Mineração de Dados Educacionais aplicada a performance de estudantes: uma revisão sistemática da literatura

Messias Rafael Batista, PPGEC - Universidade de Pernambuco,
mrb@ecomp.poli.br, https://orcid.org/0000-0002-7893-8838
Roberta Andrade de Araújo Fagundes, PPGEC - Universidade de Pernambuco,
roberta.fagundes@upe.br, https://orcid.org/0000-0002-7172-4183

Resumo: Tomada de decisão orientada a dados é parte dos processos institucionais contemporâneos, o campo educacional, influenciado por este contexto, utiliza-se do *educational data mining* para ampliar sua capacidade de criar soluções e extrair informações existentes em grandes volumes de dados. Utilizando-se das técnicas de Revisão Sistemática da Literatura, buscou-se compreender no período de 2010 a 2022 a produção científica que analisou a *performance* de estudantes em concursos educacionais fazendo uso de *educational data mining* e técnicas machine *learning*. Esta pesquisa resultou em um conjunto de dados que demonstram as principais técnicas e problemas resolvidos com tecnologia.

Palavras-chave: mineração de dados educacionais, machine learning, performance de estudantes

Educational Data Mining to apply a student performance: a systematic literature review

Abstract: Data-driven decision-making is part of contemporary institutional processes, and the educational field, influenced by this context, utilizes educational data mining to enhance its ability to create solutions and extract insights from large volumes of data. Using techniques for Systematic Literature Review, it was understood in the period 2010 to 2022 the science production objective was to study student performance subjects in educational tests, using educational data mining and machine learning. This research results in many data demonstrating the principal techniques and problems solved with technology.

Keywords: educational data mining, machine learning, student performance

1. Introdução

A performance dos estudantes em programas de avaliação educacional internacionais, demonstram a orientação para ampliar o acesso à educação de qualidade. Neste cenário de avaliações (ou concursos educacionais), o Brasil apresenta diversos desafios, como demonstrado pelo relatório Brasil no PISA 2018 (BRASIL, 2020). De acordo com este documento, a posição brasileira em relação a avaliação de Leitura resulta entre 55° e 59°, em Matemática entre 69° e 72°, e em Ciências entre 64° e 67°, quando comparado com os 79 países participantes (BRASIL, 2020).

Aumentar a qualidade da educação com objetivo de melhores resultados em concursos educacionais não configura uma tarefa elementar. Mas, um contexto orientado a dados, fazendo uso de ferramentas como mineração de dados e modelos/técnicas de *machine learning*, pode evidenciar fatores de destaque que podem impactar tais resultados. Portanto, a mineração de dados aplicada à educação, atua como um campo interdisciplinar, composto por um processo de análise de dados envolvendo extração de informações de grandes bases de dados. Esse processo de mineração de dados é

consolidado a partir da utilização de modelos/técnicas de *machine learning* e além de análises que buscam encontrar padrões, tendências e *insights* nos dados (FERRARI, 2017). Nesta perspectiva, o aprendizado de máquina processa dados históricos com o objetivo de desenvolver expertise na resolução de problemas específicos (ALPAYDIN, 2020). Assim, prever e classificar a performance dos estudantes é objeto de estudo de pesquisadores que utilizam o processo de mineração de dados, como também, modelos/técnicas de *machine learning* como metodologias aplicadas nas pesquisas.

No tocante ao recente resultado brasileiro no PISA 2018, e verificando uma tendência de soluções orientadas a dados, evidencia-se que esforços desta natureza podem promover soluções para a melhora de indicadores educacionais. Portanto, este trabalho objetiva contribuir com uma revisão sistemática da literatura que buscou compreender a aplicação do processo de mineração de dados e de modelos/técnicas de *machine learning* em problemas de performance de estudantes.

A estrutura deste artigo está dividida em: Seção 2 que descreve os Trabalhos Relacionados, no qual se busca apresentar os trabalhos relevantes que integram o tema de pesquisa; Seção 3, descreve as etapas realizadas para Revisão Sistemática da Literatura (RSL), no qual se demonstra o planejamento e condução da RSL; Seção 4, apresenta os Resultados e Discussão dos dados encontrados com a RSL, como também, a aplicação dos três principais modelos/técnicas de ML encontrados na RSL e aplicado ao conjunto de dados do PISA 2018; Seção 5, Ameaças à Validade, no qual se elucida fatores prejudiciais e estratégias aplicadas. Por fim, Seção 6, Considerações Finais, em que é sintetizada a contribuição realizada e evidencia a lacuna encontrada para o desenvolvimento de trabalhos futuros.

