Mineração de Dados Educacionais para Predição de Desempenho Profissional: uma revisão sistemática

Tarcísio Lopes de Almeida Sousa[1], Moacyr Cunha Filho[1]

[1]Universidade Federal Rural de Pernambuco

tarcisiocomp@gmail.com, moacyr.cunhafo@ufrpe.br

Abstract. *The component considered one of the most crucial for academic success is professional achievement, and correct career guidance enhances students' performance and increases their level of motivation. In this study, a systematic review was conducted to assess the state of the art regarding the possibility of automating a system for predicting students' professional performance. Based on a systematic literature review model encompassing planning, implementation, and results phases, the article search process was carried out using major scientific databases. After establishing inclusion, exclusion, and quality criteria, 455 articles were identified, and among these, 5 were selected based on the guiding questions. The studies selected in this systematic review highlight variables necessary for predicting professional performance, such as grades, absences, parents' education, school, private lessons; and the Weka framework emerged as the predominant tool in these studies.*

Resumo. *O componente que é considerado um dos mais importantes para o sucesso acadêmico é o sucesso profissional e a orientação profissional correta impulsiona o desempenho dos alunos e aumenta seu nível de motivação. Neste trabalho, realizou-se uma revisão sistemática visando verificar o estado da arte sobre a possibilidade de automação de um sistema de predição de desempenho profissional de estudantes. Baseado em um modelo de revisão sistemática da literatura abrangendo fases de planejamento, implementação e resultados, o processo de busca de artigos foi conduzido utilizando as principais bases de dados científicas. Após estabelecer critérios de inclusão, exclusão e qualidade, chegou-se a 455 artigos e destes 5 foram selecionados de acordo com as perguntas norteadoras. Os estudos selecionados nesta revisão sistemática evidenciam variáveis necessárias para a realização de predições de desempenho profissional, como: Notas, faltas, educação dos pais, escola, aulas particulares; e o framework Weka como ferramenta predominante nos estudos.*

1. Introdução

A educação passou por transformações significativas nos últimos anos devido aos avanços tecnológicos. A tecnologia desempenhou um papel fundamental na modernização do setor educacional, não apenas por meio do desenvolvimento de

softwares educacionais, mas também na gestão eficiente de registros acadêmicos e na administração de instituições de ensino virtualmente. Além disso, a internet proporcionou aos estudantes novas oportunidades de aprendizado, tornando o acesso ao conhecimento mais amplo e flexível. Essa combinação de elementos tecnológicos revoluciona a maneira como a educação é concebida, entregue e administrada [SOUZA e SANTOS 2021].

Com o crescimento do volume e armazenamento de dados educacionais, juntamente com as informações dos alunos, torna-se imperativo utilizar recursos computacionais para otimizar a eficiência e qualidade na análise desses dados. Nesse contexto, a Mineração de Dados Educacionais (MDE) desempenha um papel fundamental, uma vez que essa metodologia pode ser aplicada em várias áreas e para diversas finalidades. Ela oferece uma abordagem poderosa para extrair insights valiosos dos dados educacionais, tornando os processos de ensino e aprendizado mais eficazes. A MDE permite a identificação de tendências, padrões e áreas de melhoria, auxiliando educadores e instituições de ensino na tomada de decisões embasadas em dados [VIVIAN et al. 2022].

A MDE envolve a aplicação de técnicas de mineração de dados, e em alguns casos, a aplicação de modelos probabilísticos em dados na área da educação. Para empregar a MDE de forma eficaz, é crucial compreender quais ferramentas são apropriadas e quais técnicas melhor se adequam ao contexto de estudo. Uma maneira de adquirir esse conhecimento é realizar uma Revisão Sistemática de Literatura (RSL), que consiste em uma análise dos materiais encontrados na literatura, para identificar e expor os métodos que podem ser replicados, proporcionando assim uma visão abrangente das ferramentas, técnicas e fatores explorados no contexto em estudo. Isso facilita a tomada de decisões e a implementação eficaz da MDE na educação [MARQUES et al. 2022].

