UMA ANÁLISE DE TRABALHOS DE MINERAÇÃO DE DADOS EDUCACIONAIS NO CONTEXTO DA EVASÃO ESCOLAR NOS CENÁRIOS NACIONAL E INTERNACIONAL

AN ANALYSIS OF EDUCATIONAL DATA MINING ARTICLES IN THE CONTEXT OF SCHOOL DROPOUT IN NATIONAL AND INTERNATIONAL SCENARIOS

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

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 em seis portais científicos com o propósito de responder a nove questões de pesquisa sobre ferramentas/bibliotecas, algoritmos, bases de dados e níveis de escolaridade considerados nos trabalhos. É apresentada uma metodologia composta por três etapas: *(i)* planejamento da revisão da literatura, *(ii)* preparação dos dados e *(iii)* construção das visualizações. Considerando a grande quantidade de informações analisadas neste trabalho, foi utilizada a ferramenta de análise visual Tableau 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, linguagens, algoritmos e bases de dados em projetos na área de EDM.

Palavras-chave: Análise. Evasão Escolar. Mineração de Dados Educacionais.

ABSTRACT

This paper presents an analysis of national and international Educational Data Mining (EDM) articles that deal with the subject of school dropout. The searches for articles

* Graduando em Bacharelado em Ciência da Computação, Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Aracati, Ceará, Brasil. E-mail: vitorhugocrf16@gmail.com
** Mestre em Ciência da Computação pelo Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Pesquisador do Laboratório de Redes de Computadores e Sistemas (LAR/IFCE), Aracati, Ceará, Brasil. E-mail: dvictordvs@gmail.com
*** Doutora em Informática pela Université Joseph Fourier (UJF), Docente do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Aracati, Ceará, Brasil. E-mail: carina@lar.ifce.edu.br

were carried out in six scientific portals with the purpose of answering nine research questions about tools/libraries, algorithms, databases and educational levels considered in the articles. A methodology consisting of three steps is presented: *(i)* literature review planning, *(ii)* data preparation and *(iii)* views construction. Considering 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 achieved can be used as a reference by students, researchers and/or professionals in general interested in the area of EDM, as well as can be used as a subsidy to optimize the decision making of data analysts regarding the use of tools, languages, algorithm and databases in projects in the EDM area.

Keywords: Analysis. School Dropout. Educational Data Mining.

1 INTRODUÇÃO

Com o advento da Constituição brasileira decretada em 1988 (BRASIL, 1988), também denominada Constituição Cidadã, a educação foi universalizada como direito social, estabelecendo-se como um dos princípios para o firmamento da democracia em território nacional, segundo os artigos 6° e 205°. Assim, a educação se estabeleceu como um dever do Estado que, em conjunto com as famílias, passaram a buscar soluções para promover e desenvolver pessoas no exercício da cidadania e qualificação profissional, visando a construção de uma sociedade mais igualitária e justa (BRASIL, 1988).

Diante disso, no Brasil, a Lei de Diretrizes e Base da Educação (LDB 9.394 de 1996) (BRASIL, 1996) foi criada com o objetivo de regulamentar o sistema educacional, permitindo aos estudantes adquirirem conhecimentos e habilidades imprescindíveis para qualquer profissional. Entretanto, mesmo com a LDB reafirmando o direito à educação garantido pela Constituição (BRASIL, 1988), a alta evasão observada desde a educação básica até o ensino superior impacta negativamente na eficiência dos sistemas educacionais (FILHO et al., 2007).

Por exemplo, segundo o Relatório do Desenvolvimento Humano de 2018 elaborado pelo Programa das Nações Unidas para o Desenvolvimento (*United Nations Development Programme* - UNDP) (UNDP, 2018), o Brasil possui uma taxa de permanência de estudantes até a última série do ensino médio de 74%. Entre os doze países da América do Sul, o Brasil fica à frente apenas da Colômbia, que possui uma taxa de permanência de 71%.

Ainda, conforme os dados do Censo da Educação Superior do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2021), em 2010, as instituições de ensino públicas brasileiras, em sua totalidade, receberam 2.576.304

novos estudantes. Em contrapartida, ao final de dez anos de acompanhamento desses estudantes, 40% concluíram, 59% desistiram do seu curso de ingresso durante esse período e 1% dos estudantes ainda permanecem matriculados.

No âmbito da Rede Federal de Educação Profissional, Ciência e Tecnologia, o Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) oferece cursos que abrangem desde o ensino básico até a pós-graduação por meio da tríade Ensino, Pesquisa e Extensão (IFCE, 2017). No IFCE, os índices de evasão têm sido preocupantes. Por exemplo, entre os períodos de 2009/1 até 2020/2 (IFCE, 2021) foram realizadas 7.713 matrículas no curso Técnico em Informática (concomitante, integrado e subsequentes) na instituição. Do total de ingressantes no curso, 1.868 (24,21%) conseguiram concluir seus estudos (egresso com êxito) e 3.281 (42,53%) dos discentes não concluíram seus estudos (egresso sem êxito), ou seja, a taxa de evasão é bastante elevada, maior que a de concludentes.

Em relação às consequências do fenômeno da evasão, interromper o ciclo de estudos em qualquer nível de ensino pode ocasionar em um grande prejuízo, tanto acadêmico, quanto socioeconômico (BAGGI; LOPES, 2010). No setor público, por exemplo, as perdas provocadas por discentes que evadem do sistema escolar representam recursos públicos investidos na educação sem o retorno apropriado para a sociedade, enquanto no setor privado acarreta perda de receitas (FILHO et al., 2007).

Diante dessas problemáticas, muitas instituições de ensino (ex: escolas, universidades e institutos) têm dedicado recursos consideráveis para acompanhar o progresso acadêmico de seus estudantes com o objetivo de tomar decisões administrativas embasadas nas análises dessas informações. Além do acompanhamento do progresso, algumas instituições deram um passo além ao usar os registros dos históricos dos estudantes para prever o seu desempenho, de modo a intervir e ajudar os estudantes que são identificados em situação de risco (ZHANG; LI, 2018).

Uma solução computacional benéfica para estudar as causas da evasão escolar é a utilização da descoberta de conhecimento através da Mineração de Dados (MD), conhecida como Mineração de Dados Educacionais (em inglês, *Educational Data Mining* - EDM) (JÚNIOR et al., 2019). Esse área de estudo é 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 (ex: socioeconômicos, acadêmicos, demográficos, pessoais etc.). Com os resultados obtidos pela EDM, professores e gestores escolares podem melhorar ou criar ações pedagógicas e administrativas dentro/fora das salas de aula direcionadas à permanência e ao êxito estudantil.

Deste modo, para atuar na área de EDM é fundamental desvendar, por exemplo, as principais ferramentas/bibliotecas, algoritmos e bases de dados para aplicar a MD de forma adequada no contexto da evasão escolar. Sendo assim, um primeiro

passo importante que pode ser dado para acelerar esse processo de aprendizado é identificando, avaliando e interpretando toda a pesquisa já disponível na temática.

Mediante a esse contexto, este trabalho propõe uma análise de trabalhos nacionais e internacionais de EDM que tratam da temática da evasão escolar. Considerando essa temática, os principais objetivos a serem alcançados neste trabalho são: identificar os principais 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); identificar os níveis de escolaridade mais analisados nos trabalhos; dentre outros.

Para alcançar tais objetivos, a metodologia da proposta foi dividida em três etapas principais. 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 (versão *Desktop* 2021.1) 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.

O restante deste trabalho está organizado da seguinte maneira: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 apresenta a proposta desta pesquisa; a Seção 4 apresenta os resultados obtidos; por fim, a Seção 5 expõe a conclusão desta pesquisa e os direcionamentos para trabalhos futuros.

2 TRABALHOS RELACIONADOS

Essa seção apresenta trabalhos atuais que realizam revisão da literatura focados na EDM e na evasão de estudantes.

Em (MARQUES et al., 2019), os autores apresentam um mapeamento sistemático da literatura sobre a evasão escolar, buscando identificar as tecnologias de MD utilizadas e os fatores que causam a evasão estudantil. Os portais de trabalhos científicos utilizados para as pesquisas foram: *ACM Digital Library*, *IEEE Xplore*, *Science Direct* e *Scopus*. Com o resultado da busca, foram selecionados 14 artigos. Essa pesquisa tinha o objetivo de responder ao seguinte questionamento proposto pelos autores: "*Quais ferramentas, técnicas e fatores indutores vêm sendo utilizados para desvendar possíveis causas da evasão escolar?*". Dentre os resultados obtidos nesse trabalho, os autores constataram que a ferramenta *Waikato Environment for Knowledge*

Analysis (*Weka*) foi a mais utilizada para facilitar a descoberta das causas da evasão escolar. Eles observaram também que as técnicas de classificação se destacam dentre as demais. Por fim, o mapeamento identificou que os principais trabalhos se concentram em estudar fatores indutores associados às características individuais dos alunos. Apesar de os autores serem brasileiros, nenhum trabalho nacional foi analisado na revisão sistemática.

No trabalho de (ALBAN; MAURICIO, 2019) é proposta uma revisão sistemática sobre a predição da evasão de estudantes universitários através de técnicas de MD. O estudo foi desenvolvido utilizando diversas bibliotecas digitais, dentre elas: *IEEE Xplore, ACM Digital Library, Science Direct, Springer*, *Directory of Open Access Journals* (DOAJ), *Taylor and Francis, Emerald, Proquest* e *Ebsco*. As buscas por trabalhos relevantes resultaram em 67 artigos nos critérios de inclusão e exclusão estabelecidos. Essa pesquisa visou responder cinco questionamentos feitos pelos autores. Dentre os resultados obtidos, foi possível afirmar que as ferramentas *Waikato Environment for Knowledge Analysis* (*Weka*) e *Statistical Package for the Social Sciences (SPSS)* foram as mais utilizadas dentre os trabalhos analisados. Em relação às técnicas de MD, foram encontradas 14 técnicas de classificação, sendo que as mais utilizadas foram árvore de decisão, seguida por regressão logística, redes neurais e máquina de vetores de suporte. Além disso, os fatores indutores investigados nos trabalhos resultaram em 112 fatores que influenciam em uma possível evasão. Esses fatores foram classificados em categorias como: fatores pessoais, acadêmico, econômico, social e institucional.

Em (AGRUSTI; BONAVOLONTÀ; MEZZINI, 2019) é feita uma revisão sistemática com foco na investigação de estudos que utilizam técnicas de mineração de dados educacionais para prever o abandono de universitários em cursos tradicionais. As fontes de dados usadas para escolha de trabalhos relevantes foram as bibliotecas digitais do *Scopus* e *Web of Science* (*WoS*). Nas pesquisas foram selecionados 73 trabalhos após a aplicação dos critérios de inclusão e exclusão pré-estabelecidos pelos autores. No trabalho, foram identificadas 6 técnicas de classificação, 53 algoritmos e 14 ferramentas de MD. Concluiu-se que as técnicas de MD mais utilizadas foram árvore de decisão, seguida por classificação bayesiana, redes neurais e regressão logística. Dentre as 14 ferramentas de MD investigadas, destacaram-se: *Waikato Environment for Knowledge Analysis* (*Weka*), *Statistical Package for the Social Sciences (SPSS)* e *R*.

A Tabela 1 apresenta um comparativo entre os trabalhos relacionados, além de um comparativo com este trabalho. Um grande diferencial do presente trabalho é abordar trabalhos de EDM na temática da evasão no cenário brasileiro, além do cenário internacional. Para tanto, foram selecionados artigos tanto em idioma inglês quanto português. Além disso, o período considerado na busca é amplo e atual (o único que considera o ano de 2020). Em relação aos portais científicos consultados, este trabalho utilizou 6, dentre os quais 2 são portais brasileiros. Os outros trabalhos não usam nenhuma base brasileira. Por fim, ao contrário dos outros artigos, neste

trabalho é utilizada uma ferramenta de análise de dados para uma melhor compreensão e comparação dos dados, gerando variadas visualizações amigáveis.

Tabela 1 – Comparativo dos Trabalhos Relacionados.