2. Trabalhos Relacionados

Os trabalhos destacados nesta seção, foram selecionados por estarem presentes nos resultados da RSL e por manterem coerência com a linha de pesquisa.

Cortez e Silva (2008) aplicaram mineração de dados para prever as notas de alunos do terceiro ano médio (G3). Utilizando dados socioeconômicos e de desempenho escolar (notas do segundo - G2 - e primeiros anos - G1), os autores apresentaram os melhores resultados com os algoritmos *Random Forest* e *Decision Tree*. Eles também ressaltaram a importância de estudos adicionais sobre a influência de outras áreas sociais nos resultados dos estudantes, usando técnicas de seleção de atributos.

Baker et al. (2011) conduziram uma análise das oportunidades de aplicação da mineração de dados na educação brasileira. Suas conclusões destacam que o cenário oferece perspectivas promissoras devido ao volume de dados gerados.

Shahiri et al. (2015) conduziram uma RSL para identificar lacunas na existência de métodos de predição e atributos para analisar o desempenho dos estudantes. Eles destacaram a importância dos atributos e os métodos de predição mais utilizados, como *neural network* e *decision tree*, especialmente para prever a média de notas (GPA) devido à sua tangibilidade. Outros conjuntos de atributos, como dados de interação social, demográficos e comportamentais, também foram considerados para prever a GPA.

Alamri e Alharbi (2021) analisaram produções científicas recentes que investigavam o desempenho do estudante usando modelos preditivos e explicáveis (*Explainable Models*). O estudo categorizou trabalhos por problemas computacionais, níveis educacionais e tipos de predições. Os autores enfatizaram a necessidade de criação de modelos explicáveis e propuseram mais pesquisas científicas para desenvolver métricas que permitam a comparação entre os modelos dessa técnica.

Al-Fairouz e Al-Hagery (2020) comparou técnicas de classificação em uma base de dados de 72.259 registros, do Departamento de Administração da Faculdade de

Economia e Negócios da Grécia. Os autores contribuíram demonstrando que dentre outros algoritmos o *Random Forest* apresentou melhor acurácia para prever a performance dos estudantes por meio de classes (categorias) construídas a partir de intervalos de notas, tais quais: excelente, muito bom, bom, aceitável, reprovado.

Mengash (2020) aplicou mineração de dados para melhorar a tomada de decisão na admissão de candidatos ao ensino superior. O estudo usou 2.039 registros e obteve resultados satisfatórios, destacando o sucesso do algoritmo *Artificial Neural Networks*. Esses modelos podem ser úteis para o planejamento de otimização de recursos limitados pelos *stakeholders*.

Dien et al. (2020), aplicando *Deep Learning* com *Data Transformation*. A contribuição dos autores está direcionada em uma transformação baseada em *Quantile Transformation* e *MinMaxScaler*, recursos utilizados para converter os valores em padrão de intervalo que permitem algoritmos de *deep learning* apresentarem melhores resultados.

Cechinel (2020) realizou um mapeamento da produção científica em *learning analytics* na América Latina, comparando-a com outras regiões. A conclusão do estudo revela que, embora tenha sido observado crescimento, ainda há espaço para maior exploração nesse campo de estudo na região.

Cagliero et al. (2021), fazendo uso de *Associative Classification Models*, estuda os dados de 5.000 alunos com objetivo de abordar o modelo de *machine learning* com um aprendizado explicativo. Elucidando que os modelos associativos tem resultados de qualidade tais quais os modelos de classificação, e fornecem informações importantes sobre a taxa de sucesso dos alunos.

Queiroga (2022) enriquece a pesquisa ao enfatizar as discrepâncias e convergências entre mineração de dados educacionais e *learning analytics*. Seu estudo resulta na proposição de três metodologias aplicadas à área, sendo uma delas desenvolvida com a utilização de algoritmos genéticos.