Uma revisão sistemática da literatura (muitas vezes referida como revisão sistemática) é um meio de identificar, avaliar e interpretar todas as pesquisas disponíveis relevantes para uma determinada questão de pesquisa, ou área temática, ou fenômeno de interesse. Os estudos individuais que contribuem para uma revisão sistemática são chamados de estudos primários; uma revisão sistemática é uma forma de estudo secundário [KEELE et al. 2007].

O método MDE tem como objetivo primordial na área acadêmica a predição do desempenho dos alunos ao longo de seu curso. Isso permite que as instituições de ensino realizem intervenções quando necessário, visando aprimorar o processo de aprendizagem dos alunos. Essa investigação é de grande importância não apenas para a vida acadêmica e profissional dos estudantes, mas também para a própria universidade. A qualidade do trabalho da instituição é diretamente refletida no desempenho de seus alunos, o que torna essas análises essenciais para a melhoria constante do ensino e para a avaliação da qualidade educacional [SOUZA e SANTOS 2021].

O sucesso dos estudantes possui um papel fundamental nas instituições de ensino, pois frequentemente serve como um indicador do desempenho da própria

instituição. A identificação precoce de alunos em situação de risco, juntamente com a implementação de medidas preventivas, pode ter um impacto significativo em seu êxito acadêmico. Técnicas de aprendizado de máquina têm sido amplamente adotadas para fins de predição.

No entanto, a aplicação eficaz e eficiente de métodos de mineração de dados envolve muitas decisões, que vão desde como definir o sucesso do estudante, passando por quais atributos do aluno devem ser enfatizados, qual o método de aprendizado de máquina é mais apropriado para o problema em questão, se a solução é passível de automação e quais as tecnologias podem ser utilizadas em sua implementação.

De acordo com a definição do problema apresentado, a motivação para esta pesquisa reside na necessidade de investigar quais são as técnicas de MDE, com um foco especial nas determinísticas, que tenham demonstrado melhor acurácia nos resultados e se é passível de automação um sistema de avaliação e predição de desempenho de estudantes com previsibilidade de desempenho profissional.

O objetivo deste estudo é dar uma contribuição no *review* da área de avaliação de desempenho acadêmico com o auxílio de ferramentas sintáticas, compreendendo o estado da arte e identificando as técnicas de MDE com melhor acurácia.

O artigo está estruturado em cinco seções: a primeira comporta a introdução ao tema, expondo a problemática e uma proposta para esta ser minimizada; a segunda apresenta uma contextualização sobre predição de desempenho acadêmico; a terceira aborda a metodologia utilizada para a seleção dos artigos para responder às questões de pesquisa; a quarta expõe os resultados; e a quinta aponta as considerações finais e os trabalhos futuros.

2. Referencial Teórico

2.1 Sucesso Acadêmico

O sucesso do aluno é caracterizado pelo seu desempenho acadêmico, participação ativa em atividades de valor educacional, satisfação pessoal, aquisição de conhecimentos desejados, desenvolvimento de habilidades e competências, continuidade nos estudos, realização de metas educacionais e desempenho após a conclusão do curso [KUH et al. 2006].

O sucesso dos estudantes é um elemento de extrema importância nas instituições de ensino superior, uma vez que é considerado um critério fundamental para a avaliação da qualidade dessas instituições.

Embora existam diversas definições na literatura, seis componentes são considerados mais importantes para o sucesso acadêmico: Desempenho acadêmico, satisfação, aquisição de aptidões e competências, persistência, cumprimento de objetivos de aprendizagem e sucesso profissional [YORK et al. 2015].

A orientação profissional correta impulsiona o desempenho dos alunos e aumenta seu nível de motivação. Os sistemas baseados em orientação profissional são essenciais para avaliar as habilidades dos alunos e, consequentemente, os recrutadores atribuem o cargo certo a eles [VIHERÄKOSKI 2020]. A orientação profissional é necessária para solucionar os problemas que surgem na mente dos pais em relação aos seus filhos, desempenhando um papel importante ao guiar os pais na escolha do curso certo para seus filhos [YANG et al. 2020].

2.2 Algoritmos Classificadores em MDE

A classificação é um dos métodos mais comumente usados para categorizar grandes conjuntos de dados. Essa abordagem envolve adaptação de dados e algoritmos de aprendizado supervisionado. O objetivo é estabelecer a relação entre a variável de interesse de forma qualitativa e outras variáveis observadas [HUSSAIN et al. 2018].