Trabalho	**Período Considerado na busca**	**Qtd de trabalhos**	**Idioma dos trabalhos analisados**	**Qtd de Portais Científicos utilizados**	**Utiliza Ferramenta de Análise Visual de Dados**
(MARQUES et al., 2019)	2008 - 2018	14	Inglês	4	Não
(ALBAN; MAURICIO, 2019)	2006 - 2017	67	Inglês	9	Não
(AGRUSTI; BONAVOLONTÀ; MEZZINI, 2019)	1999 - 2019	73	Inglês	2	Não
Este trabalho	2008 - 2020	50	Inglês Português	6	Sim

Fonte: Elaborada pelos autores (2021).

3 PROPOSTA

Neste trabalho é proposta uma análise de trabalhos nacionais e internacionais de EDM que tratam da temática da evasão escolar. A metodologia adotada para implementação da proposta é detalhada nesta seção, tendo sido dividida em três etapas principais, conforme apresentado na sequência.

3.1 Etapa 1 - Planejamento da Revisão da Literatura

3.1.1 Questões de Pesquisa

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

- **QP1**: *Quais são os trabalhos científicos nacionais e internacionais que utilizam EDM no contexto da evasão escolar?*
- **QP2**: *Como estão temporalmente distribuídos os trabalhos?*
- **QP3**: *Quais as ferramentas/bibliotecas utilizadas pelos trabalhos?*
- **QP4**: *Quais os algoritmos de aprendizagem de máquina utilizados nos trabalhos?*
- **QP5**: *Quais os algoritmos com melhor desempenho?*
- **QP6**: *Qual o nível de escolaridade considerado nos trabalhos?*
- **QP7**: *Quantos registros existem nas bases de dados analisadas?*

- **QP8**: *Quantos atributos estão presentes nas bases de dados analisadas?*
- **QP9**: *Quais tipos de dados são analisados nos trabalhos selecionados?*

3.1.2 Portais Científicos

A seguir, partiu-se para a busca de trabalhos científicos capazes de responder às questões de pesquisa. As buscas realizadas para esta revisão de literatura foram feitas nos principais portais científicos (bibliotecas digitais) internacionais e nacionais.

Os portais científicos internacionais considerados foram:

- ACM Digital Library[1];
- IEEE Xplore[2];
- ScienceDirect[3];
- SpringerLink[4].

Os portais científicos nacionais considerados foram:

- Sociedade Brasileira de Computação (SBC) OpenLib[5];
- Revista Brasileira de Informática na Educação (RBIE)[6].

3.1.3 Protocolos de Busca

Como cada repositório possui sua própria sintaxe, a Tabela 2 mostra o protocolo de busca utilizado para cada motor de busca.

3.1.4 Critérios de Inclusão e Exclusão

Os critérios de inclusão e exclusão são definidos para auxiliar na condução da pesquisa, com o intuito de apoiar a classificação de relevância dos estudos (FUZETO; BRAGA, 2016). Os critérios estão familiarmente relacionados às questões de pesquisa em análise.

3.1.4.1 Critérios de inclusão

- Trabalhos nacionais e internacionais que utilizam Mineração de Dados (Educacional) com ênfase na evasão escolar em qualquer nível de ensino.

[1] https://dl.acm.org/
[2] https://IEEE Xplore.ieee.org/
[3] https://www.sciencedirect.com/
[4] https://link.springer.com/
[5] https://sol.sbc.org.br/
[6] https://www.br-ie.org/

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

Portais Cientificos	**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

Fonte: Elaborada pelos autores (2021).

3.1.4.2 Critérios de exclusão

- Trabalhos não escritos em inglês ou português;
- Trabalhos publicados antes de 2008;
- Trabalhos em andamento, que não tenham sido concluídos;
- Trabalhos duplicados;
- Estudos fora do contexto desta pesquisa.

3.1.5 Trabalhos Selecionados

A Tabela 3 apresenta os 50 trabalhos selecionados resultantes das buscas realizadas. Os trabalhos são apresentados em ordem cronológica de publicação. Vale ressaltar que nesta pesquisa buscou-se por trabalhos entre os anos de 2008 e 2020. Dessa forma, trabalhos publicados fora deste intervalo não foram analisados. Dentre os 50 trabalhos, 22 são nacionais e 28 internacionais.

Tabela 3 – Trabalhos Selecionados.

ID	**Referências**
E1	(MANSUR, 2008)
E2	(DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009)
E3	(OYELADE; OLADIPUPO; OBAGBUWA, 2010)
E4	(PAL, 2012)
E5	(MARQUEZ-VERA; MORALES; SOTO, 2013)
E6	(MANHÃES et al., 2012)
E7	(COSTA; CAZELLA; RIGO, 2014)
E8	(JAMESMANOHARAN et al., 2014)
E9	(YUKSELTURK; OZEKES; TUREL, 2014)
E10	(BRITO et al., 2015)
E11	(FONSECA, 2015)
E12	(PRADEEP; DAS; KIZHEKKETHOTTAM, 2015)
E13	(SANTANA et al., 2015)
E14	(BARBOSA, 2016)
E15	(MARBOUTI; DIEFES-DUX; MADHAVAN, 2016)
E16	(MEEDECH; IAM-ON; BOONGOEN, 2016)
E17	(PEREIRA, 2016)
E18	(AHUJA; KANKANE, 2017)
E19	(ARAÚJO, 2017)
E20	(PAZ; CAZELLA, 2017)
E21	(PEREIRA; ZAMBRANO, 2017)
E22	(ROCHA et al., 2017)
E23	(RODRIGUEZ-MAYA et al., 2017)
E24	(ROVIRA; PUERTAS; IGUAL, 2017)
E25	(ADIL; TAHIR; MAQSOOD, 2018)
E26	(ALCÂNTARA, 2018)
E27	(DHARMAWAN; GINARDI; MUNIF, 2018)
E28	(GONÇALVES; SILVA; CORTES, 2018)
E29	(HEGDE; PRAGEETH, 2018)
E30	(LIMSATHITWONG; TIWATTHANONT; YATSUNGNOEN, 2018)
E31	(MURAKAMI et al., 2018)
E32	(PEREZ; CASTELLANOS; CORREAL, 2018)
E33	(RODRIGUES, 2018)
E34	(SOLIS et al., 2018)
E35	(ALBAN; MAURICIO, 2019)
E36	(FERRERO, 2019)

E37	(GONÇALVES; BELTRAME, 2019)
E38	(LI; GOU; FAN, 2019)
E39	(NETO, 2019)
E40	(SARAIVA et al., 2019)
E41	(SOUTO, 2019)
E42	(VALENTIM, 2019)
E43	(BARROS et al., 2020)
E44	(FILHO; SIQUEIRA; LEAL, 2020)
E45	(LOTTERING; HANS; LALL, 2020)
E46	(SANTOS et al., 2020)
E47	(SOARES et al., 2020)
E48	(UTARI; WARSITO; KUSUMANINGRUM, 2020)
E49	(VILORIA et al., 2020)
E50	(YAACOB et al., 2020)

Fonte: Elaborada pelos autores (2021).

3.2 Etapa 2 - Preparação dos Dados

Cada questão desta pesquisa motiva a extração dos dados. Assim, foi fundamental realizar a leitura completa de todos os artigos e analisá-los. Consequentemente, foram extraídas informações gerais, como: ano de publicação, nome da ferramenta, linguagem de programação, algoritmos usados e com melhor desempenho, quantidade de registros e atributos, tipos de dados e nível de escolaridade considerado.

Inicialmente, para a preparação dos dados utilizou-se a ferramenta *MS Excel* para armazenar todas as informações retiradas dos artigos analisados. De tal modo, foi possível criar um conjunto de dados com informações consideradas valiosas para a análise.

Dada a vasta quantidade de informações extraídas dos trabalhos selecionados, optou-se por criar 3 planilhas com os dados. Cada planilha foi montada de forma diferente, de modo a responder a questões específicas. A Tabela 4 resume a característica/conteúdo de cada planilha e as questões de pesquisa que responde.

Tabela 4 – Visão geral das planilhas criadas na etapa de preparação dos dados.

Planilha	**Característica**	**Questão de Pesquisa (QP)**
P1	Todos os dados extraídos	QP1, QP2, QP3, QP5, QP6, QP7 e QP8
P2	Apenas algoritmos usados nos trabalhos selecionados	QP4
P3	Tipo das bases de dados (Acadêmico, Socioeconômicos etc.)	QP9

Fonte: Elaborada pelos autores (2021).

A planilha P1, apresentada parcialmente na Figura 1, é composta por 11 colunas referentes às informações extraídas nos artigos estudados, sendo elas:

- **Autores:** indica a referência bibliográfica do trabalho;
- **Ano:** indica o ano de publicação do trabalho;
- **Cenário:** indica se o trabalho considera uma base de dados nacional ou internacional;
- **Ferramentas/Bibliotecas:** indica a ferramenta ou biblioteca utilizada no trabalho;
- **Linguagens:** indica a linguagem de programação utilizada no desenvolvimento da ferramenta ou biblioteca;
- **Melhor algoritmo:** indica o algoritmo que obteve o melhor desempenho;
- **Instâncias:** indica a quantidade de registros presentes na base de dados utilizada no trabalho;
- **Tamanho Instância:** indica o intervalo que a quantidade de registros se enquadra;
- **Quantidade de atributos:** indica a quantidade de atributos presente na base de dados utilizada no trabalho;
- **Atributos:** indica o intervalo que a quantidade de atributos se enquadra;
- **Modalidade de Ensino:** indica os níveis de ensino analisados no trabalho.

Figura 1 – Visão parcial da Planilha 1.

	C	D	E	G	H	I	L	N	O	P	Q
1	**Autores**	**Ano**	**Cenário**	**Ferramentas/ Bibliotecas**	**Linguagens**	**Melhor Algoritmo**	**Instâncias**	**Tamanho Instância**	**Quant. de atributos**	**Atributos**	**Modalidade de Ensino**
2	SARAIVA et al., 2019	2019	Nacional	Scikit-learn	Python	SVM (Support Vector Machine)	500	De 100 a 500	10	Até 10 atributos	Curso Técnico
3	PAZ; CAZELLA, 2017	2017	Nacional	WEKA	Java	Decision Tree	4.601	De 1000 a 5000	7	Até 10 atributos	Graduação
4	ALCâNTARA, 2018	2018	Nacional	WEKA	Java	Decision Tree	916	De 500 a 1000	8	Até 10 atributos	Graduação
5	PEREIRA, 2016	2016	Nacional	Genie	Python	Naive Bayes	666	De 500 a 1000	17	Entre 10 e 20 atributos	Curso Técnico
6	GONÇALVES; SILVA; CORTES, 2018	2018	Nacional	Weka	Java	Decision Tree	574	De 500 a 1000	40	Entre 30 e 40 atributos	Graduação
7	FILHO; SIQUEIRA; LEAL, 2020	2020	Nacional	Scikit-learn	Python	SVM (Support Vector Machine)	1.942	De 1000 a 5000	8	Até 10 atributos	Curso Técnico e graduação

Fonte: Elaborada pelos autores (2021).

A planilha P2, apresentada parcialmente na Figura 2, possui 6 colunas referentes aos algoritmos utilizados nos trabalhos estudados. Essa planilha será utilizada para

indicar os algoritmos usados nos trabalhos. No entanto, servirá também para exibir o algoritmo com melhor desempenho. Sendo assim, as 6 colunas da P2 são:

- **Autores:** indica a referência bibliográfica do trabalho;
- **Ano:** indica o ano de publicação do trabalho;
- **Algoritmos:** indica todos os algoritmos utilizados no trabalho;
- **Algoritmos abreviados:** indica a abreviação do algoritmo;
- **Tipo de algoritmo:** indica a classificação do algoritmo;
- **Cenário:** indica se o trabalho usa uma base de dados nacional ou internacional.

Figura 2 – Visão parcial da Planilha 2.

