Análise da construção de modelos preditivos sob a perspectiva de indicadores de evasão

Patricia Mariotto Mozzaquatro Chicon, UNICRUZ/UNIJUI, patriciamozzaquatro@gmail.com
Leo Natan Paschoal, ICMC–USP, paschoalln@usp.br
Fabricia Carneiro Roos Frantz, UNIJUÍ, frfrantz@unijui.edu.br
Rafael Zancan Frantz, UNIJUÍ, rzfrantz@unijui.edu.br
Sandro Sawicki, UNIJUÍ, sawicki@unijui.edu.br

Resumo: *A predição de evasão de alunos com base em modelos que utilizam dados oriundos de ambientes virtuais de aprendizagem é um tema de pesquisa que desperta interesse por auxiliar na gestão acadêmica. Foi constatado em um estudo anterior que três indicadores principais sinalizam a evasão: comportamento do aluno, desempenho do aluno e aspectos demográficos. Apesar de que os modelos de predição de evasão encontrados na literatura fazem uso de alguns dados que caracterizam esses indicadores, não está claro o quanto os indicadores orientam a construção dos modelos preditivos. Uma análise de modelos preditivos reportados em estudos primários foi realizada para identificar como os indicadores de desempenho, comportamental e demográfico são utilizados na construção dos modelos de predição de evasão. Os resultados obtidos revelam que a maioria dos modelos preditivos, independente do indicador de evasão, faz uso da mineração de dados. Adicionalmente, os indicadores de evasão abordados pelos modelos preditivos, implementam em sua maioria métodos de árvore de decisão. Por fim, a métrica mais usada para avaliar os modelos preditivos é a precisão/acurácia, também independente dos indicadores de evasão.*

Palavras-chave: *Evasão escolar, Indicadores de evasão, Modelos de predição.*

Analysis of the construction of predictive models from the perspective of evasion indicators

Abstract: *Predicting student dropout based on models that use data from virtual learning environments is a research topic that arouses interest in helping with academic management. It was found in a previous study that three main indicators signal evasion: student behavior, student performance and demographic aspects. Although the evasion prediction models found in the literature make use of some data that characterize these indicators, it is not clear how much the indicators guide the construction of predictive models. An analysis of predictive models reported in primary studies was carried out to identify how performance, behavioral and demographic indicators are used in the construction of evasion prediction models. The results obtained reveal that most predictive models, regardless of the evasion indicator, make use of data mining. Additionally, the evasion indicators addressed by the predictive models, implement mostly decision tree methods. Finally, the most used metric to evaluate predictive models is precision/accuracy, also independent of the evasion indicators.*

Keywords: *School dropout, Dropout indicators, Prediction models.*

1. Introdução

Em se tratando do contexto educacional, verifica-se que o índice de evasão é alto e preocupa os educadores, uma vez que a evasão não se apresenta como realidade apenas de uma modalidade de ensino e atinge, além da educação presencial, também a Educação a Distância

(EaD) e fere o princípio da educação como direito social (BAGGI; LOPES, 2011; BITTENCOURT; MERCADO, 2014; MARTINHO, 2014). A evasão é responsável por importantes perdas sociais, acadêmicas e econômicas (MACHADO *et al.*, 2015; SILVA *et al.*, 2015).

Descobrir as causas e soluções para a evasão tem sido objeto de estudo para os pesquisadores envolvidos com a educação presencial e a distância. Constitui-se um desafio para estes educadores desenvolver modelos que possibilitem prever o comportamento dos estudantes, de modo a propiciar aos professores e demais envolvidos a possibilidade de intervenção com o objetivo de resgatar o aluno antes que ele evada ou reprove (BAKER; ISOTANI; CARVALHO, 2011).

Modelos preditivos têm sido desenvolvidos com a finalidade de acompanhar as ações dos alunos nos Ambientes Virtuais de Aprendizagem (AVAs). Apesar da proposição de diversos modelos teóricos para compreender o comportamento dos estudantes e fenômenos decorrentes de determinadas práticas educativas, introduzidas e/ou apoiadas por meio dos AVAs, observa-se que a maioria das pesquisas foca na problemática de um contexto específico (KAMPFF *et al.*, 2014; QUEIROGA; CECHINEL; ARAÚJO, 2015; SILVA *et al.*, 2015).

