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: 17/Jun/2023; 1st round notif.: 01/Aug/2023; New version: 31/Aug/2023; 2nd round notif.: 02/Dec/2023;
Camera ready: 04/Mar/2024; Edition review: 10/Mar/2024; Available online: 10/Mar/2024; Published: 10/Mar/2024;

Investigação da Evasão Estudantil por meio da Mineração de Dados e Aprendizagem de Máquina: Um Mapeamento Sistemático

Title: Investigation of Student Dropout through Data Mining and Machine Learning: A Systematic Mapping

Título: Investigación de la Evasión Estudiantil mediante la Minería de Datos y el Aprendizaje Automático: Un Mapeo Sistemático

Jeferson Andrade de Jesus
Universidade Federal de Sergipe
ORCID: 0009-0005-9436-4861
andrade-jeferson@hotmail.com

Renê Pereira de Gusmão
Universidade Federal de Sergipe
ORCID: 0000-0002-4806-6506
renepgusmao@gmail.com

Resumo

A evasão dos alunos nas escolas e universidades é um problema recorrente na educação, tanto é danoso para o aluno em termos de aprendizagem, como gera prejuízos financeiros para as instituições, sejam públicas ou privadas. Estudos que utilizam técnicas de mineração de dados (MD) e aprendizado de máquina (AM) para investigar problemas na educação estão em ascensão. A evasão estudantil é um desses problemas. Por meio dessas técnicas, é possível identificar padrões em indivíduos ou grupos que possam vir a abandonar os estudos. Este artigo tem como objetivo mapear sistematicamente artigos no estado da arte sobre a aplicação de DM e ML na classificação de dados em estudos sobre evasão escolar. A busca foi realizada em 5 bases de dados bibliográficas, ACM Digital Library, IEEE Xplore, Scopus, ScienceDirect e Web of Science, e retornou um total de 336 estudos primários. Após a aplicação dos critérios de exclusão e inclusão, restaram 71 estudos relevantes. Após a extração de dados desses estudos, identificou-se que, as experiências com estudantes do ensino superior e na modalidade presencial são as mais recorrentes nesses artigos, o ano que mais se destacou em termos de publicação foi 2020, e os algoritmos mais frequentemente utilizados para construção dos modelos de classificação são algoritmos baseados em árvores de decisão.

Palavras-chave: *Predição; Classificação; Evasão do estudante; Aprendizagem de máquina; Mineração de dados.*

Abstract

Dropping out of students in schools and universities is a recurrent problem in education, it is both harmful for the student in terms of learning, and generates financial expenses for educations institutions, whether public or private. Studies using data mining (DM) and machine learning (ML) techniques to investigate problems in education are on the rise. Student dropout is one such problem. Through these techniques, it is possible to identify patterns in individuals or groups that may drop out of studies. This article aims to systematically map state-of-the-art articles on the application of DM and ML in data classification in studies on school dropout. The search was carried out in 5 bibliographic databases, ACM Digital Library, IEEE Xplore, Scopus, ScienceDirect, and Web of Science, and returned a total of 336 primary studies. After applying the exclusion and inclusion criteria, 71 relevant studies remained. After extracting data from these studies, it was identified that the experiences with higher education students and in the face-to-face modality are the most recurrent in these articles, the year that most stood out in terms of publication was 2020, and the most frequently used algorithms for building the classification models are algorithms based on decision trees.

Cite as: Jesus, J. A. J. & Gusmão, R. P. G. (2024). Investigação da Evasão Estudantil por meio da Mineração de Dados e Aprendizagem de Máquina: Um Mapeamento Sistemático. Revista Brasileira de Informática na Educação, 32, 157-191. https://doi.org/10.5753/rbie.2024.3466.

Keywords: *Prediction; Classification; Student dropout; Machine learning; Data mining.*

Resumen

La deserción estudiantil en escuelas y universidades es un problema recurrente en la educación, tanto perjudicial para el estudiante en términos de aprendizaje como generador de pérdidas financieras para las instituciones, ya sean públicas o privadas. Los estudios que utilizan técnicas de minería de datos (MD) y aprendizaje automático (AM) para investigar problemas en la educación están en aumento. La deserción estudiantil es uno de esos problemas. A través de estas técnicas, es posible identificar patrones en individuos o grupos que puedan abandonar los estudios. Este artículo tiene como objetivo mapear sistemáticamente los artículos en el estado del arte sobre la aplicación de MD y AM en la clasificación de datos en estudios sobre deserción escolar. La búsqueda se realizó en 5 bases de datos bibliográficas, ACM Digital Library, IEEE Xplore, Scopus, ScienceDirect y Web of Science, y arrojó un total de 336 estudios primarios. Después de aplicar los criterios de exclusión e inclusión, quedaron 71 estudios relevantes. Tras la extracción de datos de estos estudios, se identificó que las experiencias con estudiantes de educación superior y en la modalidad presencial son las más recurrentes en estos artículos, el año que más se destacó en términos de publicación fue 2020, y los algoritmos más frecuentemente utilizados para la construcción de los modelos de clasificación son algoritmos basados en árboles de decisión.

Palabras clave: *Predicción; Clasificación; Deserción estudiantil; Aprendizaje automático; Minería de datos.*

1 Introdução

O problema da evasão estudantil tem muitas faces e afeta instituições em diferentes níveis de ensino em todo o mundo. Trata-se de um problema crítico com características especiais, que requer abordagens inovadoras, como a adoção das técnicas de mineração de dados educacionais (*Educational Data Mining* - EDM). Com o uso da EDM é possível ter um panorama dos índices de evasão escolar. No Brasil, as pesquisas sobre a questão da evasão usando EDM ainda estão em seu início. Contudo, se considerarmos o cenário internacional, diversas pesquisas expressivas utilizando EDM têm sido realizadas (Manhães et al., 2014).

Com o grande avanço da Inteligência Artificial (IA) na educação, mais especificamente a aplicação da Mineração de Dados (do inglês, *Data Mining* - DM) e o Aprendizado de Máquina (do inglês, *Machine Learning* - ML), é crescente a tentativa das instituições de ensino usarem da quantidade enorme de dados sobre seus alunos para desenvolverem soluções robustas com o objetivo de solucionar alguns problemas, e a evasão de alunos é um dos principais, pois gera muitos prejuízos, econômicos, acadêmicos e sociais (Santos Baggi & Lopes, 2011). EDM e ML são recursos relevantes para identificar os motivos da evasão de um aluno. Algoritmos clássicos de aprendizagem de máquina, como Regressão Logística, Árvores de Decisão, Naive Bayes e Máquinas de Vetores de Suporte, são comumente utilizados em tarefas de predição. Com alguns estudos identificando os benefícios desses recursos, alavancaram-se fortemente as pesquisas nesse tema específico. A procura por melhores algoritmos, técnicas e modelos que possam classificar se, futuramente, aquele aluno vai abandonar ou não os estudos. Tendo essa visão prévia, as instituições poderiam planejar a melhor forma de evitar a evasão desse aluno. Essa tomada de decisão também é um problema ao qual a aplicação de IA tem sido investigada, mas que não será tratado neste artigo.

O objetivo deste trabalho é mapear estudos relevantes sobre a aplicação de DM e ML na

classificação de dados em estudos sobre evasão escolar, de forma a caracterizar a evolução dessa área de pesquisa. Assim, este trabalho está organizado da seguinte forma: na seção 2 são apresentados os trabalhos relacionados. A seção 3 apresenta a metodologia aplicada nesse mapeamento. Na seção 4 é apresentada a análise dos resultados. Na seção 5 são apresentadas as ameaças à validade desse estudo. Na seção 6 é apresentada a conclusão.

2 Trabalhos Relacionados

Com o objetivo de buscar e identificar estudos primários e selecionar estudos relevantes sobre a aplicação de DM e ML na classificação de dados em estudos sobre evasão escolar, de forma a caracterizar a evolução dessa área de pesquisa, foram detectados revisões e mapeamentos da literatura, mas que não foram classificados como estudos relevantes baseados nas questões de pesquisa e os critérios de inclusão e exclusão que são apresentados nas seções seguintes. No entanto, alguns desses trabalhos se relacionam com a proposta deste estudo, sendo estes apresentados abaixo.

No trabalho realizado por Colpo et al. (2020), é apresentada uma revisão sistemática de estudos que utilizam técnicas de EDM no contexto da previsão de evasão estudantil e que foram publicados no Congresso Brasileiro de Informática na Educação (CBIE), principal evento da área no Brasil. O trabalho ressalta o enfoque em regiões como Sul, Nordeste e Sudeste, notadamente nos estados do Rio Grande do Sul e Rio de Janeiro. As pesquisas variaram em quantidade e tipos de dados. Árvores de decisão e Weka foram amplamente empregadas. O interesse crescente em EDM, especialmente na educação básica e técnica, destaca a necessidade de maior aplicação prática e disponibilização eficaz das previsões.

O segundo trabalho relacionado, realizado por Tamada et al. (2019), tem como objetivo principal identificar soluções que utilizem técnicas de ML para reduzir as altas taxas de desistência em Ambientes Virtuais de Aprendizagem (AVA). A pesquisa identificou que as soluções dos estudos utilizam principalmente ML supervisionado para previsão, mais especificamente os algoritmos de Regressão Logística e Máquina de Vetores de Suporte, com alta precisão em dados de cliques e fóruns. Três sistemas com *Graphical User Interface* (GUI) gerenciam aprendizado. Poucos usam ML não supervisionado para agrupar comportamentos. A precisão dos classificadores varia conforme o ambiente dinâmico de aprendizagem.

Posto isso, o presente estudo diferencia-se dos trabalhos relacionados no escopo da busca, apesar do primeiro trabalho realizar uma busca em todos os níveis de ensino, ele se limita aos trabalhos publicados no CBIE; o segundo trabalho relacionado restringe-se a modalidade de ensino que é Educação à Distância (EAD) e, mais especificamente, nos AVA's. Com isso, este trabalho apresenta um escopo maior de estudos com esse contexto da evasão estudantil.

3 Metodologia

Na realização desse mapeamento foi aplicado o método de mapeamento sistemático da literatura (do inglês, *Systematic Literature Mapping* - SLM) (Petersen et al., 2008). O método consiste nas etapas definidas abaixo.

1. “Questões de Pesquisa” são questões que quando respondidas nos ajudam a alcançar os objetivos do mapeamento sistemático.
2. “Identificação de Estudos” que são as palavras-chave, a *string* de busca, os critérios de seleção das fontes de busca, lista das fontes de busca e estratégia de busca.
3. “Seleção e Avaliação dos Estudos” são os critérios de inclusão e de exclusão dos estudos primários.
4. “Síntese dos Dados e Apresentação dos Resultados” é apresentada na seção 4 (Resultados e Discussões). É nessa seção que estão as informações sobre a estratégia de extração de dados, estratégia de sumarização dos dados e as informações sobre os dados extraídos e sumarizados.

3.1 Questões de Pesquisa

Foram definidas questões que se mostraram importantes para melhor extração dos dados obtidos, essas questões foram resultado de um processo meticuloso. Isso incluiu revisão ampla da literatura, análise de tendências recentes e exploração das demandas práticas no campo da "Aplicação de DM e ML na classificação de dados em estudos sobre evasão escolar". A seleção considerou a relevância para os objetivos do mapeamento e a capacidade de orientar a coleta precisa e abrangente de dados e dessa forma alcançar os objetivos do mapeamento sistemático. Abaixo, na Tabela 1, estão definidas as questões de pesquisa junto aos seus respectivos dados a serem obtidos:

Tabela 1: Questões de pesquisa.

Questões de pesquisa	**Dados a serem obtidos**
Como pode ser caracterizada a evolução do estado da arte sobre “Aplicação de DM e ML na classificação de dados em estudos sobre evasão escolar” em termos de publicações, autores e grupos de pesquisa?	Para extrair os dados serão analisados os anos das publicações dos artigos, os autores, os países e grupos de pesquisa.
Como contribuir para o amadurecimento dessa área de pesquisa preenchendo as lacunas encontradas?	Para extrair os dados serão analisados os métodos utilizados nas predições, quais fatores influenciam na evasão, informações sobre as bases de dados, metodologias utilizadas na mineração de dados, aprendizagem de máquina e seleção de atributos, e resultados dos experimentos feitos nos artigos.

3.2 Identificação de Estudos

Para formar as *strings* de busca foram escolhidos assuntos pelo reconhecimento científico e associação com a área de tecnologia da informação e educação, frequentemente apresentadas no contexto desta pesquisa. Após análise e experimentos com palavras-chave que envolvem o tema de

pesquisa, as seguintes palavras-chave foram escolhidas: *"PREDICTION"*, *"CLASSIFICATION"*, *"STUDENT DROPOUT"*, *"MACHINE LEARNING"*, *"DATA MINING"* e *"DATA SCIENCE"*. Assim, a *string* de busca que é o conjunto de termos utilizadas para a busca dos estudos, ficou da seguinte forma: *((PREDICTION OR CLASSIFICATION) AND "STUDENT DROPOUT" AND ("MACHINE LEARNING" OR "DATA MINING" OR "DATA SCIENCE"))*, em termos de quantidade e escopo essa *string* de busca abrange um escopo grande, trazendo estudos sobre a aplicação de DM e ML na classificação de dados em estudos sobre evasão escolar em diferentes níveis escolar e categorias de ensino distintas.

Para a realização da busca foram selecionadas as seguintes bases bibliográficas: *ACM Digital Library* (ACM)[1], *IEEE Xplore* (IEEE) [2], *Scopus* [3], *Web of Science* [4] e *ScienceDirect* [5]. As pesquisas dos termos de busca foram realizadas em 30 de outubro de 2022.

