Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar

Vitor Hugo Barbosa dos Santos, Daniel Victor Saraiva, Carina Teixeira de Oliveira

[1] Laboratório de Redes de Computadores e Sistemas (LAR)
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

vitorhugocrf16@gmail.com, victordvs@hotmail.com, carina@lar.ifce.edu.br

Abstract. *This paper presents an analysis of national and international Educational Data Mining (EDM) articles in the context of school dropout. The searches for articles were carried out with the purpose of answering research questions about tools/libraries, algorithms and databases considered in the articles. Due to the large amount of information analyzed in this paper, the visual analysis tool Tableau was used for a better understanding and comparison of the analyzed factors. The results serve as a subsidy to optimize the decision making of data analysts in the planning and development of EDM projects.*

Resumo. *Este artigo apresenta uma análise de trabalhos de mineração de dados educacionais (Educational Data Mining - EDM) nacionais e internacionais que tratam da temática da evasão escolar. As buscas por trabalhos foram realizadas com o propósito de responder a sete questões de pesquisa sobre ferramentas/bibliotecas, algoritmos e bases de dados considerados nos trabalhos. Dada a grande quantidade de informações analisadas, foi utilizada a ferramenta Tableau para uma melhor compreensão e comparação dos fatores estudados. Os resultados servem como subsídio para otimizar a tomada de decisão de analistas de dados no planejamento e desenvolvimento de projetos de EDM.*

1. Introdução

Muitas instituições de ensino têm dedicado recursos consideráveis para acompanhar o progresso acadêmico de seus estudantes para prever o seu desempenho, de modo a intervir e ajudar os estudantes que são identificados em situação de risco [Zhang and Li 2018]. As pesquisas são geralmente baseadas na extração de bases de dados educacionais. Entretanto, analisar uma grande quantidade de dados para encontrar informações úteis de forma resumida é uma tarefa custosa para o ser humano. Assim, essa análise deve estar associada a um processo analítico, sistemático e, até onde possível, automatizado [Silva et al. 2016].

A aplicação de métodos de Mineração de Dados e Aprendizagem de Máquina na educação é um campo interdisciplinar emergente. Essa área de pesquisa é chamada de Mineração de Dados Educacionais (*Educational Data Mining* - EDM) [Hegde and Prageeth 2018], sendo encarregada do desenvolvimento e/ou aplicação de ferramentas, métodos e técnicas capazes, por exemplo, de delinear o perfil dos estudantes com maiores riscos de evasão [Saraiva et al. 2019], assim como detectar o impacto de fatores no entorno do estudante que podem influenciar na evasão, como fatores pessoais, acadêmico, econômico, social e institucional [Alban and Mauricio 2019].

DOI: 10.5753/sbie.2021.218167

Para atuar na área de EDM, é importante pesquisar, por exemplo, as principais ferramentas/bibliotecas, algoritmos de Aprendizagem de Máquina e bases de dados para aplicar a Mineração de Dados de forma adequada no contexto da evasão escolar. Sendo assim, um primeiro passo importante para acelerar esse processo de aprendizado é identificar, avaliar e interpretar pesquisas que já estão disponíveis na temática.

Neste contexto, este trabalho propõe uma análise de trabalhos nacionais e internacionais de EDM que tratam da temática da evasão escolar. Os principais objetivos a serem alcançados neste trabalho são: identificar os trabalhos nacionais e internacionais que tratam dessa temática; identificar as ferramentas/bibliotecas usadas nesses trabalhos; identificar os algoritmos utilizados e indicar os que possuem melhor desempenho; extrair informações sobre as bases de dados estudadas nos trabalhos selecionados (como tamanho da base, quantidade e tipos de atributos); dentre outros.

Para alcançar tais objetivos, a metodologia da proposta foi dividida em três etapas. A primeira consistiu no planejamento da revisão da literatura, a segunda na preparação dos dados e a terceira na construção das visualizações. Considerando a grande quantidade de informações a serem analisadas neste trabalho, o software Tableau foi utilizado como ferramenta de análise visual para uma melhor compreensão e comparação dos fatores analisados. Os resultados alcançados podem ser usados como referência por estudantes, pesquisadores e/ou profissionais em geral interessados na área de EDM, assim como podem ser utilizados como subsídio para otimizar a tomada de decisão de analistas de dados quanto ao uso de ferramentas, algoritmos e bases de dados em projetos na área de EDM.

2. Trabalhos Relacionados

A Tabela 1 apresenta de forma resumida um comparativo entre trabalhos atuais que realizam revisão da literatura focada em EDM na evasão escolar e o presente trabalho.

Tabela 1. Comparativo dos Trabalhos Relacionados.

Trabalho	Período usado na busca	Qtd de trabalhos	Idioma dos trabalhos	Qtd de Portais Científicos	Usa Ferramenta de Análise Visual de Dados
[Marques et al. 2019]	2008 - 2018	14	Inglês	4	Não
[Alban and Mauricio 2019]	2006 - 2017	67	Inglês	9	Não
[Agrusti et al. 2019]	1999 - 2019	73	Inglês	2	Não
Este trabalho	2008 - 2020	50	Inglês e Português	6	Sim

Um grande diferencial deste trabalho é abordar trabalhos no cenário brasileiro, além do internacional. Além disso, o período considerado na busca é amplo e atual (o único que considera 2020). Em relação aos portais científicos consultados, este trabalho utilizou 6, sendo 2 brasileiros. Os outros trabalhos não usam nenhuma base brasileira. Por fim, este trabalho é o único que utiliza uma ferramenta de análise de dados para uma melhor compreensão e comparação dos dados, gerando variadas visualizações amigáveis.

3. Proposta

Neste trabalho é proposta uma análise de trabalhos nacionais e internacionais de EDM que tratam da temática da evasão escolar. As três etapas da metodologia adotada para implementação da proposta é detalhada nesta seção.

3.1. Etapa 1 - Planejamento da Revisão da Literatura

Primeiramente, foram definidas as Questões de Pesquisa (QP) a serem respondidas antes da busca por trabalhos na literatura. As 7 QP que conduziram esse estudo foram:

- **QP1**: *Quais trabalhos científicos utilizam EDM no contexto da evasão escolar?*
- **QP2**: *Quais as ferramentas/bibliotecas utilizadas pelos trabalhos?*
- **QP3**: *Quais os algoritmos de aprendizagem de máquina utilizados nos trabalhos?*
- **QP4**: *Quais os algoritmos com melhor desempenho?*
- **QP5**: *Quantos registros existem nas bases de dados analisadas?*
- **QP6**: *Quantos atributos estão presentes nas bases de dados analisadas?*
- **QP7**: *Quais tipos de dados são analisados nos trabalhos selecionados?*

Em seguida, partiu-se para a busca de trabalhos científicos capazes de responder as QP. As buscas realizadas para esta revisão de literatura foram feitas nos principais portais científicos (bibliotecas digitais) internacionais e nacionais: ACM Digital Library, IEEE Xplore, ScienceDirect, SpringerLink, SBC OpenLib (SOL) e Revista Brasileira de Informática na Educação (RBIE). Como cada repositório possui sua própria sintaxe, a Tabela 2 mostra o protocolo de busca utilizado para cada motor de busca.