Há estudos que analisam técnicas e métodos aplicados na predição de evasão, sem buscar uma compreensão de como os indicadores orientam a construção dos modelos preditivos (MASCHIO *et al.*, 2018; MARQUES *et al.*, 2019; COLPO *et al.*, 2020). Os estudos de Maschio *et al.* (2018), Marques *et al.* (2019), Mduma, Kalegele and Machuve (2019), Will, Kemczinski and Parpinelli (2019), Colpo *et al.* (2020) concentraram-se em revisar aspectos técnicos (*i.e.*, métodos, algoritmos e ferramentas), resultados e aplicações de Mineração de Dados Educacional (MDE), não se preocuparam portanto em analisar também a influência dos indicadores na seleção dos métodos ou algoritmos (uma das etapas da construção dos modelos preditivos). Os estudos secundários feitos até aqui não têm exteriorizado como os indicadores são abordados na construção dos modelos. Portanto, não há uma visão sistêmica se os indicadores de evasão influenciam em decisões tomadas na construção dos modelos.

Diante do exposto, este artigo tem como objetivo apresentar um entendimento sobre como os indicadores de evasão orientam a construção de modelos de predição que usam exclusivamente dados de AVAs. Para obter esse entendimento, uma análise foi realizada em modelos preditivos descritos que fazem uso de dados que têm origem nos AVAs, em estudos primários. Essa análise considerou indicadores de evasão reconhecidos recentemente por Chicon, Paschoal and Frantz (2020), são eles: demográfico [1], comportamental[2] e desempenho [3].

Para apresentar a análise realizada, o artigo está estruturado da seguinte forma. A Seção 2 apresenta os estudos anteriores que analisaram os modelos de predição de evasão. Na Seção 3, os materiais e métodos são apresentados. Na Seção 4, os resultados são analisados e discutidos. Por fim, na Seção 5, apresenta-se as principais conclusões.

2. Trabalhos relacionados

Alguns esforços têm sido realizados na tentativa de automatizar a análise da evasão. Em especial, esses empenhos utilizam a MDE e *learning analytics* (LA), abordagens que possi-

[1]Os indicadores demográficos são relacionados a fatores relativos a dados históricos como sexo, idade, estado civil, ano de graduação,status anterior de graduação, localização, profissional (HLIOUI; ALOUI; GARGOURI, 2018).

[2]Os indicadores comportamentais são relacionados a fatores relativos à aprendizagem ativa, motivação e envolvimento dos alunos, como acesso ao curso, aos materiais lidos e sua evolução com a programação do curso (MAZZA; DIMITROVA, 2007).

[3]Os indicadores de desempenho estão relacionados ao desempenho geral dos alunos, com base em seu desempenho nas atividades (*e.g.*, questionário, tarefas, fóruns de discussão, dentre outros) (MAZZA; DIMITROVA, 2007).

bilitam análise de dados educacionais com a intenção de compreender de forma mais eficaz o comportamento de alunos e outros fatores relacionados a sua aprendizagem. Ao passo que tais pesquisas emergem, estudos de revisão também surgem, na tentativa de sintetizar as contribuições estabelecidas, reunindo evidências sobre como a automatização de tais análises têm sido abordada pela comunidade. Esta seção reúne algumas dessas pesquisas de revisão e discute sobre como elas se diferem do estudo abordado ao longo deste artigo.

Maschio *et al.* (2018) apresentaram um mapeamento sistemático sobre o cenário brasileiro tendo como foco a MDE. Foram mapeadas técnicas, resultados e aplicações de MDE no contexto brasileiro. Os autores observaram que a maioria dos estudos primários têm focado na análise de desempenho dos alunos (*e.g.*, predição de alunos que tendem a reprovar). Também constataram que os estudos se inclinam a usar dados provenientes de sistemas educacionais e/ou AVAs para predizer esse desempenho. Outro aspecto contemplado pelo estudo secundário foi a identificação dos dados que são empregados durante o processo de mineração. Nesse sentido, descobriu-se que as análises de evasão tendem a ser projetadas com base em dados pessoais dos estudantes e pela quantidade de interação desses sujeitos.

