Análise comparativa de algoritmos de *machine learning* para prever a evasão escolar: Uma revisão sistemática da literatura

João A. S. Lima, PPGEC - Universidade de Pernambuco,
jasl@ecomp.poli.br, https://orcid.org/0000-0001-5010-0024
Roberta A. A. Fagundes, Universidade de Pernambuco, roberta.fagundes@upe.br
https://orcid.org/0000-0002-7172-4183

Resumo: A evasão dos alunos das instituições de ensino são questões que merecem atenção, principalmente nos países em desenvolvimento que foi agravado, por toda mudança econômica, política e educacional. Para investigar este evento, este artigo tem o objetivo de apresentar os resultados de uma Revisão Sistemática da Literatura (RSL) identificando as técnicas de Machine Learning (ML), especialmente os modelos de previsão, para prever a evasão escolar dos alunos no ensino médio utilizando os dados do Inep do ano de 2020 das escolas brasileiras. Para avaliar os principais modelos, foi utilizada a métrica de Erro Médio Absoluto (EMA). Como resultado se obteve entre os principais modelos utilizados, sendo: a regressão linear (EMA - 7,321462), árvore de decisão (EMA - 7,0665218) e regressão robusta (EMA - 6,785051), foi comprovado estatisticamente com o teste de hipótese *t-student*.

Palavras-chave: Evasão Escolar; Machine Learning; Mineração de Dados; Predição

Comparative analysis of regression models applied to school dropout: A systematic review of the literature

Abstract: The evasion of students from educational institutions are issues that deserve attention, especially in developing countries, which has been aggravated by all economic, political and educational changes. To investigate this event, this article aims to present the results of a Systematic Literature Review (SLR) identifying Machine Learning (ML) techniques, especially prediction models, to predict school dropout of students in high school using Inep data for the year 2020 for Brazilian schools. To evaluate the main models, the Mean Absolute Error (EMA) metric was used. As a result, it was obtained among the main models used, namely: linear regression (EMA - 7.321462), Decision Tree (EMA - 7.0665218) and robust regression (EMA - 6, 785051), it was statistically proven with the t-student test.

Keywords: Dropout School; Machine Learning; Data Mining; Prediction

1. Introdução

Uma instituição de ensino (IE) precisa de ter foco em algumas questões básicas que são: ter alunos devidamente matriculados e bom desempenho acadêmico, no intuito de influenciar no desenvolvimento destes alunos que são elementos chave na sociedade. Freitas (2011) entende o papel da escola como um agente de transformação, que possibilita o desenvolvimento de determinadas habilidades intelectuais fortalecendo sua capacidade de reflexão, pensamento, análise, síntese, criação, classificação,

argumentações dentre outros. Ao contextualizar a evasão deve-se observar que este termo permite diversas interpretações e em alguns casos pode ser compreendido como a desistência do estudante do curso, independentemente do número de participações deste na instituição (FERREIRA, 2014). No entanto, o abandono por parte destes alunos as IE podem ter fatores, em muitos casos, e passa a sofrer influência a partir do momento que ocorre o ingresso deste indivíduo na referida instituição, observa-se três comportamentos: a permanência (percurso), a conclusão (sucesso) ou a desistência (insucesso) (INEP, 2017, p.9). No entanto, podemos afirmar que essa desistência pode ser caracterizada como evasão escolar, ou seja, a saída antecipada do indivíduo, caracterizando o insucesso em relação ao objetivo de ampliar o conhecimento e o desenvolvimento cognitivo, além de habilidades e competências esperadas para o nível de ensino.

É neste contexto de mudanças que a utilização das ferramentas ligadas às tecnologias da informação e comunicação (TICs) passa a ser elemento crucial na implementação, acompanhamento, utilização e interação entre docentes e discentes e na formulação de planos que atendam a real necessidade do usuário. No mesmo contexto, a mineração de dados permite manipular um volume massivo de dados, oferecendo a possibilidade de extrair diversas informações complexas ou localizar regras valiosas na proposta de estudo. A evasão se dá por vários motivos, tais como: problemas familiares, oferta de vagas, distância das unidades institucionais, problema de assimilação do conteúdo, falta de interesse na disciplina, entre outros. Tanto que Santos e Perry (2023) trazem à tona o suporte por meio da tecnologia como apoio tanto à instituição quanto ao aprendizado do aluno, cabendo assim uma maior integração entre ambos.