Tabela 2. Protocolo de buscas utilizados nos motores de busca.

Portal Cientifico	**Protocolo de Busca**
ACM Digital Library	”Query”: {(prediction of students OR school dropout OR school retention OR school failure) AND (educational data mining OR knowledge discovery OR machine learning)AND (institution OR university)}
IEEE Xplore	(”Full Text & Metadata”:prediction of students OR school dropout OR school retention or school failure) AND (”Full Text & Metadata”:educational data mining OR knowledge discovery OR machine learning) AND (”Full Text & Metadata”:institution or university)
ScienceDirect	All:{(prediction of students OR school dropout OR school retention OR school failure) AND (educational data mining OR knowledge discovery OR machine learning) AND (institution OR university)}
SpringerLink	”Search”: {(prediction of students OR school dropout OR school retention OR school failure) AND (educational data mining OR knowledge discovery OR machine learning) AND (institution OR university)}
SBC OpenLib	Pesquisa Manual
RBIE	Pesquisa Manual

Os critérios de inclusão e exclusão foram definidos para auxiliar na condução da pesquisa, com o intuito de apoiar a classificação de relevância dos estudos. Para atender ao critério de inclusão, os trabalhos deveriam ser nacionais ou internacionais que utilizam EDM com ênfase na evasão escolar em qualquer nível de ensino e publicados entre os anos de 2008 e 2020. Já os critérios de exclusão foram definidos da seguinte forma: trabalhos não escritos em inglês ou português; trabalhos em andamento, que não tenham sido concluídos; trabalhos duplicados e estudos fora do contexto desta pesquisa. Ao final dessa etapa, foram selecionados 50 artigos, que são detalhados na Seção 4.

3.2. Etapa 2 - Preparação dos Dados

Cada QP motiva a extração dos dados. Através da leitura dos artigos, foram extraídas informações gerais de cada trabalho, como: ano de publicação; nome da ferramenta;

algoritmos usados e com melhor desempenho; quantidade de registros e atributos e tipos de dados considerados. Utilizou-se a ferramenta *MS Excel* para armazenar o conjunto de dados considerados valiosos para a análise.

3.3. Etapa 3 - Construção das Visualizações

A visualização de dados é a apresentação visual de informações quantitativas, capaz de auxiliar na descoberta de conhecimento, tendências e padrões de dados. Neste trabalho, foi utilizado o Tableau (versão *Desktop* 2021.1), uma ferramenta de *Business Intelligence* usada para criar visões de dados variadas de forma simples. Neste caso, a planilha resultante da Etapa 2 serviu de entrada para o Tableau. Nesse contexto, a Etapa 3 consistiu na utilização do Tableau para a criação das visualizações no contexto das QP definidas na Etapa 1. Na próxima seção são apresentadas e discutidas as respostas para as 7 QP.

4. Resultados e Discussões

4.1. QP1: *Quais trabalhos científicos utilizam EDM no contexto da evasão escolar?*

Como dito na Etapa 1, obteve-se 50 trabalhos selecionados resultantes das buscas realizadas (Tabelas 3 e 4).

Tabela 3. Trabalhos Nacionais Selecionados.

ID	Referências	ID	Referências
E1	[Amorim et al. 2008]	E12	[Gonçalves et al. 2018]
E2	[Manhães et al. 2012]	E13	[Ramos et al. 2018]
E3	[Costa et al. 2014]	E14	[Bitencourt and Ferrero 2019]
E4	[Brito et al. 2015]	E15	[Gonçalves and Beltrame 2019]
E5	[Silva et al. 2015]	E16	[Saraiva et al. 2019]
E6	[Santana et al. 2015]	E17	[Pereira et al. 2019]
E7	[Kantorski et al. 2016]	E18	[Junior et al. 2019]
E8	[Maria et al. 2016]	E19	[Barros et al. 2020]
E9	[Queiroga et al. 2017]	E20	[Filho et al. 2020]
E10	[Paz and Cazella 2017]	E21	[Santos et al. 2020]
E11	[Lanes and Alcântara 2018]	E22	[Soares et al. 2020]

Tabela 4. Trabalhos Internacionais Selecionados.

ID	Referências	ID	Referências
E23	[Dekker et al. 2009]	E37	[Adil et al. 2018]
E24	[Oyelade et al. 2010]	E38	[Dharmawan et al. 2018]
E25	[Pal 2012]	E39	[Hegde and Prageeth 2018]
E26	[Marquez-Vera et al. 2013]	E40	[Limsathitwong et al. 2018]
E27	[Jamesmanoharan et al. 2014]	E41	[Murakami et al. 2018]
E28	[Yukselturk et al. 2014]	E42	[Perez et al. 2018]
E29	[Pradeep et al. 2015]	E43	[Solis et al. 2018]
E30	[Marbouti et al. 2016]	E44	[Alban and Mauricio 2019]
E31	[Meedech et al. 2016]	E45	[Li et al. 2019]
E32	[Ahuja and Kankane 2017]	E46	[Beltran et al. 2019]
E33	[Pereira and Zambrano 2017]	E47	[Lottering et al. 2020]
E34	[Rocha et al. 2017]	E48	[Utari et al. 2020]
E35	[Rodrıguez-Maya et al. 2017]	E49	[Viloria et al. 2020]
E36	[Rovira et al. 2017]	E50	[Yaacob et al. 2020]