Marques *et al.* (2019) realizaram um mapeamento sistemático sobre evasão escolar com o propósito de identificar ferramentas, técnicas e fatores indutores para evasão escolar. Com base nos resultados, os autores identificaram que a ferramenta *Waikato Environment for Knowledge Analysis* (Weka)[4] foi a mais citada para apoiar a MDE com propósitos de predição de evasão. Entre as técnicas de MDE, destacou-se a classificação. Em relação aos fatores envolvidos na evasão escolar, reconheceu-se que os estudos costumam abordar principalmente as características individuais dos alunos durante a predição.

O estudo de Colpo *et al.* (2020) apresenta uma revisão sistemática de estudos que utilizam técnicas de MDE no contexto da previsão de evasão. O estudo mapeou características contextuais (*e.g.*, modalidades e níveis de ensino), técnicas (*e.g.*, tarefas, métodos, algoritmos e ferramentas) e dados (*e.g.*, tipos, abrangência e volume) com o intuito de identificar como esse tema está sendo abordado no cenário de pesquisa brasileira. Como resultado, reconheceu-se que os estudos primários apontam uma maior exploração da evasão no nível de graduação, na modalidade presencial e na rede pública de ensino. Quanto aos tipos de dados, os estudos primários tendem a utilizar dados acadêmicos, sociais e econômicos dos estudantes. Considerando as técnicas empregadas, os trabalhos indicam a unanimidade da tarefa de classificação e predominância do algoritmo de árvores de decisão, geralmente aplicados com o auxílio da ferramenta Weka.

Campos *et al.* (2020) por meio de um mapeamento sistemático buscaram verificar as contribuições de LA e MDE no contexto brasileiro. Os autores constataram que há maior ocorrência de estudos com finalidade de analisar desempenho acadêmico e previsão de evasão escolar. A maioria dos estudos são voltados ao ensino superior e na modalidade de EaD. Há variedade de tecnologias e recursos utilizados no desenvolvimento de soluções, o uso das linguagens R e MySQL para soluções baseadas em LA, e a ferramenta Weka para modelos baseados em MDE.

Por último, destaca-se o estudo de Chicon, Paschoal and Frantz (2020), que teve como objetivo reconhecer na literatura os estudos de pesquisa sobre indicadores de evasão no contexto da EaD e a relação destes indicadores com recursos e dados de AVAs. Para tanto, os autores conduziram um mapeamento sistemático e obtiveram como resultado os

[4]Mais informações disponíveis em: <https://www.cs.waikato.ac.nz/ml/weka/>.

principais indicadores de evasão no contexto da EaD, são eles: indicador comportamental, desempenho e demográfico. Além disso, observaram que os dados coletados pelos AVAs e utilizados pelos indicadores, provêm de diferentes recursos, tais como envio de tarefas, participação em fórum, interações em chat, respostas e interações em questionários, dentre outros.

Frente ao que a comunidade tem investigado sobre modelos de predição de evasão, observa-se que existe na literatura um conjunto de estudos focados nas técnicas, métodos e ferramentas para a construção de modelos de predição de evasão. Apenas um artigo da literatura teve o cuidado ao observar os indicadores de evasão, sem considerar os modelos preditivos produzidos para contemplar esses indicadores. Assim, até o momento não foi possível observar esforços que mapeiam como os indicadores de evasão estão sendo considerados no desenvolvimento dos modelos preditivos. Isso indica, por exemplo, que não há um reconhecimento sistemático sobre como os indicadores de evasão guiam a construção dos modelos preditivos. Nessas perspectiva, questões como "o indicador abordado pelo modelo preditivo influência na escolha do algoritmo usado pelo modelo?" e "o indicador abordado pelo modelo preditivo influência no tipo de dado usado para treinar o modelo?" ainda estão em aberto e foram pouco discutidas pela comunidade.

3. Materiais e métodos

Para localizar e estudar os modelos preditivos de evasão, considerando os indicadores que possuem relação com AVAs, recuperou-se o conjunto de estudos primários identificados pelo estudo secundário de Chicon, Paschoal and Frantz (2020). Isso foi feito porque o estudo de Chicon, Paschoal and Frantz (2020) reconheceu os indicadores de evasão que utilizam dados registrados pelos AVAs, mas os autores não se atentaram aos modelos preditivos abordados. Além disso, tal estudo foi conduzido a partir de uma sólida busca em bases de dados internacionais e definição de critérios de apoio à seleção.