	A	B	D	E	F	G
1	Autores	Ano	Algoritmos	Algoritmos Abreviados	Tipo de Algoritmos	Cenário
2	SARAIVA et al., 2019	2019	Support Vector Machine	SVM	SUPPORT VECTOR MACHINE ALGORITHMS	Nacional
3	SARAIVA et al., 2019	2019	Naive Bayes	NB	BAYESIAN CLASSIFICATION ALGORITHMS	Nacional
4	SARAIVA et al., 2019	2019	Decision Tree	DT	DECISION TREE ALGORITHMS	Nacional
5	SARAIVA et al., 2019	2019	Random Forest	RF	DECISION TREE ALGORITHMS	Nacional
6	SARAIVA et al., 2019	2019	K-nearest Neighbors Algorithm	KNN	NEAREST NEIGHBORS	Nacional
7	SARAIVA et al., 2019	2019	Neural Network	NN	NEURAL NETWORK ALGORITHMS	Nacional
8	PAZ; CAZELLA, 2017	2017	Decision Tree	DT	DECISION TREE ALGORITHMS	Nacional
9	ALCâNTARA, 2018	2018	Decision Tree	DT	DECISION TREE ALGORITHMS	Nacional
10	PEREIRA, 2016	2016	Naive Bayes	NB	BAYESIAN CLASSIFICATION ALGORITHMS	Nacional
11	GONÇALVES; SILVA; CORTES, 2018	2018	Naive Bayes	NB	BAYESIAN CLASSIFICATION ALGORITHMS	Nacional
12	GONÇALVES; SILVA; CORTES, 2018	2018	Support Vector Machine	SVM	SUPPORT VECTOR MACHINE ALGORITHMS	Nacional
13	GONÇALVES; SILVA; CORTES, 2018	2018	Decision Tree	DT	DECISION TREE ALGORITHMS	Nacional
14	FILHO; SIQUEIRA; LEAL, 2020	2020	Decision Tree	DT	DECISION TREE ALGORITHMS	Nacional
15	FILHO; SIQUEIRA; LEAL, 2020	2020	Naive Bayes	NB	BAYESIAN CLASSIFICATION ALGORITHMS	Nacional
16	FILHO; SIQUEIRA; LEAL, 2020	2020	K-nearest Neighbors Algorithm	KNN	NEAREST NEIGHBORS	Nacional
17	FILHO; SIQUEIRA; LEAL, 2020	2020	Support Vector Machine	SVM	SUPPORT VECTOR MACHINE ALGORITHMS	Nacional
18	FILHO; SIQUEIRA; LEAL, 2020	2020	GradientBoosting	GB	MISCELLANEA ALGORITHMS	Nacional

Fonte: Elaborada pelos autores (2021).

Por fim, a planilha P3 tem apenas 3 colunas relacionadas aos tipos de bases de dados analisados pelos trabalhos selecionados. Essa planilha pode conter repetições de trabalho para indicar que foi utilizada mais de um tipo de base. Sendo assim, se planilha exibe repetidamente o mesmo trabalho, significa que cada linha possui uma base de dados diferente. Portanto, na planilha P3 temos:

- **Autores:** indica a referência bibliográfica do trabalho;
- **Ano:** indica o ano de publicação do trabalho;
- **Tipo de dados:** indica o tipo de base de dados analisada pelo trabalho.

Figura 3 – Visão parcial da Planilha 3.

	C	D	L
1	**Autores**	**Ano**	**Tipo de dados**
2	SARAIVAet al., 2019	2019	Socioeconômicos
3	SARAIVAet al., 2019	2019	Acadêmicos
4	PAZ; CAZELLA, 2017	2017	Acadêmicos
5	ALCâNTARA, 2018	2018	Acadêmicos
6	PEREIRA, 2016	2016	Acadêmicos
7	PEREIRA, 2016	2016	Socioeconômicos
8	GONÇALVES; SILVA; CORTES, 2018	2018	Acadêmicos
9	FILHO; SIQUEIRA; LEAL, 2020	2020	Acadêmicos
10	SOARES et al., 2020	2020	Acadêmicos
11	BRITO et al., 2015	2015	Acadêmicos
12	GONçALVES; BELTRAME, 2019	2019	Socioeconômicos
13	BARBOSA, 2016	2016	Socioeconômicos
14	FERRERO, 2019	2019	Acadêmicos

Fonte: Elaborada pelos autores (2021).

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. Existem diversas ferramentas capazes de criar visualizações intuitivas e interativas.

Neste trabalho foi utilizado o Tableau, uma ferramenta de *Business Intelligence* (BI) usada para criar visões de dados variadas de forma simples. O Tableau provem das melhores práticas de visualização, assim pode recomendar excelentes gráficos, disposição de informações, cores, criação de filtros, imagens, painéis interativos, entre outras funcionalidades que facilitam a exploração e a análise de dados. Sua principal tecnologia é o *VizQL*, que se baseia em ações *Drag and Drop* para gerar análises de dados completas através de uma interface essencialmente visual.

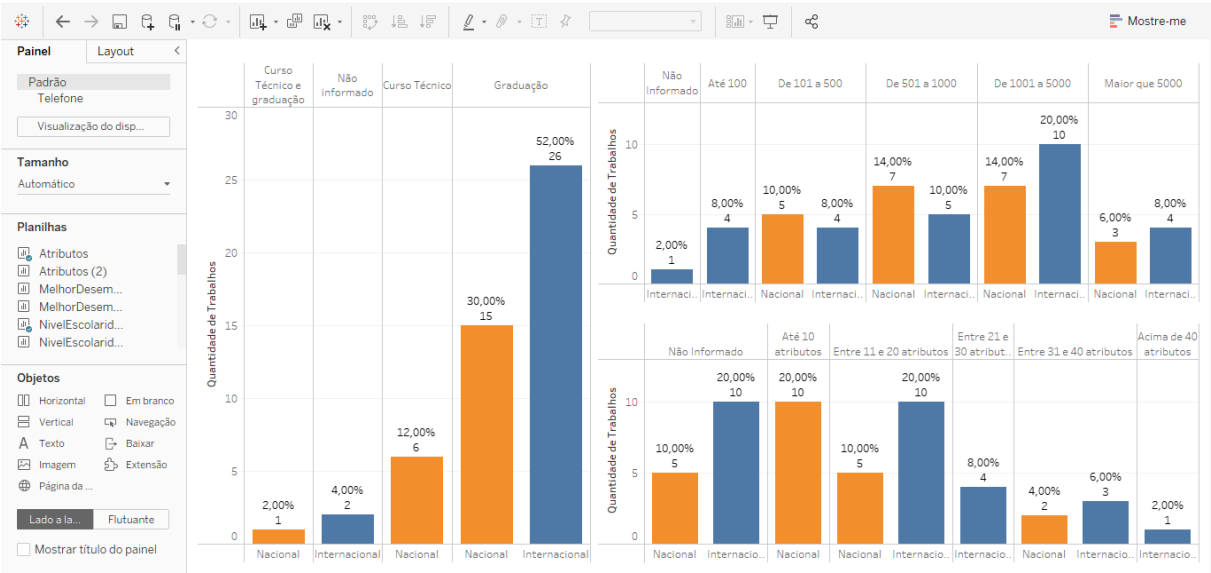
Em 2021, o Tableau completou 9 anos consecutivos como líder no Quadrante Mágico do *Gartner Group* para *Analytics and Business Intelligence Platforms*[7]. O Tableau Desktop 2021.1 (versão para universitários) foi utilizado neste trabalho.

A Figura 4 ilustra um exemplo do painel de trabalho concedido pelo Tableau Desktop para interação com as planilhas criadas. No caso específico da figura, o painel apresenta os níveis de escolaridade, o tamanho das instâncias e atributos.

O Tableau proporciona como entrada diferentes tipos de arquivos, como *Comma-separated values* (CSV), arquivos Excel (xlsx, xls), PDFs, arquivos de texto (txt), JSON e variados tipos de fonte de dados. Todas as três planilhas criadas para esta pesquisa

[7] https://www.tableau.com/reports/gartner

Figura 4 – Exemplo de painel de trabalho do Tableau.



Fonte: Elaborada pelos autores no Tableau (2021).

(Tabela 4) têm como extensão .xlsx e serviram como entrada para o Tableau.

Nesse contexto, a última etapa da metodologia consistiu na utilização do Tableau para a criação das visualizações no contexto das Questões de Pesquisa (QP) definidas na Seção 3.1.1. Na próxima seção são apresentadas e discutidas as respostas para as 9 QP.

4 RESULTADOS

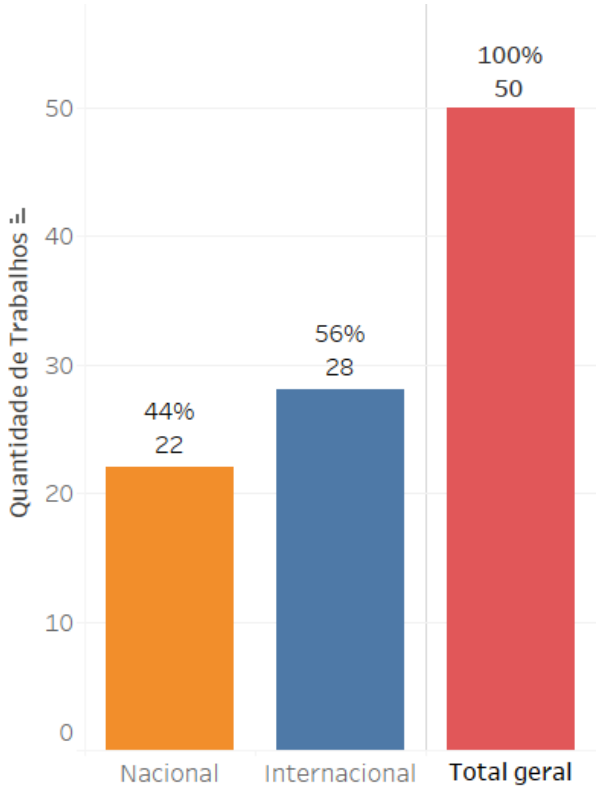
4.1 QP1: *Quais são os trabalhos científicos nacionais e internacionais que utilizam EDM no contexto da evasão escolar?*

Como dito na Seção 3.1.5, obteve-se 50 trabalhos selecionados resultantes das buscas realizadas seguindo a metodologia descrita nas Seções 3.1.1 a 3.1.4. Vale relembrar que nesta pesquisa buscou-se por trabalhos entre os anos de 2008 e 2020, conforme definido nos *critérios de exclusão* da Seção 3.1.4.

Primeiramente, a Figura 5 apresenta como os 50 trabalhos selecionados estão divididos segundo o cenário (nacional ou internacional). Assim, a figura mostra que dentre os 50 trabalhos selecionados, 22 são nacionais e 28 internacionais. De forma geral, pode-se dizer que há um equilíbrio entre as quantidades de trabalhos encontrados/selecionados para cada cenário.

Como resposta direta para a QP1, a Figura 6 apresenta a lista dos 50 trabalhos selecionados em ordem alfabética e o respectivo cenário de aplicação (nacional ou internacional) de cada um.

Figura 5 – Total de trabalhos selecionados por cenário.



Fonte: Elaborada pelos autores no Tableau (2021).

4.2 QP2: *Como estão temporalmente distribuídos os trabalhos?*

A Figura 7 apresenta a quantidade de trabalhos publicados em função do ano de publicação. Considerando os 50 trabalhos selecionados, a figura sugere que do ano 2008 a 2012 os estudos acerca dessa temática eram pouco relevantes. A partir do ano de 2013 pode-se notar um crescimento significativo do número de trabalhos na área de EDM com ênfase na evasão escolar. Merece um destaque especial o ano de 2018 com um total de 10 trabalhos publicados.

Porém, ao separarmos os trabalhos por tipo de cenário (Figura 8), pode-se observar que no cenário internacional a temática começou a ter uma maior quantidade de trabalhos publicados a partir de 2017, mas sofreu uma queda brusca de 2018 para 2019. Já no cenário nacional, percebe-se que houve um maior interesse pela temática somente nos últimos anos.

4.3 QP3: *Quais as ferramentas/bibliotecas utilizadas pelos trabalhos?*

Para responder a QP3, a Figura 9 apresenta as ferramentas/bibliotecas em função da quantidade de trabalhos que as utilizam.