3.3 Seleção e Avaliação dos Estudos

Com o propósito de analisar quais artigos poderão ser utilizados para responder às questões de pesquisa, foram definidos critérios de inclusão e exclusão. Os critérios de inclusão definidos foram:

1. Estudos que foquem em predição de evasão do aluno por meio da mineração de dados e aprendizagem de máquina.
2. Estudos que apresentem inovação na mineração dos atributos utilizados na predição de evasão do aluno.
3. Estudos que apresentem comparativos entre algoritmos de classificação trabalhando com predição de evasão do aluno.

Foram também selecionados os seguintes critérios de exclusão dos artigos, com a finalidade de retirar estudos não relevantes para o mapeamento:

1. Estudos duplicados.
2. Estudos não disponibilizados na íntegra.

Após a aplicação desses critérios, 71 artigos se mantiveram na pesquisa.

4 Resultados e Discussões

Nessa seção são apresentados os resultados iniciais da busca, denominados Estudos Primários (EP), Tabela 2; os estudos resultantes após aplicação dos critérios de inclusão e exclusão, deno-

[1] http://dl.acm.org
[2] http://ieeexplore.ieee.org
[3] https://www.scopus.com
[4] https://www.webofscience.com/
[5] https://www.sciencedirect.com

minados Estudos Relevantes (ER), Tabela 3; e os resultados da análise dos estudos relevantes, respondendo às questões de pesquisa definidas na seção 3.1 deste mapeamento sistemático.

4.1 Resultados encontrados com a string de busca

A Tabela 2 detalha a distribuição dos estudos em relação às diferentes bases de dados consultadas para o mapeamento sistemático. No total, foram identificados 336 EP nas diferentes bases de dados. Após a aplicação rigorosa dos critérios de inclusão e exclusão estabelecidos, 265 estudos foram rejeitados, enquanto 71 estudos satisfizeram esses critérios e foram classificados como ER. Esse processo de seleção rigorosa garantiu que apenas os estudos mais pertinentes e alinhados aos objetivos deste mapeamento sistemático fossem considerados para análise.

Os números apresentados na tabela revelam variações interessantes entre as diferentes bases de dados. Nota-se que a base Scopus apresentou o maior número de EP (126), porém, após a aplicação dos critérios de seleção, 98 estudos foram rejeitados, resultando em 28 ER. Por outro lado, a base IEEE teve 37 Estudos Primários, dos quais 31 foram rejeitados, restando 6 ER.

Esses resultados preliminares ressaltam a importância da seleção criteriosa de estudos para garantir a qualidade e a relevância das informações analisadas. Os ER selecionados nesta etapa representam uma base sólida para a análise aprofundada que visa responder às questões de pesquisa delineadas na seção 3.1 deste mapeamento sistemático.

Tabela 2: Resultados da Seleção dos ER.

Bases	**EP**	**Rejeitados**	**Aceitos**
ACM	50	41	9
IEEE	37	31	6
ScienceDirect	71	64	7
Scopus	126	98	28
Web of Science	52	31	21
Total	**336**	**265**	**71**

4.2 Estudos resultantes após aplicação de critérios de inclusão e exclusão

A Tabela 3 apresenta a lista completa dos ER após a aplicação dos critérios de inclusão e exclusão aos EP. Cada ER é identificado por um número de referência único (ER1, ER2, ER3, etc.), associado à sua respectiva fonte bibliográfica. A lista de ER abrange uma variedade de fontes, incluindo artigos de conferências, revistas acadêmicas e outros recursos importantes. As referências incluídas nessa tabela fornecem um panorama abrangente dos trabalhos selecionados para análise posterior. Esses estudos representam contribuições significativas para a investigação abordada neste mapeamento sistemático.

Ao analisar a lista, é possível observar a diversidade de fontes e autores envolvidos na área de pesquisa examinada. Os ER foram selecionados com base em sua relevância para os objetivos deste estudo e servirão como base para a análise detalhada das questões de pesquisa propostas. É importante ressaltar que a escolha desses estudos foi guiada pela necessidade de garantir a representatividade e a confiabilidade das informações utilizadas nesta pesquisa. A análise subsequente

Tabela 3: Lista de Estudos Relevantes.

Id.	Referência	Id.	Referência
ER1	(Kang & Wang, 2018)	ER37	(Tenpipat & Akkarajitsakul, 2020)
ER2	(Wu et al., 2019)	ER38	(Pradeep et al., 2015)
ER3	(Meca et al., 2020)	ER39	(Naseem et al., 2019)
ER4	(Baranyi et al., 2020)	ER40	(Bello et al., 2020)
ER5	(Chen et al., 2018)	ER41	(Sorensen, 2019)
ER6	(Ameri et al., 2016)	ER42	(Şahin, 2021)
ER7	(Manhães et al., 2014)	ER43	(Lottering et al., s.d.)
ER8	(Lottering et al., 2020)	ER44	(Pereira & Zambrano, 2017)
ER9	(Hegde, 2016)	ER45	(Perez et al., 2018)
ER10	(Nagy & Molontay, 2018)	ER46	(Pérez et al., 2018)
ER11	(Kuo et al., 2017)	ER47	(Pérez-Gutiérrez, 2020)
ER12	(Fu et al., 2021)	ER48	(Burgos et al., 2018)
ER13	(Mubarak et al., 2021)	ER49	(Xing & Du, 2019)
ER14	(Chung & Lee, 2019)	ER50	(Limsathitwong et al., 2018)
ER15	(Lykourentzou et al., 2009)	ER51	(Berens et al., 2018)
ER16	(Viloria & Lezama, 2019)	ER52	(Panagiotakopoulos et al., 2021)
ER17	(Santana et al., 2015)	ER53	(Figueroa-Cañas & S.Vinuesa, 2020)
ER18	(Selvan et al., 2019)	ER54	(Hegde & Prageeth, 2018)
ER19	(Nuankaew et al., 2019)	ER55	(de la Peña et al., 2017)
ER20	(Rovira et al., 2017)	ER56	(Coussement et al., 2020)
ER21	(Radovanović et al., 2020)	ER57	(Tan & Shao, 2015)
ER22	(Dewan et al., 2015)	ER58	(Kotsiantis et al., 2003)
ER23	(Maksimova et al., 2020)	ER59	(Heredia et al., 2015)
ER24	(Yaacob et al., 2020)	ER60	(Fei & Yeung, 2015)
ER25	(Moseley & Mead, 2008)	ER61	(Li et al., 2022)
ER26	(Costa et al., 2020)	ER62	(Su et al., 2022)
ER27	(Quishpe-Morales et al., 2020)	ER63	(Revathy et al., 2022)
ER28	(Martins et al., 2020)	ER64	(Hossain et al., 2022)
ER29	(SALLAN & BEHAL, s.d.)	ER65	(Niyogisubizo et al., 2022)
ER30	(Mubarak et al., 2020)	ER66	(Guzmán-Castillo et al., 2022)
ER31	(ALBAN & MAURICIO, 2018)	ER67	(Nuanmeesri et al., 2022)
ER32	(Gamao & Gerardo, 2019)	ER68	(Mnyawami et al., 2022)
ER33	(Aguirre & Pérez, 2020)	ER69	(Jarbou et al., 2022)
ER34	(Sivakumar et al., 2016)	ER70	(Vega et al., 2022)
ER35	(Hutagaol & Suharjito, 2019)	ER71	(Naseem et al., 2022)
ER36	(Del Bonifro et al., 2020)		

dos ER permitirá a extração de *insights*, identificação de tendências e a obtenção de respostas substanciais para as questões delineadas.

Essa lista completa de ER é um componente crucial do processo de mapeamento sistemático, fornecendo uma base sólida para a investigação e discussão dos resultados obtidos a partir desses estudos selecionados.

4.3 Como pode ser caracterizada a evolução do estado da arte sobre aplicação de DM e ML na classificação de dados em estudos sobre evasão escolar?

Essa questão tem como finalidade identificar como se deu a evolução da aplicação de DM e ML na classificação de dados em estudos sobre evasão escolar do aluno, e quais as perspectivas de futuro para a área. A Figura 1 apresenta a evolução da quantidade de publicações ao longo dos anos. Não foi estabelecido um critério de busca baseado no ano de publicação, o que resultou em um intervalo de tempo abrangendo as publicações de 2003 a 2022. É notado um crescimento médio de estudos relacionados à área na última década, com destaque para o ano de 2020.

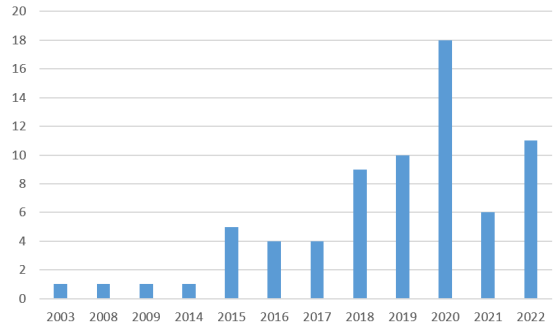

Figura 1: Anos das Publicações dos Artigos.

O gráfico representado pela Figura 2 visa mostrar quais países se destacam na quantidade de artigos publicados sobre a aplicação de DM e ML na classificação de dados em estudos sobre evasão escolar. O país destaque em publicações foi a China, seguida por Estados Unidos e Espanha. Foram encontrados também três estudos brasileiros. Os estados desses estudos foram: Rio de Janeiro (ER7), Rio Grande do Sul (ER26) e Alagoas (ER17), cada um deles com uma publicação.

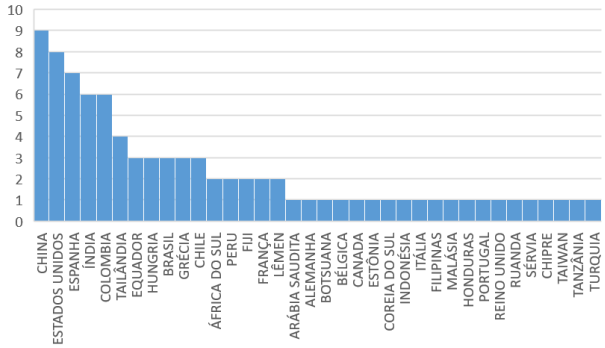

Figura 2: Países dos artigos.

4.4 Como contribuir para o amadurecimento dessa área de pesquisa preenchendo as lacunas encontradas?

Essa questão busca responder quais os métodos utilizados nas predições, quais fatores influenciam na evasão, informações sobre as bases de dados, metodologias utilizadas na mineração de dados e seleção de atributos, e resultados dos experimentos feitos nos artigos. A Tabela 4 apresenta uma síntese geral dos ER com relação as metodologias de DM, técnicas de seleção de atributos e redução de dimensionalidade, e algoritmos utilizados.

Tabela 4: Tabela de Síntese Geral dos Dados Extraídos.

Id.	**Assunto**	**Estudos**
A1	Metodologia CRISP-DM.	ER3, ER24, ER26, ER39, ER45, ER50 e ER71.
A2	Metodologia KDD-DM.	ER27, ER29, ER33, ER43 e ER59.
A3	Redução de Dimensionalidade.	ER8, ER9, ER43 e ER54.
A4	Seleção de Atributos com técnica especificada.	ER3, ER10, ER28, ER31, ER32, ER33, ER35, ER39, ER50, ER58, ER66, ER67, ER70 e ER71.
A5	Seleção de Atributos com técnica não especificada.	ER2, ER17, ER26, ER30, ER34, ER38, ER40, ER46, ER52, ER53, ER54, ER59, ER63, ER65, ER68 e ER69.
A6	Estudos que utilizam pelo menos uma dessas técnicas/algoritmos: Deep Learning, Adaboost, K Nearest Nerighbor, Naive Bayes, Support Vector Machine, Regressão Logística, Redes Neurais, Floresta Aleatória e Árvore de Decisão.	ER1, ER2, ER3, ER4, ER5, ER6, ER7, ER8, ER10, ER11, ER12, ER13, ER14, ER15, ER17, ER18, ER19, ER20, ER21, ER22, ER23, ER24, ER26, ER27, ER28, ER29, ER30, ER31, ER32, ER34, ER35, ER36, ER37, ER38, ER39, ER40, ER41, ER42, ER43, ER44, ER45, ER46, ER47, ER48, ER49, ER50, ER52, ER53, ER54, ER55, ER56, ER57, ER58, ER59, ER60, ER63, ER64, ER65, ER66, ER67, ER68, ER69, ER70 e ER71.
A7	Estudos que utilizam nenhuma das técnicas citadas no A6.	ER9, ER16, ER25, ER33, ER51, ER61 e ER62.

4.4.1 Metodologia de Mineração de Dados

Apenas 12 estudos especificaram a metodologia de DM utilizada em seus experimentos, sendo 7 deles utilizando CRISP (A1) e 5 KDD (A2).

4.4.2 Seleção de atributos e/ou redução de dimensionalidade

Dos 71 artigos, 34 explicitaram a aplicação de seleção de atributos ou redução de dimensionalidade em seus experimentos, enquanto 37 não o fizeram. Desses 34, 4 aplicaram redução de

dimensionalidade (A3), enquanto os outros 30 utilizaram seleção de atributos (A4 e A5). As técnicas utilizadas para a seleção de atributos estão apresentadas na Figura 3.

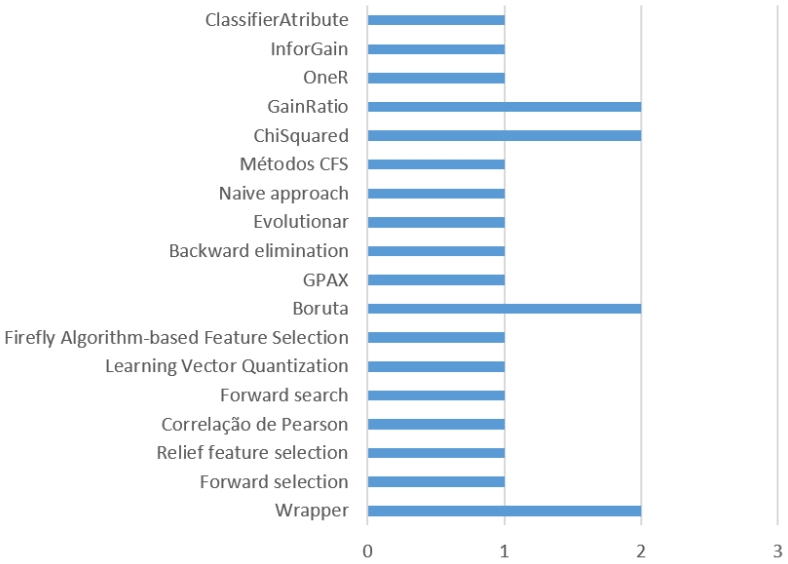

Figura 3: Técnicas de Seleção de Atributos.