Tendo em vista os estudos primários, almejando apoiar a descoberta de como os indicadores são abordados na concepção dos modelos preditivos, questões de pesquisa (QP) foram pré-definidas. Ao total foram delineadas cinco questões, a saber:

QP1 – Qual a relação entre os indicadores de evasão com os tipos de abordagens utilizadas para gerar as predições?
QP2 – Qual a relação entre os indicadores de evasão com os tipos de dados contemplados pelos modelos de predição?
QP3 – Qual a relação entre os indicadores de evasão com os métodos e algoritmos utilizados pelos modelos de predição?
QP4 – Qual a relação entre os indicadores de evasão com as métricas adotadas para avaliar os modelos de predição?
QP5 – Os indicadores de evasão que guiaram a construção do modelo preditivo foram fundamentados por algum aporte teórico?

A partir da definição das questões de pesquisa, critérios para apoiar a seleção dos modelos preditivos foram estipulados. Dado que os modelos são reportados nos estudos primários, os critérios foram construídos para averiguar o conteúdo descrito ao longo de estudos primário. Em particular, estabeleceu-se que somente seriam incluídos estudos que descrevessem ao menos uma solução para predizer evasão, desde que a solução fosse baseada nos indicadores de evasão que persistem nos AVAs. O outro critério, de característica exclusiva, consiste em descartar estudos que não apresentam uma solução para predizer evasão.

Ao passo que a fonte de busca foi estabelecida (*i.e.*, 22 estudos primários reconhecidos no mapeamento sistemático apresentado por Chicon, Paschoal and Frantz (2020)), questões de pesquisa foram definidas e os critérios de seleção determinados, o processo de estudo e análise dos modelos de predição de evasão foi inicializado. Primeiramente, uma leitura completa nos 22 estudos primários foi realizada. Com base na leitura, os critérios foram sendo aplicados e o conjunto base de estudos que aborda modelos de predição de evasão foi identificado, todos listados na Tabela 1. A partir disso, dados que auxiliam na obtenção de respostas para as questões de pesquisa foram extraídos. Os resultados obtidos com esse processo são apresentados ao longo da próxima seção.

Tabela 1. Lista dos estudos primários que abordam soluções preditivas de evasão

Título do estudo primário	ID do estudo
An Infographics-based Tool for Monitoring Dropout Risk on Distance Learning in Higher Education	[e01]
Classification and predictive analysis of educational data to improve the quality of distance learning courses	[e02]
Co-embeddings for Student Modeling in Virtual Learning Environments	[e03]
Identificação de Perfis de Evasão e Mau Desempenho para Geração de Alertas num Contexto de Educação a Distância	[e04]
Learning Analytics in Practice: Providing E-Learning Researchers and Practitioners with Activity Data	[e05]
Learning to Identify At-Risk Students in Distance Education Using Interaction Counts	[e06]
Multi-agent System Based on Fuzzy Logic for Elearning Collaborative System	[e07]
OULAD MOOC Dropout and Result Prediction using Ensemble, Deep Learning and Regression Techniques	[e08]
Predicting Students Success in Blended Learning—Evaluating Different Interactions Inside Learning Management Systems	[e09]
Predicting students' final performance from participation in on-line discussion forums	[e10]
Students' Success Predictive Models Based on Selected Input Parameters Set	[e11]
Um estudo do uso de contagem de interações semanais para predição precoce de evasão em educação a distância	[e12]
Um modelo preditivo para diagnóstico de evasão baseado nas interações de alunos em fóruns de discussão	[e13]
Uma Abordagem Genérica de Identificação Precoce de Estudantes com Risco de Evasão em um AVA utilizando Técnicas de Mineração de Dados	[e14]
Understanding Learner Engagement in a Virtual Learning Environment	[e15]
Using data mining as a strategy for assessing asynchronous discussion forums	[e16]

4. Resultados

Esta seção foi estruturada de modo a responder cada questão de pesquisa do estudo. As respostas foram organizadas a partir de dados coletados nos estudos que estão listados na Tabela 1. Vale salientar que para responder cada questão de pesquisa, indicou-se na apresentação dos resultados o identificador (ID) do estudo que aborda o modelo analisado.

QP1 - Qual a relação entre os indicadores de evasão com os tipos de abordagens utilizadas para gerar as predições?