A análise de classificação pode ser usada para consultar, determinar ou prever o comportamento por meio de um algoritmo. Funciona gerando um conjunto de dados de treinamento contendo várias características essenciais e resultados potenciais. A tarefa do algoritmo de classificação é descobrir como o conjunto de características chega ao seu final, o que determina e atribui categorias a um grupo de dados a ser avaliado com mais precisão. Essa abordagem utiliza métodos matemáticos como árvores de decisão, J48, Floresta Aleatória, SVM, programação linear, redes neurais e estatísticas [JASSIM 2021]

A classificação é uma tecnologia de mineração de dados que atribui classes a diferentes conjuntos de dados para apoiar predições e análises. A classificação é uma das maneiras de processar eficientemente grandes conjuntos de dados. O objetivo da classificação é prever corretamente a população alvo nos dados para cada evento. A técnica de classificação envolve a aprendizagem e a classificação dos dados. É o método de mineração de dados mais amplamente utilizado e usado para desenvolver classes e atribuir conjuntos de dados a cada uma delas. Em resumo, se a variável alvo for discreta, então isso é um problema de classificação, e se a variável alvo for contínua, é uma tarefa de regressão [JASSIM 2021].

2.2.1 Árvore de Decisão J48

O algoritmo J48 é usado para classificar várias aplicações e obter resultados de classificação corretos. O J48 é uma extensão do ID3 [LEE et al. 2020]. Recursos adicionais do J48 são suas características ausentes e a escolha da poda da árvore; é um algoritmo para criar uma árvore de decisão gerada pelo C4.5 (uma extensão do ID3) [KAUR et al. 2020]. Também é conhecido como um classificador estatístico. Para classificar a árvore de decisão, é necessário um banco de dados.

2.2.2 Rede Neural

Uma rede neural é uma coleção de algoritmos que têm como objetivo identificar relacionamentos fundamentais em vários dados por meio de uma estrutura que imita o funcionamento do cérebro humano [QIU et al. 2020]. Nesse contexto, as redes neurais se referem a sistemas neurais artificiais. Redes neurais podem ser adaptadas a entradas em evolução. A estrutura, assim, alcança o melhor resultado sem a necessidade de redesenhar os parâmetros de saída [HU e RANGWALA 2020]. A ideia de redes neurais, que tem suas raízes em inteligência artificial, tornou-se muito comum no desenvolvimento de sistemas de negociação, como bolsa de valores.

2.2.3 Algoritmo ID3

O algoritmo ID3 foi desenvolvido por Ross Quinlan [AHMED et al. 2020] e é usado para gerar uma árvore de decisão a partir de um conjunto de dados. O ID3 é uma introdução ao C4.5 e é comumente usado para processar aprendizado de máquina e linguagens naturais. A árvore resultante é usada para identificar possíveis amostras.

2.2.4 Floresta Aleatória

A Floresta Aleatória é uma teoria de aprendizado estatístico e o método é usado com múltiplas árvores de decisão para fazer predições e usar a votação para obter os resultados da predição [HUNG et al. 2020]. O tamanho do subconjunto é sempre o mesmo que o tamanho da amostra de entrada original, mas a substituição é usada para retirar exemplos.

2.2.5 Algoritmo C4.5

O algoritmo C4.5 é usado para mineração de dados como classificação de árvore de decisão para tomar decisões com base em um conjunto de dados fornecido (predições univariadas ou multivariadas) [MAHAJAN e SAINI 2020]. O termo "classificador" aqui se refere a um método de mineração de dados que leva os dados que precisamos para identificar a nova categoria de dados e prever essa categoria.

2.2.6 Algoritmo de Naïve Bayes

Um grupo de algoritmos de classificação baseados no teorema de Bayes são os classificadores Naïve Bayes. Não é um único algoritmo; é um grupo de algoritmos. Todos eles compartilham um princípio comum: cada par de características é tratado como independente. O Naïve Bayes é fornecido com um conjunto de dados de treinamento padronizado [MAHESHWARI et al. 2020]. Portanto, é tratado como um algoritmo de aprendizado supervisionado. Ele ajuda a prever a probabilidade de ocorrência com base nas circunstâncias que são conhecidas.