Os trabalhos desta seção elucidam, além de suas particularidades, a relevância da execução de uma Revisão Sistemática da Literatura (RSL) no sentido de compreender quais conjuntos de técnicas e dificuldades foram encontrados nas recentes pesquisas da área. Verifica-se que há uma busca nos trabalhos recentes por modelos que possam ser explicados (Cagliero et al., 2021), bem como por processos refinados no pré-processamento de dados (Dien et al., 2020), além da consolidação de técnicas tradicionais (Mengash et al., 2020; Al et al., 2020; Alamri et al., 2021) e do uso de algoritmos genéticos (Queiroga, 2020; 2022).

3. Revisão Sistemática da Literatura

A Revisão Sistemática da Literatura (RSL) é o procedimento pelo qual se busca sistematizar evidências de pesquisa (Nakagawa & Cannavino, 2017). Corrobora o que foi dissertado por Kitchenham e Charters, afirmando que a RSL é um estudo de caráter secundário, que tem como objetivo identificar, avaliar ou interpretar as pesquisas relevantes disponíveis em um campo de estudo (Kitchenham & Charters, 2007).

As RSL desempenham um papel significativo no campo da educação ao identificar lacunas em contextos educacionais específicos. Um exemplo notável é o campo de *learning analytics*, como evidenciado pelo estudo de Wilker Pereira Luz et al. (2021), ou o trabalho de Torres Marques et al. (2022) que versou uma análise sobe conhecimento relacionado à retenção acadêmica.

Portanto, o objetivo da RSL deste trabalho foi analisar a produção científica que aplicasse mineração de dados em conjunto com técnicas de aprendizado de máquina para verificar o desempenho dos estudantes. Para isso, utilizou-se como estratégia de revisão

a seleção de estudos primários, nos quais foram verificadas produções científicas a partir de fontes de estudos anteriores, palavras-chave e período da pesquisa, com base nos critérios de inclusão e exclusão.

O protocolo da revisão sistemática previu as fases de planejamento, condução e apresentação dos resultados. No planejamento, foram definidos os objetivos, a estratégia de busca, as fontes de pesquisa, a estratégia de busca e os critérios de inclusão, exclusão e qualidade, concluindo com uma etapa de avaliação do protocolo. Na fase de condução, foram identificados e selecionados os estudos primários, e os dados foram extraídos e sintetizados para apresentação.

3.1 Planejamento

O planejamento da RSL é derivado da etapa inicial de construção das Perguntas de Pesquisa.

3.1.1 Perguntas de Pesquisa

Para esta pesquisa foram desenvolvidas a Pergunta Geral (PG) e três Perguntas Específicas (PE).

PG: Quais técnicas de mineração de dados educacionais utilizam técnicas de *machine learning* para avaliar a performance de estudantes em concursos educacionais?

PE1: Como a mineração dados educacionais auxilia na explicação da *perfomance* de estudantes em concursos educacionais?

PE2: Por que usar mineração de dados para avaliar a *perfomance* de estudantes em concursos educacionais?

PE3: Quais as técnicas de *machine learning* são mais utilizadas para identificar a performance de estudantes em concursos educacionais?

3.1.2 Engenhos e String de Busca

Conduziu-se buscas automáticas e manuais com o objetivo de coletar Estudos Primários (EPs) em Engenhos de Busca selecionados com base em seu reconhecimento na comunidade científica e em sua associação com o campo de estudo (Tecnologia da Informação e Educação).

Foram selecionados cinco engenhos para realização de busca automática: ACM *Digital Library*, IEEE, *Science Direct*, Scopus e *Web of Science*. Além disso, foram selecionados três engenhos (revistas/conferências) para busca manual: Simpósio Brasileiro de Informática na Educação (SBIE), Revista Novas Tecnologias na Educação (RENOTE) e *Fourteenth International Conference on Educational Data Mining*.

A etapa seguinte consistiu na definição da *string* de busca, que tinha como objetivo sintetizar os argumentos representativos do contexto da pesquisa, com a inclusão de termos relacionados à PG. Dessa forma, foi gerada a seguinte *string* de busca:

(*"educational data mining" OR "data mining" OR "machine learning") AND (classification OR predict) AND ("educational performance" OR "academic performance" OR "student performance")*

3.1.3 Critérios para Seleção

Em uma Revisão Sistemática da Literatura (RSL), os critérios de seleção são utilizados para identificar os Estudos Primários (EPs) que se adequam às perguntas de pesquisa. Esses critérios são divididos em critérios de inclusão e exclusão.