4.4.3 Bases de dados utilizadas nos experimentos

As bases de dados descritas nos artigos são referentes aos anos de 1994 a 2021. A menor amostra utilizada nos experimentos foi de 26 registros, enquanto a maior foi de 220 mil. Os gráficos a seguir destacam a proporção dos níveis de ensino e das modalidades trabalhadas nos experimentos dos estudos relevantes, respectivamente. Percebe-se uma predominância de estudos voltados ao ensino superior e de estudos voltados a modalidade presencial.

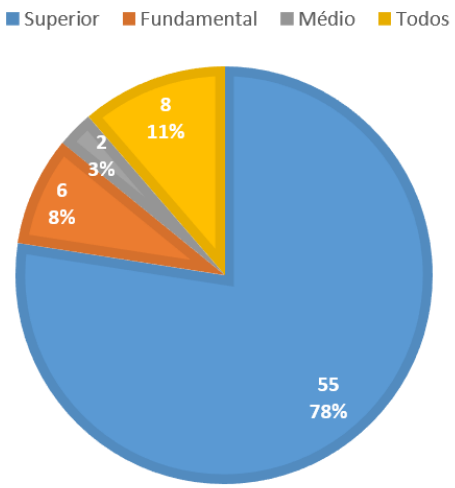

Figura 4: Proporção dos Níveis de Escolaridade.

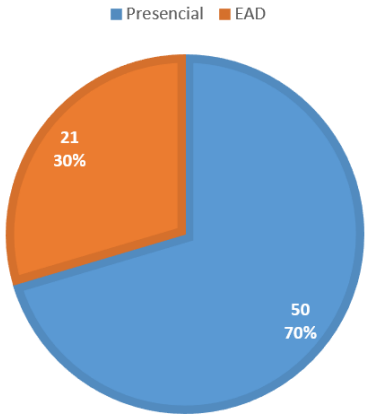

Figura 5: Proporção das Modalidades de Ensino.

Por fim, a Figura 6 apresenta a frequência de utilização de métodos de aprendizagem de máquina nos experimentos dos ER (A6), a categoria "Outros"presente no gráfico aglomera algoritmos e modelos que foram utilizados em 2 estudos, ou menos, como é o caso dos estudos A7.

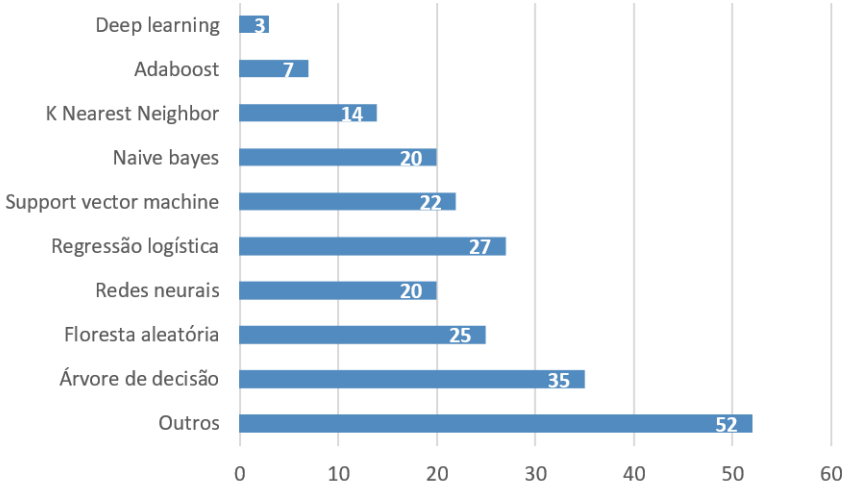

Figura 6: Frequência dos Algoritmos.

4.5 Síntese Geral e Sumarização dos Trabalhos Selecionados

Nesta etapa serão apresentados breves resumos dos trabalhos selecionados e uma síntese geral das publicações, os resumos estão organizados de acordo com os assuntos já apresentadas na seção 4.4. Destacando os pontos importantes de cada estudo, em termos de algoritmos, sistemas, em qual nível escolar a classificação foi aplicada, em qual modalidade de ensino, conjunto de dados utilizado, etc.

4.5.1 Metodologia CRISP-DM

ER3: Essa pesquisa aborda a categorização de desistência de estudantes universitários, dando foco na seleção dos atributos que melhoram essa predição de acordo com cada área acadêmica. O conjunto é formado por 24.894 participantes do ensino superior presencial, foi aplicada a abordagem Wrapper com o algoritmo RPart para seleção de atributos. As variáveis destacadas são adaptadas com base no curso específico. (Meca et al., 2020).

ER24: Esse trabalho trata da classificação de abandono do aluno universitário, especificamente no curso de ciência da computação, os algoritmos utilizados foram árvore de decisão, regressão logística, floresta aleatória, K-vizinho mais próximo e algoritmo de rede neural. O conjunto do estudo é composto por 64 participantes do ensino superior presencial, a variável mais importante é o desempenho em disciplinas específicas. (Yaacob et al., 2020).

ER26: Essa pesquisa aborda a categorização de desistência do estudante universitário, sendo empregados 3 algoritmos, destacando-se a Floresta aleatória, com uma taxa de acurácia de 95,12% e Recall de 91,41%. O estudo englobou 1.516 participantes em nível superior, com ensino presencial. Houve emprego de seleção de atributos, utilizando Regressão Logística, Árvores de Decisão e Floresta Aleatória como algoritmos de classificação. As variáveis destacadas foram a Média do terceiro semestre e a Média dos três primeiros semestres. (Costa et al., 2020).

ER39: Esse trabalho aborda a classificação de abandono do aluno universitário, especificamente no curso de Ciências da computação. O algoritmo utilizado é Árvore de decisão, aplica-se também seleção de recurso para identificar quais os melhores atributos para a classificação. A pesquisa incluiu 963 participantes no nível superior, com aulas presenciais. A seleção de atributos foi realizada por meio do algoritmo Boruta. As variáveis em destaque abrangeram GRADE, A1SCORE e QATTEMPT. (Naseem et al., 2019).

ER45: Essa pesquisa tem como foco a classificação de abandono do aluno universitário, especificamente no curso de engenharia de sistemas. Os algoritmos utilizados nos experimentos foram: Árvore de decisão, Regressão logística e Naive Bayes. Tendo Árvore de decisão como algoritmo que obteve melhor resultado, AUC de 0.94. O estudo envolveu 802 indivíduos no nível superior com aulas presenciais. Não foi realizada seleção de atributos. Não houve destaque específico de variáveis. (Perez et al., 2018).

ER50: Esse trabalho aborda a classificação de abandono do alunos do ensino médio de um instituto de tecnologia da Tailândia, os algoritmos utilizados foram Árvores de decisão e Floresta aleatória, Árvores de decisão obteve os melhores resultados. A pesquisa abrangeu 28.801 participantes em nível superior com ensino presencial. Utilizou-se o critério GPAX para seleção de atributos, as notas nos cursos de idiomas foram identificadas como variáveis de destaque. (Limsathitwong et al., 2018).

ER71: Nesse trabalho é feita uma análise da evasão no curso de Ciência da Computação, os algoritmos utilizados foram *Random Forest*, *Decision Tree*, *Naive Bayes*, *Logistic Regression* e *K-Nearest Neighbour*. No primeiro estágio o algoritmo que obteve o melhor resultado foi o Naive Bayes com uma AUC de 0.6123, nos estágios 2 e 3 o algoritmo que obteve o melhor resultado foi o Logistic Regression com AUC de 0.7523 e 0.8902, respectivamente. O tamanho do conjunto de dados não foi mencionado no artigo, mas trata-se de dados de uma única Instituição de Ensino Superior da região do Pacífico Sul, que é o único campus regional no Pacífico Sul. (Naseem et al.,

2022).

4.5.2 Metodologia KDD-DM

ER27: Essa pesquisa aborda a categorização de desistência do estudante universitário, sendo empregados os algoritmos KNN, árvore de decisão, floresta aleatória, SVM e redes neurais. O modelo de redes neurais demonstrou o melhor desempenho, alcançando uma acurácia de 92%. O estudo contou com 26 participantes no nível superior, com aulas presenciais. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Quishpe-Morales et al., 2020).

ER29: Nesse estudo, aborda-se a categorização da evasão de estudantes universitários. Como parte da pesquisa, são empregados os algoritmos *Decision Stump*, NDTREE e *Enhanced Machine Learning* (EMLA), os quais são fundamentados em técnicas de aprendizado de máquina. O estudo envolveu 407 participantes no nível superior, com ensino presencial. Não foi realizada seleção de atributos. Não houve destaque específico de variáveis. (SALLAN & BEHAL, s.d.).

ER33: Nesse trabalho, trata-se da classificação de abandono dos alunos universitários. O modelo aplicado é a regressão linear e o estudo também enfoca a seleção de recursos. O estudo incluiu 530 indivíduos no ensino superior, com aulas presenciais. A seleção de atributos foi conduzida por meio da Correlação de Pearson. Não houve destaque específico de variáveis. (Aguirre & Pérez, 2020).

ER43: Nesse estudo, explora-se a categorização do abandono de alunos universitários, utilizando diferentes algoritmos, como Floresta aleatória, máquinas de vetores de suporte, árvores de decisão, Naive Bayes, K-vizinho mais próximo e regressão logística. Destaca-se que o algoritmo de Floresta aleatória apresentou os melhores resultados, com uma acurácia de 94.14%. O estudo abordou o ensino superior em formato presencial, envolvendo 12.293 participantes. O artigo não menciona a aplicação de uma abordagem específica para seleção de atributos. As variáveis importantes englobaram a idade, as disciplinas e o curso dos participantes. (Lottering et al., s.d.).

ER59: Nesse trabalho, é examinada a classificação do abandono de estudantes universitários, os algoritmos utilizados foram árvores de decisão C4.5 e ID3. O estudo abrangeu 201 participantes no ensino superior com formato presencial. Foi aplicada seleção de atributos, porém detalhes específicos da seleção de atributos não foram especificados. (Heredia et al., 2015).

4.5.3 Redução de Dimensionalidade

ER8: Esse trabalho trata da classificação do abandono de alunos universitários, explorando a adequação da diminuição dimensional e concentrando os dados significativos encobertos nas informações de alunos com risco de abandono. Embora todos os algoritmos de aprendizado de máquina utilizados nesse trabalho tenham alcançado uma acurácia de mais de 75%, *Support Vetor Machine* e *Naive Bayes* tiveram melhor desempenho ao prever o abandono. O estudo envolveu 4.417 participantes no ensino superior presencial. Para análise, foram utilizados os métodos SVM, Naïve Bayes, Árvore de Decisão, KNN e Floresta Aleatória. Não houve destaque específico de variáveis. (Lottering et al., 2020).

ER9: Esse trabalho foca na redução de dimensionalidade dos atributos para melhorar a classificação de abandono de alunos universitários. O estudo envolveu 150 participantes no ensino superior

com aulas presenciais. Foi realizada redução dimensional utilizando a técnica PCA (Análise de Componentes Principais). Não houve destaque específico de variáveis. (Hegde, 2016).

ER43: Nesse estudo, explora-se a categorização do abandono de alunos universitários, utilizando diferentes algoritmos, como Floresta aleatória, máquinas de vetores de suporte, árvores de decisão, Naive Bayes, K-vizinho mais próximo e regressão logística. Destaca-se que o algoritmo de Floresta aleatória apresentou os melhores resultados, com uma acurácia de 94.14%. O estudo abordou o ensino superior em formato presencial, envolvendo 12.293 participantes. O artigo não menciona a aplicação de uma abordagem específica para seleção de atributos. As variáveis importantes englobaram a idade, as disciplinas e o curso dos participantes. (Lottering et al., s.d.).

ER54: Nesse estudo, aborda-se a classificação de abandono de alunos, tanto em escolas quanto em universidades. O algoritmo utilizado é o *Naive Bayes*, com a linguagem R. O trabalho também foca na seleção de recurso para escolha dos melhores atributos. O estudo envolveu 50 participantes no ensino superior em formato presencial. As variáveis relevantes incluíram fatores acadêmicos, demográficos, psicológicos e de saúde. (Hegde & Prageeth, 2018).

4.5.4 Seleção de Atributos

Técnica de Seleção de Atributos especificada

ER3: Essa pesquisa aborda a categorização de desistência de estudantes universitários, dando foco na seleção dos atributos que melhoram essa predição de acordo com cada área acadêmica. O conjunto é formado por 24.894 participantes do ensino superior presencial, foi aplicada a abordagem Wrapper com o algoritmo RPart para seleção de atributos. As variáveis destacadas são adaptadas com base no curso específico. (Meca et al., 2020).

ER10: Esse trabalho trata da classificação do abandono de alunos em programas universitários, foram utilizados algoritmos baseados em árvore de decisão, *Naive Bayes*, K-NN, modelos lineares e aprendizado profundo. Os modelos com melhores resultados foram: *Gradient Boosted Trees* e *Deep Learning*, com uma AUC de 0.808 e 0.811, respectivamente. O estudo abrangeu 15.825 participantes do ensino superior em formato presencial. Não houve destaque específico de variáveis. (Nagy & Molontay, 2018).