Para responder a primeira questão de pesquisa, os modelos foram analisados individualmente e classificados de acordo com a abordagem descrita pelos estudos. As abordagens que têm sido utilizadas para predizer a evasão estão listadas na Tabela 2. Destaca-se que a MDE foi utilizada para construir 14 modelos que contemplam indicadores de desempenho, indicador comportamental e indicador demográfico. Observou-se que quando os modelos de predição são construídos para contemplar o indicador demográfico apenas a abordagem da MDE foi utilizada. A abordagem LA foi utilizada em menor proporção pelos modelos. Quando essa abordagem é utilizada na construção de um modelo, os indicadores de desempenho e comportamental são contemplados.

Tabela 2. Relação entre os indicadores de evasão e as abordagens usadas na construção dos modelos preditivos

Indicador	Abordagem	
	Mineração de dados	Learning analytics
Comportamental	[e02],[e04],[e06],[e07],[e08],[e09],[e10],[e11],[e12],[e13],[e15],[e16]	[e01], [e05]
Demográfico	[e04],[e11],[e15]	
Desempenho	[e03], [e08],[e09],[e10],[e11],[e12],[e13],[e14],[e15],[e16]	[e01], [e05]

QP2 - Qual a relação entre os indicadores de evasão com os tipos de dados contemplados pelo modelo de predição?

Para identificar a relação entre os indicadores e os tipos de dados, a variância dos dados[5] foi considerada. Um esquema de classificação foi organizado, neste caso por indicador- tipo de dado–variância (Tabela 3). Pode-se observar que a maioria dos modelos identificados utilizaram dados relacionados ao indicador comportamental, que faz uso de dados de interação e participação dos alunos. Em relação ao indicador comportamental, os dados foram classificados como variantes no tempo pela maioria dos modelos, reportados em 14 estudos que descrevem tais modelos. O indicador de desempenho, por sua vez, utilizou dados de notas obtidas pelos alunos em tarefas disponibilizadas nos AVAs. Observou-se que a maioria dos modelos que abordam o indicador de desempenho foi concebido com dados que tendem a sofrer variações no tempo, uma vez que tais informações foram reconhecidas na descrição de 12 estudos. Finalmente, o indicador demográfico utiliza dados históricos classificados como variantes e invariantes. Ele aparece em menor proporção nos modelos apresentados pelos estudos primários analisados. Portanto, entende-se que os modelos que são construídos para abordar indicadores comportamentais e de desempenho tendem a utilizar dados que sofrem variação no tempo.

Tabela 3. Relação entre os indicadores de evasão e a variância dos dados

Indicador	Tipos de dados	Variância dos dados	
		Variantes	Invariantes
Comportamental	Interação e participação	[e01],[e02],[e04],[e05],[e06],[e07],[e08],[e09],[e010],[e11],[e12],[e13],[e15],[e16]	[e11],[e15]
Demográfico	Dados históricos	[e11],[e15]	[e04],[e11],[e15]
Desempenho	Nota	[e01],[e03],[e05],[e08],[e09],[e10],[e11],[e12],[e13],[e14],[e15],[e16]	[e11],[e15]

QP3 - Qual a relação entre os indicadores de evasão com os métodos e algoritmos utilizados pelo modelo de predição?

Para responder a terceira questão de pesquisa, inicialmente foram relacionados os indicadores de evasão com os métodos utilizados na construção dos modelos (Tabela 4). Pode-se observar que os modelos que contemplam tanto o indicador comportamental quanto o indicador de desempenho foram implementados pelos seguintes métodos: agrupamento hierárquico, agrupamento particional, árvore de decisão, classificador bayesiano, classificador baseado em regras, classificador vizinho mais próximo, functions e redes neurais. A partir da análise, contatou-se também que o método árvore de decisão pode ser usado para construir modelos que contemplam quaisquer indicadores. Por conta disso, é o método mais utilizado na construção dos modelos reportados pelos estudos primários.

[5]Variância dos dados refere-se à alteração do estado dos dados em tempo real. Os dados podem ser classificados em variantes ou invariantes. Os dados variantes no tempo são os que podem ser modificados/atualizados em tempo real (*e.g.*, notas dos alunos). Os dados invariantes no tempo são os que não podem ser modificados no tempo (*e.g.*, dados socioeconômicos e demográficos) (SANTOS; ALBURQUEQUE; SOARES, 2014).