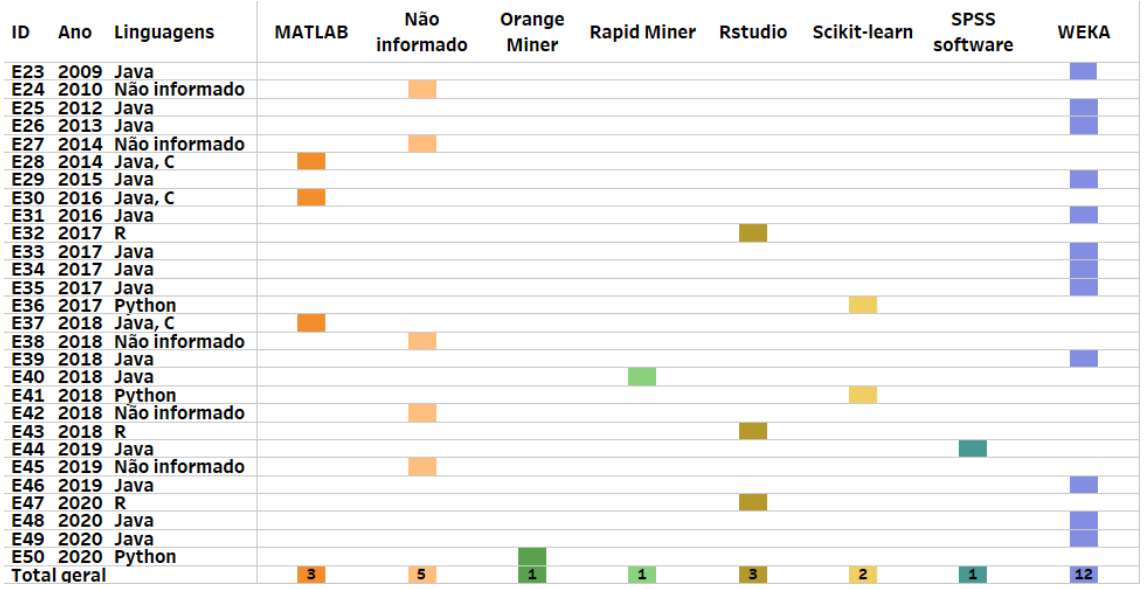
Como mostram as Tabelas 3 e 4, dentre os trabalhos selecionados, 22 são nacionais e 28 internacionais. Pode-se dizer que há um equilíbrio entre as quantidades de trabalhos encontrados para cada cenário.

4.2. QP2: *Quais as ferramentas/bibliotecas utilizadas pelos trabalhos?*

Para responder a QP2, são apresentadas as visões das Figuras 1 e 2 com detalhes do uso das ferramentas/bibliotecas de MD para os trabalhos dos cenários nacional e internacional, respectivamente. Estes trabalhos estão ordenados por ano de publicação nas figuras.

Figura 1. Ferramentas e bibliotecas utilizadas no cenário nacional.

Figura 2. Ferramentas e bibliotecas utilizadas no cenário internacional.



No total, foram identificadas 6 ferramentas/bibliotecas usadas pelos trabalhos no cenário nacional. De imediato, é possível verificar na Figura 1 que a ferramenta *Weka* e a biblioteca *Scikit-learn* destacam-se dentre as demais. Dentre os 22 trabalhos nacionais, elas foram utilizadas em 13 e 5 trabalhos, respectivamente. As ferramentas/bibliotecas *Genie, Orange Miner, Rstudio* e *Spyder* foram utilizadas nos outros estudos. Observa-se também que o *Weka* se destaca de 2008 a 2018, sendo usada em praticamente todos os estudos deste intervalo. A partir de 2019 utilizou-se com mais frequência o *Scikit-learn*.

Na Figura 2, percebe-se que no cenário internacional foram identificadas 7 ferramentas/bibliotecas. Um total de 4 trabalhos não informaram a ferramenta/biblioteca utilizada. Novamente, a ferramenta *Weka* prevalece nas pesquisas (12 dos 24 trabalhos). A ferramenta *MATLAB* é utilizada em 3 trabalhos, assim como o *Rstudio*. Na sequência, a biblioteca *Scikit-learn* aparece em 2 trabalhos. Por fim, as outras ferramentas (*Orange Miner, Rapid Miner* e *SPSS Software*) estão presentes em 1 trabalho. De forma similar ao cenário nacional, o uso do *Weka* domina nos primeiros anos considerados no cenário internacional. No entanto, entre os anos de 2018 e 2020 o uso do *Weka* vai diminuindo.

4.3. QP3: *Quais os algoritmos de aprendizagem de máquina utilizados nos trabalhos?*

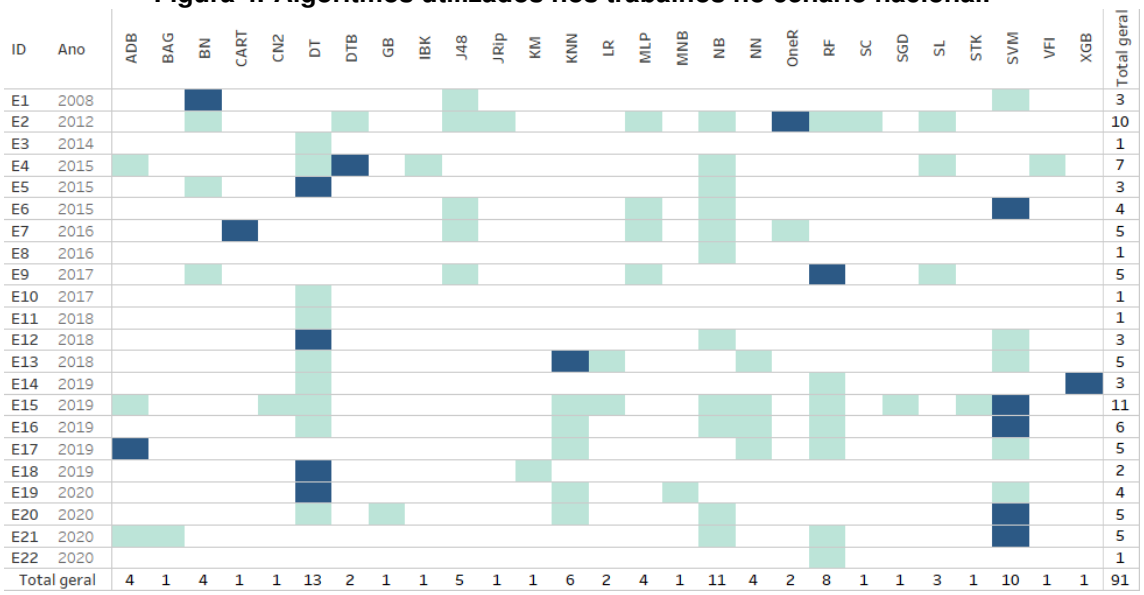
Nesta pesquisa, foram identificados vários algoritmos dentre os 50 trabalhos selecionados. Isso ocorre porque em muitos casos os trabalhos selecionados aplicaram bem mais que um algoritmo em seus experimentos. Por exemplo, somente no trabalho [Gonçalves and Beltrame 2019], 11 algoritmos foram utilizados. Assim, para facilitar a visualização dos resultados neste documento, optou-se por abreviar os nomes dos algoritmos conforme mostra a Figura 3.