ER28: Nesse estudo, é explorada a classificação de abandono do aluno universitário, ele utiliza uma combinação de 3 algoritmos de mineração de dados populares, floresta aleatória, máquinas de vetores de suporte e redes neurais artificiais. O estudo considerou o ensino superior presencial, abrangendo 3.373/3.344 participantes. Foi aplicada a técnica de seleção de atributos forward search. Para o 1º semestre, variáveis destacadas incluíram ects_aprov_s, cod_escola, média_s, sexo, bolseiro_s, entre outras; para o 2º semestre, variáveis como ects_reprov_s, cod_prof_mae, n10_11_acesso, idade, nacionalidade, entre outras, foram relevantes. (Martins et al., 2020).

ER31: Nessa pesquisa, explora-se a classificação de abandono do aluno universitário, os algoritmos utilizados foram regressão logística e árvores de decisão. Como resultado, obteve-se que a técnica com maior percentual de acurácia de abandono foi a árvore de decisão com 91.70%. O estudo englobou 1.178 participantes no ensino superior em formato presencial. Houve seleção de atributos utilizando a técnica Relief Feature Selection. Não houve destaque específico de variáveis. (ALBAN & MAURICIO, 2018).

ER32: Esse trabalho analisa modelos de classificação de abandono do aluno universitário, foram utilizados os algoritmos *Decision Tree* e *Naive Bayes*, o *Decision Tree* obteve melhores resultados. A pesquisa abrangeu 1.862 participantes no ensino superior com formato presencial. Foi aplicada a seleção de atributos baseada no algoritmo Firefly. Não houve destaque específico de variáveis. (Gamao & Gerardo, 2019).

ER33: Nesse trabalho, trata-se da classificação de abandono dos alunos universitários. O modelo aplicado é a regressão linear e o estudo também enfoca a seleção de recursos. O estudo incluiu 530 indivíduos no ensino superior, com aulas presenciais. A seleção de atributos foi conduzida por meio da Correlação de Pearson. Não houve destaque específico de variáveis. (Aguirre & Pérez, 2020).

ER35: Essa pesquisa trata da classificação de abandono do aluno universitário, os algoritmos utilizados foram *K-Nearest Neighbor* (KNN), *Naïve Bayes* (NB) e *Decision Tree* (DT). A pesquisa abordou 17.432 participantes no ensino superior presencial. Utilizou a técnica Learning Vector Quantization para seleção de atributos. As variáveis de destaque incluíram frequência, pontuações de tarefas, créditos totais, pontuações UTS, UAS pontuações, GPA, renda dos pais, educação dos pais, sexo e idade dos estudantes. (Hutagaol & Suharjito, 2019).

ER39: Esse trabalho aborda a classificação de abandono do aluno universitário, especificamente no curso de Ciências da computação. O algoritmo utilizado é Árvore de decisão, aplica-se também seleção de recurso para identificar quais os melhores atributos para a classificação. A pesquisa incluiu 963 participantes no nível superior, com aulas presenciais. A seleção de atributos foi realizada por meio do algoritmo Boruta. As variáveis em destaque abrangeram GRADE, A1SCORE e QATTEMPT. (Naseem et al., 2019).

ER50: Esse trabalho aborda a classificação de abandono do alunos do ensino médio de um instituto de tecnologia da Tailândia, os algoritmos utilizados foram Árvores de decisão e Floresta aleatória, Árvores de decisão obteve os melhores resultados. A pesquisa abrangeu 28.801 participantes em nível superior com ensino presencial. Utilizou-se o critério GPAX para seleção de atributos, as notas nos cursos de idiomas foram identificadas como variáveis de destaque. (Limsathitwong et al., 2018).

ER58: Nesse estudo, aborda-se a classificação de abandono do aluno universitário, os métodos de classificação utilizados incluíram *Naive Bayes*, C4.5, *BackPropagation*, SMO, 3NN e MLE, é implementando também um protótipo de sistema web utilizando o *Naive Bayes*. A pesquisa envolveu 354 participantes no ensino superior na modalidade EaD. A seleção de atributos foi realizada por meio da abordagem Wrapper. Não houve destaque específico de variáveis. (Kotsiantis et al., 2003).

ER66: Nesse trabalho foi desenvolvido um sistema que permite o cálculo do risco de evasão por aluno e utiliza um procedimento de geração de alertas para coordenar as intervenções. A plataforma permitiu mensurar o impacto das estratégias de intervenção na permanência dos alunos. O estudo abrangeu 15.805 participantes no ensino superior em formato presencial. Foi empregada a abordagem ingênua (naive approach) para seleção de atributos. As técnicas de classificação utilizadas incluíram o algoritmo *AdaBoost*, *Bayesian GLM*, *Decision Trees*, *LogitBoost*, *Random Forest* (RF) e *Stochastic Gradient Boosting*. Não houve destaque específico de variáveis. (Guzmán-Castillo et al., 2022).

ER67: Esse estudo criou um modelo combinando seleção de atributos com um método de rede neural perceptron multicamadas. O modelo foi comparado com modelos baseados nos algoritmos de *Logistic regression*, *Decision Tree*, *Random Forest*, *Naive Bayes*, *Support Vector Machine* e *Multilayer Perceptron Neural Network*, e obteve melhores resultados. A pesquisa considerou 1.650 participantes no ensino superior presencial. Foram aplicadas as técnicas de seleção de atributos GR, CS e Métodos CFS. As variáveis destacadas englobaram a média cumulativa de notas (GPA), média cumulativa de notas para assuntos não docentes (GPAnone) e adesão ao grupo de mídia social em assuntos (*Socialclass*). (Nuanmeesri et al., 2022).

ER70: O objetivo nesse trabalho é reduzir a taxa de evasão dos alunos da Faculdade de Engenharia de Sistemas e Informática da Universidade Nacional Mayor de San Marcos (FISI-UNMSM), o modelo de predição utilizado é baseado em árvores de decisão, obteve uma acurácia de 90.34%. A pesquisa abrangeu 1.938 participantes no ensino superior presencial. Foram aplicadas as técnicas de seleção de atributos ChiSquared, OneR, GainRatio, InforGain e ClassifierAttribute. As variáveis destacadas incluíram a média ponderada histórica de notas, a média ponderada de notas do último ciclo e o número de créditos de cursos aprovados. (Vega et al., 2022).

ER71: Nesse trabalho é feita uma análise da evasão no curso de Ciência da Computação, os algoritmos utilizados foram *Random Forest*, *Decision Tree*, *Naive Bayes*, *Logistic Regression* e *K-Nearest Neighbour*. No primeiro estágio o algoritmo que obteve o melhor resultado foi o Naive Bayes com uma AUC de 0.6123, nos estágios 2 e 3 o algoritmo que obteve o melhor resultado foi o Logistic Regression com AUC de 0.7523 e 0.8902, respectivamente. O tamanho do conjunto de dados não foi mencionado no artigo, mas trata-se de dados de uma única Instituição de Ensino Superior da região do Pacífico Sul, que é o único campus regional no Pacífico Sul. (Naseem et al., 2022).

Técnica de Seleção de Atributos não especificada

ER2: Esse trabalho propõe um modelo de rede neural profunda, que é uma combinação de *Convolutional Neural Network, Long Short-Term Memory network e Support Vector Machine*, para ter um melhor desempenho na predição do abandono do aluno nos cursos online abertos massivos (MOOCs). O estudo englobou 120.542 participantes no ensino superior na modalidade EaD. Foi aplicada seleção de atributos, mas os detalhes específicos não foram especificados. As variáveis relevantes foram ID de inscrição, hora, fonte do evento, evento e dispositivo de acesso. (Wu et al., 2019).

ER17: Esse trabalho trata da classificação de abandono do aluno universitário em ambiente online, foram utilizados 4 algoritmos, o que se destacou entre eles foi o SVM com acurácia de 92,03%. O estudo considerou 162 participantes no ensino superior na modalidade EaD. Houve seleção de atributos, mas os detalhes específicos não foram mencionados. A variável destacada foi o desempenho nas disciplinas iniciais. (Santana et al., 2015).

ER26: Essa pesquisa aborda a categorização de desistência do estudante universitário, sendo empregados 3 algoritmos, destacando-se a Floresta aleatória, com uma taxa de acurácia de 95,12% e *recall* de 91,41%. O estudo englobou 1.516 participantes em nível superior, com ensino presencial. Houve emprego de seleção de atributos, utilizando Regressão Logística, Árvores de Decisão e Floresta Aleatória como algoritmos de classificação. As variáveis destacadas foram a Média do terceiro semestre e a Média dos três primeiros semestres. (Costa et al., 2020).

ER30: Esse estudo trata da classificação de abandono de alunos em ambiente online, os modelos utilizados foram Regressão Logística adicionando um termo de regularização e o Modelo de Markov Oculto Input-Output (IOHMM). O estudo abrangeu 32.593 participantes no ensino superior na modalidade EaD. Foi aplicada seleção de atributos, mas os detalhes específicos não foram fornecidos. Não houve destaque específico de variáveis. (Mubarak et al., 2020).

ER34: Esse trabalho trata da classificação de abandono do aluno universitário, o algoritmo utilizado é árvore de decisão aprimorado com base no ID3. A pesquisa englobou 240 participantes do ensino superior em formato presencial. Houve seleção de atributos, porém os detalhes específicos não foram especificados. As variáveis relevantes incluíram problemas familiares, doença em casa, ambiente do campus, baixa taxa de colação, mudança de objetivo pessoal e problema de ajuste. (Sivakumar et al., 2016).

ER38: Nessa pesquisa, trata-se da classificação de abandono do aluno em qualquer nível escolar, são utilizadas várias técnicas de classificação, como regras de indução e árvore de decisão. O estudo abordou 670 participantes do ensino médio em formato presencial. Foi aplicada seleção de atributos, embora os detalhes específicos não tenham sido mencionados. Não houve destaque específico de variáveis. (Pradeep et al., 2015).

ER40: Nesse estudo, aborda-se a classificação de abandono de alunos universitários, o algoritmo utilizado é Árvore de decisão. A pesquisa incluiu 206 participantes no ensino superior em formato presencial. Foi aplicada seleção de atributos, embora os detalhes específicos não tenham sido mencionados. As variáveis de destaque incluíram WAG, AI, ASR, SFR, FSE e PF. (Bello et al., 2020).

ER46: Esse pesquisa explora a classificação de abandono do aluno universitário, onde são comparado dois modelos, um baseado em técnica de regressão logística e o outro modelo baseado em árvores de decisão. A pesquisa envolveu 4.519 participantes no ensino superior em formato presencial. Houve seleção de atributos, embora os detalhes específicos não tenham sido mencionados. Na regressão logística, as variáveis relevantes foram a pontuação de matemática da PSU e as notas do ensino médio (NEM). Na Árvore de Decisão, as variáveis de destaque incluíram NEM, residência do estudante e a pontuação de matemática da PSU. (Pérez et al., 2018).

ER52: Esse estudo aborda a classificação de abandono do aluno em MOOCs, são utilizados vários modelos, o melhor desempenho é do modelo *LightGBM*, com acurácia de 96%. O tamanho do conjunto não foi explicitado nesse trabalho. O estudo envolveu níveis de ensino superior, médio e fundamental na modalidade EaD. Houve seleção de atributos, embora os detalhes específicos não tenham sido mencionados. A variável mais relevante foi os dados de interação dos alunos. (Panagiotakopoulos et al., 2021).

ER53: Esse trabalho trata da classificação de abandono de alunos universitários, são usados métodos de aprendizagem de máquina baseados em árvore para a predição. O estudo considerou 197 participantes do ensino superior na modalidade EaD. Houve seleção de atributos, embora os detalhes específicos não tenham sido mencionados. As variáveis relevantes incluíram assuntos difíceis de matemática e estatística. (Figueroa-Cañas & S.Vinuesa, 2020).

ER54: Nesse estudo, aborda-se a classificação de abandono de alunos, tanto em escolas quanto em universidades. O algoritmo utilizado é o *Naive Bayes*, com a linguagem R. O trabalho também foca na seleção de recurso para escolha dos melhores atributos. O estudo envolveu 50 participantes

no ensino superior em formato presencial. As variáveis relevantes incluíram fatores acadêmicos, demográficos, psicológicos e de saúde. (Hegde & Prageeth, 2018).

ER59: Nesse trabalho, é examinada a classificação do abandono de estudantes universitários, os algoritmos utilizados foram árvores de decisão C4.5 e ID3. O estudo abrangeu 201 participantes no ensino superior com formato presencial. Foi aplicada seleção de atributos, porém detalhes específicos da seleção de atributos não foram especificados. (Heredia et al., 2015).

ER63: Esse artigo se concentra na descoberta precoce de variáveis de abandono como um avanço pela redução de dimensionalidade usando seleção de atributos e métodos de extração, é utilizada a *Synthetic Minority Oversampling Technique* (SMOTE). O estudo envolveu 1.243 participantes no ensino superior em formato presencial. Foi aplicada seleção de atributos, mas os detalhes específicos não foram mencionados. Não houve destaque específico de variáveis. (Revathy et al., 2022).

ER65: Nesse artigo foi proposto um novo modelo baseado em um híbrido de *Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB) e Feed-forward Neural Networks (FNN)* para prever a evasão de alunos universitários, o modelo se saiu melhor nas métricas de acurácia e AUC, se comparado com modelos clássicos. O estudo abordou o ensino superior em formato presencial, mas o tamanho da amostra não foi especificado. Foi aplicada seleção de atributos, mas os detalhes específicos não foram mencionados. Não houve destaque específico de variáveis. (Niyogisubizo et al., 2022).

ER68: Esse estudo investiga a evasão no ensino médio em alguns países, é também criado um modelo chamado de "AutoML", que foi usado para melhorar a acurácia da previsão, selecionando os hiperparâmetros, recursos e algoritmo ML correspondentes para o conjunto de dados adquirido. O modelo proposto obteve melhores resultados quando comparado a outros modelos. O estudo considerou o ensino fundamental em formato presencial, abrangendo um total de 206.885 participantes. Foi aplicada seleção de atributos, mas os detalhes específicos não foram mencionados. As variáveis relevantes para a análise incluíram notas dos alunos (57%), idade do aluno (18%), distância (7%) e número de crianças (5%). (Mnyawami et al., 2022).

