Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 31/Mar/2023; 1st round notif.: 09/Aug/2023; New version: 11/Nov/2023; 2nd round notif.: 15/Feb/2024;
Camera ready: 20/Feb/2024; Edition review: 25/Apr/2024; Available online: 25/Apr/2024; Published: 25/Apr/2024;

Técnicas de Mineração de Dados e Aprendizado de Máquina aplicados à Evasão Estudantil: um mapeamento sistemático da literatura

Title: Data Mining and Machine Learning techniques applied to student dropout: a systematic literature mapping

Título: Técnicas de minería de datos y aprendizaje automático aplicadas a la deserción estudiantil: un mapeo sistemático de la literatura

Fernanda Ferreira do Nascimento
Universidade Federal Rural do Semi-Árido - UFERSA
ORCID: 0000-0001-9824-2152
nascimentofernandaf@outlook.com

Lucas Cesar de Oliveira Dantas
Universidade Federal Rural do Semi-Árido - UFERSA
ORCID: 0000-0002-0503-5953
lucascodantas@hotmail.com

Angélica Félix de Castro
Universidade Federal Rural do Semi-Árido - UFERSA
ORCID: 000-0002-2250-9305
angelica@ufersa.edu.br

Paulo Gabriel Gadelha Queiroz
Universidade Federal Rural do Semi-Árido - UFERSA
ORCID: 0000-0003-3993-0208
pgabriel@ufersa.edu.br

Resumo

Este trabalho apresenta um Mapeamento Sistemático da Literatura sobre a evasão estudantil, a partir do qual buscou-se responder à seguinte questão de pesquisa: quais ferramentas, técnicas de aprendizado de máquina, fatores indutores, bases de dados abertas e métricas de avaliação de algoritmos têm sido utilizados para identificar as causas da evasão estudantil? O protocolo do mapeamento foi elaborado com base nas diretrizes apresentadas por Petersen (2008) e Kitchenham (2004). Portanto, ele consistiu na definição de questões de pesquisa, critérios de seleção, definição de strings de busca, definição das fontes de busca, entre outros elementos. Entre os resultados ressalta-se que a ferramenta R foi a mais utilizada, a classificação obteve destaque entre as técnicas de aprendizagem de máquina e os principais trabalhos da área se concentraram em estudar fatores relacionados às características individuais do aluno. Adicionalmente, foram encontradas 15 bases bases de dados abertas. Por fim, às métricas de avaliação de algoritmos que se destacaram são: Recall, Accuracy e Precision. Os resultados deste mapeamento fornecem uma visão abrangente do estado da arte da pesquisa em evasão estudantil, incluindo as ferramentas e técnicas mais populares e os fatores indutores mais investigados. Pesquisadores podem utilizar os resultados deste trabalho para direcionar esforços de pesquisa para a criação de modelos utilizando os três tipos de fatores indutores e disponibilização de bases abertas.

Palavras-chave: *Mapeamento sistemático da literatura; Evasão estudantil; Mineração de dados; Aprendizado de máquina.*

Abstract

This work presents a Systematic Mapping of the Literature on student dropout, from which we sought to answer the following research question: What tools, machine learning techniques, inducing factors, open databases, and algorithm evaluation metrics have been used to identify the possible causes of student dropout? The mapping protocol

Cite as: Nascimento, F. F., Dantas, L. C. O., Castro, A. F. & Queiroz, P. G. G. (2024). Técnicas de Mineração de Dados e Aprendizado de Máquina aplicados à Evasão Estudantil: um mapeamento sistemático da literatura. Revista Brasileira de Informática na Educação, 32, 270-294. https://doi.org/10.5753/rbie.2024.3296.

was developed based on the guidelines of Petersen (2008) and Kitchenham (2004). Thus, it consisted of defining research questions, selection criteria, search strings, and search sources, among other elements. Among the results, it is worth noting that the R tool was the most widely used, classification stood out among the machine learning techniques and the main works in the area focused on studying factors related to individual student characteristics. Additionally, 15 open databases were identified. Finally, regarding algorithm evaluation metrics, the following stand out: Recall, Accuracy, and Precision. The results of this mapping provide a comprehensive view of the state of the art from research on student dropout, including the most popular tools and techniques, and the most investigated inducing factors. Researchers can use the results of this study to direct research efforts toward the creation of models using the three types of inducing factors and the provision of open bases.
Keywords: *Systematic literature mapping; Student dropout; Data mining; Machine learning.*

Resumen

Este trabajo presenta un Mapeo Sistemático de la Literatura sobre la deserción estudiantil, a partir del cual buscamos responder la siguiente pregunta de investigación: ¿qué herramientas, técnicas de aprendizaje automático, factores inductores, bases de datos abiertas y métricas de evaluación de algoritmos se han utilizado para identificar las causas de la deserción estudiantil? El protocolo de mapeo se desarrolló con base en los lineamientos presentados por Petersen (2008) y Kitchenham (2004). Por tanto, consistió en definir preguntas de investigación, criterios de selección, definir cadenas de búsqueda, definir fuentes de búsqueda, entre otros elementos. Entre los resultados se destaca que la herramienta R fue la más utilizada, se destacó la clasificación entre técnicas de aprendizaje automático y los principales trabajos en el área se centraron en estudiar factores relacionados con las características individuales del estudiante. Además, se encontraron 15 bases de datos abiertas. Finalmente, las métricas de evaluación del algoritmo que destacaron son: Recall, Accuracy, and Precision. Los resultados de este mapeo brindan una visión integral del estado del arte en la investigación de la deserción estudiantil, incluidas las herramientas y técnicas más populares y los factores inductores más investigados. Los investigadores pueden utilizar los resultados de este trabajo para dirigir los esfuerzos de investigación hacia la creación de modelos utilizando los tres tipos de factores inductores y el suministro de bases de datos abiertas.
Palabras clave: *Mapeo sistemático de la literatura; Abandono estudiantil; Procesamiento de datos; Aprendizaje automático.*

1 Introdução

Um dos maiores desafios da educação no Brasil é a evasão estudantil, que é definida por Riffel e Malacarne (2010) como o ato de evadir, fugir, abandonar, sair, desistir e não permanecer no ambiente escolar. Esse problema afeta a qualidade de ensino, impactando os âmbitos econômicos e ambientais, além de refletir negativamente no desenvolvimento social.

Pinto (2021) apresentou um estudo, no qual afirma que o custo contábil da evasão foi de mais de 19 bilhões, e o custo econômico superior aos 5 bilhões, apenas no ano de 2019 e contabilizando, somente, o ensino superior presencial. Esse problema encontrado em nível nacional afeta diversas universidades, tanto públicas quanto privadas.

Diante dessa realidade preocupante, surgiu a motivação de direcionar esforços de pesquisa com o objetivo de reduzir a evasão no ensino superior. O início desses esforços consistiu em tentar responder a seguinte questão: quais medidas, técnicas ou abordagens podem ser adotadas para mitigar o risco de evasão estudantil?

Nessa direção, o trabalho de Lobo (2012) sugere algumas ações, como a criação de grupos de trabalho, avaliação estatística, programa de aconselhamento, entre outros. A autora também

aponta a determinação das causas da evasão como ação fundamental. Acredita-se que essa determinação pode levar a identificação dos perfis de estudantes considerados como possíveis evasores, e assim, tornar as ações citadas mais efetivas.

Portanto, este trabalho parte da premissa que a predição de evasão estudantil é um fator fundamental para a criação de estratégias efetivas para mitigar a evasão estudantil no ensino superior. Nessa direção, uma técnica promissora para desenvolver modelos preditivos de estudantes com possibilidade de evasão é a Mineração de Dados Educacionais (MDE), que é uma área emergente de Mineração de Dados (MD), na qual são desenvolvidos métodos e algoritmos, com o objetivo de compreender os dados em contextos educacionais (Costa et al., 2013).

Dentre as áreas da computação que se aliam a MDE, se destaca o Aprendizado de Máquina (AM), definido por Faria (2014) como o campo da Ciência da Computação que se concentra no uso dos algoritmos para imitar a maneira como os humanos aprendem, melhorando gradualmente sua precisão. Os algoritmos de AM são usados para realizar predições, baseadas em características conhecidas e aprendidas por meio dos dados.

Diante do contexto e motivação apresentados, este trabalho aborda a seguinte questão de pesquisa: quais as principais ferramentas, modelos de AM, métricas e bases de dados abertas, utilizados para a aplicação do AM na predição de evasão estudantil?

Para tentar responder a esta questão de pesquisa, faz-se necessário recorrer a uma metodologia de pesquisa confiável e repetível, como um Mapeamento Sistemático da Literatura (MSL). Um MSL apresenta-se como o processo para coleta, avaliação e sistematização de trabalhos primários, apresentando uma abordagem quantitativa, capaz de fornecer dados para identificação de tendências de estudos, bem como lacunas a serem exploradas (de Campos et al., 2020).

Portanto, este artigo apresenta o planejamento e execução de um MSL que buscou trabalhos entre os anos de 2018 a 2021, de modo a atender aos seguintes objetivos: (I) identificar as ferramentas mais utilizadas na MD para a descoberta das causas da evasão estudantil no ensino presencial; (II) identificar quais as técnicas AM têm sido utilizadas na predição de estudantes com alto risco de evasão estudantil no ensino presencial, identificando suas vantagens e desvantagens; (III) Desvendar quais fatores indutores têm sido investigados para a descoberta das causas da evasão estudantil no ensino presencial; (IV) Investigar a existência de bases abertas que forneçam dados escolares de alunos; (V) Identificar quais as métricas de avaliação de modelos de AM foram aplicadas nos experimentos realizados.