Entender os fatores que apresentam forte ligação com a evasão dos alunos à luz das técnicas de *machine learning* fomenta discussões capazes de mudar uma cultura educacional. Como proposta, este artigo tem como objetivo verificar quais as técnicas de mineração de dados que utilizam técnicas de ML para identificar os índices de evasão dos discentes das instituições de ensino, bem como, de mensurar essa evasão escolar comparando modelos de previsão/predição mais eficiente levando em consideração as características físicas das instituições de ensino pública, privada e rural. Este documento foi dividido em três momentos, sendo: a fundamentação teórica, trazendo as considerações a respeito das pesquisas realizadas no tema; a metodologia, que traz como foi realizada a condução da RSL, análises, utilização e avaliação dos resultados; e, finalizando com as conclusões dos autores.

Do mesmo modo, foi necessário mensurar essa evasão escolar através de técnicas de *machine learning* comparando modelos de previsão/predição levando em consideração as características físicas das instituições de ensino pública, privada e rural. Dessa maneira, a escolha do estudo de base de dados escolares, que antes apenas era visto como um repositório, passou a ser valioso, transformando em informações úteis e gerando uma vasta gama de possibilidade para a aplicação de mineração de dados educacionais, o que contribui não somente para o entendimento teórico, mas apoiando a adoção de ações práticas de combate à evasão por parte das instituições de ensino (pública, privada e rural).

2. Fundamentação teórica

Sabe-se que o fechamento prolongado das instituições de ensino na pandemia da Covid-19 promoveu mudanças profundas na percepção (POTRA et al., 2021). Do mesmo

modo, pesquisas como a de Potra et al. (2021), Tamada, Netto e Lima (2019) visam entender este novo cenário no intuito de conseguir gerir de forma eficiente o processo de aprendizagem destes alunos. A pandemia da COVID-19 provocou o afastamento de discentes e docentes do ensino presencial e, com isso, o uso das Tecnologias da Informação e Comunicação (TIC) tem aumentado (VILLEGAS-CH et al., 2021). Falar sobre evasão é olhar para uma discussão que trata do fracasso escolar dentro do contexto educacional. Silva Filho e Araújo (2017) citam que "o Brasil tem a terceira maior taxa de abandono escolar entre os 100 países com maior IDH e no PNUD e a menor média de anos de estudo entre os países da América do Sul" (p.36).

O problema da evasão escolar pode ser representado de várias formas, desde a saída do discente da instituição de ensino ou, até mesmo, dos fatores que influenciam no processo como estar alheio ao que ocorre na aula, estar desatento, chegar atrasado ou ir embora antes do término do horário previsto. A evasão é um fenômeno que atinge o mundo todo, e está presente em países como Japão (OZAWA; HIRATA, 2019), Hungria (PUSZTAI; KOCSIS, 2019), Romênia (POTRA et al., 2021), Bangladesh (AHMED; KHAN, 2019), Chile (BELLO et al., 2020), México (VIZCAINO, et al., 2020), entre outros países. Ainda, é válido ressaltar que esse fenômeno atinge todos os níveis da educação. Pesquisas recentes apontam evasão na educação básica, no ensino médio (VIZCAINO, et al., 2020) e no ensino superior (BELLO et al., 2020), demonstrando a relevância de estudos e novas metodologias que diminuam esse fenômeno. Assim se faz necessário manter esforços e pesquisas com relação *Data Mining* (DM), por apresentarem técnicas de mineração de dados que visem analisar padrões de dados no âmbito educacional (MAI, et al. 2019; NESPEREIRA; VILAS; REDONDO, 2015). Visão esta complementada nas palavras de Kostopoulos et al. (2017) ao afirmar que a DM é uma ferramenta adequada para analisar e prever o desempenho do aluno no contexto acadêmico, por se valer das técnicas utilizadas em mineração de dados. Ressaltando que Silva, Souza e Fagundes (2020) cita que frente a massiva geração de dados o campo educacional passou a incorporar análises mais estratégicas para tentar compreender e trazer eficácia na resolução dos problemas educacionais. Visão complementada por Marques Carvalho da Silva e Imran (2015) por entender que a questão chave está na extração de informações significativas que possam auxiliar no entendimento destes fatores como também no desempenho dos alunos na instituição.