Figura 3. Lista dos Algoritmos e suas Abreviaturas.

Abreviatura	Nome do Algoritmo	Abreviatura	Nome do Algoritmo
ADB	AdaBoost	MLP	Multilayes Perceptron
ADT	ADTree	MNB	Multinomial Naive Bayes
BAG	Bagging	NB	Naive Bayes
BN	Bayes Net	NF	Neuro-Fuzzy
C4.5	C4.5	NN	Neural Network
CART	CART	NNge	NNge
CN2	CN2 Rule Inducer	OneR	OneR
Ctree	Ctree	Prism	Prism
DT	Decision Tree	RF	Random Forest
DTB	DecisionTable	Ridor	Ridor
FCM	Fuzzy C-means	Rpart	Rpart
GB	GradientBoosting	RPT	REPTree
GNB	Gaussian Naive Bayes	RT	Random Tree
IBK	IBK	SC	SimpleCart
ID3	ID3	SGD	SGD
J48	J48	SL	Simple Logistic
JRip	JRip	STK	Stacking
KM	K-Means	SVM	Support Vector Machine
KNN	K-Nearest Neighbors	VFI	VFI
LR	Logistic Regression	XGB	XGBoost

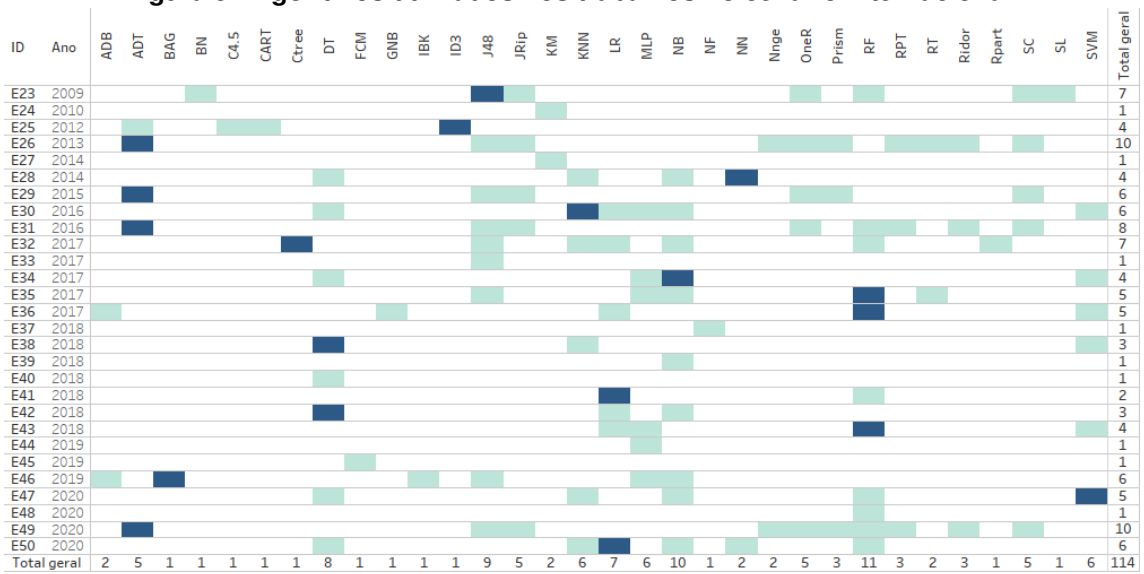
Em seguida, a resposta para a QP3 é apresentada para cada cenário (Figuras 4 e 5). Ao final de cada linha é indicada a quantidade de algoritmos utilizado por cada trabalho. Os algoritmos em destaque (em uma cor mais escura) são os que possuem o melhor desempenho para os trabalhos que possuem no mínimo 2 algoritmos em seus experimentos.

Figura 4. Algoritmos utilizados nos trabalhos no cenário nacional.



ID	Ano	ADB	BAG	BN	CART	CN2	DT	DTB	GB	IBK	J48	JRip	KM	KNN	LR	MLP	MNB	NB	NN	OneR	RF	SC	SGD	SL	STK	SVM	VFI	XGB	Total geral
E1	2008			■							▒															▒			3
E2	2012			▒				▒			▒	▒				▒		▒		■	▒	▒		▒					10
E3	2014						▒																						1
E4	2015	▒					▒	■		▒								▒						▒			▒		7
E5	2015			▒			■											▒											3
E6	2015										▒					▒		▒								■			4
E7	2016				■						▒					▒		▒		▒									5
E8	2016																	▒											1
E9	2017			▒							▒					▒					■			▒					5
E10	2017						▒																						1
E11	2018						▒																						1
E12	2018						■											▒								▒			3
E13	2018						▒							■	▒				▒							▒			5
E14	2019						▒														▒							■	3
E15	2019	▒				▒	▒							▒	▒			▒	▒		▒		▒		▒	■			11
E16	2019						▒							▒				▒	▒		▒					■			6
E17	2019	■												▒					▒		▒					▒			5
E18	2019						■						▒																2
E19	2020						■							▒			▒									▒			4
E20	2020						▒		▒					▒				▒								■			5
E21	2020	▒	▒															▒			▒					■			5
E22	2020																				▒								1
Total geral		4	1	4	1	1	13	2	1	1	5	1	1	6	2	4	1	11	4	2	8	1	1	3	1	10	1	1	91

Figura 5. Algoritmos utilizados nos trabalhos no cenário internacional.