Ao alcançar esses objetivos, este trabalho oferece como resultados principais: uma síntese do estado da arte das contribuições da computação na predição da evasão estudantil, de modo que esses resultados possam impulsionar a definição de políticas para mitigar a evasão; a identificação de lacunas de pesquisas, para que a comunidade científica possa direcionar esforços voltados a resolução desse tema tão relevante; e, o mapeamento de grupos de trabalho que se empenham na utilização de técnicas de computação na área de evasão escolar, para possíveis parcerias de trabalho entre pesquisadores com o objetivo de promover o avanço e a qualidade das pesquisas na área de MD aplicado à evasão estudantil.

O restante deste artigo está organizado conforme se detalha a seguir. Na Seção II, são apresentados os trabalhos relacionados. Na Seção III apresenta-se o protocolo deste MSL, que consiste nos procedimentos e métodos utilizados na sua execução. Na Seção IV, são apresentadas

as respostas encontradas para a questão de pesquisa levantada. Na Seção V apresenta-se uma discussão sobre os resultados encontrados e suas contribuições para a área. Por fim, na Seção VI são apresentadas as considerações finais deste MSL.

2 Trabalhos relacionados

A busca por trabalhos semelhantes, realizada nas mesmas bases de dados apresentadas na Seção III, encontrou dois outros estudos secundários com propósito semelhante, ambos na forma de MSL e com seus resultados comentados a seguir.

O primeiro mapeamento encontrado foi o trabalho de Sousa et al. (2021), no qual os autores buscaram identificar e analisar os estudos primários que aplicaram MD no problema de evasão escolar de cursos presenciais, com o objetivo de responder as seguintes questões de pesquisa: I - Quais níveis de educação foram explorados nos trabalhos? II - Qual o tamanho e como foram gerados os conjuntos de dados utilizados nas pesquisas? III - Quais fatores indutores foram utilizados para modelar o problema da evasão? IV - Que tipo de pré-processamento foi aplicado às amostras? V - Quais famílias de algoritmos foram utilizados? VI Quais métricas foram utilizadas para validar os padrões extraídos?

O objetivo desse trabalho foi fornecer uma visão geral dos aspectos relacionados às etapas de MD no contexto apresentado, sem entrar em detalhes sobre as técnicas específicas. Os autores selecionaram 118 trabalhos primários considerando um período de 10 anos (01/01/2010 a 31/12/2020) com buscas realizadas nas bases *Scopus, Compendex, Web of Science, IEEE Xplore, ACM Digital Library e Science Direct* . Os principais resultados dessa pesquisa foram: o nível de educação mais explorado é o superior, os conjuntos de dados trabalhados são pequenos; as principais variáveis estudadas são de fatores relacionados as características acadêmicas; a maioria dos estudos adere aos algoritmos de classificação; diversas técnicas de pré-processamento e validação pós-processamento são aplicadas.

O segundo mapeamento encontrado foi o de Marques et al. (2019) que serviu como base para a definição do protocolo de pesquisa deste mapeamento. O trabalho considerou o período compreendido entre 2008 e 2018 (10 anos), realizou buscas nas bases *IEEEXplore, ACM Digital Library, Science Direct e Scopus*, selecionou 14 trabalhos primários e objetivou responder as seguintes questões de pesquisa: Quais as melhores ferramentas de MD utilizadas para a descoberta das causas da evasão escolar? Quais técnicas de MD têm sido utilizadas para a descoberta das causas da evasão escolar? Que tipos de fatores indutores têm sido investigados para a descoberta das causas da evasão escolar?

Os principais resultados desse trabalho foram: as ferramentas *Waikoto environment for Knowledge Analysis (Weka)* e *Mplus*, se destacaram como as mais utilizadas; a técnica de Classificação foi a mais utilizada, seguida pela técnica de Equações Estruturais; com relação aos fatores indutores, os trabalhos encontrados utilizaram às características individuais do estudante, coletadas a partir dos dados armazenados nas instituições ou por meio de questionários. Também foi um importante resultado observar que poucos trabalhos exploraram, em conjunto, os fatores indutores relacionados as características internas à instituição e características externas à instituição.

Diferente dos mapeamentos encontrados, a presente pesquisa tem um escopo mais amplo,

de modo a englobar a busca por: técnicas de aprendizado de máquina aplicados a EDM, bases de dados abertas e métricas de avaliação de algoritmos. Dessa forma, foi possível encontrar lacunas de pesquisa com o propósito de fornecer direcionamento a comunidade científica que busca resolver o problema da evasão escolar. Adicionalmente, este é o único trabalho que utiliza critérios de qualidade para *rankear* os estudo selecionados. Também é válido destacar que este mapeamento inclui estudos primários mais recentes, publicados no ano de 2021.

3 Procedimentos e métodos

Este MSL é fundamentado pelas diretrizes de Petersen (2008) e Kitchenham (2004), segundo os quais, as etapas de um MSL são: definição do protocolo, seleção de estudos primários, extração de dados e análise dos resultados, conforme ilustrado na figura 1 e descritos a seguir.

Figura 1: Etapas do MSL.

- Na etapa de definição do protocolo do MSL são elaborados: as questões de pesquisa, os critérios e seleção das bases de dados, os termos de busca, a *string* de busca, os critérios de inclusão, exclusão e de qualidade, o formulário de extração de dados e os passos para seleção de estudos primários.
- A seleção de estudos primários consiste na execução do protocolo do MSL. Essa etapa envolve a coleta de trabalhos nas bases selecionadas e a filtragem desses trabalhos, de acordo com os critérios definidos no protocolo.
- A extração de dados é etapa de leitura dos artigos selecionados para a extração de elementos capazes de auxiliarem nas respostas das questões de pesquisa definidas no protocolo do mapeamento.
- Por fim, na etapa de análise dos resultados, os dados extraídos são consolidados em forma de gráficos e tabelas e utilizados para responder as questões de pesquisa definidas.

3.1 Questões de pesquisa

Por meio desta pesquisa, buscou-se compreender as causas de evasão estudantil que já foram investigadas por outros pesquisadores e os artefatos desenvolvidos como algoritmos e técnicas de MD e AM. Dessa forma, as questões de pesquisa que nortearam este trabalho são:

- **Q1**: Quais ferramentas de MD têm sido utilizadas para a descoberta das causas da evasão estudantil?

- **Q2**: Quais técnicas de AM têm sido utilizadas para a descoberta das causas da evasão estudantil?
- **Q3**: Quais fatores indutores têm sido investigados para a descoberta das causas da evasão estudantil?
- **Q4**: Quais bases abertas fornecem dados escolares de estudantes?
- **Q5**: Quais métricas de avaliação de modelos de AM foram aplicadas nos experimentos realizados?

O horizonte desta pesquisa considera estudos realizados após o período já analisado por Marques et al. (2019), apresentado na Seção II. Dessa forma, no presente estudo, foram mapeadas as pesquisas desenvolvidas entre os anos de 2018 e 2021. Observa-se que as buscas foram conduzidas no mês de janeiro de 2022. Além do horizonte de pesquisa, também foi definido o idioma dos trabalhos buscados. Para este mapeamento, foram considerados apenas trabalhos em Inglês, por se tratar do idioma oficial para a divulgação de pesquisas científicas.

Adicionalmente, foram selecionados 3 trabalhos como artigos de controle. Esses trabalhos eram conhecidos e já foram estudados pela equipe de pesquisa, antes da definição do protocolo do MSL, e foram considerados importantes. Dessa forma, a *string* de busca seria considerada eficiente se conseguisse retornar, pelo menos, esses trabalhos já conhecidos. Os artigos de controle são apresentados a seguir

- RODRÍGUEZ-MUÑIZ, L. J.; BERNARDO, A. B.; ESTEBAN, M.; DÍAZ, I. Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? Plos one, Public Library of Science San Francisco, CAUSA, v. 14, n. 6, p. e0218796, 2019.
- FREITAS, F. A. d. S.; VASCONCELOS, F. F.; PEIXOTO, S. A.; HASSAN, M. M.; DEWAN, M.; ALBUQUERQUE, V. H. C. d. et al. Iot system for school dropout prediction using machine learning techniques based on socioeconomic data. Electronics, Multidisciplinary Digital Publishing Institute, v. 9, n. 10, p. 1613, 2020.
- SARI, E. Y.; SUNYOTO, A. et al. Optimization of weight backpropagation with particle swarm optimization for student dropout prediction. In: IEEE. 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE). [S.l.], 2019. p. 423–428.

3.2 Critérios de seleção

Os critérios de inclusão (CI) e exclusão (CE) são estabelecidos para auxiliar na condução de um MSL, com a finalidade de contribuir na seleção dos estudos (Fuzeto & Braga, 2016). Para um trabalho ser selecionado é suficiente atender a, pelo menos, um critério de inclusão. Na seleção dos estudos para este MSL, foram aplicados os seguintes critérios de inclusão (CI):

- Critério de Inclusão (CI1): o trabalho discute ou aplica ferramentas de MD relacionadas ao problema da evasão estudantil no ensino presencial;

- Critério de Inclusão (CI2): o trabalho aplica técnicas de AM relacionadas ao problema da evasão estudantil no ensino presencial;
- Critério de Inclusão (CI3): o trabalho apresenta fatores indutores ou dados relacionados ao problema da evasão estudantil no ensino presencial;
- Critério de Inclusão (CI4): o trabalho apresenta bases de dados abertas que forneçam dados escolares de estudantes;
- Critério de Inclusão (CI5): o trabalho apresenta as métricas de avaliação de modelos de AM que foram aplicadas nos experimentos realizados durante a pesquisa.

Por outro lado, para se excluir um trabalho é suficiente que ela atenda a um dos critérios de exclusão definidos a seguir.