Nota-se que na área educacional existe o interesse em compreender as variáveis que envolvem o problema da evasão, estando ela no contexto educacional ou não. Nas palavras de Silva, Souza e Fagundes (2020) como de Batista e Fagundes (2023) esta abordagem tem sua discussão acadêmica de forma segregada em áreas afins, como: *Data Mining* (DM) e *Machine Learning* (ML). Na mesma premissa Tamada, Netto e Lima (2019), Noetzold e Pertile (2021) versam sobre a aplicação de DM e ML utilizando algoritmos estatísticos como suporte para o entendimento deste processo. Ainda que a proposta educacional passe a apresentar possibilidades de aprendizado presencial, virtual ou mista vão existir fatores que precisam ser revistos, principalmente no que diz respeito a interação entre professores e alunos, entre os próprios discentes e o desenvolvimento de um modelo que consiga atender a real necessidade do aluno no processo de ensino-aprendizagem (SÁIZ-MANZANARES, M.C. et al., 2021). Observando a relevância do tema, este documento procurou evidenciar os principais estudos com relação a evasão, suas técnicas e principais percepções dos autores no contexto educacional que possam servir de base para uma melhor compreensão do assunto e promover discussões a ponto de melhorar a gestão dos dados educacionais e fomentar propostas para diminuir este

movimento. Assim, este trabalho diferencia-se dos demais, pois se valeu da RSL como *insigths* para realização experimental, identificando os principais fatores que impactam a evasão dos alunos das escolas no ensino médio e quantificar tal movimento. Para tal, foi utilizado um protocolo para execução RSL, como metodologia de pesquisa, com o objetivo de identificar os fatores que influenciam a evasão dos alunos e de quantificar através do uso de algoritmos/técnicas de *machine learning* para avaliar/quantificar a evasão escolar. Por fim, foi possível realizar a RSL como *input* para a realização de experimentos, como também, a detecção dos fatores que influenciam através do uso de técnica de correlação.

3. Metodologia

A RSL é o procedimento onde se busca sistematizar e organizar evidências de pesquisa (NAKAGAWA et al., 2017). Neste caso a RSL tem como objetivo identificar pesquisas relevantes disponíveis em um determinado campo de estudo. Para aprofundamento dessa discussão e desenvolvimento deste estudo, foi realizada uma revisão sistemática da literatura por meio de um levantamento bibliográfico de busca automática a partir do banco de dados da ACM, Scopus, IEEE e MDPI, onde as buscas foram realizadas no período de 20 de abril de 2021 a 02 de fevereiro de 2022.

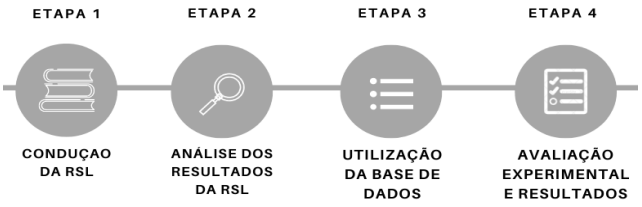

Figura 1. Etapas da revisão sistemática de literatura

3.1 Condução da Revisão Sistemática da Literatura

Nessa foram definidas as etapas iniciais para condução da RSL descrevendo as perguntas de pesquisa, *string* busca e engenhos de busca. Para esta pesquisa foram desenvolvidas três perguntas de pesquisa, tendo a pergunta geral (**PP**): Como a mineração de dados é utilizada para avaliar a evasão de alunos das escolas no ensino médio? Como pergunta secundária (PS), foram desenvolvidas: **PS1**: Como a mineração de dados auxilia a identificar os fatores que influenciam a evasão dos alunos das escolas no ensino médio? **PS2**: Quais algoritmos/técnicas de *machine learning* são mais utilizadas para identificar a evasão dos alunos das escolas no ensino médio?