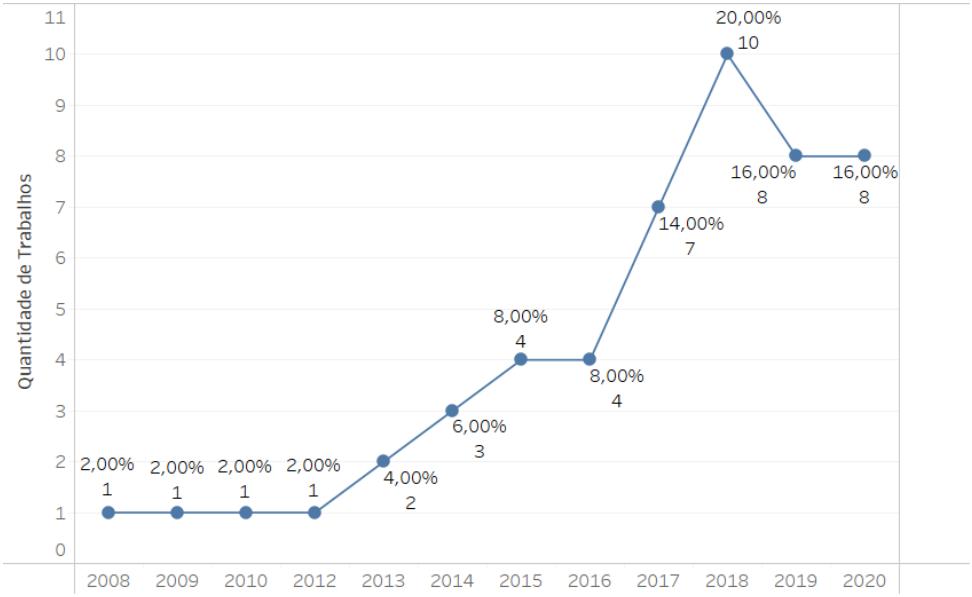
Como respondido na QP1, foram selecionados 50 trabalhos. Porém, desde já é importante dizer que os autores de 4 deles não informaram a ferramenta/biblioteca utilizada na pesquisa, conforme indicado na Figura 9.

Figura 6 – Lista de Trabalhos Selecionados.

Autores	Cenário
ADIL; TAHIR; MAQSOOD, 2018	Internacional
AHUJA; KANKANE, 2017	Internacional
ALBÁN; MAURICIO, 2019	Internacional
ALCÂNTARA, 2018	Nacional
ARAÚJO, 2017	Nacional
BARBOSA, 2016	Nacional
BARROS et al., 2020	Nacional
BRITO et al., 2015	Nacional
COSTA; CAZELLA; RIGO, 2014	Nacional
DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009	Internacional
DHARMAWAN; GINARDI; MUNIF, 2018	Internacional
FERRERO, 2019	Nacional
FILHO; SIQUEIRA; LEAL, 2020	Nacional
FONSECA, 2015	Nacional
GONÇALVES; BELTRAME, 2019	Nacional
GONÇALVES; SILVA; CORTES, 2018	Nacional
HEGDE; PRAGEETH, 2018	Internacional
JAMESMANOHARAN et al., 2014	Internacional
LI; GOU; FAN, 2019	Internacional
LIMSATHITWONG; TIWATTHANONT; YATSUNGNOEN, 2018	Internacional
LOTTERING; HANS; LALL, 2020	Internacional
MANSUR, 2008	Nacional
MARBOUTI; DIEFES-DUX; MADHAVAN, 2016	Internacional
MARQUEZ-VERA; MORALES; SOTO, 2013	Internacional
MEEDECH; IAM-ON; BOONGOEN, 2016	Internacional
MURAKAMI et al., 2018	Internacional
NETO, 2019	Internacional
OYELADE; OLADIPUPO; OBAGBUWA, 2010	Internacional
PAL, 2012	Internacional
PAZ; CAZELLA, 2017	Nacional
PEREIRA, 2016	Nacional
PEREIRA; ZAMBRANO, 2017	Internacional
PEREZ; CASTELLANOS; CORREAL, 2018	Internacional
PRADEEP; DAS; KIZHEKKETHOTTAM, 2015	Internacional
ROCHA et al., 2017	Internacional
RODRIGUES, 2018	Nacional
RODRIGUEZ-MAYA et al., 2017	Internacional
ROVIRA; PUERTAS; IGUAL, 2017	Internacional
SANTANA et al., 2015	Nacional
SANTOS et al., 2020	Nacional
SARAIVA et al., 2019	Nacional
SOARES et al., 2020	Nacional
SOLIS et al., 2018	Internacional
SOUTO, 2019	Nacional
UTARI; WARSITO; KUSUMANINGRUM, 2020	Internacional
VALENTIM, 2019	Nacional
VILORIA et al., 2020	Internacional
YAACOB et al., 2020	Internacional
YUKSELTURK; OZEKES; TUREL, 2014	Internacional
ZIMBRÃO, 2012	Nacional

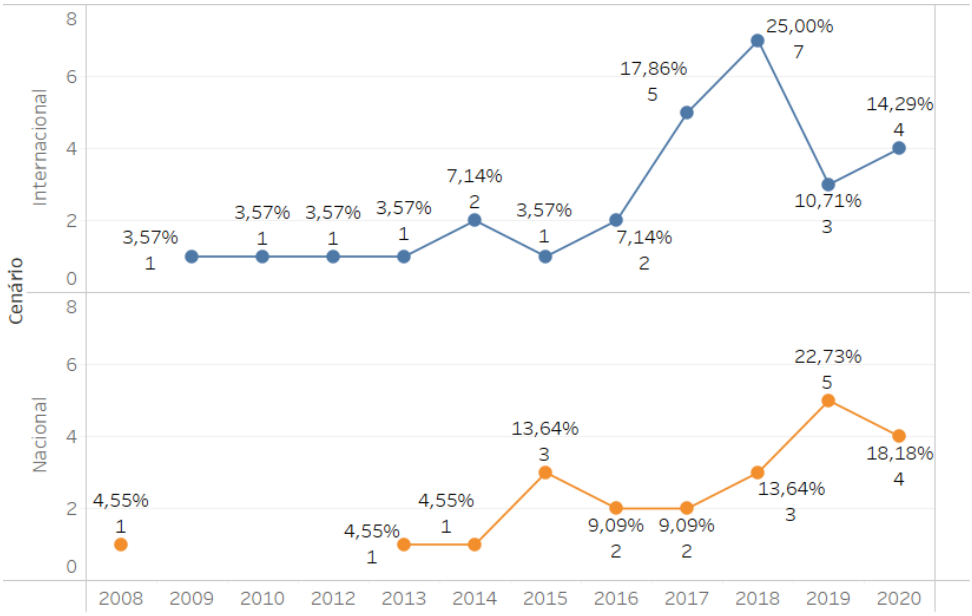
Fonte: Elaborada pelos autores no Tableau (2021).

Figura 7 – Quantidade de trabalhos publicados por Ano.



Fonte: Elaborada pelos autores no Tableau (2021).

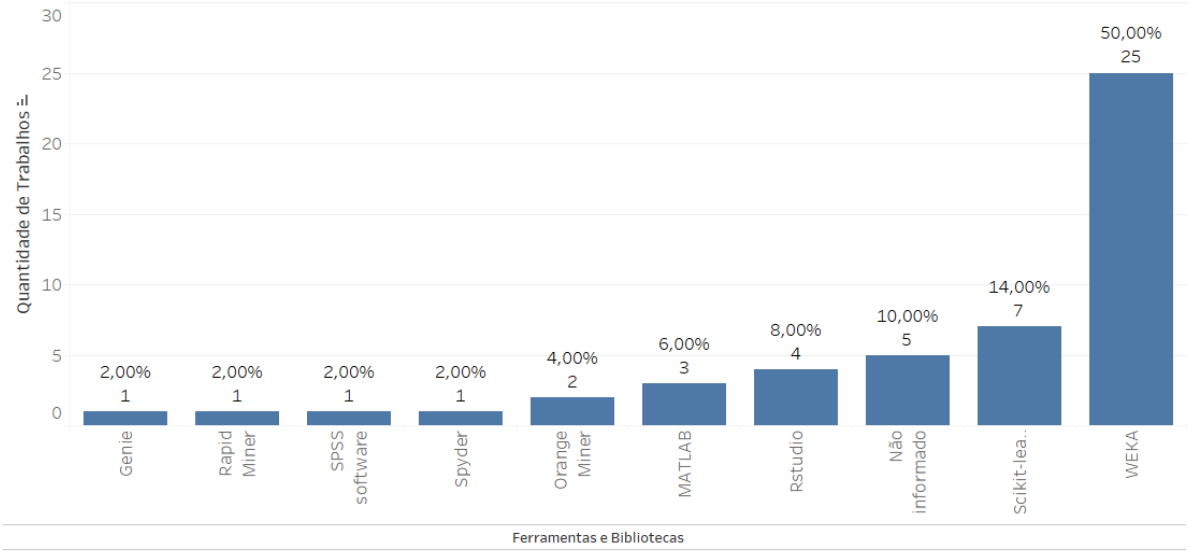
Figura 8 – Quantidade de trabalhos publicados por Ano e Cenário.



Fonte: Elaborada pelos autores no Tableau (2021).

Assim, dentre os 46 trabalhos, foram identificadas 10 ferramentas/bibliotecas sendo utilizadas para MD no contexto da evasão escolar. De imediato, é possível ver pela figura que a ferramenta *Waikato Enviroment for Knowledge Analysis* (*Weka*) se destaca entre as demais, tendo sido utilizada em 25 trabalhos. A ferramenta *WEKA* é um software livre, desenvolvido em Java, que contem um conjunto de implementações de algoritmos de diversas técnicas de Mineração de Dados (WAIKATO, 2010). O *Scikit-learn* aparece em segundo lugar, tendo sido utilizado em 7 trabalhos. O *Scikit-learn* é uma biblioteca de aprendizagem de máquina de código aberto para a linguagem de programação *Python*.

Figura 9 – Ferramentas/bibliotecas utilizadas nos Trabalhos.



Fonte: Elaborada pelos autores no Tableau (2021).

Para responder a QP3 de forma detalhada, foram acrescentadas as visões das Figuras 10 e 11 com detalhes do uso das ferramentas/bibliotecas de MD nos trabalhos selecionados nos cenários nacional e internacional, respectivamente. Estes trabalhos estão ordenados por ano de publicação nas Figuras 10 e 11.

4.3.1 Cenário Nacional

No total, foram identificadas 6 ferramentas/bibliotecas usadas pelos trabalhos no cenário nacional. De imediato, é possível verificar na Figura 10 que a ferramenta *Weka* e a biblioteca *Scikit-learn* destacam-se dentre as demais, pois 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 a ferramenta *Weka* tem um destaque de 2008 a 2018, sendo usada em praticamente todos os estudos do intervalo considerado. Porém, a partir de 2019 utilizou-se com mais frequência o *Scikit-learn*.

Figura 10 – Ferramentas e bibliotecas utilizadas no cenário nacional

Autores	Ano	Linguagens	Genie	Orange Miner	Rstudio	Scikit-learn	Spyder	WEKA
MANSUR, 2008	2008	Java						1
ZIMBRÃO, 2012	2013	Java						1
COSTA; CAZELLA; RIGO, 2014	2014	Java						1
BRITO et al., 2015	2015	Java						1
FONSECA, 2015	2015	Java						1
SANTANA et al., 2015	2015	Java						1
BARBOSA, 2016	2016	Java						1
PEREIRA, 2016	2016	Python	1					
ARAÚJO, 2017	2017	Java						1
PAZ; CAZELLA, 2017	2017	Java						1
ALCÂNTARA, 2018	2018	Java						1
GONÇALVES; SILVA; CORTES, 2018	2018	Java						1
RODRIGUES, 2018	2018	R			1			
FERRERO, 2019	2019	Python				1		
GONÇALVES; BELTRAME, 2019	2019	Python		1				
SARAIVA et al., 2019	2019	Python				1		
SOUTO, 2019	2019	Java						1
VALENTIM, 2019	2019	Java						1
BARROS et al., 2020	2020	Python				1		
FILHO; SIQUEIRA; LEAL, 2020	2020	Python				1		
SANTOS et al., 2020	2020	Python				1		
SOARES et al., 2020	2020	Python					1	
Total geral			1	1	1	5	1	13

Fonte: Elaborada pelos autores no Tableau (2021).