ER69: Nesse trabalho é realizada uma investigação e criação de modelo de predição para alunos que possuem o transtorno do espectro autista, o modelo de predição desenvolvido é baseado em aprendizagem profunda, usando os algoritmos *Long Short-Term Memory (LSTM) e Multilayer Perceptron (MLP)*. O estudo abordou o ensino fundamental em formato presencial, envolvendo 120 participantes. Foi aplicada seleção de atributos, mas os detalhes específicos não foram mencionados. Não houve destaque específico de variáveis. (Jarbou et al., 2022).

4.5.5 Estudos que utilizaram métodos clássicos

ER1: Esse trabalho trata da classificação do abandono de alunos universitários, os algoritmos abordados foram, *KNN, Naive Bayes, SVM, Decision Tree e Random Forest*, tiverem resultados bem próximos na maioria das métricas. O estudo abrangeu 1.211 participantes do ensino superior na modalidade EaD. Não houve seleção de atributos. As variáveis de destaque incluíram etnia, sexo, idade, status de tempo, graduação do estudante, classificação do estudante, GPA cumulativo e GPA de termo. (Kang & Wang, 2018).

ER2: Esse trabalho propõe um modelo de rede neural profunda, que é uma combinação de *Con-*

volutional Neural Network, Long Short-Term Memory network e Support Vector Machine, para ter um melhor desempenho na predição do abandono do aluno nos cursos online abertos massivos (MOOCs). O estudo englobou 120.542 participantes no ensino superior na modalidade EaD. Foi aplicada seleção de atributos, mas os detalhes específicos não foram especificados. As variáveis relevantes foram ID de inscrição, hora, fonte do evento, evento e dispositivo de acesso. (Wu et al., 2019).

ER3: Essa pesquisa aborda a categorização de desistência de estudantes universitários, dando foco na seleção dos atributos que melhoram essa predição de acordo com cada área acadêmica. O conjunto é formado por 24.894 participantes do ensino superior presencial, foi aplicada a abordagem Wrapper com o algoritmo RPart para seleção de atributos. As variáveis destacadas são adaptadas com base no curso específico. (Meca et al., 2020).

ER4: São usados nesse trabalho modelos avançados de aprendizagem de máquina, como redes neurais e árvores aumentadas de gradiente para prever o desempenho do aluno universitário e o abandono do mesmo, que baseia-se nos dados disponíveis no momento da inscrição. O melhor desempenho foi da *Fully Connected Deep Neural Network* (FCNN), com 72.4% de acurácia e 0.771 de AUC. O estudo envolveu 8.319 participantes do ensino superior em formato presencial. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Baranyi et al., 2020).

ER5: Esse trabalho trata da classificação de abandono de alunos universitários, dando foco aos cursos de Ciência, Tecnologia, Engenharia e Matemática. Ele compara técnicas de aprendizagem de máquina (AM) com análise de sobrevivência, os algoritmos de AM obtiveram melhores resultados. O estudo considerou 12.293 participantes do ensino superior em formato presencial. Não houve seleção de atributos. As variáveis de destaque foram idade, disciplinas e curso. (Chen et al., 2018).

ER6: São usados nesse estudo os seguintes algoritmos e modelos para predição de abandono na universidade, *Logistic Regression (LR), Adaboost (AB) e Decision tree (DT)* com Cox e TD-Cox. De forma que os melhores resultados foram para *Decision tree (DT)* com Cox e TD-Cox, o Cox com 71,9% de acurácia e 0.751 de AUC, e o TD-Cox com 71,9% e 0.705, no primeiro experimento, e se mantiveram com os melhores resultados no segundo. O estudo envolveu 11.121 participantes do ensino superior em formato presencial. Não houve seleção de atributos. As variáveis relevantes incluíram GPA do ensino médio, informações semestrais como GPA e porcentagem de crédito reprovado, além de atributos financeiros. (Ameri et al., 2016).

ER7: Essa pesquisa explora o abandono de alunos universitários, usando os seguintes algoritmos, *Naive Bayes, Multilayer Perception, Support Vetor Machine* e Tabela de Decisão. Tendo o Naive Bayes como algoritmo que obteve os melhores resultados. O estudo abrangeu 1.359 participantes do ensino superior em formato presencial. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Manhães et al., 2014).

ER8: Esse trabalho trata da classificação do abandono de alunos universitários, explorando a adequação da diminuição dimensional e concentrando os dados significativos encobertos nas informações de alunos com risco de abandono. Embora todos os algoritmos de aprendizado de máquina utilizados nesse trabalho tenham alcançado uma acurácia de mais de 75%, *Support Vetor Machine* e *Naive Bayes* tiveram melhor desempenho ao prever o abandono. O estudo envolveu 4.417 participantes no ensino superior presencial. Para análise, foram utilizados os métodos SVM,

Naïve Bayes, Árvore de Decisão, KNN e Floresta Aleatória. Não houve destaque específico de variáveis. (Lottering et al., 2020).

ER10: Esse trabalho trata da classificação do abandono de alunos em programas universitários, foram utilizados algoritmos baseados em árvore de decisão, *Naive Bayes*, K-NN, modelos lineares e aprendizado profundo. Os modelos com melhores resultados foram: *Gradient Boosted Trees* e *Deep Learning*, com uma AUC de 0.808 e 0.811, respectivamente. O estudo abrangeu 15.825 participantes do ensino superior em formato presencial. Não houve destaque específico de variáveis. (Nagy & Molontay, 2018).

ER11: Nessa pesquisa, aborda-se a classificação de alunos universitários, utilizando rede neural profunda. Obtendo uma acurácia de 92,41% com pré-treinamento. O estudo considerou 9.153 participantes do ensino fundamental em formato presencial. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Kuo et al., 2017).

ER12: Esse estudo trata da classificação de abandono de alunos em MOOCs, por meio de um modelo denominado CLSA, o modelo usa uma rede neural convolucional. O estudo não especificou o nível de ensino e foi conduzido na modalidade EaD, englobando um total de 79.186 participantes. Não houve seleção de atributos. As variáveis relevantes englobaram características de comportamento dos alunos durante as primeiras 5 semanas. (Fu et al., 2021).

ER13: Esse estudo explora da classificação de abandono de alunos em MOOCs, por meio de um modelo denominado CONV-LSTM, o modelo usa redes neurais convolucionais e memória de longo prazo. O estudo considerou o ensino superior na modalidade EaD e envolveu um total de 120.542 participantes. Não houve seleção de atributos. Não houve destaque específico de variáveis (Mubarak et al., 2021).

ER14: Esse trabalho trata da classificação de abandono do aluno do ensino médio, o algoritmo utilizado é Floresta Aleatória aplicado a um big data, com amostra de 165.715 alunos. Não houve seleção de atributos. As variáveis relevantes incluíram ausência, atraso, tempo de atividade autorregulada, tempo de desenvolvimento de carreira e saída não autorizada. (Chung & Lee, 2019).

ER15: Esse modelo trata da classificação de abandono de alunos universitários, são usadas as seguintes técnicas de aprendizagem de máquina, redes neurais *feed-forward*, *support vector machine* e conjuntos probabilísticos simplificados Fuzzy ARTMAP. Esses algoritmos são combinados, a fim de obter uma única classificação baseada nessa combinação. O estudo considerou o ensino superior na modalidade EaD, com um total de 193 participantes. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Lykourentzou et al., 2009).

ER17: Esse trabalho trata da classificação de abandono do aluno universitário em ambiente online, foram utilizados 4 algoritmos, o que se destacou entre eles foi o SVM como acurácia de 92,03%. O estudo considerou 162 participantes no ensino superior na modalidade EaD. Houve seleção de atributos, mas os detalhes específicos não foram mencionados. A variável destacada foi o desempenho nas disciplinas iniciais. (Santana et al., 2015).

ER18: Esse estudo explora a classificação de abandono de alunos no ensino fundamental e médio, o algoritmo utilizado foi árvores de decisão. O estudo abordou o ensino fundamental em formato presencial, englobando 220 participantes. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Selvan et al., 2019).

ER19: Nessa pesquisa, trata-se da classificação de abandono do aluno universitário, especificamente de um curso de computação empresarial, o algoritmo utilizado foi Árvores de decisão. O estudo considerou 1.888 participantes do ensino superior em formato presencial. Não houve seleção de atributos. As variáveis de destaque incluíram curso e período. (Nuankaew et al., 2019).

ER20: Nesse estudo, explora-se a classificação de abandono de alunos universitários, é utilizado um sistema que utiliza várias técnicas de aprendizagem de máquina, esse sistema foi feito com intuito de auxiliar os tutores na Universidade de Barcelona. O estudo abrangeu 4.434 participantes do ensino superior em formato presencial. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Rovira et al., 2017).

ER21: Essa pesquisa trata da classificação de abandono do aluno universitário em ambiente online, foi utilizado regressão logística lasso e ridge para criar um modelo de previsão de abandono. O estudo envolveu 32.593 participantes do ensino superior na modalidade EaD. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Radovanović et al., 2020).

ER22: Nessa pesquisa, trata-se da classificação de abandono do aluno universitário, ele utiliza uma combinação de algoritmos para aprimorar a classificação. O estudo considerou 200 participantes do ensino superior na modalidade EaD. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Dewan et al., 2015).

ER23: Esse trabalho explora a classificação de abandono do aluno universitário, especificamente do curso de ciência da computação, os algoritmos utilizados foram *Naive Bayes*, árvores de decisão, Regressão Logística, Máquinas de Vetores de Suporte e Redes Neurais. O estudo considerou 367 participantes do ensino superior em formato presencial. Não houve seleção de atributos. As variáveis de destaque foram ECTS do primeiro semestre (ECTS_1_sem) e SGPA do primeiro semestre (SGPA_1_sem). (Maksimova et al., 2020).

ER24: Esse trabalho trata da classificação de abandono do aluno universitário, especificamente no curso de ciência da computação, os algoritmos utilizados foram Árvore de decisão, regressão logística, floresta aleatória, K-vizinho mais próximo e algoritmo de rede neural. O conjunto do estudo é composto por 64 participantes do ensino superior presencial, a variável mais importante é o desempenho em disciplinas específicas. (Yaacob et al., 2020).

ER26: Essa pesquisa aborda a categorização de desistência do estudante universitário, sendo empregados 3 algoritmos, destacando-se a Floresta aleatória, com uma taxa de acurácia de 95,12% e *recall* de 91,41%. O estudo englobou 1.516 participantes em nível superior, com ensino presencial. Houve emprego de seleção de atributos, utilizando Regressão Logística, Árvores de Decisão e Floresta Aleatória como algoritmos de classificação. As variáveis destacadas foram a Média do terceiro semestre e a Média dos três primeiros semestres. (Costa et al., 2020).

ER27: Essa pesquisa aborda a categorização de desistência do estudante universitário, sendo empregados os algoritmos KNN, árvore de decisão, floresta aleatória, SVM e redes neurais. O modelo de redes neurais demonstrou o melhor desempenho, alcançando uma acurácia de 92%. O estudo contou com 26 participantes no nível superior, com aulas presenciais. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Quishpe-Morales et al., 2020).

ER28: Nesse estudo, é explorada a classificação de abandono do aluno universitário, ele utiliza uma combinação de 3 algoritmos de mineração de dados populares, floresta aleatória, máquinas de vetores de suporte e redes neurais artificiais. O estudo considerou o ensino superior presencial,

abrangendo 3.373/3.344 participantes. Foi aplicada a técnica de seleção de atributos forward search. Para o 1º semestre, variáveis destacadas incluíram ects_aprov_s, cod_escola, média_s, sexo, bolseiro_s, entre outras; para o 2º semestre, variáveis como ects_reprov_s, cod_prof_mae, n10_11_acesso, idade, nacionalidade, entre outras, foram relevantes. (Martins et al., 2020).

ER29: Nesse estudo, aborda-se a categorização da evasão de estudantes universitários. Como parte da pesquisa, são empregados os algoritmos *Decision Stump*, NDTREE e *Enhanced Machine Learning* (EMLA), os quais são fundamentados em técnicas de aprendizado de máquina. O estudo envolveu 407 participantes no nível superior, com ensino presencial. Não foi realizada seleção de atributos. Não houve destaque específico de variáveis. (SALLAN & BEHAL, s.d.).

ER30: Esse estudo trata da classificação de abandono de alunos em ambiente online, os modelos utilizados foram Regressão Logística adicionando um termo de regularização e o Modelo de Markov Oculto Input-Output (IOHMM). O estudo abrangeu 32.593 participantes no ensino superior na modalidade EaD. Foi aplicada seleção de atributos, mas os detalhes específicos não foram fornecidos. Não houve destaque específico de variáveis. (Mubarak et al., 2020).

ER31: Nessa pesquisa, explora-se a classificação de abandono do aluno universitário, os algoritmos utilizados foram regressão logística e árvores de decisão. Como resultado, obteve-se que a técnica com maior percentual de acurácia de abandono foi a árvore de decisão com 91.70%. O estudo englobou 1.178 participantes no ensino superior em formato presencial. Houve seleção de atributos utilizando a técnica Relief Feature Selection. Não houve destaque específico de variáveis. (ALBAN & MAURICIO, 2018).

ER32: Esse trabalho analisa modelos de classificação de abandono do aluno universitário, foram utilizados os algoritmos *Decision Tree* e *Naive Bayes*, o *Decision Tree* obteve melhores resultados. A pesquisa abrangeu 1.862 participantes no ensino superior com formato presencial. Foi aplicada a seleção de atributos baseada no algoritmo Firefly. Não houve destaque específico de variáveis. (Gamao & Gerardo, 2019).