Como estratégia de busca, os descritores (*strings* de busca) utilizados para esta pesquisa foram: *university educational*, *data mining*, *machine learning*, *university dropout* e *dropout rate*. Esses descritores foram combinados e acrescidos do operador boleando “AND” e/ou “OR” no *abstract* com intuito de encontrar estudos relevantes relacionados com a pergunta da pesquisa.

Para os artigos estarem incluídos na análise, foram considerados seis critérios de inclusão **(CI)** para garantir que a pesquisa está em conformidade com o tema proposto.

Os critérios estão descritos a seguir: **(CI1)** Artigo publicado em português ou inglês; **(CI2)** Artigo publicado entre 01/01/2011 a 31/12/2021; **(CI3)** Artigo que aborda sobre mineração de dados para gestão dos dados educacionais dos alunos; **(CI4)** Estudos primários que respondam a no mínimo uma questão de pesquisa; **(CI5)** Artigo que oferece dados sobre a evasão de alunos; e, **(CI6)** Artigo que aborda sobre o estudante no ensino médio. Durante a leitura do abstract também foram aplicados os critérios de exclusão **(CE)**. Assim, os artigos que foram excluídos apresentavam um ou mais critérios de exclusão e/ou apresentavam informações repetidas que foram encontradas em artigos mais recentes. Os critérios de exclusão estão descritos a seguir: **(CE1)** Artigo publicado em língua estrangeira que não fosse inglês; **(CE2)** Artigo publicado antes ou depois de 2021; **(CE3)** Resumos de artigos primários; **(CE4)** Artigo duplicados; e, **(CE5)** Artigo que sejam de acesso aberto. Os artigos selecionados foram lidos na íntegra e, em seguida, foram aplicados os critérios de qualidade com finalidade de aumentar a confiabilidade na pesquisa, foram apresentados na **Tabela 1,** a quantidade de artigos para cada critério de qualidade definido na pesquisa.

Critério de qualidade	Relatório	Rigor	Credibilidade	Relevância
CQ1: Pesquisa se relaciona com a aplicação de técnicas de machine learning para prever a evasão dos alunos?	51			
CQ2: Os objetivos se apresentam de forma clara?	62			
CQ3: Existe descrição detalhada do contexto que a pesquisa foi desenvolvida?	57			
CQ4: Houve análise criteriosa dos dados apresentados no documento?		62		
CQ5: O tipo de pesquisa conduzida foi claramente expressa?		58		
CQ6: Os resultados são claramente descritos?			31	
CQ7: A contribuição está claramente expressa?			62	
CQ8: A pesquisa deixa claro a quem contribui?				62

Tabela 1. Critério de qualidade

A **Tabela 1** mostra que 50% (31) dos artigos resultantes do estudo primário da RSL não apresentam a descrição de como os resultados foram obtidos, dificultando a credibilidade dos parâmetros dos métodos utilizados nesses artigos (CQ6). Por outro lado, 82% dos textos pesquisados deixam explicitadas as técnicas utilizadas (CQ1 - 51), apresentam a descrição detalhada do contexto (CQ 3 - 57) e a forma como foi conduzida a pesquisa (CQ8 - 58). Do mesmo modo, 100% (62) artigos apresentarem de forma clara, seus objetivos (CQ2), critérios adotados para uso dos dados (CQ4), bem como as contribuições executadas (CQ7 e CQ8).