ID	Ano	ADB	ADT	BAG	BN	C4.5	CART	Ctree	DT	FCM	GNB	IBK	ID3	J48	JRip	KM	KNN	LR	MLP	NB	NF	NN	Nnge	OneR	Prism	RF	RPT	RT	Ridor	Rpart	SC	SL	SVM	Total geral
E23	2009				▒									■	▒									▒		▒					▒	▒		7
E24	2010															▒																		1
E25	2012		▒			▒	▒						■																					4
E26	2013		■											▒	▒								▒	▒	▒		▒	▒	▒		▒			10
E27	2014															▒																		1
E28	2014								▒									▒		▒		■													4
E29	2015		■											▒	▒									▒	▒						▒			6
E30	2016								▒									■	▒	▒														▒	6
E31	2016		■											▒	▒									▒		▒	▒		▒		▒			8
E32	2017							■						▒				▒		▒						▒				▒				7
E33	2017													▒																				1
E34	2017								▒										▒	■														▒	4
E35	2017													▒					▒	▒						■		▒							5
E36	2017	▒									▒							▒								■								▒	5
E37	2018																				▒													1
E38	2018								■									▒																▒	3
E39	2018																			▒														1
E40	2018								▒																									1
E41	2018																	■								▒								2
E42	2018								■									▒		▒															3
E43	2018																	▒	▒							■								▒	4
E44	2019																		▒																1
E45	2019									▒																									1
E46	2019	▒		■								▒		▒					▒	▒															6
E47	2020								▒									▒		▒						▒								■	5
E48	2020																																		1
E49	2020		■											▒	▒								▒	▒	▒	▒	▒		▒		▒			10
E50	2020								▒									▒	■	▒		▒				▒									6
Total geral		2	5	1	1	1	1	1	8	1	1	1	1	9	5	2	6	7	6	10	1	2	2	5	3	11	3	2	3	1	5	1	6	114

A Figura 4 mostra o(s) algoritmo(s) utilizado(s) por cada um dos 22 trabalhos do cenário nacional. A maioria dos trabalhos usa 5 algoritmos em seus experimentos. No total, foram identificados 27 algoritmos diferentes. O algoritmo *Decision Tree* (DT) é o mais utilizado, aparecendo em 13 dos 22 trabalhos. Além do mais usado, também se destaca como um dos algoritmos com melhor desempenho em 4 trabalhos. Em seguida, aparece o *Naive Bayes* (NB) sendo usado por 11 trabalhos. O *Support Vector Machine* (SVM) aparece na sequência em 10 dos 22 trabalhos, sendo também destaque com melhor desempenho em 5 trabalhos.

A Figura 5 mostra o(s) algoritmo(s) utilizado(s) por cada um dos 28 trabalhos do

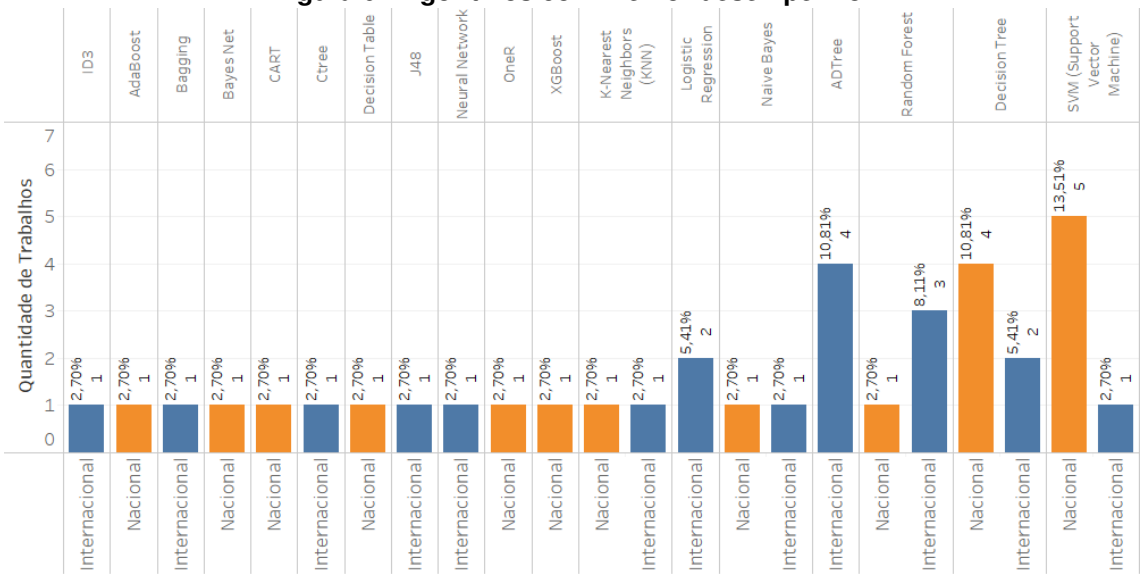
cenário internacional. A maioria dos trabalhos usa 1 algoritmo em seus experimentos. No total, foram identificados 33 algoritmos diferentes. O algoritmo *Random Forest* (RF) é o mais utilizado, aparecendo em 11 dos 28 trabalhos. Além de ser o mais usado, também se destaca como um dos algoritmos com melhor desempenho em 3 trabalhos. Em seguida, aparece o NB sendo usado por 10 trabalhos. Já os algoritmos *J48* e DT aparecem em terceiro e quarto lugares como algoritmos mais usados, respectivamente, aparecendo em 9 dos 28 trabalhos e em 8 dos 28 trabalhos. Por fim, na Figura 5, cabe destacar o resultado do *ADTree* (ADT), que apesar de ser usado em apenas 5 dos 28 trabalhos, apresentou melhor desempenho em 4 deles.

4.4. QP4: *Quais os algoritmos com melhor desempenho?*

A resposta para a QP4 já foi discutida na QP3. No cenário nacional, o SVM e o DT aparecem como os algoritmos com melhores desempenhos, respectivamente. Já no cenário internacional, o ADT e o RF aparecem com os melhores desempenhos.

A Figura 6 também responde a QP4 sob outra perspectiva, apresentando somente os algoritmos que tiveram melhor desempenho em trabalhos que possuem no mínimo 2 algoritmos em seus experimentos. Assim, dentre os 40 algoritmos identificados, a figura mostra que 18 têm melhor desempenho em pelo menos um trabalho. Se for retirada a visão por cenário, pode-se notar que os algoritmos DT e SVM com 6 trabalhos estão entre os melhores.

Figura 6. Algoritmos com melhor desempenho.



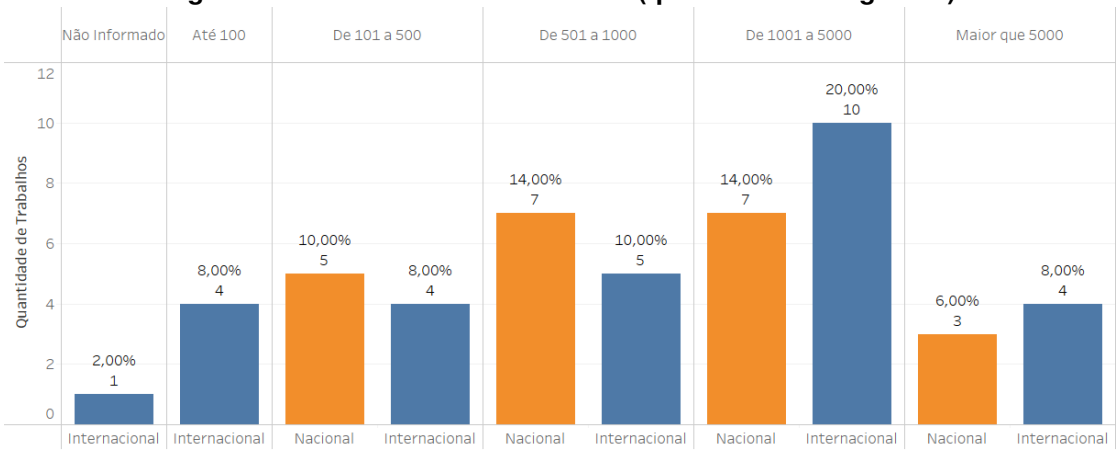
4.5. QP5: *Quantos registros existem nas bases de dados analisadas?*