- Critério de Exclusão (CE1): trabalhos que utilizam MD ou AM no contexto escolar, mas não abordam a evasão estudantil;
- Critério de Exclusão (CE2): estudos que não tratem das causas da evasão estudantil no ensino presencial;
- Critério de Exclusão (CE3): artigos em outros idiomas que não o Inglês;
- Critério de Exclusão (CE4): em casos de estudos duplicados apenas o mais recente será aceito.

3.3 Critérios de qualidade

Os Critérios de Qualidade (CQ) têm como finalidade avaliar aspectos metodológicos dos estudos, tais como, a relevância do tema de pesquisa e a adoção de métodos que conduzam aos objetivos propostos (Kitchenham, 2004). Observa-se que os critérios de qualidade não excluem trabalhos. Os artigos selecionados foram avaliados de acordo com 02 critérios de qualidade, apresentados a seguir.

- Critério de Qualidade (CQ1): trabalhos publicados em veículos de divulgação científica com Fator de Impacto (FI) entre 0,1 e 0,8 (meio ponto), entre 0,9 e 1,9 (1 ponto), entre 2,0 e 2,9 (2 pontos), entre 3,0 e 5,0 (3 pontos) e maior que 5,0 (5 pontos).
- Critérios de Qualidade (CQ2): o trabalho responde às questões de pesquisa (QP). Será adicionado 1 ponto para cada QP que o trabalho responder.

3.4 Definição das Palavras-chave, *String* de Busca e Fontes

A partir das questões de pesquisa, foram definidos as palavras-chave e respectivos sinônimos dispostos no Quadro 1. Com base nas palavras-chave e seus respectivos sinônimos, definiu-se a seguinte *string* de busca:

Palavra	Sinônimo em Inglês
Data Mining	*Knowledge Discovery, Data Extraction, Predicting, Prediction, Detection, Detecting, Database, Machine Learning, Deep Learning, Artificial Intelligence*
School Dropout	*School Drop-Out, School Evasion, School Retention, School Truancy, School Failure*

Quadro 1: Termos de busca

((“Data Mining” OR "Machine Learning"OR "artificial intelligence"OR "Deep Learning"OR “Knowledge discovery” OR “data extraction” OR prediction OR detection OR "database") AND (“school dropout” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”)).

Para se chegar a *string* utilizada, foram realizados testes, nos quais foram aplicadas diferentes combinações dos termos supracitados e verificados quais destas combinações retornavam o maior número de trabalhos para cada uma das fontes de busca. Os testes resultaram na adição de *“Machine Learning, Artificial Intelligence e Deep Learning”* por se tratarem de técnicas que vêm sendo utilizadas em trabalhos relacionados a evasão estudantil e que retornaram uma maior quantidade de trabalhos.

As bases de dados eletrônicas selecionadas para este MSL são: IEEE Xplore, ACM Digital Library, Science Direct, Scopus e Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Essas bases foram escolhidas, pois têm se consolidado e tornaram-se referências no campo da Ciência da Computação, em especial na área de educação em Ciência da Computação conforme reportado por Valente et al. (2022).

3.5 Processo de seleção

O processo de seleção dos estudos foi dividido em três fases: coleta de trabalhos, seleção inicial e seleção final. As fases de seleção dos estudos são descritas a seguir.

- **Fase 1 - Coleta de trabalhos**: coleta dos trabalhos relacionados aos termos de busca, retornados em cada base de dados eletrônicas.
- **Fase 2 - Seleção Inicial**: leitura de títulos e resumos das publicações coletadas na fase anterior e eleição daquelas que satisfazem aos critérios de inclusão. Trabalhos que não se enquadram em nenhum critério de inclusão são excluídos com base em alguns critérios de exclusão.
- **Fase 3 - Seleção Final**: leitura integral dos artigos selecionados na fase anterior e reavaliação dos trabalhos com base na aplicação dos critérios de inclusão e exclusão.

3.6 Condução do Estudo

Inicialmente, em janeiro de 2021, foram realizadas buscas nas bases de dados selecionadas. Observa-se que as *strings* de busca foram adaptadas para cada uma das bases, de modo a se

aquedar as suas particularidades, mas sem apresentar diferença semântica em relação à *string* apresentada na seção anterior.

A principal alteração aconteceu na base *Science Direct*, por aceitar apenas 8 palavras por pesquisa. Dessa forma, fez-se necessária a realização de três buscas com as seguintes *Strings*:

- *((“Data Mining” OR “Knowledge discovery” OR “data extraction”) AND (“school dropout” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”))*;
- *((prediction OR detection OR "database") AND (“school dropout” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”))*;
- *(("Machine Learning"OR "Deep Learning"OR "artificial intelligence") AND (“school dropout” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”))*.

Para cada retorno (buscas 1, 2 e 3), coletaram-se os trabalhos retornados e, ao final, houve a junção desses trabalhos, excluindo-se as repetições. As buscas nas bases de dados selecionadas com as strings adaptadas resultaram em um total de 4453 trabalhos retornados. Na fase 2, os artigos foram selecionados aplicando-se os critérios de inclusão e exclusão, a partir da leitura dos títulos e resumos dos artigos coletados na fase anterior. Ao final desta etapa, foram incluídos 201 trabalhos. Na última fase da seleção, os artigos selecionados na fase anterior foram lidos integralmente e os critérios de inclusão e exclusão foram reaplicados. Na Figura 2 são exibidos os números de artigos incluídos e excluídos em cada fase e agrupados por base de dados.

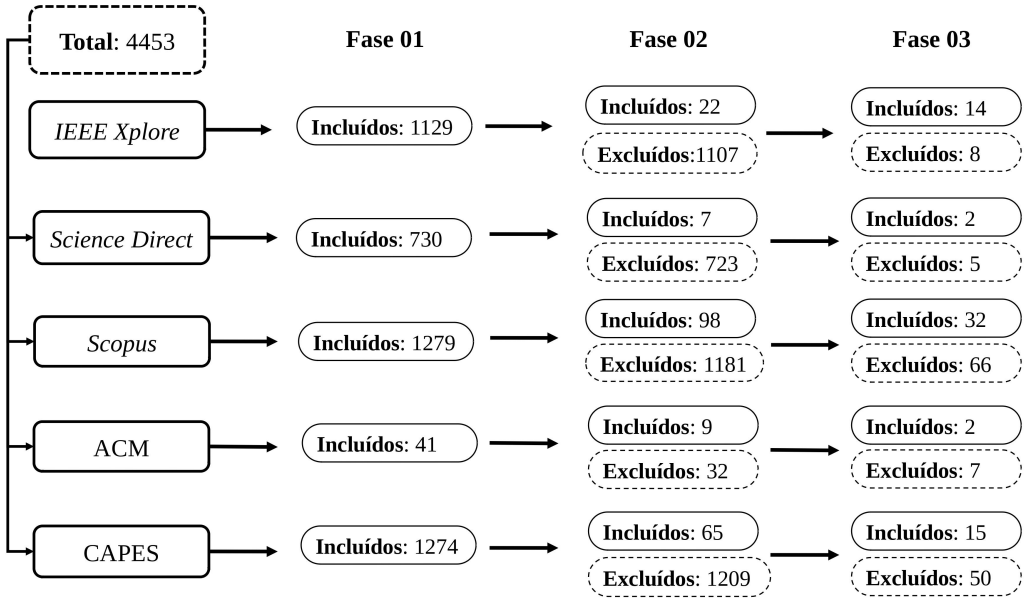

Figura 2: Quantidade de trabalhos aceitos em cada etapa do mapeamento.

Após a seleção final, os dados apresentados pelos trabalhos foram coletados e usados para responder às questões de pesquisa, conforme apresenta-se na Seção 4.

O processo de avaliação dos trabalhos ocorreu da seguinte forma: 4 pesquisadores participaram do processo de avaliação de trabalhos. Na etapa de avaliação dos resumos, o filtro inicial foi feito por dois pesquisadores que apresentaram os dados para debate e discussão com os outros pesquisadores. Na etapa de leitura completa, um pesquisador fez a leitura de todos os trabalhos e apresentou a tabela de seleção com todas as justificativas para discussão com os outros pesquisadores. Além disso, 72 trabalhos geraram dúvidas na sua avaliação e foram lidos pelos 4 pesquisadores. Todos os trabalhos incluídos foram debatidos e adicionados apenas após a confirmação da unanimidade por parte dos 4 pesquisadores.

Os trabalhos que geraram dúvidas foram debatidos pela equipe de pesquisa e enquadrados como incluído ou excluído, após decisão por consenso. Nesta fase, foram selecionados 65 artigos apresentados na Figura 3.