3.2 Análise dos Resultados da RSL

Na busca inicial foram encontrados 1424 artigos, sendo dividido nos seguintes engenhos com seus respectivos percentuais: ACM (766 - 53,79 %), Scopus (10 - 0,70%), IEEE Xplore (610 - 42,84%), MDPI (38 - 2,67%). Após a leitura dos abstracts, os artigos foram submetidos aos critérios de inclusão e exclusão, resultando em 216 artigos (duzentos e dezesseis) para leitura na íntegra. Posteriormente, durante a leitura dos artigos, eles foram submetidos aos critérios de qualidade, resultando em **62 artigos** (sessenta e dois), que foram utilizados na composição do corpus de análise. Após a aplicação dos Critérios de Qualidade, temos: ACM (18 - 29,03 %), Scopus (7 - 11,29%), IEEE Xplore (30 - 48,39%), MDPI (7 - 11,29%). O engenho do IEEE Xplore concentra a maior quantidade de artigos relevantes para esta pesquisa.

Após a leitura completa dos 62 artigos foi respondida a pergunta principal. **PP**: **Como a mineração de dados é utilizada para avaliar a evasão de alunos das escolas no ensino médio**? Em função do volume de dados se faz necessário utilizar técnicas que possam ajudar na análise e encontrar os fatores que mais influenciam nos problemas educacionais. Daí a necessidade de se valer de técnicas de *machine learning* que estejam estreitamente relacionadas, como a correlação por resumir o grau de relacionamento entre duas variáveis, sendo X e Y; e, regressão, tendo por resultado a equação matemática capaz de descrever o relacionamento entre estas variáveis, neste caso da evasão com as outras existentes na base de dados.

PS1: Como a mineração de dados auxilia a identificar os fatores que influenciam a evasão dos alunos das escolas no ensino médio? Tamada, Netto e Lima (2019) entendem que o DM ganha destaque por conseguir se valer da Mineração de Dados, Machine Learning e algoritmos estatísticos com foco na melhoria do processo de ensino e aprendizagem. E tal aplicação já é algo corriqueiro para os pesquisadores e estudiosos deste campo de pesquisa. Estudos como os de Dharmawan, Ginardi e Munif (2018), Ahmed e Khan (2019) e Palacios et al. (2021) apontam que uma análise preliminar dos dados e conseguem mostrar não apenas as tendências ou possibilidades da evasão, mas informam ainda os principais motivos que levam a reprovação ou ao baixo aproveitamento do ensino. Fatores como falta de acessos aos dados, falta de profissionais capacitados para análises e ausência de colaboração no manuseio em torno dos dados dificultam as pesquisas, análises e o desenvolvimento de meios para consolidar aquelas que já estão em execução (GKONTZIS et al., 2018). Daí a necessidade em considerar os fatores acadêmicos e não acadêmicos para se ter um panorama que consiga evidenciar os reais fatores para o abandono (DHARMAWAN, GINARDI e MUNIF, 2018).

PS2: Quais algoritmos/técnicas de *machine learning* são mais utilizadas para identificar a evasão dos alunos das escolas no ensino médio? Foi identificado o uso de **97 técnicas** no processo todo. Desta forma identificou-se a preponderância de algumas técnicas em função dos algoritmos, sendo listada as principais técnicas de *machine learning* utilizadas, sendo agrupadas nas seguintes técnicas: Previsão/Predição (38) - 39,18%, Árvore de Decisão (28) - 28,87%, Rede Neural (16) - 16,49%, *Support Vector Machine* (11) - 11,34%, *KNN* (4) - 4,12%. Percebe-se que as técnicas de Previsão/Predição e Árvore de Decisão foram as mais mencionadas nos artigos com 68,05% do total das técnicas mencionadas para o problema de evasão, elicitando que uma das técnicas pode ser considerada relevante para prever a evasão. Assim, uma avaliação experimental foi realizada baseada nesses conjuntos de técnicas na **Etapa 3.4**.

3.3 Utilização da Base de Dados

Para este estudo foram utilizados os dados escolares de alunos do ensino médio disponíveis no Inep do ano de 2020 (INEP, 2020) das escolas brasileiras aplicando as técnicas mais citadas pela RSL na **Etapa 3.2** que responde a **PP**. Inicialmente será feita a pesquisa nos engenhos de pesquisa com base nos principais termos que tratam do assunto, no intuito de realizar uma busca ampla de documentos que abordam tal temática. Posteriormente, será realizada a leitura do resumo para identificação inicial dos assuntos relacionados, para que em seguida seja feita a leitura integral do documento e com isso quantificar as técnicas e abordar os fatores que são usados.

