecção do risco de reprovação. Foram identificados dois tópicos que carecem de mais pesquisas: Sistemas de Alerta e Previsão Comportamental, que são áreas fundamentais para que o aluno possa acompanhar seu desempenho e obter sucesso em seu aprendizado. Além disso, com esses resultados, os docentes podem desenvolver estratégias para o acompanhamento do progresso do aluno e identificar os aspectos essenciais para melhorias no ensino.

Sobre as técnicas utilizadas nos modelos de Previsão Precoce, é importante destacar a grande frequência de uso das seguintes técnicas: *Random Forest*, *Naives Bayes* e *Support Vector Machine*, pois estas são destacadas pelos autores como eficazes e que produzem resultados mais expressivos através do uso de dados prévios dos alunos. Além disso, os trabalhos direcionados ao mapeamento do comportamento do aluno em seu percurso curricular adotaram a técnica de *Deep Learning* por apresentar um melhor desempenho com a mesclagem de dados acadêmicos, demográficos e interativos dos alunos nos ambientes de aprendizagem. Para o aperfeiçoamento das estratégias de PP, o uso das técnicas de MDE é combinado com os métodos de Aprendizagem Autorregulada e *Learning Analytics* para a extração de características dos discentes encontradas fora do ambiente acadêmico. Outro campo de estudo identificado no MSL é a utilização de técnicas de interpretabilidade de Inteligência Artificial no detalhamento e esclarecimento dos resultados de desempenho acadêmico, com o objetivo de criar as bases para a personalização do ensino do aluno.

Para futuras pesquisas, planeja-se realizar uma investigação sobre a Previsão Precoce em ambientes de aprendizagem híbridos e quais metodologias são utilizadas para a extração dos dados de interação dos alunos nesse cenário. Além disso, é necessário analisar a interpretabilidade das técnicas de MDE utilizadas nas abordagens de Previsão Precoce. Isso é importante para determinar a eficácia e a transparência dos resultados transmitidos aos docentes, permitindo que eles detectem os alunos em risco de reprovação.

Referências

Batista, M. R.; Fagundes, R. A. d. A. Mineração de dados educacionais aplicada a performance de estudantes: uma revisão sistemática da literatura. **Revista Novas Tecnologias na Educação**, v. 21, n. 1, p. 271–280, jul. 2023. Disponível em: ⟨https://seer.ufrgs.br/index.php/renote/article/view/134355⟩.

Berens, J.; Schneider, K.; Gortz, S.; Oster, S.; Burghoff, J. Early detection of students at risk - predicting student dropouts using administrative student data from german universities and machine learning methods. **Journal of Educational Data Mining**, v. 11, n. 3, p. 1–41, Dec. 2019. Disponível em: ⟨https://jedm.educationaldatamining.org/index.php/JEDM/article/view/389⟩.

Dybå, T.; Dingsøyr, T.; Hanssen, G. Applying systematic reviews to diverse study types: An experience report. In: . [S.l.: s.n.], 2007. p. 225 – 234. ISBN 978-0-7695-2886-1.

Kitchenham, B. *et al.* Refining the systematic literature review process-two participant-observer case studies. **Empirical Software Engineering**, v. 15, p. 618–653, 12 2010.

López-Zambrano, J.; Torralbo, J. A. L.; Romero, C. Early prediction of student learning performance through data mining: A systematic review. **Psicothema**, v. 33, 07 2021.

Pelima, L.; Sukmana, Y.; Rosmansyah, Y. Predicting university student graduation using academic performance and machine learning: A systematic literature review. **IEEE Access**, PP, p. 1–1, 01 2024.

Silva, B.; Pimentel, E.; Botelho, W. Predição de desempenho de estudantes: Uma revisão sistemática de literatura. In: **Anais do XXXIII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2022. p. 1040–1052. ISSN 0000-0000. Disponível em: ⟨https://sol.sbc.org.br/index.php/sbie/article/view/22480⟩.