Figura 3: Trabalhos selecionados

ID	autor(es)
#1	(Deepika & Sathvanaravana, 2018)
#2	(Hasan, 2019)
#3	(Barros, Silva et al., 2019)
#4	(Pašić & Kučak, 2020)
#5	(Sari, Sunyoto et al., 2019)
#6	(Nagy & Molontay, 2018)
#7	(Kiss et al., 2019)
#8	(Orooji & Chen, 2019)
#9	(Dombrovskaia et al., 2020)
#10	(Tenpipat & Akkarajitsakul, 2020)
#11	(Chung & Lee, 2019)
#12	(Cruz-Jesus et al., 2020)
#13	(Meca et al., 2020)
#14	(Shiau, 2020)
#15	(Sandoval-Palis et al., 2020)
#16	(Ilieva & Yankova, 2020)
#17	(Freitas et al., 2020)
#18	(Mduma et al., 2019)
#19	(Barros, Souza Neto et al., 2019)
#20	(Lima et al., 2018)
#21	(Urbina Nájera et al., 2020)
#22	(Agrusti et al., 2019)
#23	(Lee, 2005)
#24	(Kelly et al., 2019)
#25	(Acero et al., 2019)
#26	(Bittencourt et al., 2020)
#27	(Von Hippel & Hofflinger, 2021)
#28	(ERIC, 2019)
#29	(Ripamonti, 2018)
#30	(Jimenez et al., 2019)
#31	(Al Amin Biswas et al., 2019)
#32	(Venkatesan et al., 2019)
#33	(Serra et al., 2018)

ID	autor(es)
#34	(Adelman et al., 2018)
#35	(Bahel et al., 2019)
#36	(Ortigosa et al., 2019)
#37	(da Silva et al., 2019)
#38	(do Nascimento et al., 2018)
#39	(Limsathitwong et al., 2018)
#40	(Sansone, 2018)
#41	(Jaiswal et al., 2019)
#42	(TIMBAL, 2019)
#43	(Mduma et al., 2019)
#44	(Sorensen, 2019)
#45	(Brandão et al., 2019)
#46	(Gunawan, 2019)
#47	(Nuankaew et al., 2020)
#48	(Afia et al., 2019)
#49	(Vaughn et al., 2020)
#50	(MORTAGY et al., 2018)
#51	(Thornburg, 2001)
#52	(Lemkin et al., 2018)
#53	(Lemkin et al., 2020)
#54	(Schwab, 2018)
#55	(Gilbert & Hamid, 2019)
#56	(Meens et al., 2018)
#57	(Buss et al., 1987)
#58	(BOYACI, 2019)
#59	(Rodrıguez-Muñiz et al., 2019)
#60	(Pertiwi et al., 2017)
#61	(Haugan et al., 2019)
#62	(Zuilkowski et al., 2019)
#63	(Fernández-Garcıa et al., 2021)
#64	(Opazo et al., 2021)
#65	(Bengesai & Pocock, 2021)

4 Resultados

Nesta seção são apresentadas as respostas para as questões de pesquisa, baseadas no estudo dos trabalhos selecionados. A análise dessas questões é apresentada a seguir.

1 - Quais ferramentas de MD têm sido utilizadas para a descoberta das causas da evasão estudantil?

Na Figura 4 é representado o quantitativo de uso das ferramentas de MD nos trabalhos selecionados. É possível verificar que o software R, Python e Weka destacam-se pela utilização em 10 (15%), 7 (11%) e 6 (10%) trabalhos, respectivamente. As ferramentas, *RapidMiner, Excel, SPSS*, foram utilizadas 2 (6%) trabalhos cada. Enquanto as ferramentas *MATLAB e Octave* foram utilizadas em apenas 1 (3%) trabalho, respectivamente.

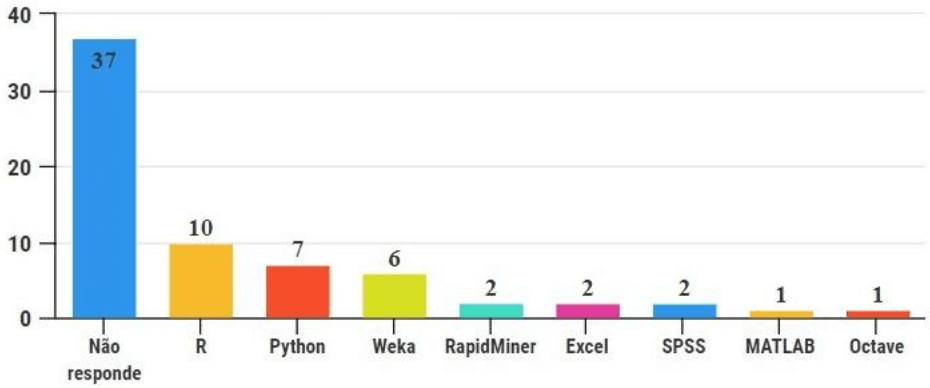

Figura 4: Ferramentas de MD.

O *software* R[1] foi utilizado nos trabalhos: #2, #8, #11, #13, #20, #22, #23, #25, #36, e #42. O R é um ambiente de *software* de licença livre, projetado para executar computação estatística e gráficos. O R é amplamente utilizado na academia, pesquisa, engenharia e aplicações industriais. O *software* possui um grande número de pacotes com modelos, fórmulas e testes estatísticos, auxiliando na análise de dados (German et al., 2013).

A linguagem de programação *Python*[2] foi utilizada nos trabalhos #12, #17, #19, #24, #26, #63 e #64. Esta linguagem é de licença livre e além de ser utilizada para MD, pode ser utilizada para o uso geral, como: *data science*, *machine learning*, desenvolvimento de aplicativos, automação de *scripts* e desenvolvimento de aplicações *web* (Stančin & Jović, 2019).

A ferramenta *Weka*[3] foi utilizada nos trabalhos #1, #22, #32, #45, #46, #59 e consiste em uma coleção de algoritmos de aprendizado de máquina para tarefas de MD, de licença livre. A ferramenta *Weka* é aplicada para MD, para processar *big data* e *deep learning* (Ratra & Gulia, 2020).

O *Statistical Package for the Social Sciences (SSPS)*[4] foi utilizado pelos trabalhos #22 e #65

[1]https://www.r-project.org/
[2]https://www.python.org/
[3]https://www.cs.waikato.ac.nz/ml/weka/
[4]https://www.ibm.com/account/reg/br-pt/signup?formid=urx-19774

e consiste em um pacote de software aplicado à análise estatística interativa (Sofyan & Kurniawan, 2009).

Ainda vale destacar a plataforma de ciência de dados baseada na nuvem para profissionais de dados *RapidMiner* (Naik & Samant, 2016), que foi utilizada nos trabalhos #6 e #39; o editor de planilhas *Microsoft Excel* que foi utilizado pelos trabalhos #20 e #50; o MATLAB, que foi utilizado pelo trabalho #5 e é um software interativo de alto desempenho voltado para o cálculo numérico (Elhorst, 2014); e, o *Octave*, que é uma linguagem computacional desenvolvida para computação matemática (Silva & Moody, 2014) e foi utilizada pelo trabalho #4.

Constata-se que as ferramentas R, *Weka* e *Python* são as mais populares na área de MD, pois além de disponibilizarem licença livre, apresentam ampla documentação e resolução de problemas por meio de comunidades online. Além do que, ambas são de fácil utilização e aprendizado para pessoas que não são cientistas de dados ou programadores, como é o caso de profissionais da área de matemática, estatística, dentre outras. Em suma, optar por uma determinada ferramenta é resultado de uma combinação de fatores, incluindo versatilidade, comunidade ativa, tradição e contexto do projeto.

2 - Quais técnicas de AM têm sido utilizadas para a descoberta das causas da evasão estudantil?

As técnicas de predição frequentemente utilizadas pela EDM e que predominaram nesta pesquisa são a classificação e a regressão (Costa et al., 2013).

A técnica de classificação foi utilizada em 67% dos trabalhos selecionados. Os trabalhos que utilizaram esta técnica são: #1, #2, #4, #5, #6, #7, #8, #9, #10, #11, #12, #13, #14, #15, #17, #19, #20, #21, #22, #23, #24, #25, #26, #27, #30, #31, #32, #33, #35, #36, #39, #40, #41, #42, #43, #44, #45, #46, #50, #58, #59, #60, #63 e #65. Essa técnica consiste em associar objetos a determinadas classes, objetivando prever automaticamente a qual classe um novo dado pode ser associado (Ignacio, 2021). Um exemplo aplicado à evasão estudantil, consiste em analisar múltiplas variáveis que estão relacionadas as características dos estudantes e associá-las a classes de estudantes que evadem ou permanecem no curso.

A Regressão foi explorada em 16% dos trabalhos avaliados e identificados pelos números: #1, #7, #13, #16, #18, #34, #35, #37, #38, #41 e #48. A Regressão é uma técnica de MD que busca modelar o relacionamento de variáveis independentes (chamadas preditoras) com uma variável dependente (chamada resposta). As variáveis preditoras são os atributos dos registros, e a resposta é o que se quer predizer (Bonaccorso, 2017).

A análise de correspondência, abordada pelos trabalhos #3 e #14 é uma técnica de análise multivariada, adequada para dados categóricos, que permite analisar graficamente as relações existentes por meio da redução de dimensionalidade do conjunto de dados.

A análise fatorial, utilizada pelo trabalho #50, é uma técnica da estatística multivariada que objetiva a redução do número de variáveis iniciais em uma determinada análise, minimizando a perda de informação. Essa técnica cria um grupo reduzido de fatores comuns não observados a partir da interpretação das inter-relações existentes no conjunto de variáveis observadas.

Parte dos trabalhos analisados apresentam uma combinação de técnicas de classificação e regressão, conforme observa-se na Figura 5. Essa técnica, conhecida como abordagem híbrida, é interessante e relevante, pois pode permitir uma análise mais abrangente e flexível dos dados.

Observa-se que cada técnica apresenta particularidades, que podem se adequar melhor a diferentes problemas. Dessa forma, a combinação de métodos pode permitir a modelagem de relações complexas nos dados, que não seriam totalmente capturadas por apenas uma técnica. Com o objetivo de auxiliar na compreensão das diferenças entre essas duas técnicas, apresenta-se no Quadro 2, algumas características que as distinguem.

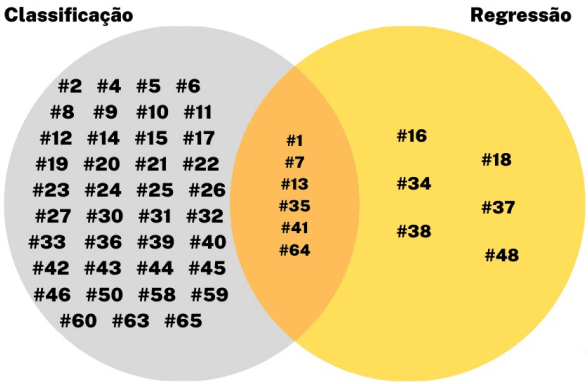

Figura 5: Técnicas de MD.