Os dados estão presentes em duas bases de pesquisa disponíveis pelo INEP para apresentar os dados com relação à estrutura das escolas e das áreas disponíveis à pesquisa e, em outra, apresentar os índices de evasão e rendimento das escolas disponibilizados em suas séries. E para tratar dos assuntos relacionados à evasão foi necessário correlacionar os dados presentes nos documentos com a taxa de evasão, originando o documento alvo desta pesquisa. Com base no pré-processamento, a base de dados passou a ter 20 colunas, o que antes estavam divididos em 19 colunas em um documento e 1 coluna em outro, e tendo 1651 linhas ao todo presentes em ambos os documentos. As características dessas colunas estão relacionadas à estrutura física das escolas públicas, privadas e rurais do ensino médio do Brasil. Com isso, foi analisado quais desses fatores foram mais influentes para explicar a Taxa de Abandono (aqui chamado de evasão escolar). Para utilização das técnicas de Previsão/Predição e Árvore de Decisão foram definidas as seguintes características de variável resposta (dependente) e variáveis explicativas (independentes) para utilização. Assim, foram calculados os valores da correlação entre essas variáveis, como variável resposta (Y) foi utilizada Taxa de Abandono em relação variáveis independentes (X), observou-se que algumas colunas se destacam com correlação mais altas para essas técnicas: aquelas que trazem a informação da dependência seja de âmbito federal, estadual, municipal ou privada (0.2901); Se possuem distribuição das séries do processo educacional na instituição (0.1139); Se possuem órgão de conselho escolar na instituição (-0.2328); Se são divididas entre urbanas e rurais (-0.1655); Se possuem espaços para atividades escolares (-0.1744), quadra de esportes (-0.1801) e laboratório de ciências (-0.1070); Se disponibiliza material pedagógico científico (-0.1135); Se disponibilizam atendimento em caráter especial ao discente (-0.1374) e, se dispõe de acesso a internet nas salas (-0.1137).

3.4 Avaliação Experimental e Resultados

Neste ponto é de vital importância para que seja mantido o foco nas técnicas de maior impacto na pesquisa, como resultado da RSL. Aqui ficou evidenciado as técnicas de Previsão/Predição, sendo: Regressão Linear e Regressão Robusta e Árvore de Decisão. Nessa etapa de modelagem foi dividida a base de dados em treino e teste, sendo na proporção de 70% e 30%. Para tal, utilizou-se técnicas de Previsão/Predição: regressão linear, linear robusta, como também, técnica de árvore de decisão para verificar a comparação dos resultados obtidos na Base de Dados do Inep (INEP, 2020). Os resultados serão apresentados com base na aplicação das técnicas mais relevantes mencionadas nesta RSL. Assim, o experimento foi utilizado por meio da simulação de Monte Carlo (MC) com 600 interações. Desta forma, os dados a mostram os resultados das simulações de MC para de três principais técnicas de *machine learning* utilizando os valores de erro

médio absoluto e o desvio padrão entre parênteses, são elas: Regressão Linear (7,3214 - (0,0107)), **Regressão Robusta** (**6,7850** - (**0,0065**)) e Árvore de Decisão (7,0652 - (7,0652)).

Dentre as técnicas utilizadas, a regressão linear e árvore de decisão analisa apenas relações lineares entre variáveis dependentes e independentes. Isto é, pressupõe que existe uma relação direta entre eles, ratificando que essa relação não foi modelada tão bem quanto na regressão robusta, entre a Taxa de Abandono e as características das instituições de ensino avaliadas. Por sua vez, a técnica de regressão robusta encontrou o modelo que se ajusta melhor à maioria dos dados. Pois, essa técnica, além de uma relação linear entre variáveis, também leva em consideração os pontos aberrantes, também conhecidos como *outliers*. Para comprovar estatisticamente esse melhor desempenho foi realizado teste *t-student*, o que ratificou o melhor desempenho dela com 95% de confiança.