4.3.2 Cenário Internacional

A Figura 11 apresenta as ferramentas/bibliotecas utilizadas nos trabalhos de cenário internacional. Os 4 trabalhos que não informaram a ferramenta/biblioteca estão nesse cenário.

Na Figura 11 percebe-se também que neste cenário internacional há uma grande diversificação de ferramentas/bibliotecas, tendo sido identificadas 7 no total, um número um pouco maior que o total identificado no cenário nacional (6).

Novamente, a ferramenta *Weka* prevalece nas pesquisas. Dentre os 24 trabalhos considerados, 12 utilizam o *Weka*. A ferramenta *MATLAB* é utilizada em 3 trabalhos, assim como o *Rstudio*. Na sequência, a biblioteca *Scikit-learn* aparece sendo utilizada 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.

Figura 11 – Ferramentas e bibliotecas utilizadas no cenário internacional.

Autores	A..	Linguagens	MATLAB	Não informado	Orange Miner	Rapid Miner	Rstudio	Scikit-learn	SPSS software	WEKA
DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009	2009	Java								1
OYELADE; OLADIPUPO; OBAGBUWA, 2010	2010	Não informado		1						
PAL, 2012	2012	Java								1
MARQUEZ-VERA; MORALES; SOTO, 2013	2013	Java								1
JAMESMANOHARAN et al., 2014	2014	Não informado		1						
YUKSELTURK; OZEKES; TUREL, 2014	2014	Java, C	1							
PRADEEP; DAS; KIZHEKKETHOTTAM, 2015	2015	Java								1
MARBOUTI; DIEFES-DUX; MADHAVAN, 2016	2016	Java, C	1							
MEEDECH; IAM-ON; BOONGOEN, 2016	2016	Java								1
AHUJA; KANKANE, 2017	2017	R					1			
PEREIRA; ZAMBRANO, 2017	2017	Java								1
ROCHA et al., 2017	2017	Java								1
RODRIGUEZ-MAYA et al., 2017	2017	Java								1
ROVIRA; PUERTAS; IGUAL, 2017	2017	Python						1		
ADIL; TAHIR; MAQSOOD, 2018	2018	Java, C	1							
DHARMAWAN; GINARDI; MUNIF, 2018	2018	Não informado		1						
HEGDE; PRAGEETH, 2018	2018	Java								1
LIMSATHITWONG; TIWATTHANONT; YATSUNG..	2018	Java				1				
MURAKAMI et al., 2018	2018	Python						1		
PEREZ; CASTELLANOS; CORREAL, 2018	2018	Não informado		1						
SOLIS et al., 2018	2018	R					1			
ALBÁN; MAURICIO, 2019	2019	Java							1	
LI; GOU; FAN, 2019	2019	Não informado		1						
NETO, 2019	2019	Java								1
LOTTERING; HANS; LALL, 2020	2020	R					1			
UTARI; WARSITO; KUSUMANINGRUM, 2020	2020	Java								1
VILORIA et al., 2020	2020	Java								1
YAACOB et al., 2020	2020	Python			1					
Total geral			3	5	1	1	3	2	1	12

Fonte: Elaborada pelos autores no Tableau (2021).

4.4 QP4: *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 autores dos trabalhos selecionados aplicaram bem mais que um algoritmo em seus experimentos. Por exemplo, somente no trabalho (GONÇALVES; BELTRAME, 2019), 11 algoritmos foram utilizados. Já nos trabalhos de (MANHÃES et al., 2012), (MARQUEZ-VERA; MORALES; SOTO, 2013) e (VILORIA et al., 2020), 10 algoritmos são utilizados. Assim, para facilitar a visualização dos resultados neste documento, optou-se por abreviar os nomes dos algoritmos nas figuras que respondem a QP5. A Figura 12 mostra uma tabela com o nome completo dos 40 algoritmos identificados e sua respectiva abreviatura.

Além disso, também optou-se por apresentar os algoritmos em categorias. Assim, utilizou-se as 7 categorias:

- *Bayesian Classification Algorithms*,
- *Decision Tree Algorithms*,
- *Logistic Regression Algorithms*,
- *Miscellanea Algorithms*,
- *Nearest Neighbors*,
- *Neural Network Algorithms*,

Figura 12 – Lista dos Algoritmos e suas Abreviaturas.

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

Fonte: Elaborada pelos autores no Tableau (2021).

- *Support Vector Machine Algorithms*.

Em seguida, a resposta para a QP5 é apresentada para cada cenário com o objetivo de oferecer uma melhor visualização das informações. Nas Figuras 13 e 14, para cada trabalho (cada linha) há um algoritmo destacado em uma cor mais escura. Esse algoritmo em destaque é o que obteve o melhor desempenho no trabalho em questão. Ao final de cada linha também é indicada a quantidade de algoritmos utilizada por cada trabalho.

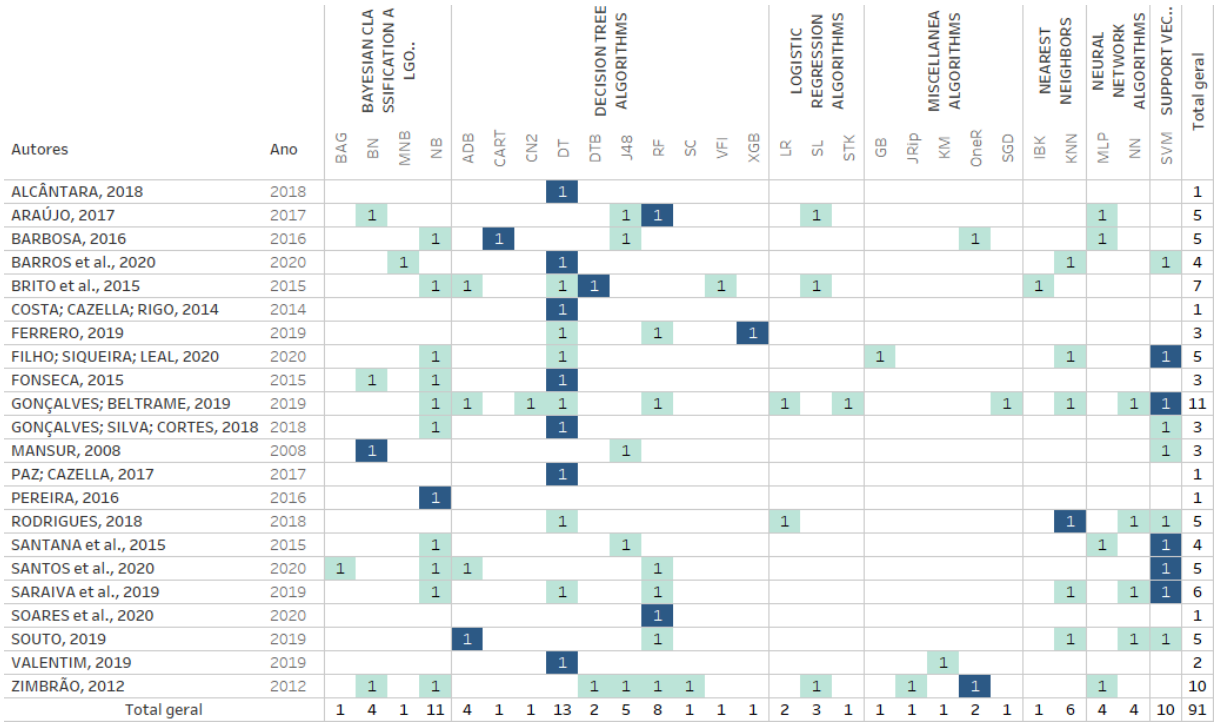
4.4.1 Cenário Nacional

A Figura 13 mostra o(s) algoritmo(s) utilizado(s) por cada um dos 22 trabalhos do cenário nacional. No total, foram identificados 27 algoritmos diferentes. O algoritmo *Decision Tree* (DT) é o mais utilizado, aparecendo em 13 dos 22 trabalhos. Além de ser o mais usado, também se destaca como o algoritmo com melhor desempenho (ver destaque em azul-escuro). Em seguida, aparece o *Naive Bayes* (NB) sendo usado por 11 trabalhos, mas com somente 1 resultado de melhor desempenho. O *Support Vector Machine* (SVM) aparece em terceiro lugar como algoritmo mais usado, aparecendo em 10 dos 22 trabalhos. O SVM também tem destaque como melhor desempenho em muitos trabalhos.

4.4.2 Cenário Internacional

A Figura 14 mostra o(s) algoritmo(s) utilizado(s) por cada um dos 28 trabalhos do cenário internacional. 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 (ver destaque em azul-escuro). Em seguida, aparece o *Naive Bayes* (NB) sendo usado por 10 trabalhos, mas com somente 1 resultado de melhor desempenho. O *J48* e o *Decision Tree* (DT) aparecem em terceiro e quarto lugares como algoritmos mais usados, respectivamente, aparecendo em 9 dos 28 trabalhos e em 8 dos 28 trabalhos. O DT apresentou melhor desempenho em apenas 3 dos 8 trabalhos que o utilizaram. Por fim, na Figura 14, cabe destacar o resultado do *Adtree* (ADT), que apesar de ser usado em apenas 5 dos 28 trabalhos, apresentou melhor desempenho em 4 deles.

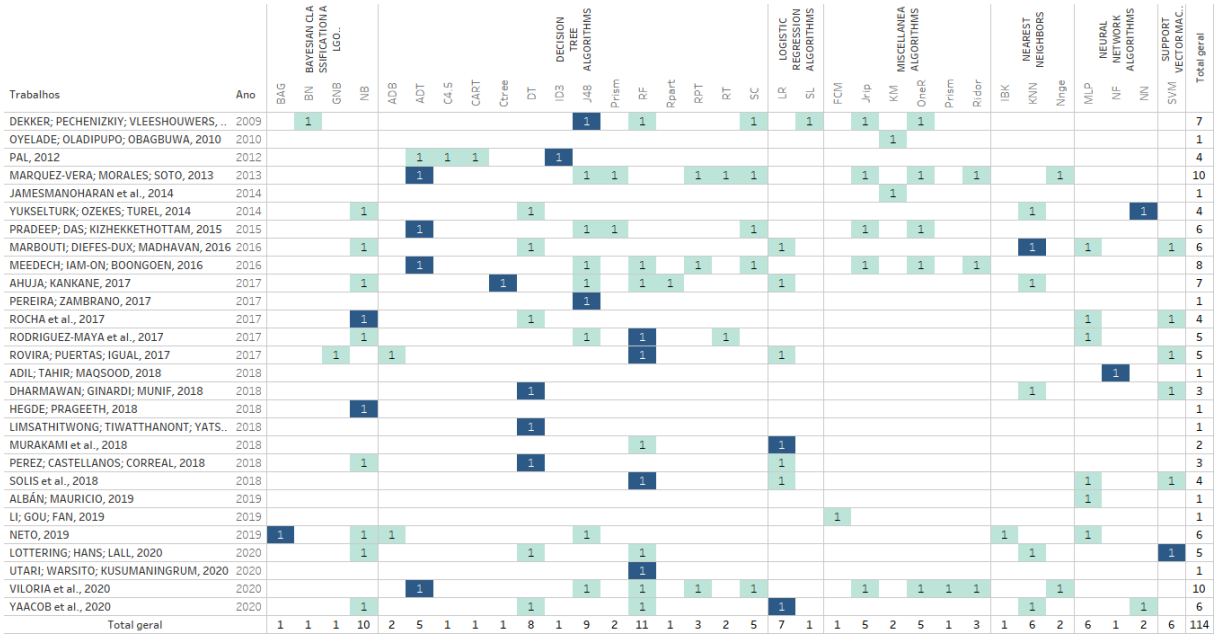
Figura 13 – Algoritmos utilizados nos trabalhos no cenário nacional.