Tabela 4. Relação entre os indicadores de evasão e os métodos usados no modelo preditivo

Métodos	Indicadores de evasão		
	Comportamental	Demográfico	Desempenho
Agrupamento hierárquico	[e10]		[e10]
Agrupamento particional	[e02],[e10]		[e10]
Árvore de decisão	[e02],[e04],[e08],[e09],[e10], [e11], [e12],[e13],[e15]	[e04], [e11], [e15]	[e08],[e09],[e10],[e11],[e12], [e13], [e14],[e15]
Classificador bayesiano	[e06],[e09],[e10],[e12],[e13]		[e09],[e10],[e12],[e13]
Classificador baseado em regras	[e04],[e10],[e15]	[e04],[e15]	[e10],[e15]
Classificador vizinho mais próximo	[e09]		[e09]
Functions	[e10]		[e10]
Redes neurais	[e04],[e07],[e10],[e12]	[e04]	[e03],[e10],[e12]

Para complementar a análise, foram especificados os algoritmos que implementam os métodos citados na Tabela 4. A Tabela 5 apresenta a relação entre os métodos, os algoritmos e os estudos que descrevem os modelos de predição. Observou-se que os algoritmos J48, Random Forest, Naive Bayes e Bayesnet foram utilizados pela maioria dos modelos de predição de evasão, todos algoritmos de classificação. Nota-se que os algoritmos mais citados (*i.e.*, J48 e Random Forest) implementam o método árvore de decisão, que por sua vez pode contemplar os três indicadores: demográfico, comportamental e desempenho. A possibilidade de uso desses algoritmos pode estar relacionada com o tipo de dado que caracterizam estes indicadores, uma vez que o tipo de dado registrado pelo AVA tende a ser numérico e variante.

Tabela 5. Relação entre os métodos e algoritmos usados no modelo preditivo

Tarefa	Método	Algoritmos	ID dos estudos
Agrupamento	Agrupamento particional	FarthestFirst	[e10]
		K-means	[e02],[e10]
		Xmeans	[e10]
	Agrupamento hierárquico	Hierarchical clusterer	[e10]
	Classificador vizinho mais próximo	K-Nearest Neighbor	[e09]
	Functions	SMO	[e10]
Classificação	Árvore de decisão	AdaBoost	[e02],[e09]
		ADTree	[e15]
		BFTree	[e13]
		CART	[e02],[e13],[e14]
		Decision table	[e04],[e12]
		Decision tree	[e04]
		Distributed random forest	[e08]
		Gradient boosting	[e08]
		J48	[e10],[e11],[e12],[e13],[e14]
		Random forest	[e02],[e08],[e09],[e10],[e12]
		Simple logistic	[e10],[e12]
	Classificador baseado em regras	Jrip	[e10],[e15]
		Rule learner	[e04]
	Classificador bayesiano	Bayesnet	[e06],[e10],[e12],[e13]
		Naive bayes	[e09],[e10],[e12],[e13]
	Redes neurais	Back propagation through time	[e03]
		Lógica fuzzy	e07]
		Multilayer perceptron	[e04],[e10],[e12]
		RBF Network	[e10],[e12]

Em relação aos algoritmos Naive bayes e Bayesnet, integrantes do método classificador bayesiano, reconhecido como o segundo método mais usado nos modelos preditivos, acredita-se que o seu uso pode ser justificado por eles suportarem dados numéricos (*e.g.*, notas dos alunos) e nominais (*e.g.*, registros históricos). Diante desses resultados, conclui-se

que: (i) os modelos que consideram os indicadores comportamental e desempenho tendem a usar o método árvore de decisão e geralmente são implementados pelos algoritmos J48 e Random Forest; (ii) modelos que consideram o indicador demográfico fazem uso de diferentes métodos, assim, não foi possível observar uma tendência de uso de métodos e algoritmos para concebê-los; por fim, (iii) modelos que contemplam mais de um indicador tendem a usar mais de um método e algoritmo.

QP4 - Qual a relação entre os indicadores de evasão com as métricas adotadas para avaliar os modelo de predição?