Critérios de Inclusão: **(CI1)** Estudos primários que respondam a pelo menos uma questão de pesquisa. **(CI2)** Estudos primários publicados entre janeiro de 2010 e março de 2022. **(CI3)** Estudos primários escritos em inglês ou Português do Brasil. **(CI4)** Estudos primários publicados em congressos, conferências ou revistas científicas. **(CI5)**

Estudos que tenham como objeto de estudo a análise de desempenho de estudantes utilizando técnicas de aprendizado de máquina.

Critérios de Exclusão: **(CE1)** Estudos que não sejam primários. **(CE2)** Estudos publicados apenas como resumo. **(CE3)** Estudos descritos como pôster, editorial, resumo ou relatório parcial de pesquisa. **(CE4)** Estudos que apareçam em mais de uma base de dados (duplicados). **(CE5)** Estudos que não possuam resumo. **(CE6)** Estudos inacessíveis.

Os critérios de inclusão definiram as características que um estudo deveria apresentar para ser incluído na revisão. Por outro lado, os critérios de exclusão estabeleceram os limites para classificar um estudo como irrelevante para a revisão em andamento (Nakagawa & Cannavino, 2017).

3.1.4 Critérios de Qualidade

Com a finalidade de aumentar a confiabilidade do resultado da RSL, a aplicação dos critérios de qualidade resultou em um processo necessário. Busca-se nesta etapa qualificar os EP selecionados na fase anterior. O processo de desenvolvimento e aplicação dos critérios de qualidade se fundamentou no trabalho de Dyba *et al.* (2007), que sugere onze critérios categorizados em quatro conceitos de qualidades.

Nesta fase, foram adaptados os critérios de qualidade, e cada um foi relacionado com uma categoria de qualidade. Sua aplicação foi realizada após a leitura íntegra dos estudos, aplicando uma pontuação para cada critério de qualidade. Cada estudo poderia receber uma pontuação [0], quando não aderente; ou [1], caso o estudo aderente ao Critério de Qualidade.

Como demonstrado na Tabela 1, os critérios versaram em quatro grupos, **relatório**, referindo-se à qualidade das informações descritas no estudo; **rigor**, referente a adequação dos métodos utilizados; **credibilidade**, alusivo à significância das descobertas; **relevância**, relativo à utilidade das descobertas para a indústria e comunidade científica.

Tabela 1 – Critérios de qualidade por tipo

ID	Critérios de Qualidade	Tipo
CQ1	A pesquisa está relacionada com a aplicação de técnicas de *machine learning* para prever a *performance* dos estudantes?	Relatório
CQ2	Os objetivos se apresentam de forma clara?	Relatório
CQ3	Existe uma descrição adequada do contexto em que a pesquisa foi realizada?	Relatório
CQ4	O desenho da pesquisa foi apropriado para abordar os objetivos da pesquisa?	Rigor
CQ5	Os dados coletados são coerentes com a questão de pesquisa?	Rigor
CQ6	A análise dos dados foi suficientemente rigorosa?	Rigor
CQ7	Os resultados são descritos de forma clara?	Credibilidade
CQ8	O estudo expõe com clareza o local de publicação da pesquisa?	Credibilidade
CQ9	O estudo evidencia sua contribuição para a área ou campo de pesquisa?	Credibilidade
CQ10	A pesquisa deixa claro a quem contribui?	Relevância

Nesta Revisão Sistemática da Literatura (RSL), foram adaptados e aplicados dez critérios de qualidade, reduzidos a partir dos onze propostos por Dyba et al. (2007). Todos os estudos foram classificados, sem remoção de nenhum deles. Ao analisar a aderência aos critérios, observou-se que 77,8% dos estudos apresentaram conformidade nos critérios do tipo relatório, sendo o CQ3 o menos aderente, com apenas 36,4% de conformidade.

Em relação aos critérios de rigor, todos os estudos alcançaram o máximo de conformidade. Quanto à credibilidade, 99% dos estudos se adequaram aos critérios, com o CQ9 sendo o único menos aderente, com 97% de conformidade. Já em relação à relevância, o critério CQ10 apresentou aderência de 42,4%, sendo o mais afetado.