2.2.7 Máquina de Vetores de Suporte

A Máquina de Vetores de Suporte (SVM) é um algoritmo de aprendizado de máquina supervisionado para classificação, regressão e detecção de variações. O algoritmo SVM tem como objetivo encontrar uma separação entre dois grupos de objetos, presumindo que quanto maior a separação, mais robusta a classificação [WAHEED et al. 2020]. A Máquina de Vetores de Suporte (SVM) opera de forma semelhante ao algoritmo C4.5, exceto que a SVM não utiliza árvores de decisão. Para classificar o conhecimento em dois grupos, a SVM aprende conjuntos de dados e determina o hiperlink.

2.2.8 Algoritmo dos K Vizinhos Mais Próximos

Os K Vizinhos Mais Próximos (KNN) é um algoritmo simples que armazena todos os estados existentes e classifica novos casos com base em uma escala semelhante. No início da década de 1970, o KNN já havia sido usado como um método não paramétrico em estimação estatística e reconhecimento de padrões [YULIANTO et al. 2020]. O KNN é um algoritmo de aprendizado supervisionado e classificação de padrões que ajuda a determinar a classe à qual a nova entrada pertence (valor de teste) ao selecionar os K Vizinhos Mais Próximos e calcular a distância entre eles.

2.3 Algoritmos Determinísticos

Algoritmos determinísticos são procedimentos ou conjuntos de regras que, quando seguidos, produzem um resultado previsível e consistente. Isso significa que, dadas as mesmas entradas iniciais e as mesmas condições iniciais, um algoritmo determinístico sempre produzirá o mesmo resultado [ARAUJO et al. 2001].

Os algoritmos determinísticos são usados em uma ampla variedade de aplicações, desde cálculos matemáticos simples até processos complexos em ciência da computação, engenharia, física e muitos outros campos. Eles são contrastados com algoritmos não determinísticos, nos quais as saídas podem variar mesmo com as mesmas entradas devido a fatores aleatórios ou probabilísticos. A previsibilidade e consistência dos algoritmos determinísticos são muitas vezes uma vantagem em situações em que a precisão e a repetitividade são essenciais, como em sistemas de controle, criptografia e em muitos aspectos da computação [REZENDE 2003].

Algoritmos de MDE podem ser tanto determinísticos quanto não determinísticos, dependendo da abordagem e da técnica utilizada. Alguns algoritmos de MDE, como árvores de decisão, regressão logística e muitos algoritmos estatísticos, são determinísticos. Eles produzem resultados previsíveis e consistentes com base nas entradas e dados de treinamento fornecidos [ARAÚJO et al. 2001].

No entanto, também existem abordagens de MDE que podem incorporar elementos não determinísticos. Por exemplo, o uso de técnicas de aprendizado de máquina, como redes neurais, pode envolver aspectos estocásticos durante o treinamento, tornando o processo de aprendizado não estritamente determinístico. Além

disso, a aleatoriedade em alguns modelos pode tornar o processo de MDE menos determinístico [ARAÚJO et al. 2001].

3. Metodologia

Os trabalhos apresentados nesta seção são o resultado de uma revisão sistemática da literatura de publicações relevantes dos últimos 5 anos (2018-2023), com o propósito de identificar pesquisas semelhantes a esta e explorar o que está sendo pesquisado sobre predição de desempenho acadêmico no intuito de responder às questões de pesquisa, relacionadas na Tabela 1.

Tabela 1 – Questões norteadoras

Questões	Descrição
Q1:	É passível de automação um sistema de avaliação e predição de desempenho de estudantes deterministicamente com previsibilidade de desempenho profissional?
Q2:	Quais tecnologias atuais poderiam ser usadas em um protótipo nesse sentido?
Q3:	Ranks (como TOEFL, GRE, ENEM, PISA) são eficazes na elaboração de diretrizes e guias para as instituições de ensino terem melhores egressos?

Fonte: Elaborado pelo autor.

3.1 Procedimentos Metodológicos

A revisão sistemática da literatura (RSL) adere a uma abordagem metodológica claramente delineada, demonstrando contribuições para um tema ou fenômeno de maneira imparcial e passível de repetição, enquanto examina perguntas específicas de pesquisa [KITCHENHAM 2012].