Para responder a quarta questão de pesquisa, analisou-se a relação dos indicadores com as métricas adotadas para avaliar os modelos de predição por meio de um esquema de classificação. Foram identificadas 11 métricas usadas durante a avaliação dos modelos preditivos, todas listadas na Tabela 6. Pode-se observar que a maioria dos modelos de predição utilizou a métrica Precisão/Acurácia, independente do indicador abordado. Além disso, reconhece-se que há modelos que são avaliados por mais de uma métrica, geralmente por duas métricas, com exceção ao modelo descrito no estudo [e03] que utilizou quatro métricas.

Tabela 6. Relação entre os indicadores de evasão e as métricas usadas na avaliação de modelos de predição

Métricas	Indicador de evasão		
	Comportamental	Demográfico	Desempenho
Precisão/Acurácia	[e02],[e10],[e11],[e12],[e15]	[e11],[e15]	[e10],[e11],[e12],[e14],[e15]
Area under the roc curve	[e06],[e08],[e09]		[e03],[e08],[e09]
Recall	[e02]		
Root mean square error			[e03]
F-measure	[e06],[e01]		[e10]
Mean absolute error			[e03]
Coeficiente kappa	[e15]	[e15]	[e15]
Verdadeiros positivos e falsos positivos	[e06],[e12]		[e12]
Teste de cálculo de Z	[e04]	e04]	

QP5 - Os indicadores de evasão que guiaram a construção do modelo preditivo foram fundamentados por algum aporte teórico?

Existem na literatura vários modelos teóricos que abordam os indicadores de evasão na construção de modelos de predição (CASTRO; TEIXEIRA, 2014; RAMOS; BICALHO; SOUSA, 2014). Buscando compreender se os estudos que reportam os modelos preditivos descrevem o uso de algum modelo teórico, constatou-se que os indicadores que orientaram a construção dos modelos de predição não seguiram um aporte teórico. Portanto, é possível que os modelos, ao considerarem os indicadores, deixaram de abordar aspectos importantes que caracterizam os indicadores. Nesse sentido, pode-se fazer alguns questionamentos e alertas a respeito da completude dos indicadores por parte dos modelos. Por exemplo, pode-se questionar como o modelo que utiliza o indicador demográfico garante que os dados coletados e utilizados na construção deste modelo são suficientes para caracterizar esse indicador.

5. Conclusões

Neste estudo foi apresentada uma análise de modelos preditivos reportados em estudos primários a fim de identificar como os indicadores de evasão são utilizados na construção desses modelos. A MDE foi a abordagem mais utilizada nos modelos que contemplam esses indicadores. Os modelos que são construídos para abordar indicadores comportamentais e de desempenho tendem a utilizar dados que sofrem variação no tempo. O método

árvore de decisão pode ser usado para construir modelos que contemplam quaisquer indicadores, sendo o método mais utilizado na construção dos modelos reportados pelos estudos primários. Constatou-se ainda que os modelos que consideram os indicadores comportamental e desempenho tendem a usar o método árvore de decisão e geralmente são implementados pelos algoritmos J48 e Random Forest. A maioria dos modelos de predição utilizou a métrica de avaliação Precisão/Acurácia, independente do indicador abordado. Por fim, constatou-se que os indicadores que orientaram a construção dos modelos de predição não seguiram um aporte teórico. Diante disso, acredita-se que os resultados identificados neste estudo podem ser utilizados para apoiar a construção de novos modelos preditivos, uma vez que foram apresentados diferentes aspectos relacionados a construção dos modelos e como os indicadores exercem diferentes influências em sua concepção. Também, espera-se que esta análise não sirva apenas para destacar soluções/abordagens, tipos dados e técnicas utilizadas para a construção de modelos de predição de evasão, mas sirva para atrair pesquisadores e profissionais para descobrir um corpo de conhecimento que identifica como os indicadores de evasão orientam/caracterizam a construção dos modelos preditivos. Além disso, como sugestão de trabalho futuro espera-se que modelos preditivos não listados pelo estudo secundário de Chicon, Paschoal and Frantz (2020) também sejam investigados, buscando caracterizar como os indicadores de evasão abordados por tais modelos foram usados para guiar a construção dos mesmos.

Referências

BAGGI, C. A. D. S.; LOPES, D. A. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, v. 16, n. 2, p. 355–374, 2011.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 02, p. 03–13, 2011.