Comparação	**Classificação**	**Regressão**
Comportamento	Modela uma função por meio da qual os dados são previstos em rótulos de classe discretos	Modelo no qual o mapeamento de objetos é feito em valores contínuos
Valores de previsão	Valores discretos	Valores contínuos
Algoritmos	Árvore de decisão, regressão logística, etc.	Árvore de regressão (floresta aleatória), regressão linear, etc
Natureza dos dados	Não ordenado.	Pedido
Método de cálculo	Precisão de medição.	Medição do erro quadrático médio da raiz

Quadro 2: Características dos algoritmos mais utilizados nos trabalhos selecionados

A escolha dos algoritmos de classificação e regressão apropriados para serem combinados é fundamental, exigindo conhecimento especializado e experimentação cuidadosa para determinar quais algoritmos funcionam bem juntos. Dessa forma, alguns algoritmos de AM podem ser utilizados tanto para problemas de Classificação quanto para problemas de Regressão, conforme apresentado no Quadro 3. Neste quadro os algoritmos que mais se destacaram em relação aos trabalhos selecionados são apresentados, definidos e enquadrados de acordo com a técnicas utilizada.

3 - Quais fatores indutores têm sido investigados para a descoberta das causas da evasão estudantil?

Um dos pioneiros no desenvolvimento de pesquisas acerca da evasão estudantil e dos fatores que a desencadeiam foi Tinto (1975) , consolidando-se como principal estudo para compreender as causas da evasão estudantil. No Brasil, a principal referência dessa temática é a Comissão Especial de Estudos da Evasão, que destaca as características individuais do aluno, os fatores internos

às instituições e os fatores externos às instituições, como os principais fatores relacionados a desistência de alunos (MEC, 2019).

Algoritmo	Definição	Técnica
Decision Tree	Usado para prever a classe ou o valor da variável de destino, aprendendo regras de decisão simples, inferidas de dados anteriores (dados de treinamento)	Regressão e Classificação
Logistic Regression	Usado para classificar os dados em duas ou mais classes	Classificação
Randon Forest	Combina a saída de várias árvores de decisão para chegar a um único resultado	Regressão e Classificação
Artificial Neural Network	Imita as operações do cérebro humano para reconhecer as relações entre grandes quantidades de dados	Classificação e Regressão
Naive Bayes	Funciona com base no teorema de probabilidade de Bayes para prever a classe de conjuntos de dados desconhecidos	Classificação
Support Vector Machine	Funciona criando uma linha ou um hiperplano que separa os dados em classes	Regressão e Classificação

Quadro 3: Algoritmos que mais se destacaram em relação aos trabalhos selecionados

As características individuais, representam dados inerentes ao contexto individual de cada aluno. Exemplos dessas características são: fatores acadêmicos (desempenho do estudante nas avaliações, faltas, participação em atividades extracurriculares), falta de habilidade de estudo, repetência, formação anterior precária, dentre outros (Tinto, 1975). Esses, além de serem os mais pesquisados no campo da evasão, também são os mais relevantes diante dos modelos de AM.

Fatores internos à instituição estão relacionados à estrutura da instituição, grade curricular, corpo docente desinteressado ou com qualificação insuficiente, ausência de professores ou até falta de refeitório no campus (Marques, 2020). Geralmente esses fatores são investigados combinados aos dados de características individuais. Os Fatores Externos à instituição são fatores relacionados à vulnerabilidade socioeconômica, como por exemplo, o estudante não conseguir conciliar estudo e trabalho, falta de políticas públicas adequadas para a educação, entre outros (Tinto, 1975).

No Quadro 4 são exibidos os fatores indutores e principais variáveis relacionadas ao risco de evasão estudantil abordados pelos trabalhos selecionados neste mapeamento. Adicionalmente, na Figura 6, são ilustradas as categorias dos fatores investigados em cada trabalho. Por meio da figura é possível perceber que, embora alguns trabalhos abordem a investigação de mais de uma categoria de fatores indutores, apenas 1 trabalho investiga a evasão considerando as três categorias.

Os resultados desta questão constatam que os principais fatores indutores investigados para a descoberta das causas da evasão estudantil ainda são relacionados às características individuais do estudante, por meio da análise de dados armazenados nas instituições ou levantando-se dados por meio de questionários, sob a perspectiva do estudante. Fatores relacionados às características internas à instituição vêm sendo combinadas com as características individuais para traçar um perfil do estudante.

Fator	ID	Variáveis
Individuais do estudante	#1, #2, #3, #7, #8, #9, #10, #12, #13, #14, #15, #17, #18, #19, #20, #23, #25, #26, #28, #29, #31, #33, #34, #36, #37, #38, #42, #43, #45, #46, #47, #48, #49, #50, #51, #53, #54, #55, #56, #57, #58, #59, #60, #62, #63	gênero, idade, cor, renda, quantidade de pessoas na mesma casa, se vive com os pais, escolaridade dos pais, estado civil, possuir filhos, uso de substâncias, doenças, oriundo de escola pública, frequência nas aulas, faltas, reprovações, notas de português e matemática no ensino médio, médias no ensino médio, engajamento na escola, realiza atividades extracurriculares
Interno a instituição	#10, #12, #13, #15, #16, #21, #26, #33, #34, #36, #37, #39, #42, #51, #54, #58, #62	recursos do corpo docente, presença de mentoria ou aconselhamento, ambiente escolar adequado, acompanhamento acadêmico, programas de apoio, oferece atividades extracurriculares
Externo a Instituição	#4, #10, #16, #31	situação profissional, ensino anterior de baixa qualidade, pouca oferta no mercado de trabalho

Quadro 4: Fatores indutores e principais variáveis encontradas nos trabalhos selecionados.

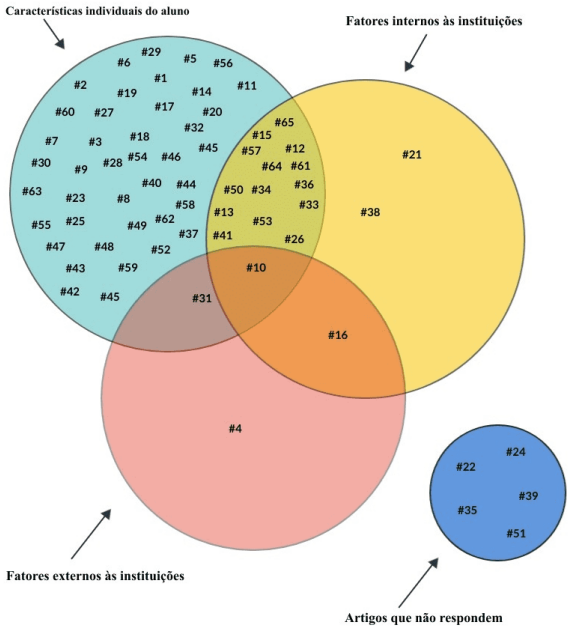

Figura 6: Combinação de categorias de fatores nos trabalhos encontrados.

Fatores externos à instituição são pouco considerados quando comparados aos demais fatores, indicando uma área que pode ser explorada com mais profundidade. Fatores como condições socioeconômicas, influências familiares e características da comunidade também podem desempenhar um papel importante na evasão.

Estes resultados destacam a importância de manter esforços na investigação e compreensão

dos fatores que contribuem para a evasão estudantil. A ênfase nas características individuais dos estudantes e nas características internas à instituição indica a necessidade de testar e validar abordagens mais amplas, que também levem em consideração os fatores externos à instituição, a fim de obter uma compreensão mais completa e robusta desse fenômeno complexo.

4 - Quais bases abertas fornecem dados escolares de estudantes?

Entre os 65 trabalhos selecionados, a disponibilidade de dados escolares de estudantes se apresenta da seguinte maneira:

- 15 Bases Abertas: Isso representa 23% dos trabalhos analisados, evidenciando que uma parcela significativa da pesquisa em evasão estudantil utiliza fontes de dados acessíveis ao público.
- 37 Não Informam: Um número considerável de trabalhos (57%) não fornece informações sobre a origem dos dados utilizados, o que pode indicar uma falta de transparência na fonte dos dados.
- 13 Bases Privadas - Bases Universitárias: Treze trabalhos (20%) fazem uso de bases de dados privadas, geralmente mantidas por instituições de ensino superior.

Nesse sentido, observa-se que a privacidade dos dados dos estudantes é uma consideração crítica na pesquisa em evasão estudantil. A falta de disponibilidade de dados específicos pode apresentar desafios, mas também incentiva os pesquisadores a desenvolver abordagens éticas e transparentes para abordar o problema. A colaboração com instituições educacionais e o uso de metodologias gerais são estratégias que podem ser adotadas para superar essas limitações. No Quadro 5 apresentam-se os trabalhos e as bases abertas utilizadas pelos mesmos.