Autores	Ano	BAG	BAYESIAN CLASSIFICATION ALGO.. BN	MNB	NB	ADB	CART	CN2	DECISION TREE ALGORITHMS DT	DTB	J48	RF	SC	VFI	XGB	LOGISTIC REGRESSION ALGORITHMS LR	SL	STK	MISCELLANEA ALGORITHMS GB	JRip	KM	OneR	SGD	NEAREST NEIGHBORS IBK	KNN	NEURAL NETWORK ALGORITHMS MLP	NN	SUPPORT VEC.. SVM	Total geral
ALCÂNTARA, 2018	2018								1																				1
ARAÚJO, 2017	2017		1								1	1					1									1			5
BARBOSA, 2016	2016				1		1				1												1			1			5
BARROS et al., 2020	2020			1					1																	1		1	4
BRITO et al., 2015	2015				1	1			1	1				1			1							1					7
COSTA; CAZELLA; RIGO, 2014	2014								1																				1
FERRERO, 2019	2019								1			1			1														3
FILHO; SIQUEIRA; LEAL, 2020	2020				1				1										1							1		1	5
FONSECA, 2015	2015		1		1				1																				3
GONÇALVES; BELTRAME, 2019	2019				1	1		1	1			1				1		1					1			1		1	11
GONÇALVES; SILVA; CORTES, 2018	2018				1				1																			1	3
MANSUR, 2008	2008		1								1																	1	3
PAZ; CAZELLA, 2017	2017								1																				1
PEREIRA, 2016	2016				1																								1
RODRIGUES, 2018	2018								1							1									1		1	1	5
SANTANA et al., 2015	2015				1						1															1		1	4
SANTOS et al., 2020	2020	1			1	1						1																1	5
SARAIVA et al., 2019	2019				1				1			1													1		1	1	6
SOARES et al., 2020	2020											1																	1
SOUTO, 2019	2019					1						1													1		1	1	5
VALENTIM, 2019	2019								1													1							2
ZIMBRÃO, 2012	2012		1		1					1	1	1	1				1			1		1				1			10
Total geral		1	4	1	11	4	1	1	13	2	5	8	1	1	1	2	3	1	1	1	1	2	1	1	6	4	4	10	91

Fonte: Elaborada pelos autores no Tableau (2021).

Figura 14 – Algoritmos utilizados nos trabalhos no cenário internacional.



Trabalhos	Ano	BAG	BAYESIAN CLASSIFICATION ALGO.. BN	GNB	NB	ADB	ADT	C4.5	CART	Ctree	DT	DECISION TREE ALGORITHMS ID3	J48	Prism	RF	Rpart	RPT	RT	SC	LOGISTIC REGRESSION ALGORITHMS LR	SL	MISCELLANEA ALGORITHMS FCM	Jrip	KM	OneR	Prism	Ridor	NEAREST NEIGHBORS IBK	KNN	Nnge	NEURAL NETWORK ALGORITHMS MLP	NF	NN	SUPPORT VECTOR MAC.. SVM	Total geral
DEKKER; PECHENIZKIY; VLEESHOUWERS, ..	2009		1										1		1				1		1		1		1										7
OYELADE; OLADIPUPO; OBAGBUWA, 2010	2010																							1											1
PAL, 2012	2012						1	1	1			1																							4
MARQUEZ-VERA; MORALES; SOTO, 2013	2013						1						1	1			1	1	1				1		1		1			1					10
JAMESMANOHARAN et al., 2014	2014																							1											1
YUKSELTURK; OZEKES; TUREL, 2014	2014				1						1																	1			1			4
PRADEEP; DAS; KIZHEKKETHOTTAM, 2015	2015						1						1	1					1				1		1										6
MARBOUTI; DIEFES-DUX; MADHAVAN, 2016	2016				1						1									1								1		1			1	6
MEEDECH; IAM-ON; BOONGOEN, 2016	2016						1						1		1		1		1				1		1		1								8
AHUJA; KANKANE, 2017	2017				1					1			1		1	1				1								1							7
PEREIRA; ZAMBRANO, 2017	2017												1																						1
ROCHA et al., 2017	2017				1						1																			1			1		4
RODRIGUEZ-MAYA et al., 2017	2017				1								1		1			1												1					5
ROVIRA; PUERTAS; IGUAL, 2017	2017			1		1									1					1													1	5
ADIL; TAHIR; MAQSOOD, 2018	2018																														1				1
DHARMAWAN; GINARDI; MUNIF, 2018	2018										1																		1				1	3
HEGDE; PRAGEETH, 2018	2018				1																														1
LIMSATHITWONG; TIWATTHANONT; YATS..	2018										1																								1
MURAKAMI et al., 2018	2018														1					1															2
PEREZ; CASTELLANOS; CORREAL, 2018	2018				1						1									1														3
SOLIS et al., 2018	2018														1					1										1			1		4
ALBÁN; MAURICIO, 2019	2019																													1					1
LI; GOU; FAN, 2019	2019																					1													1
NETO, 2019	2019	1			1	1							1														1			1				6
LOTTERING; HANS; LALL, 2020	2020				1						1				1													1					1	5
UTARI; WARSITO; KUSUMANINGRUM, 2020	2020														1																				1
VILORIA et al., 2020	2020						1						1		1		1		1				1		1	1	1			1					10
YAACOB et al., 2020	2020				1						1				1					1								1				1			6
Total geral		1	1	1	10	2	5	1	1	1	8	1	9	2	11	1	3	2	5	7	1	1	5	2	5	1	3	1	6	2	6	1	2	6	114

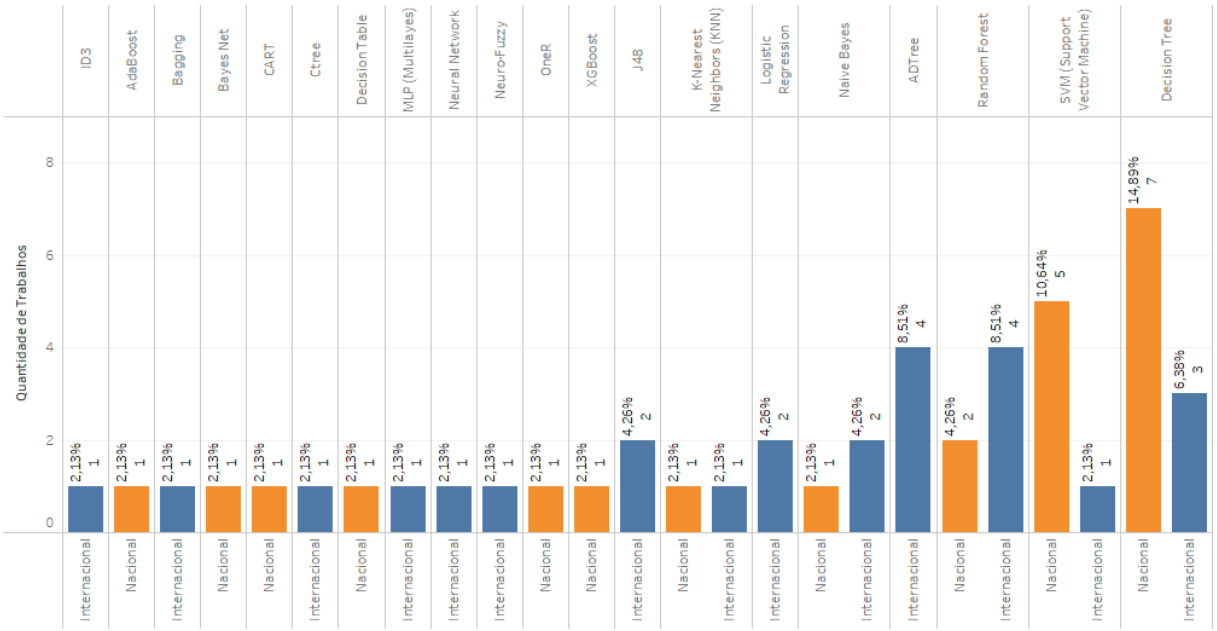
Fonte: Elaborada pelos autores no Tableau (2021).

4.5 QP5: *Quais os algoritmos com melhor desempenho?*

A resposta para a QP6 já foi discutida na QP5. No cenário nacional, o DT e o SVM aparecem como os algoritmos com melhores desempenhos, respectivamente. Já no cenário internacional, o RF e o ADT aparecem empatados com melhor desempenho.

A Figura 15 também responde a QP6 sob outra perspectiva. A figura apresenta somente os algoritmos que tiveram melhor desempenho pelo menos uma vez em um trabalho. Assim, dentre os 40 algoritmos identificados nos 50 trabalhos selecionados, a Figura 15 mostra que 20 apresentaram melhor desempenho em pelo menos um trabalho. Se for retirada a visão por cenário, podemos observar que os algoritmos DT, SVM e RF estão entre os melhores em 10, 6 e 6 trabalhos, respectivamente.

Figura 15 – Algoritmos com melhor Desempenho.



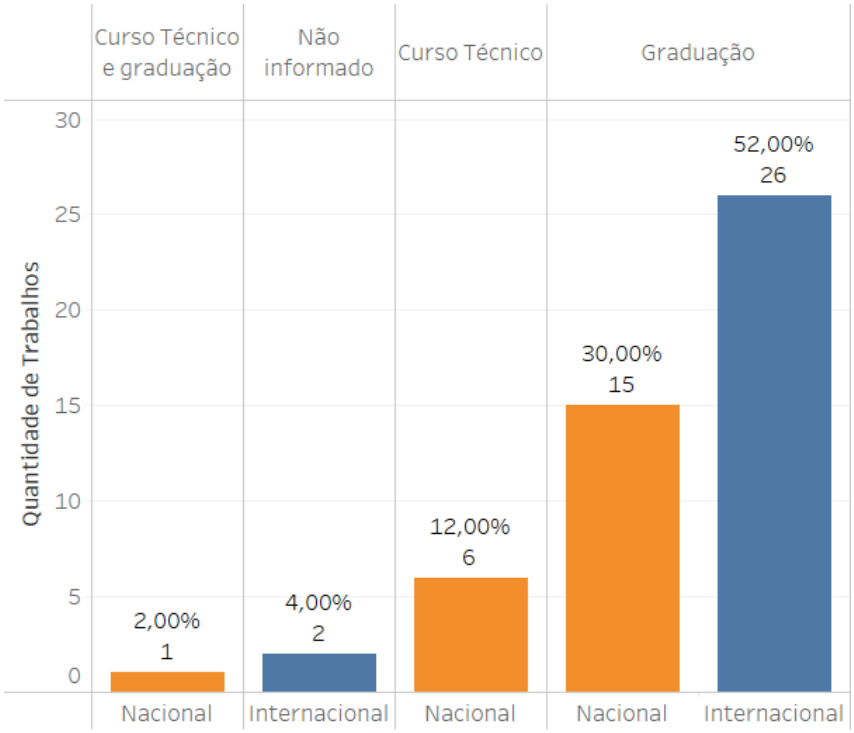
Fonte: Elaborada pelos autores no Tableau (2021).

4.6 QP6: *Quais o nível de escolaridade considerado nos trabalhos?*

A Figura 16 responde a QP7 mostrando os níveis de escolaridade considerados pelos trabalhos selecionados nos cenários nacional e internacional. Nota-se que os trabalhos selecionados avaliaram cursos de nível técnico e/ou graduação. Dentre os 50 trabalhos, 2 não informaram o nível de escolaridade analisado.

Pela Figura 16, percebe-se que a Graduação é o nível mais investigado, com 42 trabalhos, sendo 16 do cenário nacional e 26 do cenário internacional. Todos os trabalhos do cenário internacional tiveram como foco a Graduação. Já no cenário nacional, além da Graduação, também há o estudo de cursos de nível técnico realizado por 7 trabalhos.

Figura 16 – Níveis de Escolaridade.



Fonte: Elaborada pelos autores no Tableau (2021).