BITTENCOURT, I. M.; MERCADO, L. P. L. Evasão nos cursos na modalidade de educação a distância: estudo de caso do curso piloto de administração da UFAL/UAB. *Ensaio: Avaliação e políticas públicas em educação*, v. 22, n. 83, p. 465–504, 2014.

CAMPOS, A. de *et al.* Mineração de dados educacionais e learning analytics no contexto educacional brasileiro: um mapeamento sistemático. *Informática na educação: teoria & prática*, v. 23, n. 3, 2020.

CASTRO, A. K. d. S. S.; TEIXEIRA, M. A. P. Evasão universitária: modelos teóricos internacionais e o panorama das pesquisas no brasil. *Psicol. argum*, v. 32, p. 9–17, 2014.

CHICON, P. M. M.; PASCHOAL, L. N.; FRANTZ, F. C. R. Indicadores de evasão em ambientes virtuais de aprendizagem no contexto da educação a distância: Um mapeamento sistemático. *RENOTE*, v. 18, n. 2, p. 111–120, 2020.

COLPO, M. P. *et al.* Mineração de dados educacionais na previsão de evasão: uma rsl sob a perspectiva do congresso brasileiro de informática na educação. In: *Simpósio Brasileiro de Informática na Educação*, 2020. p. 1102–1111.

HLIOUI, F.; ALOUI, N.; GARGOURI, F. Understanding learner engagement in a virtual learning environment. In: *International Conference on Intelligent Systems Design and Applications*, 2018. p. 709–719.

KAMPFF, A. J. C. *et al.* Identificação de perfis de evasão e mau desempenho para geração de alertas num contexto de educação a distância. *RELATEC*, v. 13, p. 61–76, 2014.

MACHADO, R. D. *et al.* Estudo bibliométrico em mineração de dados e evasão escolar. In: *Congresso Nacional de Excelência em Gestão*, 2015. p. 1–21.

MARQUES, L. T. *et al.* Mineraçao de dados auxiliando na descoberta das causas da evasao escolar: Um mapeamento sistemático da literatura. *RENOTE*, v. 17, n. 3, p. 194–203, 2019.

MARTINHO, V. R. de C. *Sistema inteligente para a predição de grupo de risco de evasão discente*. Tese (Doutorado) — Universidade Estadual Paulista em Franca, Ilha Solteira, São Paulo, 2014.

MASCHIO, P. *et al.* Um panorama acerca da mineração de dados educacionais no brasil. In: *Brazilian Symposium on Computers in Education*, 2018. p. 1936–1940.

MAZZA, R.; DIMITROVA, V. Coursevis: A graphical student monitoring tool for supporting instructors in web-based distance courses. *International Journal of Human-Computer Studies*, Elsevier, v. 65, n. 2, p. 125–139, 2007.

MDUMA, N.; KALEGELE, K.; MACHUVE, D. A survey of machine learning approaches and techniques for student dropout prediction. p. 1–10, 2019.

QUEIROGA, E.; CECHINEL, C.; ARAÚJO, R. Um estudo do uso de contagem de interaçoes semanais para prediçao precoce de evasao em educaçao a distância. In: *Workshops do Congresso Brasileiro de Informática na Educação*, 2015. p. 1074–1083.

RAMOS, W. M.; BICALHO, R. N. M.; SOUSA, J. V. de. Evasão e persistência em cursos superiores a distância: o estado da arte da literatura internacional. In: *CONFERÊNCIA FORGES*, 2014. v. 5, p. 38–64.

SANTOS, R. D.; ALBURQUEQUE, C. de; SOARES, E. D. Uma abordagem genérica de identificação precoce de estudantes com risco de evasão em um ava utilizando técnicas de mineração de dados. In: *Congreso Internacional de Informática Educativa*, 2014. p. 794–799.

SILVA, F. *et al.* Um modelo preditivo para diagnóstico de evasão baseado nas interações de alunos em fóruns de discussão. In: *Brazilian Symposium on Computers in Education*, 2015. v. 26, p. 1187–1196.

WILL, N. N.; KEMCZINSKI, A.; PARPINELLI, R. Deep learning para previsão do desempenho do estudante: Um mapeamento sistemático da literatura. In: *Brazilian Symposium on Computers in Education*, 2019. p. 1798–1807.