Ao final, observou-se que um estudo apresentou aderência a sete critérios, 14 estudos aderiram a oito critérios, 13 estudos aderiram a nove critérios, e sete estudos aderiram a todos os dez critérios analisados.

3.2. Condução da RSL

A primeira tarefa conduzida foi a coleta dos estudos primários, aplicando o corte no tempo que versou entre janeiro de 2010 e março de 2022. Nesta tarefa, foram considerados apenas trabalhos completos. A busca automática resultou em um número de 1.811 EP retornados. Por outro lado, a busca manual não obteve estudos relacionados ao tema.

A etapa em sequência foi a de seleção, como expresso no planejamento (seção 3.1.3), demonstrado na Figura 1. Foram realizadas as leituras dos títulos e palavras-chave de cada um dos estudos coletados na fase inicial, e aplicado cada um dos critérios de inclusão e exclusão, em paralelo. Assim, houve uma redução de 1.710 estudos, resultando em 101 estudos selecionados pelos critérios aplicados. Ainda foram verificados os estudos duplicados e este número finalizado, em 91 estudos aptos a nova rodada de análise.

Os estudos foram submetidos a nova rodada de leitura, na qual além do título e palavras-chaves, foram lidos resumo, introdução, resultados e conclusões, aprovando um número final de 35 estudos primários.

Figura 1 – Etapas da revisão sistemática da literatura por resultados de EPs
Fonte – Elaborado pelos autores.

A última fase de condução da RSL foi realizada com a leitura dos estudos na íntegra, e com a aplicação dos critérios de qualidade. Esta fase buscou avaliar e identificar a adesão dos estudos em relação aos critérios de qualidade. Devido a criticidade aplicada na fase anterior, nenhum estudo foi removido nesta etapa. Destaca-se que em cada ciclo houve a participação de dois pesquisadores, na qual o segundo atuou como especialista e revisor da RSL, com o objetivo de manter sólidos fundamentos científicos.

4. Resultados e Discussões

Nesta seção serão demonstrados o resultado e discussões encontrados durante a condução desta RSL. A análise desta seção será suportada por análise quantitativa e discussões qualitativas.

Entre os resultados é possível verificar que no período analisado, a maior parte dos estudos se concentra nos anos de 2015, 2018, 2019, 2020, mas com produções existentes em quase todos os anos entre 2012 e 2022, o que demonstra a consistência da área de pesquisa, apesar da inexistência de estudos aderentes ao tema em 2010, 2011 e 2014.

A distribuição dos estudos ao longo do tempo demonstra a relevância dos trabalhos sobre análise de desempenho de estudantes com a aplicação de mineração de dados e técnicas de aprendizado de máquina. A pesquisa revelou que mais de 50% dos

estudos sobre o tema foram publicados nos últimos três anos, como mostrado na Figura 2. Isso evidencia a atualidade e importância do tema abordado na revisão proposta.

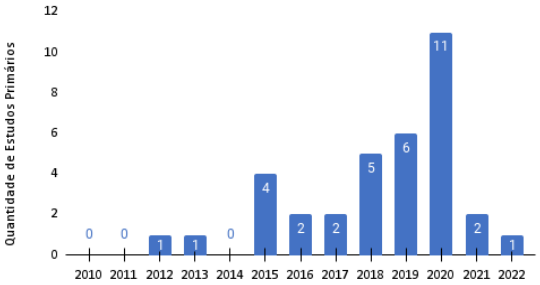

Figura 2 – Quantidade de EPs por ano de publicação
Fonte – Elaborado pelos autores.

Outra dimensão descritiva é o número de trabalhos por problema computacional proposto, especificamente Classificação e Regressão. De acordo com os resultados da revisão, foram identificados 32 Estudos Primários (EPs) que abordam problemas de classificação, enquanto apenas 3 EPs tratam de problemas de regressão. Isso indica que os métodos e técnicas de regressão devem ser explorados mais amplamente nesse contexto de desempenho dos estudantes.

Apesar do foco em classificação e regressão para a análise de desempenho dos estudantes, a revisão revelou uma diversidade de técnicas.