ID	Base de Dados
#1	Dados da *University of California at Irvine (UCI), Machine Learning Repository*
#2	Fundo das Nações Unidas para a Infância (UNICEF)
#8	*Social Research and Evaluation Center* da *Louisiana State University*
#10	*KMUTT Registrar's Office*
#11	*National Education Information System (NEIS) in Korea*
#12	Direção-Geral de Estatística da Educação e Ciência (Portugal)
#18	Dados da *George Mason University, Universidade de Minnesota* e da *Stanford University*
#23	*National Education Information System (NEIS)* da Coréia do Sul
#24	*Department of Computer Science in the UFS*
#27	Departamento de Avaliação, Medição e Registros Educacionais (DEMRE) - Chile
#34	Conjuntos de dados retirados da coleção UCI
#37	Censo da Educação Superior e seus Indicadores de Fluxo (INEP 2013)
#38	INEP - microdados do Censo Escolar e Taxa de Desempenho Escolar - 2016
#44	Dados administrativos do Departamento de Instrução Pública da Carolina do Norte
#58	Dados do *Family Structure Survey* - Instituto de Estatística da Turquia (2016)

Quadro 5: Bases de Dados abertas

Algumas das bases de dados mencionadas incluem:

- *UCI - Machine Learning Repository: Uma fonte amplamente reconhecida que é utilizada por alguns trabalhos: #1, #16, #32.*
- INEP: Embora o INEP seja mencionado como fonte de dados, é importante notar que existem duas bases distintas, uma de 2013 e outra de 2016 (referidas como bases 37 e 38). Além disso, é relevante destacar que a base do INEP é diferente da *UCI - Machine Learning Repository, pois esta última é um repositório de diversas bases que pesquisadores podem utilizar em seus estudos.*

5 - Quais métricas de avaliação de modelos de AM foram aplicadas nos experimentos realizados?

Métricas de avaliação são utilizadas com a finalidade de avaliar a qualidade do modelo por meio da avaliação do desempenho do modelo aplicado. A avaliação do modelo pode determinar se ele terá bom desempenho na previsão do destino em dados novos e futuros. Dos 65 trabalhos selecionados, 61% (40 trabalhos) aplicaram uma ou mais métricas para avaliar os modelos de AM.

A métrica *Recall* ou *Sensibility* (Dombrovskaia et al., 2020), foi aplicada em 16 (26%) trabalhos: #4, #7, #9, #10, #12, #17, #21, #25, #31, #33, #34, #40, #41, #46 e #60. Essa métrica objetiva calcular a proporção de positivos reais identificados incorretamente (Grandini et al., 2020). O cálculo é feito pelo número de resultados positivos corretos, dividido pelo número de todas as amostras que deveriam ter sido identificadas como positivas. Quanto maior o *recall*, mais amostras positivas foram detectadas.

A métrica *Accuracy* é considerada uma das mais simples de implementar, conhecida como a métrica global, dado que seu resultado pode ser calculado como a razão entre o número de previsões corretas para o número total de previsões. Esta métrica funciona bem apenas se houver igual número de amostras pertencentes a cada classe, podendo comprometer sua avaliação, caso o conjunto de dados seja desbalanceado (Dombrovskaia et al., 2020). A *Accuracy* foi aplicada em 12 (18%) trabalhos selecionados: #7, #10, #12, #15, #17, #21, #24, #25, #31, #33, #34 e #60.

A métrica *Precision* (Freitas et al., 2020) foi aplicada nos trabalhos #4, #7, #8, #10, #17, #21, #31, #46 e #60, o que corresponde a 13% dos trabalhos selecionados. Esta métrica determina a exatidão do modelo em classificar uma amostra como positiva. Seu cálculo é dado pela razão entre o número de amostras positivas classificadas corretamente e o número total de amostras classificadas como positivas (correta ou incorretamente). When et al. (2012) pontuam que "a *precision* reflete a confiabilidade do modelo em classificar as amostras como positivas". Para tanto, além da métrica classificar todas as amostras positivas como positivas, é importante ela não classificar erroneamente uma amostra negativa como positiva.

A *F1 Score* é uma métrica de AM utilizada em modelos de classificação, que tem como base a combinação das métricas *precision* e *recall* em uma única métrica. O *F1 Score* é definido como a média harmônica de *precision* e *recall*. Desse modo, a presente métrica atribui peso igual às duas métricas, de tal forma que um modelo obterá uma pontuação alta na *F1 Score* se a *precision* e o *recall* forem altos. Por outro lado, um modelo obterá uma pontuação *F1 Score* baixa se *precision* e o *recall* forem baixos. Por fim, um modelo poderá obter uma pontuação *F1 Score* média se uma das duas métricas que formam a combinação obtiver pontuação baixa e a outra obtiver pontuação

alta, ou se as duas métricas obtiverem um valor médio(Grandini et al., 2020). A *F1 Score* foi utilizada em 10% dos trabalhos selecionados: #4, #9, #10, #17, #31, #40 e #46.

A métrica de *Confusion Matrix* foi aplicada pelos trabalhos #10, #11, #13, #21, #41, #47 e #61, correspondendo a 10% dos trabalhos selecionados. A *Confusion Matrix* é uma métrica que é extraída por meio de uma matriz que possibilita determinar o desempenho dos modelos de classificação para um conjunto de dados de teste nos quais a variável resposta é categórica. A matriz é dividida em duas dimensões: os valores previstos pelo modelo e os valores reais, que são os valores verdadeiros para as observações dadas juntamente com o número total de previsões (Freitas et al., 2020). É importante destacar que, por meio da *confusion matrix*, é possível calcular os diferentes parâmetros do modelo, como a exatidão, precisão, dentre outros. Desse modo, a*confusion matrix* pode servir como parâmetro para outras métricas como a *recall* e *precision* (Stančin & Jović, 2019).

Receiver Operating Characteristic Curve - ROC foi utilizada em 6 (9%) trabalhos: (#6, #8, #11, #12, #15 e #23). Essa é uma métrica para avaliar o desempenho do modelo de classificação por meio de um gráfico em curva. A *ROC* representa um gráfico para mostrar o desempenho do modelo em diferentes níveis de limiar. A curva *ROC* é desenhada sob dois parâmetros: a taxa de verdadeiros positivos e a taxa de falsos positivos. De forma complementar, a *AUC* calcula o desempenho em todos os limites e fornece uma medida agregada. O valor da *AUC* possui variação entre 0 (um modelo com previsão 100% errada terá uma *AUC* de 0,0) e 1 (modelos com previsões 100% corretas terão uma *AUC* de 1,0) (Muschelli, 2020).

Outras métricas foram aplicadas em um menor número de trabalhos: *Probability of Correct Classification - PCC* e *Root Mean Square Error (RMSE)* foram aplicadas apenas no trabalho #1; *Area under the receiver operating characteristic (AUROC)*, foi aplicada no trabalho #2; *Missclassification Error (ME)*, aplicada pelo trabalho #4; *Sensitivity* aplicada nos trabalhos #10, #31, #33 e #34; *Lift* utlizada nos trabalhos #12 e #27; *Specificity*, *False Positive Rating - FPR* e *False Negative Rating -FNR*, foram aplicadas pelo trabalho #31; *Overall* foi utilizada apenas pelo trabalho #34; *Absolute Mean error (MAE)*, foi aplicada nos trabalhos #38 e #39; *Mean Square Error (MSE)* foi aplicada nos trabalhos #38 e #41.

Observa-se que: (I) a maioria dos trabalhos aplicou duas ou mais métricas de avaliação nos experimentos realizados. Este é um fator positivo, pois mostra uma abordagem abrangente para avaliar o desempenho dos modelos de AM, considerando múltiplos aspectos; (II) os trabalhos #16, #22, #28, #29, #30, #35, #36, #42, #44, #48, #49, #50, #51, #52, #53, #54, #55, #56, #57, #58, #62 e #63 não responderam à pergunta de pesquisa por não aplicarem de forma prática os algoritmos de AM. Esses trabalhos tratam de estudos exploratórios ou estudos de caso. O trabalho #3 não responde à pergunta por se tratar de um estudo de visualização de dados que não aplica algoritmos de AM; (III) por fim, os trabalhos #5, #14, #20, #32, #37, #45 e #59 não informam a utilização de métricas de avaliação de modelos de AM nos experimentos realizados, o que pode denotar uma limitação em termos de transparência na pesquisa. A falta de relato sobre as métricas de avaliação dificulta avaliar o rigor dos experimentos e a validade dos resultados. Isso também destaca a importância da comunicação clara de metodologia. É fundamental salientar a importância da transparência na pesquisa, incluindo a documentação adequada das métricas de avaliação utilizadas em experimentos, para que outros pesquisadores possam avaliar e replicar os estudos com confiança.

4.1 Qualidade dos Trabalhos Selecionados

Ao final da extração de dados, coleta das respostas para as perguntas de pesquisa e aplicação dos critérios de qualidade, obteve-se um *ranking* de trabalhos selecionados, com base na pontuação aplicada aos critérios de qualidade, conforme se apresenta nesse link: ranking de trabalhos selecionados. Os comentários sobre os trabalhos mais bem ranqueados são apresentados a seguir.

- O trabalho de Barros et al. (Barros et al., 2019) no qual a pesquisa é desenvolvida por meio de técnica de análise de correspondência para relacionar os motivos de desistências de estudantes e investigar como o uso de técnicas de visualização podem identificar as principais características socioeconômicas e demográficas de estudantes evadidos. Para a validar o estudo, foram analisados dados educacionais de estudantes dos cursos de ensino médio integrado de um Instituto Federal no Brasil.

- Por meio de técnicas de análise de correspondência, o trabalho de Lima et al. (Lima et al., 2018) analisou uma amostra de 1.844 estudantes, entre graduados e evadidos, no período de 2007 a 2015 para analisar perfis e graus de estudantes com base em notas, número de tentativas e outros indicadores de desempenho. O trabalho também apresentou uma proposta de um modelo baseado em árvores de decisão, objetivando a geração de instruções padronizadas, de fácil interpretação e permitindo a adição de diversos resultados possíveis, contribuindo para o processo de tomada de decisão.