4. Conclusões

Entende-se que o objetivo de condução da RSL foi alcançado, por conseguir apresentar documentos capazes de reunir os principais estudos na temática e por elucidar as principais técnicas *machine learning* relacionadas para a problemática da evasão escolar. Porém, a apresentação dos resultados de forma clara e objetiva foi um dificultador na obtenção de informações para a construção deste artigo o que, leva, a uma nova possibilidade de pesquisa no tocante aos resultados presentes na RSL.

Mas, ainda assim, foi possível traçar um planejamento condizente no intuito de identificar os principais fatores, como elucidado neste documento, reduzindo neste caso, em 50%, sendo sinalizado apenas as 10 variáveis com maior influência com relação a taxa de evasão e com isso dar condições aos gestores educacionais de estabelecer os meios pelos quais se deseja alcançar na tratativa da redução da evasão. Assim, foi possível observar que a previsão/predição, como também árvore de decisão foram as mais mencionadas para o problema em estudo. Como também, verificou-se que a mineração de dados é utilizada para mensurar o estudo em questão através de algoritmos/técnicas de *machine learning*. Vale ressaltar que as TICs são ferramentas que permitem a potencialização da aprendizagem dos alunos, mas que não podem ser vistas como garantias de eficiência. E dentro do que foi pesquisado, identificou-se que fatores físicos podem influenciar, como: Ser unidade federal, estadual, municipal ou privada; Localizada na área urbana ou rural; Possuir ou não laboratório de ciências, quadra de esportes, espaços para atividades externas e acesso à internet; Ter um acompanhamento na escola por um órgão escolar e de possuir ou não um atendimento educacional especializado podem influenciar na vida dos discentes no seu processo de evasão, na qual investiga-se padrões de respostas obtidas dos alunos no contexto escolar. Além disso, Xiao e Yi (2020) sugerem a inclusão de variáveis relacionadas a aspectos comportamentais para aumentar a precisão dos modelos de previsão.

Desta forma ao identificar as variáveis no risco de evasão, de forma direta ou indiretamente, faz com que seja compreendido o perfil deste movimento promovendo estratégias de mitigação deste risco, sendo que o acompanhamento, período a período, pode trazer maior efetividade para o entendimento, visto que o movimento de evasão é dinâmico ao longo do tempo. Para ratificar a influências desses fatores foram utilizadas

três técnicas de *machine learning* de Previsão/Predição (regressão linear, regressão linear robusta), como também, árvore de decisão aplicada no contexto de evasão das escolas do ensino médio do ano de 2020 identificando aquele que possui melhor adequacidade aos dados. Concluiu-se e foi ratificado estatisticamente que o modelo de regressão robusta possui melhor desempenho.

REFERÊNCIAS

AHMED, S. A.; KHAN, S. I.. **A machine learning approach to Predict the Engineering Students at risk of dropout and factors behind: Bangladesh Perspective**, 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1-6, doi: 10.1109/ICCCNT45670.2019.8944511.

BATISTA, M. R.; FAGUNDES, R. A. de A. **Mineração de dados educacionais aplicada a performance de estudantes: uma revisão sistemática da literatura**. Revista Novas Tecnologias na Educação, Porto Alegre, v. 21, n. 1, p. 271–280, 2023.

DHARMAWAN, T.;GINARDI, H.; MUNIF, A.; **Dropout Detection Using Non-Academic Data, 2018 4th International Conference on Science and Technology (ICST)**, 2018, pp. 1-4, doi: 10.1109/ICSTC.2018.8528619.

FERREIRA, Luiz Antonio Miguel. **Evasão escolar**. 2014.

GKONTZIS, Andreas F.; KOTSIANTIS, Sotiris; TSONI, Rozita; VERYKIOS, Vassilios S.. 2018. **An effective LA approach to predict student achievement. In Proceedings of the 22nd Pan-Hellenic Conference on Informatics (PCI '18)**. Association for Computing Machinery, New York, NY, USA, 76–81. https://doi.org/10.1145/3291533.3291551

INEP, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Diretoria de Estatísticas Educacionais (DEED). **Metodologia de Cálculo dos Indicadores de Fluxo da Educação Superior**. Brasília, Inep, 2017.