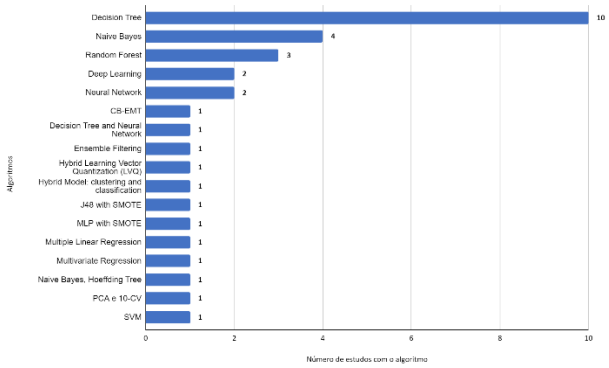

Figura 3 – Quantidade de EPs por ano de publicação
Fonte – Elaborado pelos autores.

A RSL também elucidou quais métodos/técnicas de *machine learning* tiveram maior relevância na comunidade científica quando aplicadas ao problema educacional (*performance* de estudantes). De acordo com a Figura 3, *Decision Tree* é o método/técnica que largamente é utilizado, sempre associado a problemas de classificação. Destaca-se que *Naive Bayes* está na segunda posição, mas com uma diferença de seis estudos. Por fim, a terceira posição é ocupada pela técnica *Random Forest*.

Estes estudos que compõem o "estado da arte" dos métodos/técnicas tradicionais para classificação e regressão sugerem que este cenário de pesquisa tem sido pouco explorado pelas novas metodologias, com resultados esporádicos como: SVM (Bhutto et al., 2020), PCA e 10-CV (Sokkhey & Ong, 2020), MLP e J48 com SMOTE (Thaher et al., 2020; Bujang & Zulkarnain, 2021), ou ainda com os métodos/técnicas Ensemble, como mencionado por Rahman et al. (2017), por exemplo.

No tocante à motivação dos trabalhos, os estudos em sua maioria representam experimentos entre conjuntos de metodologias, visando compreender o que resulta em melhor capacidade de acerto e menor incidência de erro. Entretanto, o trabalho de Shahiri et al. (2015) evidencia um experimento prático baseado em técnicas resultantes de uma Revisão Sistemática da Literatura (RSL), bem como a proposição de um processo de mineração de dados híbrida, fazendo uso de mais de uma metodologia de aprendizado de

máquina, proposto por Francis et al. (2019), e ainda a arquitetura de treinamento a partir de um cluster distribuído proposto por Almasri et al. (2020).

O conjunto de estudos contribuiu com esta pesquisa ao demonstrar o panorama dos últimos 10 anos em relação aos problemas solucionados e metodologias aplicadas, evidenciando a existência de lacunas de pesquisas.

5. Experimentação e Análise

Nesta seção será demonstrada a aplicação das três metodologias mais utilizadas no problema de performance de estudantes, de acordo com a Figura 1. O *Programme for International Student Assessment* (PISA) é organizado pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE), e sua última aplicação foi no ano de 2018. O PISA avalia os estudantes em dimensões como ciências, leitura e matemática, além de realizar um extenso questionário socioeconômico, social, familiar e escolar. Esse conjunto de perguntas resulta em uma base de dados com 1.119 atributos (colunas) (OCDE, 2018).

Utilizando como objeto de estudo o conjunto de dados do PISA 2018, foi aplicado um filtro para resultar apenas em dados de estudantes brasileiros. Foram executados os algoritmos *Decision Tree*, *Bayesian Ridge* e *Random Forest* no conjunto de dados resultante, implementados a partir da biblioteca *Scikit Learn* 1.2.

O problema em questão consiste em buscar compreender quais os fatores (variáveis) que influenciam a performance do estudante (média das notas de ciências, leitura e matemática). Para a seleção dos atributos (variáveis) a partir do conjunto de 1.119 atributos, foram realizadas as seguintes etapas:

Etapa 1 de compreensão da base de dados, por meio da aplicação de técnicas de estatística descritiva e estudo do dicionário de dados.

Etapa 2 de pré-processamento, na qual foi realizada a limpeza e redução da base de dados, removendo colunas não relevantes e aplicando a substituição de valores nulos pela mediana.