4.7 QP7: *Quantos registros existem nas bases de dados analisadas?*

Essa questão traz uma análise do tamanho das bases de dados investigadas nos trabalhos selecionados. Como cada base de dados tem tamanho único, optou-se por criar intervalos, como mostra a Figura 17. Os valores apresentados são relativos à quantidade de registro após a fase de preparação dos dados, ou seja, é a base usada nas fases de treinamentos e testes dos algoritmos.

Assim, percebe-se pela Figura 17 que praticamente todos os autores dos trabalhos informaram o tamanho da base de dados estudada. Somente 1 trabalho não informa. A maioria das bases estudadas estão no intervalo *De 1.001 a 5.000* registros (17 no total) ou no intervalo *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.

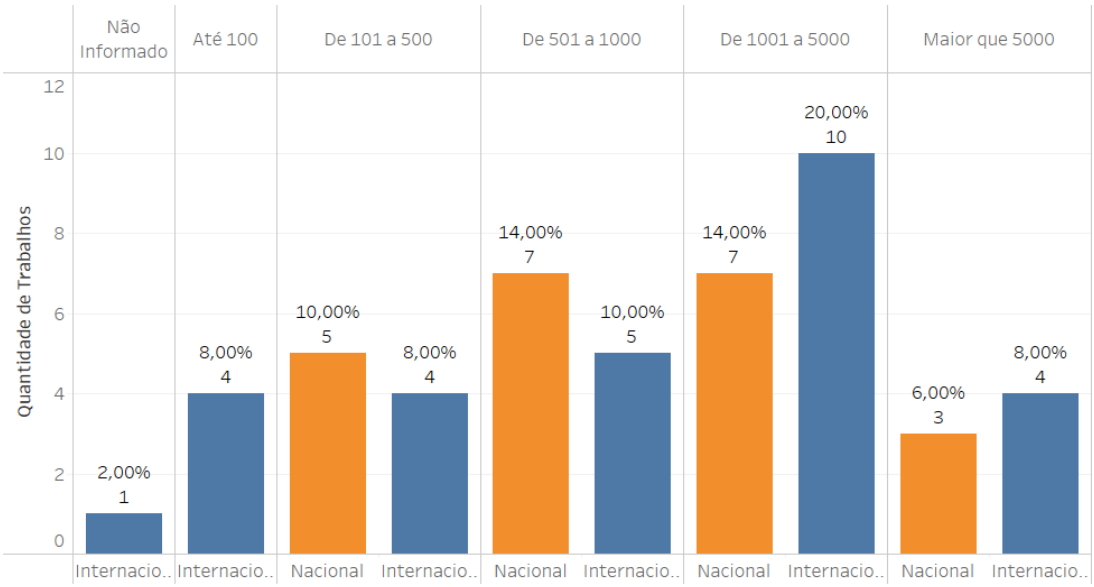
4.8 QP8: *Quantos atributos estão presentes nas bases de dados analisadas?*

A Figura 18 responde a QP9 mostrando informações sobre a quantidade de atributos presentes nas bases de dados utilizadas nos trabalhos selecionados. Assim como foi dito para os registros na QP8, cada base de dados pode conter valores únicos de atributos. Logo, seguiu-se o mesmo processo de criar intervalos para melhorar a visualização dos resultados da QP9.

Dessa forma, a Figura 18 mostra que das bases de dados estudadas, a maioria

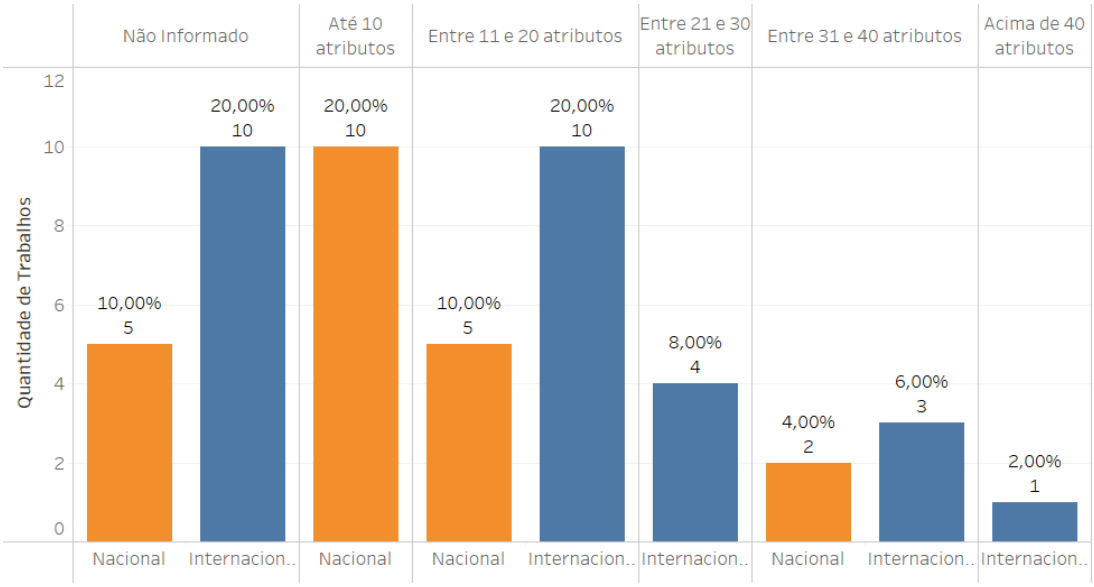
dos trabalhos utilizam *Entre 11 e 20 atributos* (15 trabalhos). Porém, há uma maior predominância para os trabalhos do cenário internacional para esse intervalo. Na verdade, no cenário nacional, a maioria dos trabalhos utilizam *Até 10 atributos*. Por fim, a figura também mostra que muitos trabalhos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

Figura 17 – Tamanho da base de dados (quantidade de registros).



Fonte: Elaborada pelos autores no Tableau (2021).

Figura 18 – Atributos da base de dados.



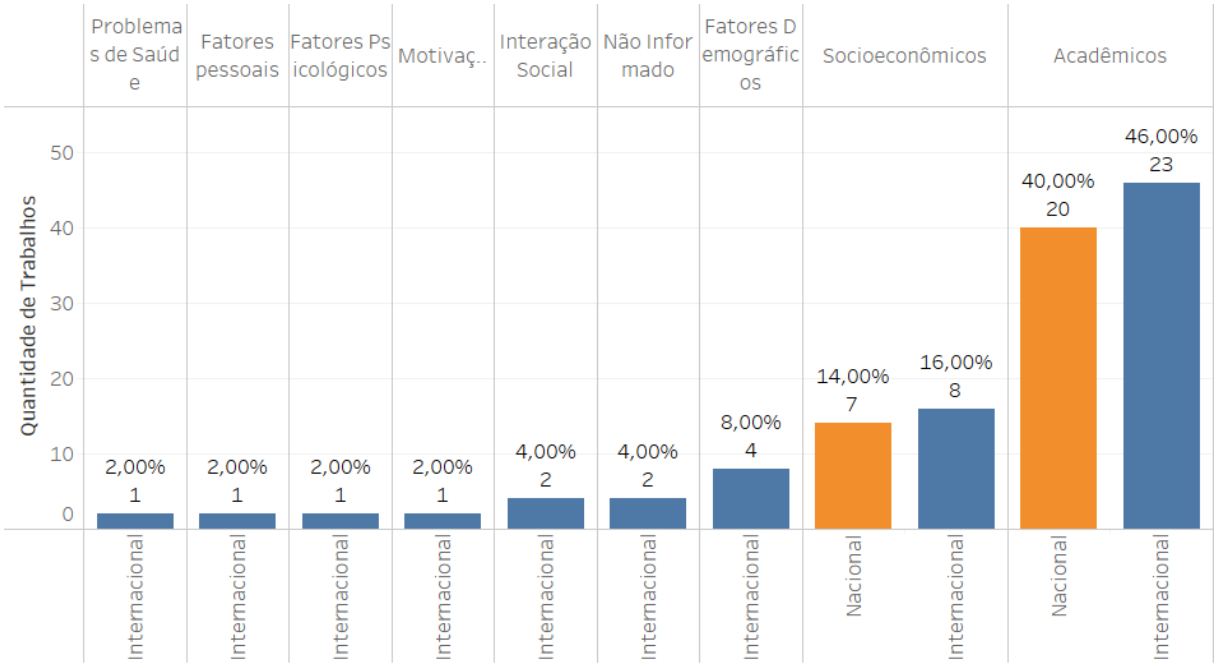
Fonte: Elaborada pelos autores no Tableau (2021).

4.9 QP9: *Quais os tipos de dados analisados nos trabalhos?*

A Figura 19 apresenta os tipos de dados que foram estudados pelos trabalhos selecionados, ou seja, os fatores que são investigados que podem contribuir para que estudantes abandonem seus cursos. Como explicado no detalhamento da planilha P3 na Seção 3.2, algumas bases de dados podem contemplar mais de um tipo. Por isso, o somatório dos trabalhos da Figura 19 é maior que 50 trabalhos.

Em resumo, podemos observar pela Figura 19 que a grande maioria das bases de dados são compostas por tipos de dados acadêmicos (43 trabalhos). Em seguida, aparecem os fatores socioeconômicos (15 trabalhos). Para os dois tipos, podemos ver que há um certo equilíbrio entre o cenário nacional e internacional. Outros 6 tipos são considerados, mas com poucos trabalhos aplicando.

Figura 19 – Tipo de dados.



Fonte: Elaborada pelos autores no Tableau (2021).

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. A metodologia adotada para elaboração do trabalho foi detalhada em três etapas principais. Na etapa final, o Tableau foi utilizado para a análise visual dos resultados.

As diversas visões disponibilizadas no trabalho facilitam a compreensão do estado da arte na área quanto às ferramentas/bibliotecas, algoritmos, bases de dados, dentre outros. Por isso, pode-se dizer que os resultados apresentados servem como

um verdadeiro guia para todos aqueles interessados em estudar e/ou desenvolver soluções/projetos para combate à evasão através da EDM.

Esta pesquisa pode ser continuada em vários aspectos como trabalhos futuros, tais como: adicionando os 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 tamanho das bases, tipos de atributos etc.), dentre outros. Como trabalhos futuros pretende-se disponibilizar as visões já alcançadas em um projeto público do Tableau para permitir consultas web.

REFERÊNCIAS

ADIL, M.; TAHIR, F.; MAQSOOD, S. Predictive analysis for student retention by using neuro-fuzzy algorithm. In: **2018 10th Computer Science and Electronic Engineering (CEEC)**. [S.l.: s.n.], 2018. p. 41–45.

AGRUSTI, F.; BONAVOLONTÀ, G.; MEZZINI, M. University dropout prediction through educational data mining techniques: A systematic review. **Journal of E-Learning and Knowledge Society**, v. 15, n. 3, p. 161–182, 2019.

AHUJA, R.; KANKANE, Y. Predicting the probability of student's degree completion by using different data mining techniques. In: **2017 Fourth International Conference on Image Information Processing (ICIIP)**. [S.l.: s.n.], 2017. p. 1–4.

ALBAN, M.; MAURICIO, D. Predicting university dropout through data mining: A systematic literature. **Indian Journal of Science and Technology**, v. 12, n. 4, p. 1–12, 2019.

ALCÂNTARA, M. L. e C. Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, v. 29, n. 1, p. 1921, 2018. ISSN 2316-6533. Disponível em: <http://br-ie.org/pub/index.php/sbie/article/view/8191>.

ARAÚJO, E. Q. e Cristian Cechinel e R. Predição de estudantes com risco de evasão em cursos técnicos a distância. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, v. 28, n. 1, p. 1547, 2017. ISSN 2316-6533. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/7686>.

BAGGI, C. A.; LOPES, D. A. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. **Avaliação: revista da avaliação da educação superior**, SciELO Brasil, v. 16, n. 2, 2010.

BARBOSA, G. K. e Evandro Flores e Jáder Schmitt e Ivan Hoffmann e F. Predição da evasão em cursos de graduação em instituições públicas. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, v. 27, n. 1, p. 906, 2016. ISSN 2316-6533. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/6776>.