- O trabalho de Deepika et al. (Deepika & Sathvanaravana, 2018) se concentra em aplicar diferentes métodos de categorização com base nas características dos estudantes para classificar os estudantes com base em informações pessoais, colhidas pela instituição de ensino. Para tanto, os pesquisadores usaram a ferramenta *Weka* com os algoritmos de classificação e regressão: *Linear Regression, Radom Forest e Support Vector Machine*. Como pontos positivos desse trabalho destaca-se a utilização de um repositório de dados abertos da *(University of California at Irvine (UCI)* chamado de *Machine Learning Repository*[5], bem como o enfoque em classificar os estudantes com base no desempenho dos estudantes nas disciplinas de Matemática e Português. Como pontos de melhoria, observam-se: a pesquisa não deixa clara as etapas de análise e preparação dos dados, dificultado a reprodutibilidade; a pesquisa não faz uso de métricas conhecidas para avaliar o desempenho dos algoritmos.

- Pašić e Kučak (Pašić & Kučak, 2020) desenvolveram um modelo de AM a partir de dados coletados de escolas de ensino médio em que os estudantes cursaram antes de ingressar na Universidade, a fim de calcular a probabilidade do estudante não concluir a Universidade. O modelo de AM ainda classifica o estudante que não concluirá e recomenda para este estudante outro programa de estudos. O trabalho foi realizado por meio da ferramenta *Octave* com algoritmos de regressão. Para avaliação dos modelos, os autores utilizaram: *Recall, Precision* e *F1-Score*. Apesar dos excelentes resultados alcançados na aplicação dos modelos de AM por meio das métricas de avaliação, dos cálculos a serem realizados pelos algoritmos de regressão e das hipóteses bem definidas, o trabalho deixa dúvidas com relação à preparação dos dados, a quais os algoritmos de regressão foram utilizados e quantas vezes cada algoritmo rodou com a validação cruzada. O número do conjunto de dados (161) e os atributos escolhidos para execução dos modelos (4 atributos) são considerados pequenos.

[5] disponível em http://UCI/machinelearning/repository

- No estudo de Sari et al. (Sari, Sunyoto et al., 2019), a previsão de evasão de estudantes foi realizada para os estudantes que tiveram a possibilidade de ultrapassar o período máximo de estudo. As previsões são feitas a partir de dados no banco de dados acadêmico do estudante, utilizando os dados do índice de desempenho acadêmico de cada semestre e frequência às aulas. Os algoritmos utilizados foram *backpropagation (BP)* otimizada com *Particle Swam Optimization (PSO)*. A avaliação do modelo de classificação é feita com validação cruzada de 10 vezes: os dados são divididos em dados de treino e teste, treinados e o teste é feito 10 vezes até obter a maior precisão. O trabalho apresenta de forma resumida uma metodologia clara, os dados, algoritmos e métricas utilizadas. Como pontos de melhorias, destaca-se que a pesquisa poderia detalhar melhor a seleção dos atributos que foram submetidos ao modelo na etapa de preparação dos dados; aplicação de uma métrica complementar à acurácia, pois essa significa apenas o acerto global do algoritmo. Acredita-se que, adicionar o resultado da matriz de confusão ao presente estudo traria uma melhor compreensão do comportamento do modelo de AM, porcentagem de previsões corretas e incorretas.

- No trabalho de Nagy e Molontay (Nagy & Molontay, 2018), foram aplicados e avaliados vários algoritmos de AM para identificar estudantes em risco e prever a evasão de estudantes em programas universitários com base nos dados disponíveis no momento da inscrição (desempenho no ensino médio, dados pessoais). Também apresentaram uma plataforma de apoio à decisão baseada em dados para a diretoria de educação e partes interessadas. Os algoritmos aplicados foram: *Naive Bayes, KNN, Linear Models e Deep Learning*, por meio da ferramenta *RapidMiner* e para avaliar a performance de cada modelo, utilizaram a curva ROC.

  Dentre os trabalhos relacionados apresentados, o trabalho de Nagy e Molontay (Nagy & Molontay, 2018) apresenta a melhor proposta: maior conjunto de dados, metodologia clara e definida, aplicação de alguns algoritmos de AM, bem como a avaliação por meio das principais métricas de avaliação de desempenho. O trabalho ainda apresenta como diferencial a disponibilização da plataforma de auxílio. Alguns pontos que poderiam ser melhorados tem relação aos resultados das métricas que alcançaram entre 50% e 80% de acerto. Acredita-se que, mais testes com os demais atributos disponíveis, bem como alterações nos parâmetros de configuração dos classificadores elevariam os resultados da pesquisa.

5 Discussão dos Resultados

Com relação as ferramentas utilizadas nas pesquisas, 37 (55%) dos trabalhos selecionados não respondem a esta pergunta. Acredita-se que os artigos selecionados não mencionam as ferramentas de MD utilizadas, pois concentraram-se apenas em apresentar os resultados obtidos. A falta de informação sobre as ferramentas dificulta ou até impossibilita replicar os resultados e compromete a validade e a confiabilidade da pesquisa, uma vez que outros pesquisadores não conseguem reproduzir os métodos ou verificar os resultados. A divulgação das ferramentas utilizadas não apenas permite a replicação, mas também ajuda outros pesquisadores a entenderem melhor o processo de pesquisa e a avaliar a qualidade do trabalho. Essa lacuna, causada pela ausência de informações sobre as ferramentas, dificulta a comparação de resultados entre diferentes estudos, o que é particularmente problemático em campos nos quais a comparação e a síntese de estudos são cruciais

para o avanço do conhecimento.

Com relação as técnicas de AM, é importante destacar que alguns trabalhos realizaram a combinação de técnicas. Entretanto, vale destacar que combinar duas técnicas como classificação e regressão, pode ocasionar em algumas desvantagens, como por exemplo a necessidade de trabalhar com modelos mais complexos, o que pode dificultar a interpretação dos resultados e exigir recursos computacionais significativos. Outra questão a ser observada ao combinar duas técnicas é o sobreajuste, no qual o modelo se ajusta muito bem aos dados de treinamento, mas não generaliza para novos dados.

As pesquisas que utilizam modelos combinados podem ser menos interpretáveis do que modelos simples de classificação ou regressão. Adicionalmente, a avaliação de modelos combinados pode ser complicada. Portanto, é necessário submeter os resultados a métricas de desempenho apropriadas e realizar validação cruzada adequada para garantir resultados confiáveis. Por fim a abordagem híbrida não garante necessariamente um melhor desempenho em todos os cenários. Em alguns casos, uma técnica simples pode ser mais eficaz do que uma combinação complexa.

Com relação aos dados utilizados nas pesquisa avaliadas é importante destacar algumas considerações: I - privacidade dos dados: a ênfase na privacidade dos dados dos estudantes é uma preocupação legítima e ética. A proteção dos dados pessoais é fundamental, especialmente em contextos educacionais, nos quais as informações sensíveis podem ser coletadas; II - limitações na acessibilidade dos dados: a falta de disponibilidade de dados pode limitar a replicabilidade e a comparação entre estudos, uma vez que outros pesquisadores não têm acesso aos mesmos conjuntos de dados para verificar os resultados; III - necessidade de dados sintéticos ou amostras representativas: quando os dados não podem ser disponibilizados devido a preocupações com a privacidade, os pesquisadores podem considerar o uso de dados sintéticos (dados gerados artificialmente que mantêm características importantes) ou amostras representativas que preservem a privacidade dos estudantes; IV - colaboração com instituições educacionais: para estudos que dependem de dados específicos de instituições, a colaboração com essas instituições pode ser uma abordagem valiosa. Isso pode envolver a obtenção de permissão para acessar e utilizar dados de forma ética e segura.

6 Considerações Finais

No presente estudo, foi apresentado o protocolo e condução de um Mapeamento Sistemático da Literatura acerca da evasão estudantil, em que se buscou elencar ferramentas, técnicas, fatores indutores e bases de dados abertas que foram utilizadas para predição das causas do problema de evasão, entre os anos de 2018 e 2021.

A busca pelos trabalhos resultou na pré-seleção de 4453 trabalhos, dentre os quais sessenta e cinco (65) foram incluídos para a extração de dados. A partir dos resultados, evidenciou-se que a ferramenta R se consolida como uma das mais utilizadas para a descoberta das causas da evasão estudantil, por meio da aplicação de técnicas e algoritmos nos conjuntos de dados. Com esse pacote, encontram-se modelos, fórmulas e testes estatísticos que auxiliam na análise de dados.

Verificou-se que, a técnica de classificação vem sendo bastante utilizada nas previsões de tendência de evasão estudantil. Os algoritmos de classificação mais utilizados nos trabalhos fo-

ram: *decision tree, logistic regression, random forest, naive bayes, support-vector machine, e multilayer perceptron*. Para selecionar o algoritmo que melhor resolve determinado problema se faz necessário analisar alguns fatores, como: complexidade, tempo e quantidade de dados. Os algoritmos mencionados mostram-se adequados para o problema da evasão estudantil.

Os principais fatores indutores para a evasão estudantil, se relacionam com as características individuais do estudante, analisando dados armazenados nas instituições ou levantando dados por meio de questionários, da perspectiva do estudante. Fatores relacionados às características internas à instituição vêm sendo combinadas com as características individuais para traçar um perfil do estudante. Fatores externos à instituição são pouco considerados quando comparados aos demais fatores. Faz-se necessária uma investigação minuciosa para descobrir se isso se dá por falta de dados relacionados a estes fatores ou a pouca influência dos mesmos na problemática da evasão estudantil.

Com relação à pergunta de utilização de bases abertas, apenas quinze (15) atuaram com bases abertas. Isso significa que existem poucas fontes de dados abertas para se trabalhar a evasão de estudantes. Boa parte dos pesquisadores atua com dados privados das instituições a qual tem ligação.