Etapa 3 de aplicação da seleção de atributos, por meio da técnica de *mutual info regression*, que utiliza o algoritmo KNN para reconhecer os dez vizinhos mais próximos. Como resultado dessa técnica, foram selecionados os seguintes atributos (variáveis): ST127Q01TA, ST127Q02TA, AGE, ISCEDL, REPEAT, BSMJ, MMINS, JOYREAD, GCSELFEFF, SENWT.

Etapa 4, denominada modelagem, envolveu a aplicação das três técnicas mais frequentes nos estudos primários da RSL: *Decision Tree*, *Naive Bayes* e *Random Forest*. As variáveis independentes (x) foram definidas na etapa 3, enquanto a variável dependente (y) selecionada foi a média das notas dos estudantes em ciências, matemática e leitura, conforme avaliação do PISA 2018, representada pelo atributo ST001D01T.

Etapa 5, denominada avaliação dos modelos, foram analisadas três métricas: erro quadrado (R2), erro médio quadrado (MSE) e erro médio absoluto (MAE). São métricas tradicionais na literatura quando aplicadas a problemas de previsão de valores (regressão). O R2 é uma medida que verifica a assertividade do modelo, portanto, quanto mais próximo de 1, mais performático se mostra o modelo. Por outro lado, o MSE e o MAE são medidas de erro, e quanto menor seu valor, mais favorável é a performance do modelo. O ambiente de desenvolvimento utilizado para o experimento foi o Google Colaboratory, com a linguagem de programação Python e a biblioteca Scikit-Learn.

Tabela 2 – Métricas por algoritmo

Técnica	R^2	MSE	MAE
Decision Tree	0.82	0.18	0.14
Naive Bayes	0.70	0.30	0.44
Random Forest	**0.89**	**0.11**	**0.22**

A partir dos dados na Tabela 2, é possível verificar que este é um problema complexo para os algoritmos experimentados. Entretanto, o ***Random Forest*** apresentou a melhor performance no tocante a sua assertividade, visualizada na coluna R2, com 0.89 de acerto, e o menor erro quando verificada a coluna MSE. Porém, a menor performance foi verificada com o algoritmo *Naive Bayes*, que obteve 0.70 de R2, e 0.30 de MSE. e a performance intermediária foi o *Decision Tree*, atingindo 0.82 na métrica R2 e 0.18 na métrica de erro MSE, mas se destacando como o menor MAE com um resultado de 0.14.

Este contexto explicita que o modelo baseado no *Random Forest* acertaria em torno de 9 em cada 10 novos registros de estudantes no tocante a prever a nota do PISA 2018 em relação ao conjunto de atributos. Desta forma, seria possível antever o resultado de estudantes a partir do conjunto de atributos, e estimular dimensões que possam impactar para um aumento da performance.

6. Ameaças à validade

A condução desta RSL buscou seguir as orientações da comunidade científica, de forma que seu resultado possa agregar a esta área de conhecimento. Entretanto, esta seção elucida possíveis limitações que possam ameaçar a validade dos resultadas, da quais se destacam: **(i)** o viés do pesquisador, pode influenciar o processo de inclusão e exclusão de EP, bem como sua categorização, classificando como uma ameaça; (**ii)** a construção da *string* de busca, pode não ter incluídos todos os termos relacionados a pesquisa desenvolvida; e (**iii)** a condução por apenas um pesquisador e um especialista, pode ter resultado em algum tipo de viés.

7. Considerações Finais

Esta revisão sistemática da literatura revelou que não há uma solução única para o problema, uma vez que cada contexto apresenta particularidades e poucos atributos em comum, dificultando a comparação entre as técnicas.

A avaliação dos atributos de entrada e saída revelou que os grupos mais comuns de atributos são os dados educacionais, demográficos e comportamentais, ou uma combinação desses fatores. A revisão sistemática contribuiu para identificar lacunas de pesquisa e destacou a importância do processamento de grandes bases de dados, além de identificar a otimização de hiperparâmetros por meio de técnicas de inteligência de enxames como um nicho de pesquisa.