ER34: Esse trabalho trata da classificação de abandono do aluno universitário, o algoritmo utilizado é árvore de decisão aprimorado com base no ID3. A pesquisa englobou 240 participantes do ensino superior em formato presencial. Houve seleção de atributos, porém os detalhes específicos não foram especificados. As variáveis relevantes incluíram problemas familiares, doença em casa, ambiente do campus, baixa taxa de colação, mudança de objetivo pessoal e problema de ajuste. (Sivakumar et al., 2016).

ER35: Essa pesquisa trata da classificação de abandono do aluno universitário, os algoritmos utilizados foram *K-Nearest Neighbor* (KNN), *Naïve Bayes* (NB) e *Decision Tree* (DT). A pesquisa abordou 17.432 participantes no ensino superior presencial. Utilizou a técnica Learning Vector Quantization para seleção de atributos. As variáveis de destaque incluíram frequência, pontuações de tarefas, créditos totais, pontuações UTS, UAS pontuações, GPA, renda dos pais, educação dos pais, sexo e idade dos estudantes. (Hutagaol & Suharjito, 2019).

ER36: Nesse estudo, trata-se da classificação de abandono de alunos no ensino base, foram utilizados vários algoritmos, apesar de tradicional o SVM mostrou bons resultados comparado com modelos mais complexos. O conjunto é formado por um total de 15.000 participantes. Foi aplicada seleção de atributos, mas os detalhes específicos não foram mencionados. Não houve destaque específico de variáveis. (Del Bonifro et al., 2020).

ER37: Esse trabalho trata da classificação do abandono de alunos universitários. Foram desenvolvidos 3 classificadores com base em uma árvore de decisão, floresta aleatória e classificação de aumento de gradiente. Os resultados mostram que a acurácia da previsão de aumento de gradiente, árvore de decisão e modelos de floresta aleatória são 93%, 92% e 92%, respectivamente. O trabalho também aponta alguns atributos relevantes na classificação, ano letivo, GPA do ensino médio e canais de admissão. O estudo envolveu 13.714 participantes do ensino superior em formato presencial. Não houve seleção de atributos. (Tenpipat & Akkarajitsakul, 2020).

ER38: Nessa pesquisa, trata-se da classificação de abandono do aluno em qualquer nível escolar, são utilizadas várias técnicas de classificação, como regras de indução e árvore de decisão. O estudo abordou 670 participantes do ensino médio em formato presencial. Foi aplicada seleção de atributos, embora os detalhes específicos não tenham sido mencionados. Não houve destaque específico de variáveis. (Pradeep et al., 2015).

ER39: Esse trabalho aborda a classificação de abandono do aluno universitário, especificamente no curso de Ciências da computação. O algoritmo utilizado é Árvore de decisão, aplica-se também seleção de recurso para identificar quais os melhores atributos para a classificação. A pesquisa incluiu 963 participantes no nível superior, com aulas presenciais. A seleção de atributos foi realizada por meio do algoritmo Boruta. As variáveis em destaque abrangeram GRADE, A1SCORE e QATTEMPT. (Naseem et al., 2019).

ER40: Nesse estudo, aborda-se a classificação de abandono de alunos universitários, o algoritmo utilizado é Árvore de decisão. A pesquisa incluiu 206 participantes no ensino superior em formato presencial. Foi aplicada seleção de atributos, embora os detalhes específicos não tenham sido mencionados. As variáveis de destaque incluíram WAG, AI, ASR, SFR, FSE e PF. (Bello et al., 2020).

ER41: Esse trabalho trata da classificação de abandono do aluno em qualquer nível escolar, ele utiliza técnicas de previsão, com foco em métodos de classificação baseados em árvores e SVM. O estudo abrangeu 220.685 participantes da modalidade presencial. Não houve seleção de atributos. As variáveis de destaque foram ausências geralmente altas e pontuações baixas em matemática. (Sorensen, 2019).

ER42: Esse estudo trata da classificação de abandono de alunos em MCCOs, o estudo utiliza um sistema de inferência neuro-difuso adaptativo (ANFIS), que é comparado com algoritmos clássicos, como árvore de decisão, regressão logística, svm, etc. Os resultados mostram que a ANFIS obteve o melhor desempenho com relação aos algoritmos. O estudo abrangeu os níveis de ensino superior, médio e fundamental na modalidade EaD, com um total de 120.542 participantes. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Şahin, 2021).

ER43: Nesse estudo, explora-se a categorização do abandono de alunos universitários, utilizando diferentes algoritmos, como Floresta aleatória, máquinas de vetores de suporte, árvores de decisão, Naive Bayes, K-vizinho mais próximo e regressão logística. Destaca-se que o algoritmo de Floresta aleatória apresentou os melhores resultados, com uma acurácia de 94.14%. O estudo abordou o ensino superior em formato presencial, envolvendo 12.293 participantes. O artigo não menciona a aplicação de uma abordagem específica para seleção de atributos. As variáveis importantes englobaram a idade, as disciplinas e o curso dos participantes. (Lottering et al., s.d.).

ER44: Esse trabalho trata da classificação de abandono do aluno universitário, o modelo de classi-

ficador é baseado em árvores de decisão. O estudo considerou 6.870 participantes da modalidade presencial. Não houve seleção de atributos. As variáveis de destaque foram média baixa nas notas e o número de cursos perdidos nos semestres iniciais do programa. (Pereira & Zambrano, 2017).

ER45: Essa pesquisa tem como foco a classificação de abandono do aluno universitário, especificamente no curso de engenharia de sistemas. Os algoritmos utilizados nos experimentos foram: Árvore de decisão, Regressão logística e Naive Bayes. Tendo Árvore de decisão como algoritmo que obteve melhor resultado, AUC de 0.94. O estudo envolveu 802 indivíduos no nível superior com aulas presenciais. Não foi realizada seleção de atributos. Não houve destaque específico de variáveis. (Perez et al., 2018).

ER46: Esse pesquisa explora a classificação de abandono do aluno universitário, onde são comparado dois modelos, um baseado em técnica de regressão logística e o outro modelo baseado em árvores de decisão. A pesquisa envolveu 4.519 participantes no ensino superior em formato presencial. Houve seleção de atributos, embora os detalhes específicos não tenham sido mencionados. Na regressão logística, as variáveis relevantes foram a pontuação de matemática da PSU e as notas do ensino médio (NEM). Na Árvore de Decisão, as variáveis de destaque incluíram NEM, residência do estudante e a pontuação de matemática da PSU. (Pérez et al., 2018).

ER47: Esse estudo trata da classificação de abandono do aluno universitário, mais especificamente dos cursos de engenharia. Os algoritmos utilizados foram Árvores de decisão, regressão logística e Naive Bayes. O estudo abrangeu 762 participantes da modalidade presencial. Não houve seleção de atributos. A variável de destaque foi o Std GPA (GPA padronizado). (Pérez-Gutiérrez, 2020).

ER48: Esse trabalho trata da classificação de abandono do aluno em ambiente online, modelos de regressão logística são usados para a classificação. O estudo considerou o ensino superior, envolvendo um total de 104 participantes. Não houve seleção de atributos. A variável de destaque foram as notas nas semanas cruciais do curso. (Burgos et al., 2018).

ER49: Essa pesquisa aborda a classificação de abandono do aluno em MOOCs, é estudado modelos de aprendizado profundo, já que segundo o estudo essa abordagem constrói modelos de previsão de abandono mais precisos. O estudo abrangeu os níveis de ensino superior, médio e fundamental, tendo como conjunto um total de 3.617 participantes. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Xing & Du, 2019).

ER50: Esse trabalho aborda a classificação de abandono do alunos do ensino médio de um instituto de tecnologia da Tailândia, os algoritmos utilizados foram Árvores de decisão e Floresta aleatória, Árvores de decisão obteve os melhores resultados. A pesquisa abrangeu 28.801 participantes em nível superior com ensino presencial. Utilizou-se o critério GPAX para seleção de atributos, as notas nos cursos de idiomas foram identificadas como variáveis de destaque. (Limsathitwong et al., 2018).

ER52: Esse estudo aborda a classificação de abandono do aluno em MOOCs, são utilizados vários modelos, o melhor desempenho é do modelo *LightGBM*, com acurácia de 96%. O tamanho do conjunto não foi explicitado nesse trabalho. O estudo envolveu níveis de ensino superior, médio e fundamental na modalidade EaD. Houve seleção de atributos, embora os detalhes específicos não tenham sido mencionados. A variável mais relevante foi os dados de interação dos alunos. (Panagiotakopoulos et al., 2021).

ER53: Esse trabalho trata da classificação de abandono de alunos universitários, são usados métodos de aprendizagem de máquina baseados em árvore para a predição. O estudo considerou 197 participantes do ensino superior na modalidade EaD. Houve seleção de atributos, embora os detalhes específicos não tenham sido mencionados. As variáveis relevantes incluíram assuntos difíceis de matemática e estatística. (Figueroa-Cañas & S.Vinuesa, 2020).

ER54: Nesse estudo, aborda-se a classificação de abandono de alunos, tanto em escolas quanto em universidades. O algoritmo utilizado é o *Naive Bayes*, com a linguagem R. O trabalho também foca na seleção de recurso para escolha dos melhores atributos. O estudo envolveu 50 participantes no ensino superior em formato presencial. As variáveis relevantes incluíram fatores acadêmicos, demográficos, psicológicos e de saúde. (Hegde & Prageeth, 2018).

ER55: Essa pesquisa trata da classificação de abandono do aluno em ambiente online, são usados modelos de regressão para a classificação. O estudo abordou 104 participantes do ensino superior na modalidade EaD. Não houve seleção de atributos. As variáveis de destaque foram as notas nas semanas cruciais do semestre. (de la Peña et al., 2017).

ER56: Nessa pesquisa, explora-se a classificação de abandono do aluno no ambiente online, ele apresenta um *benchmark* de algoritmo recentemente proposto, algoritmo de modelo de folha (LLM), comparando-o com outros 8 algoritmos, o LLM tem os melhores resultados se comparado aos outros algoritmos. O estudo envolveu níveis de ensino superior, médio e fundamental na modalidade EaD, com um total de 10.554 participantes. Não houve seleção de atributos. A variável de destaque foi acadêmico noivado. (Coussement et al., 2020).

ER57: Esse trabalho trata da classificação de abandono do aluno em ambiente online, os modelos de previsão foram desenvolvidos usando *Artificial Neural Network (ANN), Decision Tree (DT) e Bayesian Networks (BNs)*, o que obteve melhor desempenho foi DT. O estudo considerou 62.375 participantes no ensino superior na modalidade EaD. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Tan & Shao, 2015).

ER58: Nesse estudo, aborda-se a classificação de abandono do aluno universitário, os métodos de classificação utilizados incluíram *Naive Bayes*, C4.5, *BackPropagation*, SMO, 3NN e MLE, é implementando também um protótipo de sistema web utilizando o *Naive Bayes*. A pesquisa envolveu 354 participantes no ensino superior na modalidade EaD. A seleção de atributos foi realizada por meio da abordagem Wrapper. Não houve destaque específico de variáveis. (Kotsiantis et al., 2003).

ER59: Nesse trabalho, é examinada a classificação do abandono de estudantes universitários, os algoritmos utilizados foram árvores de decisão C4.5 e ID3. O estudo abrangeu 201 participantes no ensino superior com formato presencial. Foi aplicada seleção de atributos, porém detalhes específicos da seleção de atributos não foram especificados. (Heredia et al., 2015).

ER60: Esse trabalho trata da classificação de abandono do aluno em ambiente online, foi utilizado um modelo mais robustos nesse estudo, modelo de rede neural recorrente (RNN) com células de memória de curto prazo longa (LSTM). O estudo abrangeu níveis de ensino superior, médio e fundamental na modalidade EaD, com um total de 39.877/27.629 participantes. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Fei & Yeung, 2015).

ER63: Esse artigo se concentra na descoberta precoce de variáveis de abandono como um avanço pela redução de dimensionalidade usando seleção de atributos e métodos de extração, é utilizada

a *Synthetic Minority Oversampling Technique* (SMOTE). O estudo envolveu 1.243 participantes no ensino superior em formato presencial. Foi aplicada seleção de atributos, mas os detalhes específicos não foram mencionados. Não houve destaque específico de variáveis. (Revathy et al., 2022).

ER64: Esse trabalho realiza uma investigação da evasão de estudantes universitários, nessa investigação é considerado o contexto da pandemia do COVID-19 e quais fatores da pandemia influenciaram na evasão. Para a construção dos modelos de predição foi utilizado algoritmos populares, como SVM, regressão logística, floresta aleatória, árvore de decisão, etc. O estudo considerou a modalidade EaD, sem especificar o tamanho da amostra. Não houve seleção de atributos. As variáveis mais relevantes incluíram Número de Semestres com Reprovação, Motivo de Reprovação em Semestres, Tipo de Universidade, Impacto da Covid, Estudante de Engenharia, Média CGPA na Disciplina 'C', Área de Residência, Gênero, Disponibilidade de Internet e Renda Familiar. (Hossain et al., 2022).

ER65: Nesse artigo foi proposto um novo modelo baseado em um híbrido de *Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB) e Feed-forward Neural Networks (FNN)* para prever a evasão de alunos universitários, o modelo se saiu melhor nas métricas de acurácia e AUC, se comparado com modelos clássicos. O estudo abordou o ensino superior em formato presencial, mas o tamanho da amostra não foi especificado. Foi aplicada seleção de atributos, mas os detalhes específicos não foram mencionados. Não houve destaque específico de variáveis. (Niyogisubizo et al., 2022).

ER66: Nesse trabalho foi desenvolvido um sistema que permite o cálculo do risco de evasão por aluno e utiliza um procedimento de geração de alertas para coordenar as intervenções. A plataforma permitiu mensurar o impacto das estratégias de intervenção na permanência dos alunos. O estudo abrangeu 15.805 participantes no ensino superior em formato presencial. Foi empregada a abordagem ingênua (naive approach) para seleção de atributos. As técnicas de classificação utilizadas incluíram o algoritmo *AdaBoost*, *Bayesian GLM*, *Decision Trees*, *LogitBoost*, *Random Forest* (RF) e *Stochastic Gradient Boosting*. Não houve destaque específico de variáveis. (Guzmán-Castillo et al., 2022).