A QP5 traz uma análise do tamanho das bases de dados usadas nos trabalhos. Como cada base tem tamanho único, optou-se por criar intervalos, como mostra a Figura 7. Os valores apresentados são relativos à quantidade de registros após a fase de preparação dos dados, ou seja, é a base usada nas fases de treinamento e teste dos algoritmos.

A Figura 7 mostra que a maioria das bases estudadas estão nos intervalos *De 1.001 a 5.000* registros (17 trabalhos) ou *De 501 a 1.000* (12 trabalhos). Somente 7

trabalhos utilizaram bases com mais de 5 mil registros. No mais, 4 trabalhos utilizaram bases menores de *Até 100* registros. Somente 1 trabalho não informa o tamanho da base estudada.

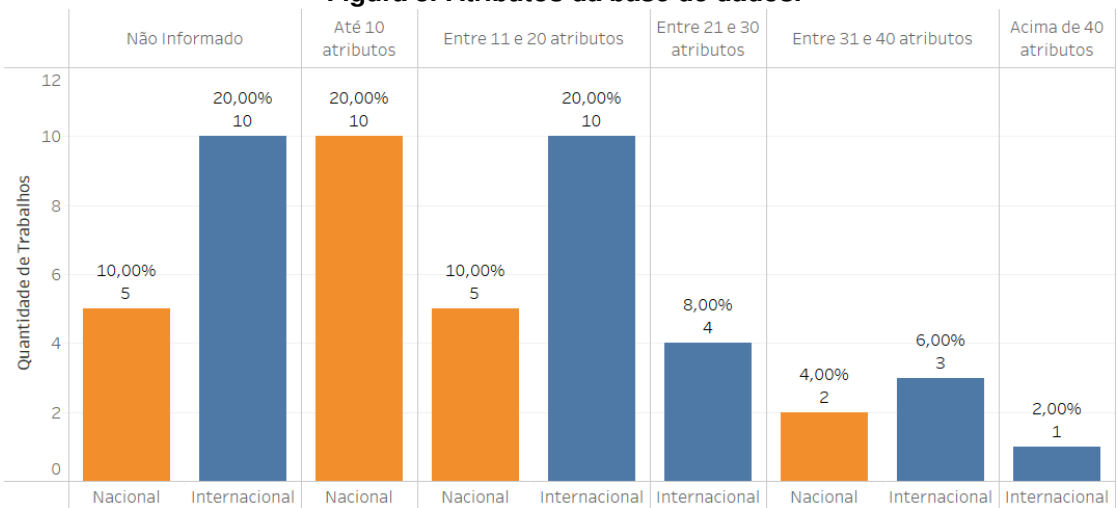
Figura 7. Tamanho da base de dados (quantidade de registros).



4.6. QP6: *Quantos atributos estão presentes nas bases de dados analisadas?*

A Figura 8 responde a QP6 mostrando informações sobre a quantidade de atributos presentes nas bases de dados utilizadas nos trabalhos selecionados. Assim como foi dito para os registros na QP5, cada base pode conter valores únicos de atributos. Logo, seguiu-se o mesmo processo de criar intervalos para melhorar a visualização dos resultados da QP6.

Figura 8. Atributos da base de dados.

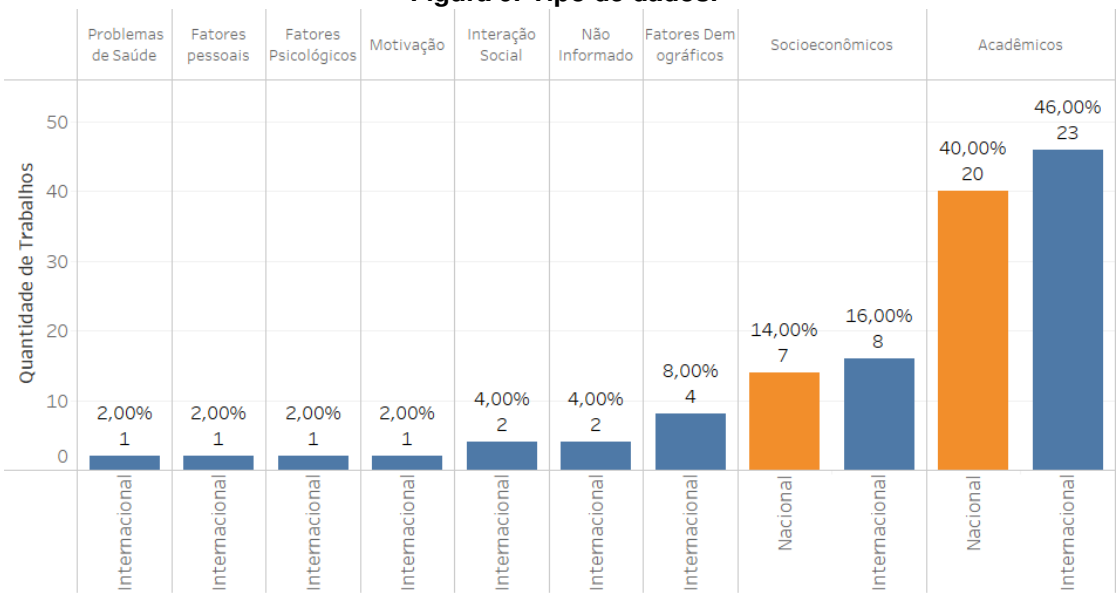


A Figura 8 mostra que das bases de dados estudadas, a maioria dos trabalhos utiliza *Entre 11 e 20 atributos* (15 trabalhos). Porém, há uma maior predominância para os trabalhos do cenário internacional para esse intervalo. Na verdade, no cenário nacional, a maioria dos trabalhos utiliza *Até 10 atributos*. Por fim, a figura também mostra que muitos trabalhos não passam informações sobre os atributos utilizados (15 trabalhos), principalmente no cenário internacional. Apenas 1 trabalho utiliza mais de 40 atributos.