BARROS, R. P. et al. 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**, v. 6, n. 1, p. 2523–2534, 2020.

BRASIL. **Constituição da República Federativa do Brasil**. Senado Federal, 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 16 jun. 2021.

BRASIL. **Lei nº 9.394, de 1996, que estabelece as diretrizes e bases da educação nacional, e legislação correlata**. 1996. Disponível em: <http://www.planalto.gov.br/ccivil_03/LEIS/L9394.htm>. Acesso em: 19 Jun. 2021.

BRITO, D. M. et al. Identificação de estudantes do primeiro semestre com risco de evasão através de técnicas de data mining. **Nuevas Ideas en Informática Educativa TISE**, p. 459–463, 2015.

COSTA, S. S. da; CAZELLA, S.; RIGO, S. J. 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. **RENOTE-Revista Novas Tecnologias na Educação**, v. 12, n. 2, 2014.

DEKKER, G. W.; PECHENIZKIY, M.; VLEESHOUWERS, J. M. Predicting students drop out: A case study. **International Working Group on Educational Data Mining**, ERIC, 2009.

DHARMAWAN, T.; GINARDI, H.; MUNIF, A. Dropout detection using non-academic data. In: **2018 4th International Conference on Science and Technology (ICST)**. [S.l.: s.n.], 2018. p. 1–4.

FERRERO, P. B. e C. 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**, v. 8, n. 1, p. 149, 2019. ISSN 2316-8889. Disponível em: <https://www.br-ie.org/pub/index.php/wcbie/article/view/8956>.

FILHO, F. H.; SIQUEIRA, D.; LEAL, B. 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**. Porto Alegre, RS, Brasil: SBC, 2020. p. 141–148. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/ercemapi/article/view/11478>.

FILHO, R. L. L. S. et al. A evasão no ensino superior brasileiro. **Cadernos de pesquisa**, SciELO Brasil, v. 37, n. 132, p. 641–659, 2007.

FONSECA, F. S. e Josenildo Silva e Reinaldo Silva e L. Um modelo preditivo para diagnóstico de evasão baseado nas interações de alunos em fóruns de discussão. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, v. 26, n. 1, p. 1187, 2015. ISSN 2316-6533. Disponível em: <http://br-ie.org/pub/index.php/sbie/article/view/5446>.

FUZETO, R.; BRAGA, R. Um mapeamento sistemático em progresso sobre internet das coisas e educação à distância. **Workshop do Congresso Brasileiro de Informática na Educação**, v. 5, n. 1, 2016. Disponível em: <http://br-ie.org/pub/index.php/wcbie/article/view/7059>.

GONÇALVES, O.; BELTRAME, W. Mineração de dados e evasão estudantil: Analisando o curso de nível superior do ifes. In: . [S.l.: s.n.], 2019.

GONÇALVES, T. C.; SILVA, J. C. da; CORTES, O. A. C. Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do instituto federal do maranhão. **Revista Brasileira de Computação Aplicada**, v. 10, n. 3, p. 11–20, 2018.

HEGDE, V.; PRAGEETH, P. P. Higher education student dropout prediction and analysis through educational data mining. In: **2018 2nd International Conference on Inventive Systems and Control (ICISC)**. [S.l.: s.n.], 2018. p. 694–699.

IFCE. **Plano estratégico para permanência e êxito dos estudantes do IFCE**. Fortaleza: IFCE, 2017. Disponível em: <https://ifce.edu.br/proen/ensino/plano-de-permanencia-e-exito.pdf>. Acesso em: 19 Jun. 2021.

IFCE. **IFCE em Numeros**. 2021. Disponível em: <https://ifceemnumeros.ifce.edu.br/>. Acesso em: 19 Jun. 2021.

INEP. **Resumo técnico do Censo da Educação Superior 2019**. Brasília-DF, 2021. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/resumo_tecnico_censo_da_educacao_superior_2019.pdf>. Acesso em: 16 Abr. 2021.

JAMESMANOHARAN, J. et al. Discovering students' academic performance based on gpa using k-means clustering algorithm. In: **2014 World Congress on Computing and Communication Technologies**. [S.l.: s.n.], 2014. p. 200–202.

JÚNIOR, O. d. G. F. et al. Melhoria da gestão escolar através do uso de técnicas de mineração de dados educacionais: um estudo de caso em escolas municipais de maceió. **RENOTE**, v. 17, n. 1, p. 296–305, 2019.

LI, Y.; GOU, J.; FAN, Z. Educational data mining for students' performance based on fuzzy c-means clustering. **The Journal of Engineering**, IET, v. 2019, n. 11, p. 8245–8250, 2019.

LIMSATHITWONG, K.; TIWATTHANONT, K.; YATSUNGNOEN, T. Dropout prediction system to reduce discontinue study rate of information technology students. In: **2018 5th International Conference on Business and Industrial Research (ICBIR)**. [S.l.: s.n.], 2018. p. 110–114.

LOTTERING, R.; HANS, R.; LALL, M. A model for the identification of students at risk of dropout at a university of technology. In: **2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)**. [S.l.: s.n.], 2020. p. 1–8.

MANHÃES, L. et al. Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. **Simpósio Brasileiro de Informática na Educação (SBIE)**, v. 1, n. 1, 2012. ISSN 2316-6533. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/1585>.

MANSUR, M. A. e Dante Barone e A. Técnicas de aprendizado de máquina aplicadas na previsão de evasão acadêmica. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação**

- SBIE), v. 1, n. 1, p. 666–674, 2008. ISSN 2316-6533. Disponível em: <http://br-ie.org/pub/index.php/sbie/article/view/755>.

MARBOUTI, F.; DIEFES-DUX, H. A.; MADHAVAN, K. Models for early prediction of at-risk students in a course using standards-based grading. **Computers & Education**, Elsevier, v. 103, p. 1–15, 2016.

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.

MARQUEZ-VERA, C.; MORALES, C. R.; SOTO, S. V. Predicting school failure and dropout by using data mining techniques. **IEEE Revista Iberoamericana de Tecnologias del Aprendizaje**, v. 8, n. 1, p. 7–14, 2013.

MEEDECH, P.; IAM-ON, N.; BOONGOEN, T. Prediction of student dropout using personal profile and data mining approach. In: **Intelligent and Evolutionary Systems**. [S.l.]: Springer, 2016. p. 143–155.

MURAKAMI, K. et al. Predicting the probability of student dropout through emir using data from current and graduate students. In: **2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI)**. [S.l.: s.n.], 2018. p. 478–481.

NETO, C. B. e João Xavier-Júnior e Cephas Barreto e C. O. Plataforma de aprendizado de maquina para detecção e monitoramento de alunos com risco de evasão. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, v. 30, n. 1, p. 1591, 2019. ISSN 2316-6533. Disponível em: <https://www.br-ie.org/pub/index.php/sbie/article/view/8892>.

OYELADE, O. J.; OLADIPUPO, O. O.; OBAGBUWA, I. C. **Application of k Means Clustering algorithm for prediction of Students Academic Performance**. 2010.

PAL, S. Mining educational data to reduce dropout rates of engineering students. **International Journal of Information Engineering and Electronic Business**, Modern Education and Computer Science Press, v. 4, n. 2, p. 1, 2012.

PAZ, F.; CAZELLA, S. 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**. [S.l.: s.n.], 2017. v. 6, n. 1, p. 624.

PEREIRA, R. T.; ZAMBRANO, J. C. Application of decision trees for detection of student dropout profiles. In: **2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)**. [S.l.: s.n.], 2017. p. 528–531.

PEREIRA, W. M. e João Damiani e M. Rede bayesiana para previsão de evasão escolar. **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**, v. 5, n. 1, p. 920, 2016. ISSN 2316-8889. Disponível em: <http://www.br-ie.org/pub/index.php/wcbie/article/view/7017>.

PEREZ, B.; CASTELLANOS, C.; CORREAL, D. Applying data mining techniques to predict student dropout: A case study. In: **2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)**. [S.l.: s.n.], 2018. p. 1–6.

PRADEEP, A.; DAS, S.; KIZHEKKETHOTTAM, J. J. Students dropout factor prediction using edm techniques. In: **2015 International Conference on Soft-Computing and Networks Security (ICSNS)**. [S.l.: s.n.], 2015. p. 1–7.

ROCHA, C. F. et al. Prediction of university desertion through hybridization of classification algorithms. In: **SIMBig**. [S.l.: s.n.], 2017. p. 215–222.

RODRIGUES, J. R. e João Silva e Leonardo Prado e Alex Gomes e R. Um estudo comparativo de classificadores na previsão da evasão de alunos em ead. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, v. 29, n. 1, p. 1463, 2018. ISSN 2316-6533. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/8107>.

RODRIGUEZ-MAYA, N. E. et al. Modeling students' dropout in mexican universities. **Research in Computing Science**, v. 139, p. 163–175, 2017.

ROVIRA, S.; PUERTAS, E.; IGUAL, L. Data-driven system to predict academic grades and dropout. **PLoS one**, Public Library of Science San Francisco, CA USA, v. 12, n. 2, p. e0171207, 2017.

SANTANA, M. A. et al. A predictive model for identifying students with dropout profiles in online courses. In: **EDM (Workshops)**. [S.l.: s.n.], 2015.

SANTOS, G. et al. 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**. Porto Alegre, RS, Brasil: SBC, 2020. p. 228–235. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/ercemapi/article/view/11489>.

SARAIVA, D. et al. 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**. Porto Alegre, RS, Brasil: SBC, 2019. p. 319–333. ISSN 2595-6175. Disponível em: <https://sol.sbc.org.br/index.php/wei/article/view/6639>.

SOARES, L. C. C. P. et al. Aplicação de técnicas de aprendizado de máquina em um contexto acadêmico com foco na identificação dos alunos evadidos e não evadidos. **Humanidades & Inovação**, v. 7, n. 8, p. 223–235, 2020.

SOLIS, M. et al. Perspectives to predict dropout in university students with machine learning. In: **2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)**. [S.l.: s.n.], 2018. p. 1–6.

SOUTO, A. P. e Leandro Carvalho e E. Predição de evasão de estudantes non-majors em disciplina de introdução à programação. **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**, v. 8, n. 1, p. 178, 2019. ISSN 2316-8889. Disponível em: <https://www.br-ie.org/pub/index.php/wcbie/article/view/8959>.

UNDP. **Human Development Indices and Indicators: 2018 Statistical Update**. 2018. Disponível em: <https://www.br.undp.org/content/dam/brazil/docs/RelatoriosDesenvolvimento/2018_human_development_statistical_update.pdf>. Acesso em: 19 Jun. 2021.

UTARI, M.; WARSITO, B.; KUSUMANINGRUM, R. Implementation of data mining for drop-out prediction using random forest method. In: **2020 8th International Conference on Information and Communication Technology (ICoICT)**. [S.l.: s.n.], 2020. p. 1–5.

VALENTIM, I. B. J. e Humberto Rabelo e Angela Naschold e Almir Ferreira e Aquiles Burlamaqui e Danieli Rabelo e R. 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**, v. 8, n. 1, p. 159, 2019. ISSN 2316-8889. Disponível em: <https://br-ie.org/pub/index.php/wcbie/article/view/8957>.

VILORIA, A. et al. Data mining applied in school dropout prediction. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2020. v. 1432, n. 1, p. 012092.

WAIKATO, U. O. **Weka 3 – Machine Learning Software in Java**. 2010. <http://www.cs.waikato.ac.nz/ml/weka>. Acesso em: 18 jun. 2021.

YAACOB, W. W. et al. Predicting student drop-out in higher institution using data mining techniques. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2020. v. 1496, n. 1, p. 012005.

YUKSELTURK, E.; OZEKES, S.; TUREL, Y. K. Predicting dropout student: An application of data mining methods in an online education program. **European Journal of Open, Distance and e-learning**, ERIC, v. 17, n. 1, p. 118–133, 2014.

ZHANG, L.; LI, K. F. Education analytics: Challenges and approaches. In: **32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)**. [S.l.: s.n.], 2018. p. 193–198.