ER67: Esse estudo criou um modelo combinando seleção de atributos com um método de rede neural perceptron multicamadas. O modelo foi comparado com modelos baseados nos algoritmos de *Logistic regression*, *Decision Tree*, *Random Forest*, *Naive Bayes*, *Support Vector Machine* e *Multilayer Perceptron Neural Network*, e obteve melhores resultados. A pesquisa considerou 1.650 participantes no ensino superior presencial. Foram aplicadas as técnicas de seleção de atributos GR, CS e Métodos CFS. As variáveis destacadas englobaram a média cumulativa de notas (GPA), média cumulativa de notas para assuntos não docentes (GPAnone) e adesão ao grupo de mídia social em assuntos (*Socialclass*). (Nuanmeesri et al., 2022).

ER68: Esse estudo investiga a evasão no ensino médio em alguns países, é também criado um modelo chamado de "AutoML", que foi usado para melhorar a acurácia da previsão, selecionando os hiperparâmetros, recursos e algoritmo ML correspondentes para o conjunto de dados adquirido. O modelo proposto obteve melhores resultados quando comparado a outros modelos. O estudo considerou o ensino fundamental em formato presencial, abrangendo um total de 206.885 participantes. Foi aplicada seleção de atributos, mas os detalhes específicos não foram mencionados. As variáveis relevantes para a análise incluíram notas dos alunos (57%), idade do aluno (18%),

distância (7%) e número de crianças (5%). (Mnyawami et al., 2022).

ER69: Nesse trabalho é realizada uma investigação e criação de modelo de predição para alunos que possuem o transtorno do espectro autista, o modelo de predição desenvolvido é baseado em aprendizagem profunda, usando os algoritmos *Long Short-Term Memory (LSTM) e Multilayer Perceptron (MLP)*. O estudo abordou o ensino fundamental em formato presencial, envolvendo 120 participantes. Foi aplicada seleção de atributos, mas os detalhes específicos não foram mencionados. Não houve destaque específico de variáveis. (Jarbou et al., 2022).

ER70: O objetivo nesse trabalho é reduzir a taxa de evasão dos alunos da Faculdade de Engenharia de Sistemas e Informática da Universidade Nacional Mayor de San Marcos (FISI-UNMSM), o modelo de predição utilizado é baseado em árvores de decisão, obteve uma acurácia de 90.34%. A pesquisa abrangeu 1.938 participantes no ensino superior presencial. Foram aplicadas as técnicas de seleção de atributos ChiSquared, OneR, GainRatio, InforGain e ClassifierAttribute. As variáveis destacadas incluíram a média ponderada histórica de notas, a média ponderada de notas do último ciclo e o número de créditos de cursos aprovados. (Vega et al., 2022).

ER71: Nesse trabalho é feita uma análise da evasão no curso de Ciência da Computação, os algoritmos utilizados foram *Random Forest*, *Decision Tree*, *Naive Bayes*, *Logistic Regression* e *K-Nearest Neighbour*. No primeiro estágio o algoritmo que obteve o melhor resultado foi o Naive Bayes com uma AUC de 0.6123, nos estágios 2 e 3 o algoritmo que obteve o melhor resultado foi o Logistic Regression com AUC de 0.7523 e 0.8902, respectivamente. O tamanho do conjunto de dados não foi mencionado no artigo, mas trata-se de dados de uma única Instituição de Ensino Superior da região do Pacífico Sul, que é o único campus regional no Pacífico Sul. (Naseem et al., 2022).

4.5.6 Estudos que não utilizaram métodos clássicos

ER9: Esse trabalho foca na redução de dimensionalidade dos atributos para melhorar a classificação de abandono de alunos universitários. O estudo envolveu 150 participantes no ensino superior com aulas presenciais. Foi realizada redução dimensional utilizando a técnica PCA (Análise de Componentes Principais). Não houve destaque específico de variáveis. (Hegde, 2016).

ER16: Esse estudo trata da classificação de abandono de alunos de todos níveis de ensino, utilizando Modelos de Equações Estruturais de Mistura (MSEM). O estudo envolveu 12.548 participantes da modalidade presencial. Não houve seleção de atributos. As variáveis de destaque foram saúde do aluno, frequência e relações interpessoais. (Viloria & Lezama, 2019).

ER25: Nesse estudo, trata-se da classificação de abandono de alunos universitários, mais especificamente do curso de enfermagem. Este estudo relata o uso do pacote Answer Tree da SPSS para esse propósito. O estudo considerou 528 participantes em modalidade presencial. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Moseley & Mead, 2008).

ER33: Nesse trabalho, trata-se da classificação de abandono dos alunos universitários. O modelo aplicado é a regressão linear e o estudo também enfoca a seleção de recursos. O estudo incluiu 530 indivíduos no ensino superior, com aulas presenciais. A seleção de atributos foi conduzida por meio da Correlação de Pearson. Não houve destaque específico de variáveis. (Aguirre & Pérez, 2020).

ER51: Essa pesquisa explora a classificação de abandono de aluno universitário, ele utiliza em seu benchmark um sistema chamado *FragSte*, um sistema de detecção precoce para universidades alemãs, o FragSte aplica métodos de aprendizagem de máquina. O estudo abrangeu 29.500 participantes em modalidade presencial. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Berens et al., 2018).

ER61: É apresentado nesse estudo um novo modelo que usa vários módulos Inception conectados por meio de uma rede residual. Operações paralelas são executadas nos dados usando a operação MaxPooling e três comprimentos diferentes de convolução dentro do módulo inicial. É feito um comparativo com outros modelos e o modelo proposto se saiu significativamente melhor. O estudo abordou os níveis de ensino superior, médio e fundamental na modalidade EaD, sem especificar o tamanho da amostra. Não houve seleção de atributos. Não houve destaque específico de variáveis. (Li et al., 2022).

ER62: Esse trabalho tem como objetivo melhorar um sistema de alerta de evasão, nele é incorporado fatores relacionados ao ambiente escolar e ao aprendizado socioemocional, diferenciando-se de outros sistemas de evasão que dependem exclusivamente dos fatores: frequência, comportamento e curso. O estudo considerou o ensino fundamental em formato presencial, com um total de 62.000 e 48.900 participantes em duas situações distintas. Não houve seleção de atributos. O algoritmo de classificação utilizado foi Regressão Linear. As variáveis de destaque foram Competência Linguística Cultural, Relacionamentos, Segurança Emocional e Segurança Física. (Su et al., 2022).

É notório o uso majoritário de algoritmos baseados em árvores de decisão na grande maioria dos experimentos, além deles mostrarem excelentes resultados em várias métricas de análise de modelos de classificação, como acurácia, curva ROC, entre outras, se comparados com algoritmos clássicos, também é notado que a modalidade de ensino presencial e alunos no nível universitário foram os mais utilizados nas predições, ainda assim foram encontrados muitos trabalhos que realizam a predição em modalidade EaD, principalmente quando se trata de MOOCs (do inglês, *Massive Open Online Course*).

Desempenho durante determinado tempo de estudo, disciplina e curso, são as variáveis mais destacadas pelos modelos de classificação. Todos os estudos tiveram uma abordagem empírica, e tratando-se dos resultados dos modelos, mais de 30% dos ER tiveram em seus resultados acurácia superiores a 95%, o estudo com menor acurácia obteve 61.68%.

5 Ameaças à Validade

Mesmo o mapeamento sistemático tendo sido realizado de forma cautelosa, ainda é possível elencar algumas possíveis limitações ou ameaças que afetem a validade dos resultados, sendo elas: (i) a decisão sobre quais estudos incluir ou excluir, pode em algum momento ter sido tendenciosa, mesmo que de forma não intencional e, portando, uma ameaça. De forma a minimizá-la, os processos de seleção foram realizados e revisados em par; (ii) a *string* de busca pode não incluir todos os termos relacionados ao tema de pesquisa. No entanto, foram realizados testes pilotos de maneira a ajustar e refinar a *string* apresentada; e (iii) este mapeamento sistemático focou totalmente em estudo empíricos, descartando estudos como revisões ou mapeamentos sistemáticos,

aumentando a ameaça de validade externa. Ainda assim esse mapeamento abrangeu o escopo ao investigar a evasão escolar em todos os níveis de ensino e modalidades de ensino, trazendo dessa forma mais detalhes sobre os experimentos de datasets variados (de Morais et al., 2021).

6 Conclusão

Esse artigo teve como objetivo mapear sistematicamente os artigos do estado da arte sobre a aplicação de DM e ML na classificação de dados em estudos sobre evasão escolar.

Para realização desse mapeamento sistemático foram utilizadas as bases ACM, IEEE, Scopus, Web of Science e ScienceDirect, citadas anteriormente. Logo depois, foram filtrados os estudos primários, através dos critérios de inclusão e exclusão, tendo como resultado os ER, que foram utilizados para responder às questões propostas. Dessa forma, percebe-se que a quantidade de publicações sobre aplicação de DM e ML na classificação de dados em estudos sobre evasão escolar é crescente, e que surgem cada vez mais modelos, algoritmos, sistemas que ajudam na evolução da área de pesquisa. Além disso, percebeu-se um uso maior de algoritmos baseados em árvores de decisão nos experimentos, e que também a modalidade de ensino presencial e alunos no nível universitário foram os mais abordados nos trabalhos.

Portanto, é perceptível os avanços de algoritmos, softwares, sistemas e outros, que tem como objetivo melhorar o entendimento sobre o fenômeno da evasão escolar, tanto na avaliação de métodos de AM em termos de acurácia, curva ROC, como também na identificação de variáveis relevantes para prever a evasão, ajudando assim na gestão das instituições de ensino, consequentemente reduzindo prejuízos para as mesmas.

Referências

Aguirre, C. E., & Pérez, J. C. (2020). Predictive data analysis techniques applied to dropping out of university studies. *2020 XLVI Latin American Computing Conference (CLEI)*, 512–521. https://doi.org/10.1109/clei52000.2020.00066 [GS Search].

ALBAN, M. S., & MAURICIO, D. (2018). Prediction of university dropout through technological factors: a case study in Ecuador. *Revista Espacios*, *39*(52). https://doi.org/10.1109/educon.2018.8363371 [GS Search].

Ameri, S., Fard, M. J., Chinnam, R. B., & Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 903–912. https://doi.org/10.1145/2983323.2983351 [GS Search].

Baranyi, M., Nagy, M., & Molontay, R. (2020). Interpretable Deep Learning for University Dropout Prediction. *Proceedings of the 21st Annual Conference on Information Technology Education*, 13–19. https://doi.org/10.1145/3368308.3415382 [GS Search].

Bello, F. A., Kóhler, J., Hinrechsen, K., Araya, V., Hidalgo, L., & Jara, J. L. (2020). Using machine learning methods to identify significant variables for the prediction of first-year Informatics Engineering students dropout. *2020 39th International Conference of the Chilean*

Computer Science Society (SCCC), 1–5. https://doi.org/10.1109/sccc51225.2020.9281280 [GS Search].

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2018). Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods. https://doi.org/10.2139/ssrn.3275433 [GS Search].

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, *66*, 541–556. https://doi.org/10.1016/j.compeleceng.2017.03.005 [GS Search].

Chen, Y., Johri, A., & Rangwala, H. (2018). Running out of stem: a comparative study across stem majors of college students at-risk of dropping out early. *Proceedings of the 8th international conference on learning analytics and knowledge*, 270–279. https://doi.org/10.1145/3170358.3170410 [GS Search].

Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, *96*, 346–353. https://doi.org/10.1016/j.childyouth.2018.11.030 [GS Search].

Colpo, M. P., Primo, T. T., Pernas, A. M., & Cechinel, C. (2020). Mineraçao de dados educacionais na previsao de evasao: uma rsl sob a perspectiva do congresso brasileiro de informática na educaçao. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, 1102–1111. https://doi.org/10.5753/cbie.sbie.2020.1102 [GS Search].

Costa, A. G., Queiroga, E., Primo, T. T., Mattos, J. C., & Cechinel, C. (2020). Prediction analysis of student dropout in a Computer Science course using Educational Data Mining. *2020 XV Conferencia Latinoamericana de Tecnologias de Aprendizaje (LACLO)*, 1–6. https://doi.org/10.1109/laclo50806.2020.9381166 [GS Search].

Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems*, *135*, 113325. https://doi.org/10.1016/j.dss.2020.113325 [GS Search].

de la Peña, D., Lara, J. A., Lizcano, D., Martínez, M. A., Burgos, C., & Campanario, M. L. (2017). Mining activity grades to model students' performance. *2017 International Conference on Engineering & MIS (ICEMIS)*, 1–6. https://doi.org/10.1109/icemis.2017.8272963 [GS Search].

Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. *International Conference on Artificial Intelligence in Education*, 129–140. https://doi.org/10.1007/978-3-030-52237-7_11 [GS Search].

de Morais, F. L., Melo, A., Moutinho, M., & Fagundes, R. (2021). Modelos de regressao aplicados na previsao da evasao escolar do ensino básico: uma revisao sistemática da literatura. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, 168–178. https://doi.org/10.5753/sbie.2021.218504 [GS Search].

Dewan, M. A. A., Lin, F., Wen, D., et al. (2015). Predicting dropout-prone students in e-learning education system. *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, 1735–1740. https://doi.org/10.1109/uic-atc-scalcom-cbdcom-iop.2015.315 [GS Search].

Fei, M., & Yeung, D.-Y. (2015). Temporal models for predicting student dropout in massive open online courses. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 256–263. https://doi.org/10.1109/icdmw.2015.174 [GS Search].