4.7. QP7: *Quais os tipos de dados analisados nos trabalhos?*

A Figura 9 apresenta os tipos de dados estudados pelos trabalhos selecionados, ou seja, os fatores que são investigados que podem contribuir para que estudantes abandonem seus cursos. O somatório dos trabalhos da Figura 9 é maior que 50 trabalhos porque algumas bases de dados contemplam mais de um tipo de dado. Assim, Pode-se observar pela Figura 9 que a maioria das bases de dados é composta por tipos de dados acadêmicos (43 trabalhos). Em seguida, aparecem os fatores socioeconômicos (15 trabalhos). Para os dois tipos, há um certo equilíbrio entre os cenários. Outros 6 tipos foram identificados, mas com poucos trabalhos aplicando.

Figura 9. Tipo de dados.



5. Conclusões

O trabalho apresentou uma análise de 22 trabalhos nacionais e 28 trabalhos internacionais que abordam a área de mineração de dados educacionais com temática central voltada para o problema da evasão escolar. Na etapa final da metodologia, o Tableau foi utilizado para a análise visual dos resultados. Observou-se que a ferramenta *Weka* e a biblioteca *Scikit-learn* destacam-se dentre as demais. Entretanto, em trabalhos mais recentes, o uso do *Weka* vai diminuindo. Entre os algoritmos mais utilizados, os modelos *Decision Tree*, *Naive Bayes*, *Random Forest* e *Support Vector Machine* são os mais utilizados. Considerando os dois cenários, os modelos *Decision Tree* e *Support Vector Machine* estão entre os melhores desempenhos. A maioria das bases estudadas está nos intervalos *De 1.001 a 5.000* registros ou *De 501 a 1.000* e com a quantidade de atributos entre 11 e 20. Em relação aos tipos de dados analisados, a maioria das bases de dados é composta por tipos de dados acadêmicos. Em seguida, aparecem os fatores socioeconômicos.

Esta pesquisa pode ser continuada em vários aspectos como trabalhos futuros, tais como: adicionando novos trabalhos publicados, investigando mais a fundo as bases de dados disponibilizadas, considerando novas relações entre as questões analisadas (ex: relação entre desempenho do algoritmo e o tamanho das bases e tipos de atributos), etc.

Referências

Adil, M., Tahir, F., and Maqsood, S. (2018). Predictive analysis for student retention by using neuro-fuzzy algorithm. In *Computer Science and Electronic Engineering (CEEC)*, pages 41–45.

Agrusti, F., Bonavolontà, G., and Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of E-Learning and Knowledge Society*, 15(3):161–182.

Ahuja, R. and Kankane, Y. (2017). Predicting the probability of student's degree completion by using different data mining techniques. In *International Conference on Image Information Processing (ICIIP)*, pages 1–4.

Alban, M. and Mauricio, D. (2019). Predicting university dropout through data mining: A systematic literature. *Indian Journal of Science and Technology*, 12(4):1–12.

Amorim, M., Barone, D., and Mansur, A. (2008). Técnicas de aprendizado de máquina aplicadas na previsão de evasão acadêmica. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 1(1):666–674.

Barros, R. P., De Santana Junior, O. V., De Medeiros Silva, I. R., Dos Santos, L. F., and Neto, V. R. C. (2020). Predição do rendimento dos alunos em lógica de programação com base no desempenho das disciplinas do primeiro período do curso de ciências e tecnologia utilizando técnicas de mineração de dados. *Brazilian Journal of Development*, 6(1):2523–2534.

Beltran, C., Xavier-Júnior, J., Barreto, C., and Neto, C. O. (2019). Plataforma de aprendizado de maquina para detecção e monitoramento de alunos com risco de evasão. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 30(1):1591.

Bitencourt, P. and Ferrero, C. (2019). Predição de risco de evasão de alunos usando métodos de aprendizado de máquina em cursos técnicos. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, 8(1):149.

Brito, D. M., Lemos, M. O., Pascoal, T. A., do Rêgo, T. G., and Araújo, J. G. G. d. O. (2015). Identificação de estudantes do primeiro semestre com risco de evasão através de técnicas de data mining. *Nuevas Ideas en Informática Educativa TISE*, pages 459–463.

Costa, S. S. D., Cazella, S., and Rigo, S. J. (2014). Minerando dados sobre o desempenho de alunos de cursos de educação permanente em modalidade EAD: Um estudo de caso sobre evasão escolar na una-sus. *Revista Novas Tecnologias na Educação (RENOTE)*, 12(2).

Dekker, G. W., Pechenizkiy, M., and Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. *International Working Group on Educational Data Mining*.

Dharmawan, T., Ginardi, H., and Munif, A. (2018). Dropout detection using non-academic data. In *International Conference on Science and Technology (ICST)*, pages 1–4.

Filho, F. H., Siqueira, D., and Leal, B. (2020). Predição de evasão utilizando técnicas de classificação: Um estudo de caso do instituto federal do ceará. In *Anais da VIII Escola*

Regional de Computação do Ceará, Maranhão e Piauí, pages 141–148, Porto Alegre, RS, Brasil. SBC.

Gonçalves, O. L. and Beltrame, W. A. R. (2019). Mineração de dados e evasão estudantil: Analisando o curso de nível superior do IFES em Guarapari. In *Anais do XII Congresso de Administração Sociedade e Inovação (CASI)*.

Gonçalves, T. C., da Silva, J. C., and Cortes, O. A. C. (2018). Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do instituto federal do maranhão. *Revista Brasileira de Computação Aplicada*, 10(3):11–20.

Hegde, V. and Prageeth, P. P. (2018). Higher education student dropout prediction and analysis through educational data mining. In *International Conference on Inventive Systems and Control (ICISC)*, pages 694–699.

Jamesmanoharan, J., Ganesh, S. H., Felciah, M. L. P., and Shafreenbanu, A. K. (2014). Discovering students' academic performance based on gpa using k-means clustering algorithm. In *World Congress on Computing and Communication Technologies*, pages 200–202.

Junior, I. B., Rabelo, H., Naschold, A., e Aquiles Burlamaqui, A. F., Rabelo, D., and Valentim, R. (2019). Uso de mineração de dados educacionais para a classificação e identificação de perfis de evasão de graduandos em sistemas de informação. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, 8(1):159.

Kantorski, G., Flores, E., Schmitt, J., Hoffmann, I., and Barbosa, F. (2016). Predição da evasão em cursos de graduação em instituições públicas. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 27(1):906.