INEP, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Microdados**. 2020. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados. Acesso em: 10 abr. 2022.

KOSTOPOULOS, Georgios; KOTSIANTIS, Sotiris; RAGOS, Omiros; GRAPSA, Theodoula N.. Early dropout prediction in distance higher education using active learning. **2017 8Th International Conference On Information, Intelligence, Systems & Applications (Iisa)**, [S.L.], p. 1-6, ago. 2017. IEEE. http://dx.doi.org/10.1109/iisa.2017.8316424.

NESPEREIRA, Celia González; VILAS, Ana Fernández; REDONDO, Rebeca P. Díaz. Am I failing this course? **Proceedings Of The 3Rd International Conference On Technological Ecosystems For Enhancing Multiculturality - Teem '15**, [S.L.], p. 271-276, 2015. ACM Press. http://dx.doi.org/10.1145/2808580.2808621.

NOETZOLD, E.; DE L. PERTILE, S. **Análise e predição de evasão dos alunos de um curso de graduação em sistemas de Informação por meio da mineração de dados educacionais.** Revista Novas Tecnologias na Educação, Porto Alegre, v. 19, n. 1, p. 351–360, 2021.

MARQUES CARVALHO DA SILVA, J.; IMRAN, H. **Um estudo sobre as variáveis para predição de alunos não concluintes em cursos suportados por Ambientes Virtuais de Ensino e Aprendizagem**. Revista Novas Tecnologias na Educação, Porto Alegre, v. 13, n. 2, 2015.

POTRA, Sabina; PUGNA, Adrian; POP, Mădălin-Dorin; NEGREA, Romeo; DUNGAN, Luisa. Facing COVID-19 Challenges: 1st-year students⁚ experience with the romanian hybrid higher educational system. **International Journal Of Environmental Research And Public Health**, [S.L.], v. 18, n. 6, p. 3058, 16 mar. 2021. MDPI AG. http://dx.doi.org/10.3390/ijerph18063058.

SÁIZ-MANZANARES, María Consuelo; RODRÍGUEZ-DÍEZ, Juan José; DÍEZ-PASTOR, José Francisco; RODRÍGUEZ-ARRIBAS, Sandra; MARTICORENA-SÁNCHEZ, Raúl; JI, Yi Peng. Monitoring of Student Learning in Learning Management Systems: an application of educational data mining techniques. **Applied Sciences**, [S.L.], v. 11, n. 6, p. 2677-2693, 17 mar. 2021. MDPI AG. http://dx.doi.org/10.3390/app11062677

SANTOS, T. C. B. dos; PERRY, G. T. Revisão sistemática sobre painéis de visualização de dados educacionais. **Revista Novas Tecnologias na Educação**, Porto Alegre, v. 21, n. 1, p. 87–96, 2023. DOI: 10.22456/1679-1916.134328. Disponível em: https://seer.ufrgs.br/index.php/renote/article/view/134328. Acesso em: 14 set. 2023.

SILVA, Paulo; SOUZA, Fernando; FAGUNDES, Roberta. 2020. Approaches to Predicting Educational Problems: **A Systematic Mapping. In XVI Brazilian Symposium on Information Systems (SBSI'20)**. Association for Computing Machinery, New York, NY, USA, Article 46, 1–8. https://doi.org/10.1145/3411564.3411657

TAMADA, Mariela Mizota; NETTO, Jose Francisco de Magalhaes; LIMA, Dhanielly Paulina R. de. Predicting and Reducing Dropout in Virtual Learning using Machine Learning Techniques: a systematic review. **2019 Ieee Frontiers In Education Conference (Fie)**, [S.L.], p. 1-9, out. 2019. IEEE. http://dx.doi.org/10.1109/fie43999.2019.9028545.

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

XIAO, M.; YI, H. Research on Adaptive Learning Prediction Based on XAPI. **International Journal of Information and Education Technology**, [*s.l.*], v. 10, n. 9, p. 679-684, 2020. DOI: 10.1002/cae.22235.