Figueroa-Cañas, J., & S.Vinuesa, T. (2020). Early prediction of dropout and final exam performance in an online statistics course. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, *15*(2), 86–94. https://doi.org/10.1109/rita.2020.2987727 [GS Search].

Fu, Q., Gao, Z., Zhou, J., & Zheng, Y. (2021). CLSA: A novel deep learning model for MOOC dropout prediction. *Computers & Electrical Engineering*, *94*, 107315. https://doi.org/10.1016/j.compeleceng.2021.107315 [GS Search].

Gamao, A., & Gerardo, B. (2019). Prediction-based model for student dropouts using modified mutated firefly algorithm. *8*(6), 3461–3469. https://doi.org/10.30534/ijatcse/2019/122862019 [GS Search].

Guzmán-Castillo, S., Körner, F., Pantoja-García, J. I., Nieto-Ramos, L., Gómez-Charris, Y., Castro-Sarmiento, A., & Romero-Conrado, A. R. (2022). Implementation of a Predictive Information System for University Dropout Prevention. *Procedia Computer Science*, *198*, 566–571. https://doi.org/10.1016/j.procs.2021.12.287 [GS Search].

Hegde, V. (2016). Dimensionality reduction technique for developing undergraduate student dropout model using principal component analysis through R package. *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 1–6. https://doi.org/10.1109/iccic.2016.7919670 [GS Search].

Hegde, V., & Prageeth, P. (2018). Higher education student dropout prediction and analysis through educational data mining. *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 694–699. https://doi.org/10.1109/icisc.2018.8398887 [GS Search].

Heredia, D., Amaya, Y., & Barrientos, E. (2015). Student dropout predictive model using data mining techniques. *IEEE Latin America Transactions*, *13*(9), 3127–3134. https://doi.org/10.1109/tla.2015.7350068 [GS Search].

Hossain, M., Azad, S. B. M. S., Hossen, M. L., Khan, S. I., & Masum, A. K. M. (2022). Predictive Analysis on University Dropout Rate of Bangladesh in Covid-19. *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, 439–444. https://doi.org/10.1109/iciset54810.2022.9775898 [GS Search].

Hutagaol, N., & Suharjito. (2019). Predictive modelling of student dropout using ensemble classifier method in higher education. *4*(4), 206–211. https://doi.org/10.25046/aj040425 [GS Search].

Jarbou, M., Won, D., Gillis-Mattson, J., & Romanczyk, R. (2022). Deep learning-based school attendance prediction for autistic students. *Scientific Reports*, *12*(1), 1–11. https://doi.org/10.1038/s41598-022-05258-z [GS Search].

Kang, K., & Wang, S. (2018). Analyze and predict student dropout from online programs. *Proceedings of the 2nd International Conference on Compute and Data Analysis*, 6–12. https://doi.org/10.1145/3193077.3193090 [GS Search].

Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. *International conference on knowledge-based and intelligent information and engineering systems*, 267–274. https://doi.org/10.1007/978-3-540-45226-3_37 [GS Search].

Kuo, J. Y., Pan, C. W., & Lei, B. (2017). Using stacked denoising autoencoder for the student dropout prediction. *2017 IEEE International Symposium on Multimedia (ISM)*, 483–488. https://doi.org/10.1109/ism.2017.96 [GS Search].

Li, Y., Cui, X., & Zhang, Z. (2022). Dropout Rate Prediction for MOOC based on Inception-time Model. *Proceedings of the 7th International Conference on Distance Education and Learning*, 54–59. https://doi.org/10.1145/3543321.3543330 [GS Search].

Limsathitwong, K., Tiwatthanont, K., & Yatsungnoen, T. (2018). Dropout prediction system to reduce discontinue study rate of information technology students. *2018 5th International Conference on Business and Industrial Research (ICBIR)*, 110–114. https://doi.org/10.1109/icbir.2018.8391176 [GS Search].

Lottering, R., Hans, R., & Lall, M. (s.d.). A Machine Learning Approach to Identifying Students at Risk of Dropout: A Case Study. https://doi.org/10.14569/ijacsa.2020.0111052 [GS Search].

Lottering, R., Hans, R., & Lall, M. (2020). A model for the identification of students at risk of dropout at a university of technology. *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 1–8. https://doi.org/10.1109/icabcd49160.2020.9183874 [GS Search].

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, *53*(3), 950–965. https://doi.org/10.1016/j.compedu.2009.05.010 [GS Search].

Maksimova, N., Pentel, A., & Dunajeva, O. (2020). Predicting First-Year Computer Science Students Drop-Out with Machine Learning Methods: A Case Study. *International Conference on Interactive Collaborative Learning*, 719–726. https://doi.org/10.1007/978-3-030-68201-9_70 [GS Search].

Manhães, L. M. B., da Cruz, S. M. S., & Zimbrão, G. (2014). WAVE: an architecture for predicting dropout in undergraduate courses using EDM. *Proceedings of the 29th annual acm symposium on applied computing*, 243–247. https://doi.org/10.1145/2554850.2555135 [GS Search].

Martins, M. P., Migueis, V. L., Fonseca, D., & Gouveia, P. D. (2020). Previsão do abandono académico numa instituição de ensino superior com recurso a data mining. [GS Search].

Meca, I., Rabasa, A., Sobrino, E., & López-Espín, J. J. (2020). Early Warning Methodology for dropping out of university degrees. *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 245–249. https://doi.org/10.1145/3434780.3436596 [GS Search].

Mnyawami, Y. N., Maziku, H. H., & Mushi, J. C. (2022). Enhanced Model for Predicting Student Dropouts in Developing Countries Using Automated Machine Learning Approach: A Case of Tanzanian's Secondary Schools. *Applied Artificial Intelligence*, *36*(1), 2071406. https://doi.org/10.1080/08839514.2022.2071406 [GS Search].

Moseley, L. G., & Mead, D. M. (2008). Predicting who will drop out of nursing courses: a machine learning exercise. *Nurse education today*, *28*(4), 469–475. https://doi.org/10.1016/j.nedt.2007.07.012 [GS Search].

Mubarak, A. A., Cao, H., & Hezam, I. M. (2021). Deep analytic model for student dropout prediction in massive open online courses. *Computers & Electrical Engineering*, *93*, 107271. https://doi.org/10.1016/j.compeleceng.2021.107271 [GS Search].

Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, 1–20. https://doi.org/10.1080/10494820.2020.1727529 [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].

Naseem, M., Chaudhary, K., & Sharma, B. (2022). Predicting Freshmen Attrition in Computing Science using Data Mining. *Education and Information Technologies*, 1–31. https://doi.org/10.1007/s10639-022-11018-3 [GS Search].

Naseem, M., Chaudhary, K., Sharma, B., & Lal, A. G. (2019). Using ensemble decision tree model to predict student dropout in computing science. *2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 1–8. https://doi.org/10.1109/csde48274.2019.9162389 [GS Search].

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, *3*, 100066. https://doi.org/10.1016/j.caeai.2022.100066 [GS Search].

Nuankaew, P., Nuankaew, W., Phanniphong, K., Fooprateepsiri, R., & Bussaman, S. (2019). Analysis dropout situation of business computer students at University of Phayao. *International Conference on Interactive Collaborative Learning*, 419–432. https://doi.org/10.1007/978-3-030-40274-7_42 [GS Search].

Nuanmeesri, S., Poomhiran, L., Chopvitayakun, S., & Kadmateekarun, P. (2022). Improving Dropout Forecasting during the COVID-19 Pandemic through Feature Selection and Multilayer Perceptron Neural Network. *International Journal of Information and Education Technology*, *12*(9). https://doi.org/10.18178/ijiet.2022.12.9.1693 [GS Search].

Panagiotakopoulos, T., Kotsiantis, S., Kostopoulos, G., Iatrellis, O., & Kameas, A. (2021). Early Dropout Prediction in MOOCs through Supervised Learning and Hyperparameter Optimization. *Electronics*, *10*(14), 1701. https://doi.org/10.3390/electronics10141701 [GS Search].

Pereira, R. T., & Zambrano, J. C. (2017). Application of decision trees for detection of student dropout profiles. *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, 528–531. https://doi.org/10.1109/icmla.2017.0-107 [GS Search].

Perez, B., Castellanos, C., & Correal, D. (2018). Applying data mining techniques to predict student dropout: a case study. *2018 IEEE 1st colombian conference on applications in computational intelligence (colcaci)*, 1–6. https://doi.org/10.1109/colcaci.2018.8484847 [GS Search].

Pérez, A., Grandón, E. E., Caniupán, M., & Vargas, G. (2018). Comparative analysis of prediction techniques to determine student dropout: Logistic regression vs decision trees. *2018 37th International Conference of the Chilean Computer Science Society (SCCC)*, 1–8. https://doi.org/10.1109/sccc.2018.8705262 [GS Search].

Pérez-Gutiérrez, B. R. (2020). Comparación de técnicas de mineria de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico. *Revista UIS Ingenierías*, *19*(1), 193–204. https://doi.org/10.18273/revuin.v19n1-2020018 [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 Soft-*

ware Engineering (EASE) 12, 1–10. https://doi.org/10.14236/ewic/ease2008.8 [GS Search].

Pradeep, A., Das, S., & Kizhekkethottam, J. J. (2015). Students dropout factor prediction using EDM techniques. *2015 International Conference on Soft-Computing and Networks Security (ICSNS)*, 1–7. https://doi.org/10.1109/icsns.2015.7292372 [GS Search].

Quishpe-Morales, S., Pillo-Guanoluisa, D., Revelo-Portilla, I., & Guerra-Torrealba, L. (2020). Modelo de predicción de la deserción universitaria mediante analítica de datos: Estrategia para la sustentabilidad. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E35), 38–47. [GS Search].

Radovanović, S., Delibašić, B., & Suknović, M. (2020). Predicting dropout in online learning environments. *Computer Science and Information Systems*, (00), 53–53. https://doi.org/10.2298/csis200920053r [GS Search].

Revathy, M., Kamalakkannan, S., & Kavitha, P. (2022). Machine Learning based Prediction of Dropout Students from the Education University using SMOTE. *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 1750–1758. https://doi.org/10.1109/icssit53264.2022.9716450 [GS Search].

Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLoS one*, *12*, e0171207. https://doi.org/10.1371/journal.pone.0171207 [GS Search].

Şahin, M. (2021). A comparative analysis of dropout prediction in massive open online courses. *Arabian Journal for Science and Engineering*, *46*(2), 1845–1861. https://doi.org/10.1007/s13369-020-05127-9 [GS Search].

SALLAN, G., & BEHAL, S. (s.d.). PREDICTION OF STUDENT DROPOUT USING ENHANCED MACHINE LEARNING ALGORITHM. https://doi.org/10.37418/amsj.9.6.61 [GS Search].

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

Santos Baggi, C. A. D., & Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, *16*, 355–374. https://doi.org/10.1590/s1414-40772011000200007 [GS Search].

Selvan, M., Navadurga, N., & Prasanna, N. (2019). An efficient model for predicting student dropout using data mining and machine learning techniques. *8*, 750–752. https://doi.org/10.35940/ijitee.i1155.0789s219 [GS Search].

Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology*, *9*(4), 1–5. https://doi.org/10.17485/ijst/2016/v9i4/87032 [GS Search].

Sorensen, L. C. (2019). "Big data" in educational administration: An application for predicting school dropout risk. *Educational Administration Quarterly*, *55*(3), 404–446. https://doi.org/10.1177/0013161x18799439 [GS Search].

Su, M., Olson, L. A., Jarratt, D. C., Varma, S., Konstan, J. A., Keller, R. J., & Chen, B. (2022). Re-envisioning a K-12 Early Warning System with School Climate Factors. *Proceedings*

of the Ninth ACM Conference on Learning Scale, 405–408. https://doi.org/10.1145/3491140.3528670 [GS Search].

Tamada, M. M., de Magalhães Netto, J. F., & de Lima, D. P. R. (2019). Predicting and reducing dropout in virtual learning using machine learning techniques: A systematic review. *2019 IEEE Frontiers in Education Conference (FIE)*, 1–9. https://doi.org/10.1109/fie43999.2019.9028545 [GS Search].

Tan, M., & Shao, P. (2015). Prediction of student dropout in e-Learning program through the use of machine learning method. *International journal of emerging technologies in learning*, *10*(1). https://doi.org/10.3991/ijet.v10i1.4189 [GS Search].

Tenpipat, W., & Akkarajitsakul, K. (2020). Student Dropout Prediction: A KMUTT Case Study. *2020 1st International Conference on Big Data Analytics and Practices (IBDAP)*, 1–5. https://doi.org/10.1109/ibdap50342.2020.9245457 [GS Search].

Vega, H., Sanez, E., De La Cruz, P., Moquillaza, S., & Pretell, J. (2022). Intelligent System to Predict University Students Dropout. *International Journal of Online & Biomedical Engineering*, *18*(7). https://doi.org/10.3991/ijoe.v18i07.30195 [GS Search].

Viloria, A., & Lezama, O. B. P. (2019). Mixture structural equation models for classifying university student dropout in latin america. *Procedia Computer Science*, *160*, 629–634. https://doi.org/10.1016/j.procs.2019.11.036 [GS Search].

Wu, N., Zhang, L., Gao, Y., Zhang, M., Sun, X., & Feng, J. (2019). CLMS-Net: dropout prediction in MOOCs with deep learning. *Proceedings of the ACM Turing Celebration Conference-China*, 1–6. https://doi.org/10.1145/3321408.3322848 [GS Search].

Xing, W., & Du, D. (2019). Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research*, *57*(3), 547–570. https://doi.org/10.1177/0735633118757015 [GS Search].

Yaacob, W. W., Sobri, N. M., Nasir, S. M., Norshahidi, N., & Husin, W. W. (2020). Predicting student drop-out in higher institution using data mining techniques. *Journal of Physics: Conference Series*, *1496*(1), 012005. https://doi.org/10.1088/1742-6596/1496/1/012005 [GS Search].