Lanes, M. and Alcântara, C. (2018). Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 29(1):1921.

Li, Y., Gou, J., and Fan, Z. (2019). Educational data mining for students' performance based on fuzzy c-means clustering. *The Journal of Engineering*, 2019(11):8245–8250.

Limsathitwong, K., Tiwatthanont, K., and Yatsungnoen, T. (2018). Dropout prediction system to reduce discontinue study rate of information technology students. In *International Conference on Business and Industrial Research (ICBIR)*, pages 110–114.

Lottering, R., Hans, R., and Lall, M. (2020). A model for the identification of students at risk of dropout at a university of technology. In *International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pages 1–8.

Manhães, L., da Cruz, S., Costa, R. M., Zavaleta, J., and Zimbrão, G. (2012). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 1(1).

Marbouti, F., Diefes-Dux, H. A., and Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, 103:1–15.

Maria, W., Damiani, J., and Pereira, M. (2016). Rede bayesiana para previsão de evasão escolar. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, 5(1):920.

Marques, L. T., De Castro, A. F., Marques, B. T., Silva, J. C. P., and Queiroz, P. G. G. (2019). Mineraçao de dados auxiliando na descoberta das causas da evasao escolar: Um mapeamento sistemático da literatura. *RENOTE*, 17(3):194–203.

Marquez-Vera, C., Morales, C. R., and Soto, S. V. (2013). Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 8(1):7–14.

Meedech, P., Iam-On, N., and Boongoen, T. (2016). Prediction of student dropout using personal profile and data mining approach. In *Intelligent and Evolutionary Systems*, pages 143–155. Springer.

Murakami, K., Takamatsu, K., Kozaki, Y., Kishida, A., Bannaka, K., Noda, I., Asahi, J., Takao, K., Mitsunari, K., Nakamura, T., and Nakata, Y. (2018). Predicting the probability of student dropout through emir using data from current and graduate students. In *International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 478–481.

Oyelade, O. J., Oladipupo, O. O., and Obagbuwa, I. C. (2010). Application of k means clustering algorithm for prediction of students academic performance. *CoRR*, abs/1002.2425.

Pal, S. (2012). Mining educational data to reduce dropout rates of engineering students. *International Journal of Information Engineering and Electronic Business*, 4(2):1.

Paz, F. and Cazella, S. (2017). Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 6, page 624.

Pereira, A., Carvalho, L., and Souto, E. (2019). Predição de evasão de estudantes non-majors em disciplina de introdução à programação. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, 8(1):178.

Pereira, R. T. and Zambrano, J. C. (2017). Application of decision trees for detection of student dropout profiles. In *IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 528–531.

Perez, B., Castellanos, C., and Correal, D. (2018). Applying data mining techniques to predict student dropout: A case study. In *IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI)*, pages 1–6.

Pradeep, A., Das, S., and Kizhekkethottam, J. J. (2015). Students dropout factor prediction using edm techniques. In *International Conference on Soft-Computing and Networks Security (ICSNS)*, pages 1–7.

Queiroga, E., Cechinel, C., and Araújo, R. (2017). Predição de estudantes com risco de evasão em cursos técnicos a distância. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 28(1):1547.

Ramos, J., Silva, J., Prado, L., Gomes, A., and Rodrigues, R. (2018). Um estudo comparativo de classificadores na previsão da evasão de alunos em ead. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 29(1):1463.

Rocha, C. F., Zelaya, Y. F., Sánchez, D. M., and Pérez, F. A. F. (2017). Prediction of university desertion through hybridization of classification algorithms. In *SIMBig*, pages 215–222.

Rodrıguez-Maya, N. E., Lara-Álvarez, C., May-Tzuc, O., and Suárez-Carranza, B. A. (2017). Modeling students' dropout in mexican universities. *Research in Computing Science*, 139:163–175.

Rovira, S., Puertas, E., and Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLoS one*, 12(2).

Santana, M. A., de Barros Costa, E., dos Santos Neto, B. F., Silva, I. C. L., and Rego, J. B. (2015). A predictive model for identifying students with dropout profiles in online courses. In *EDM*.

Santos, G., dos Santos, F., Rocha, A., and da Silva, T. (2020). Utilização de aprendizagem de máquina para a identificação de dependência em aparelhos celulares com foco em casos que possam causar reprovação e evasão. In *Anais da VIII Escola Regional de Computação do Ceará, Maranhão e Piauí*, pages 228–235, Porto Alegre, RS, Brasil. SBC.

Saraiva, D., Pereira, S., Gallindo, E., Braga, R., and Oliveira, C. (2019). Uma proposta para predição de risco de evasão de estudantes em um curso técnico em informática. In *Anais do XXVII Workshop sobre Educação em Computação*, pages 319–333, Porto Alegre, RS, Brasil. SBC.

Silva, F., Silva, J., Silva, R., and Fonseca, L. (2015). Um modelo preditivo para diagnóstico de evasão baseado nas interações de alunos em fóruns de discussão. *Simpósio Brasileiro de Informática na Educação (SBIE)*, 26(1):1187.

Silva, L. A., Peres, S. M., and Boscarioli, C. (2016). *Introdução à mineração de dados: com aplicações em R*. Elsevier, Rio de Janeiro, 1 edition.

Soares, L. C. C. P., Ronzani, R. A., de Carvalho, R. L., and da Silva, A. T. R. (2020). Aplicação de técnicas de aprendizado de máquina em um contexto acadêmico com foco na identificação dos alunos evadidos e não evadidos. *Humanidades & Inovação*, 7(8):223–235.

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., and Hernandez, M. (2018). Perspectives to predict dropout in university students with machine learning. In *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–6.

Utari, M., Warsito, B., and Kusumaningrum, R. (2020). Implementation of data mining for drop-out prediction using random forest method. In *International Conference on Information and Communication Technology (ICoICT)*, pages 1–5.

Viloria, A., Guliany, J. G., Núñez, W. N., Palma, H. H., and Núñez, L. N. (2020). Data mining applied in school dropout prediction. In *Journal of Physics: Conference Series*, volume 1432. IOP Publishing.

Yaacob, W. W., Sobri, N. M., Nasir, S. M., Norshahidi, N., and Husin, W. W. (2020). Predicting student drop-out in higher institution using data mining techniques. In *Journal of Physics: Conference Series*, volume 1496. IOP Publishing.

Yukselturk, E., Ozekes, S., and Turel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and e-learning*, 17(1):118–133.

Zhang, L. and Li, K. F. (2018). Education analytics: Challenges and approaches. In *International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pages 193–198.