Para avaliar os experimentos realizados, a maioria dos trabalhos selecionados (40 trabalhos) aplicaram duas ou mais métricas de avaliação. As métricas mais aplicadas foram: *Recall*, *Accuracy*, *Precision*, *F1 Score*, *Cross validation*, *Confusion Matrix* e *Receiver Operating Characteristic Curve - ROC*.

Os resultados deste trabalho proporcionam algumas percepções sobre a pesquisa no campo da evasão estudantil como por exemplo: (I) uma visão abrangente do estado da arte da pesquisa em evasão estudantil, incluindo as ferramentas e técnicas mais populares e os fatores indutores mais investigados; (II) ênfase na Privacidade de Dados, observada a partir da disponibilidade de poucas bases de dados abertas, o que ressalta a importância da privacidade de dados dos estudantes e da necessidade de adotar abordagens éticas na pesquisa; (III) a falta de informações sobre métricas de avaliação em alguns trabalhos destaca a necessidade de maior transparência metodológica nas pesquisas para fins de reprodutibilidade.

Adicionalmente, esta pesquisa permitiu identificar lacunas na área, como a baixa disponibilidade de bases de dados abertas, a ausência de informações sobre métricas de avaliação em alguns estudos e a necessidade de criar e avaliar modelos que utilizem os três tipos de fatores indutores. Adicionalmente, apresentam-se como limitações deste trabalho:

- Algumas pesquisas não foram consideradas por não deixar claro em sua metodologia se tratava das causas da evasão estudantil ou desempenho do estudante.
- Em algumas pesquisas, a evasão não era tratada como estudo principal e sim como uma consequência do fraco desempenho do estudante.
- Grande parte das pesquisas deixaram lacunas em seus métodos, no sentido de não justificar a escolha dos algoritmos de AM utilizados, técnicas e métricas adotadas, bem como em qual nível de ensino se tratavam os dados (educação básica, ensino médio, superior ou pós-graduação) ou qual cursos superiores eram considerados na pesquisa. Isso dificulta a replicação do trabalho, impedindo a reprodutibilidade.

Por fim, consideram-se duas principais ameaças à validade do presente estudo:

1. Não há garantias de que a *string* de busca utilizada é a melhor possível, pois alguma palavra-chave importante pode ter sido omitida. Para mitigar esta ameaça, foram realizados diversos testes com algumas opções de *strings* antes da realização da pesquisa. Além disso, o fato dos artigos de controle terem sido encontrados nas buscas, geram indícios da qualidade da *strings* utilizada;

2. Como a quantidade de trabalhos encontrada no mapeamento foi elevada, não se aplicou a técnica de *snowball* que teria potencial de encontrar outras trabalhos. Este problema foi mitigado pela quantidade considerável de trabalhos avaliados.

Referências

Barros, T. M., Silva, I., & Guedes, L. A. (2019). Determination of Dropout Student Profile Based on Correspondence Analysis Technique. *IEEE Latin America Transactions*, *17*(09), 1517–1523. https://doi.org/10.1109/TLA.2019.8931146 [GS Search].

Bonaccorso, G. (2017). *Machine learning algorithms*. Packt Publishing Ltd. [GS Search].

Costa, E., Baker, R. S., Amorim, L., Magalhães, J., & Marinho, T. (2013). Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, *1*(1), 1–29. [GS Search].

de Campos, A., Galafassi, C., Bastiani, E., Paz, F. J., Campos, R. L., Wives, L. K., Cazella, S. C., Reategui, E. B., & Barone, D. A. C. (2020). Mineração de Dados Educacionais e Learning Analytics no contexto educacional brasileiro: um mapeamento sistemático. *Informática na educação: teoria & prática*, *23*(3 Set/Dez). https://doi.org/https://doi.org/10.22456/1982-1654.102618 [GS Search].

Deepika, K., & Sathvanaravana, N. (2018). Analyze and predicting the student academic performance using data mining tools. *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 76–81. https://doi.org/10.1109/ICCONS.2018.8663197 [GS Search].

Dombrovskaia, L., José, P., & Rodríguez, P. (2020). Prediction of student's retention in first year of engineering program at a technological chilean university. *2020 39th International Conference of the Chilean Computer Science Society (SCCC)*, 1–4. https://doi.org/10.1109/SCCC51225.2020.9281195 [GS Search].

Elhorst, J. P. (2014). Matlab software for spatial panels. *International Regional Science Review*, *37*(3), 389–405. https://doi.org/10.1177/0160017612452429 [GS Search].

Faria, S. M. S. M. L. d. (2014). *Educational data mining e learning analytics na melhoria do ensino online* [tese de dout.]. [GS Search].

Freitas, F. A. d. S., Vasconcelos, F. F., Peixoto, S. A., Hassan, M. M., Dewan, M., Albuquerque, V. H. C. d., et al. (2020). IoT System for School Dropout Prediction Using Machine Learning Techniques Based on Socioeconomic Data. *Electronics*, *9*(10), 1613. https://doi.org/10.3390/electronics9101613 [GS Search].

Fuzeto, R., & Braga, R. (2016). Um mapeamento sistemático em progresso sobre internet das coisas e educação à distância. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, *5*(1), 1334. https://doi.org/10.5753/cbie.wcbie.2016.1334 [GS Search].

German, D. M., Adams, B., & Hassan, A. E. (2013). The evolution of the R software ecosystem. *2013 17th European Conference on Software Maintenance and Reengineering*, 243–252. https://doi.org/10.1109/CSMR.2013.33 [GS Search].

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*. https://doi.org/10.48550/arXiv.2008.05756 [GS Search].

Ignacio, L. F. F. (2021). Aprendizado de máquina: da teoria à aplicação. [GS Search].

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*(2004), 1–26. [GS Search].

Lima, J., Alves, P., Pereira, M., & Almeida, S. (2018). Using academic analytics to predict dropout risk in engineering courses. *17th European Conference on e-Learning ECEL 2018*, 316–321. [GS Search].

Lobo, M. B. d. C. M. (2012). Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. [GS Search].

Marques, L. T. (2020). Mateo: uma abordagem de descoberta de conhecimento para desvendar as causas da evasão escolar. [GS Search].

Marques, L. T., De Castro, A. F., Marques, B. T., Silva, J. C. P., & Queiroz, P. G. G. (2019). Mineração de dados auxiliando na descoberta das causas da evasão escolar: Um Mapeamento Sistemático da Literatura. *Revista Novas Tecnologias na Educação*, *17*(3), 194–203. https://doi.org/10.22456/1679-1916.99470 [GS Search].

MEC, M. d. E. (2019). *Censo da Educação Superior*. https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior/resultados (Acesso em: 01.03.2022).

Muschelli, J. (2020). ROC and AUC with a binary predictor: a potentially misleading metric. *Journal of classification*, *37*(3), 696–708. https://doi.org/10.1007/s00357-019-09345-1 [GS Search].

Nagy, M., & Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. *2018 IEEE 22nd international conference on intelligent engineering systems (INES)*, 000389–000394. https://doi.org/10.1109/INES.2018.8523888 [GS Search].

Naik, A., & Samant, L. (2016). Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer, Tanagra, Orange and Knime. *Procedia Computer Science*, *85*, 662–668. https://doi.org/10.1016/j.procs.2016.05.251 [GS Search].

Pašić, Đ., & Kučak, D. (2020). Machine learning model for detecting high school students as candidates for drop-out from a study program. *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 1140–1144. https://doi.org/10.23919/MIPRO48935.2020.9245405 [GS Search].

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008). Systematic mapping studies in software engineering. *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*, 1–10. https://doi.org/10.14236/ewic/EASE2008.8 [GS Search].

Pinto, S. C. (2021). Os custos da evasão de discentes das universidades brasileiras na modalidade de ensino presencial: uma perspectiva de custos contábeis e custos econômicos. [GS Search].

Ratra, R., & Gulia, P. (2020). Experimental evaluation of open source data mining tools (WEKA and Orange). *Int. J. Eng. Trends Technol*, *68*(8), 30–35. [GS Search].

Sari, E. Y., Sunyoto, A., et al. (2019). Optimization of Weight Backpropagation with Particle Swarm Optimization for Student Dropout Prediction. *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 423–428. https://doi.org/10.1109/ICITISEE48480.2019.9004032 [GS Search].

Silva, I., & Moody, G. B. (2014). An open-source toolbox for analysing and processing physionet databases in matlab and octave. *Journal of open research software*, *2*(1). https://doi.org/10.5334/jors.bi [GS Search].

Sofyan, Y., & Kurniawan, H. (2009). Teknik Analisis Statistik terlengkap dengan Software SPSS. *Salemba Infotek, Jakarta*. [GS Search].

Sousa, L. R. d., Carvalho, V. O. d., Penteado, B. E., & Affonso, F. J. (2021). A systematic mapping on the use of data mining for the face-to-face school dropout problem. *Proceedings*. [GS Search].

Stančin, I., & Jović, A. (2019). An overview and comparison of free Python libraries for data mining and big data analysis. *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 977–982. https://doi.org/10.23919/MIPRO.2019.8757088 [GS Search].

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, *45*(1), 89–125. https://doi.org/https://doi.org/10.3102/00346543045001089 [GS Search].

Valente, A., Holanda, M., Mariano, A. M., Furuta, R., & Da Silva, D. (2022). Analysis of Academic Databases for Literature Review in the Computer Science Education Field. *2022 IEEE Frontiers in Education Conference (FIE)*, 1–7. https://doi.org/10.1109/FIE56618.2022.9962393 [GS Search].

Wen, J., Li, S., Lin, Z., Hu, Y., & Huang, C. (2012). Systematic literature review of machine learning based software development effort estimation models. *Information and Software Technology*, *54*(1), 41–59. https://doi.org/10.1016/j.infsof.2011.09.002 [GS Search].