Como próximos passos, sugere-se o desenvolvimento de um sistema de informação que utilize modelos de *machine learning* e técnicas de inteligência de enxames para otimizar os hiperparâmetros desses modelos. Essa abordagem poderia aprimorar a compreensão da performance dos estudantes e possibilitar a identificação de estratégias para melhorar o desempenho acadêmico. A revisão sistemática teve um impacto relevante ao identificar lacunas na literatura e oferecer direcionamentos para futuras pesquisas nessa área.

Referências

AL-FAIROUZ, Ebtehal Ibrahim; AL-HAGERY, Mohammed Abdullah. **The most efficient classifiers for the students' academic dataset**. Int. J. Adv. Comput. Sci. Appl, v. 11, n. 9, p. 501-506, 2020.

ALAMRI, Rahaf; ALHARBI, Basma. **Explainable student performance prediction models: a systematic review**. IEEE Access, 2021.

ALPAYDIN, Ethem. **Introduction to machine learning**. MIT press, 2020.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Brasil no Pisa 2018** [recurso eletrônico]. – Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2020.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. **Mineração de dados educacionais: Oportunidades para o Brasil**. Revista Brasileira de informática na educação, v. 19, n. 02, p. 03, 2011.

CAGLIERO, Luca et al. **Predicting student academic performance by means of associative classification**. Applied Sciences, v. 11, n. 4, p. 1420, 2021.

CECHINEL, Cristian et al. **Mapping learning analytics initiatives in Latin America**. British Journal of Educational Technology, v. 51, n. 4, p. 892-914, 2020.

CORTEZ, P.; SILVA, A. M. G. **Using data mining to predict secondary school student performance**. EUROSIS-ETI, 2008.

FERRARI, Daniel Gomes; SILVA, Leandro Nunes De Castro. **Introdução a mineração de dados**. Saraiva Educação SA, 2017.

DIEN, Tran Thanh et al. **Deep learning with data transformation and factor analysis for student performance prediction**. Int. J. Adv. Comput. Sci. Appl, v. 11, n. 8, p. 711-721, 2020.

DYBA, Tore; DINGSOYR, Torgeir; HANSSEN, Geir K. **Applying systematic reviews to diverse study types: An experience report**. In: First international symposium on empirical software engineering and measurement (ESEM 2007). IEEE, 2007. p. 225-234.

KITCHENHAM, Barbara; CHARTERS, Stuart. **Guidelines for performing systematic literature reviews in software engineering**. 2007.

MENGASH, H. A. **Using data mining techniques to predict student performance to support decision making in university admission systems**. IEEE Access, IEEE, v. 8, p. 55462–55470, 2020.

MOHAMAD, Siti Khadijah; TASIR, Zaidatun. **Educational data mining: A review**. Procedia-Social and Behavioral Sciences, v. 97, p. 320-324, 2013.

NAKAGAWA, E. Y. et al. **Revisão sistemática da literatura em engenharia de software: teoria e prática**. Elsevier Brasil, 2017

QUEIROGA, Emanuel Marques et al. **A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course**. Applied Sciences, v. 10, n. 11, p. 3998, 2020.

QUEIROGA, Emanuel M. et al. **Experimenting Learning Analytics and Educational Data Mining in different educational contexts and levels**. In: 2022 XVII Latin American Conference on Learning Technologies (LACLO). IEEE, 2022. p. 1-9.

ROMERO, Cristobal; VENTURA, Sebastian. **Educational data mining: A survey from 1995 to 2005**. Expert systems with applications, v. 33, n. 1, p. 135-146, 2007.

SHAHIRI, A. M.; HUSAIN, W. et al. **A review on predicting student's performance using data mining techniques**. Procedia Computer Science, Elsevier, v. 72, p. 414–422, 2015.

TORRES MARQUES, L..; TORRES MARQUES, B..; MORAIS SILVA, C. A.. A **Descoberta das Causas da Retenção Acadêmica Utilizando Mineração de Dados: Uma Revisão Sistemática da Literatura**. Revista Novas Tecnologias na Educação, Porto Alegre, v. 20, n. 1, p. 263–272, 2022.

WILKER PEREIRA LUZ, J.; JUSSARA HEPP REHFELDT, M.; CLAUDETE SCHORR, M. **Revisão sistemática da literatura sobre o uso de learning analytics no ensino de programação**. Revista Novas Tecnologias na Educação, Porto Alegre, v. 19, n. 2, p. 203–212, 2021.