A RSL possibilita reunir e avaliar a comprovação empírica de uma área de estudo específica por meio da análise de resultados de pesquisas pertinentes ao tema de interesse, permitindo, dessa forma, tirar conclusões acerca das perguntas de pesquisa previamente definidas. Para atingir esse propósito, empregam-se mecanismos de busca com o intuito de elaborar uma revisão abrangente da pesquisa relacionada à MDE [NASCIMENTO et al. 2023].

Neste trabalho, a revisão sistemática da literatura será desenvolvida conforme o modelo proposto por Kitchenham (2012), abrangendo as seguintes fases: planejamento, implementação e resultados.

A etapa inicial compreendeu a elaboração do plano para esta pesquisa, estabelecendo o objetivo principal desta revisão sistemática: compreender quais mídias

são mais enfatizadas pela MDE, bem como os métodos utilizados para coletar e analisar esses dados, e de que forma a mineração de dados contribui para aprimorar a gestão educacional, possibilitando uma tomada de decisão mais eficaz.

Na etapa subsequente, a segunda fase, foram formuladas três questões-chave destinadas a orientar a pesquisa e a interpretação dos resultados. Isso foi realizado visando fornecer uma contextualização qualitativa abrangente do tema e, ao mesmo tempo, permitir a análise quantitativa de alguns dos dados da pesquisa relacionados a essa temática.

3.2 Estratégias de Busca

A estratégia de busca desta revisão foi dividida em duas etapas distintas. A primeira etapa envolveu a definição da string de busca utilizada nas bases de dados, conforme Tabela 2, com palavras-chave relevantes para o estudo. A segunda etapa, por sua vez, consistiu na pesquisa em bibliotecas digitais, que servirão como fontes de estudo do tema proposto [KITCHENHAM 2012].

Tabela 2 – String de busca

Operadores	OR	AND
String de busca	"academic success" OR "academic achievement" OR "student success" OR "predicting professional performance" OR "career success" OR "Employability" AND "Educational data" AND "syntactic tools" OR "classification techniques" OR "classifier algorithms" AND "automation" OR "prototype" OR "web application" OR "mobile app"	

Fonte: Elaborado pelo autor.

3.3 Bases de Dados e Processo de Extração

O processo de busca de artigos foi conduzido utilizando as principais bases de dados eletrônicas e científicas que fossem relevantes para abordar as questões Q1, Q2 e Q3. Um trabalho acadêmico relevante é um documento científico que atende a critérios como, conteúdo acadêmico de revistas científicas, teses, dissertações, conferências, relatórios técnicos e outros tipos de publicações.

São trabalhos que possuem qualidade e credibilidade, pois são revisados por pares (peerreviewed) em caso de artigos de revistas, ou avaliados e aprovados por um comitê acadêmico em caso de teses e dissertações. São trabalhos disponíveis online e

acessíveis ao público, incluindo publicações em repositórios institucionais, sites de universidades, periódicos online e outras fontes de acesso público. A relevância de um trabalho acadêmico também pode ser influenciada pelo número de citações que ele recebe de outros trabalhos acadêmicos. Trabalhos amplamente citados tendem a ser considerados mais relevantes.

As bases de dados utilizadas neste trabalho foram do Portal de Periódicos da CAPES e Google Acadêmico (Scholar Google), conforme Tabela 3. Estas bases indexam conteúdo acadêmico de várias fontes na web, incluindo artigos, teses, dissertações, relatórios técnicos, repositórios institucionais e outras fontes que estão disponíveis on-line. Muitos dos artigos que aparecem nos resultados de pesquisa dessas bases de dados podem estar disponíveis em bases de dados como Scopus, Web of Science, IEEE Xplore, e outras.

Tabela 3 – Bases de dados

Base de Dados	Endereço Eletrônico	Tipo de Busca
Periódicos CAPES	www.gov.br/capes/pt-br	Automática
Scholar Google	scholar.google.com.br/	Automática

Fonte: Elaborado pelo autor.

A Revisão Sistemática da Literatura empregou uma abordagem de busca automática, considerando critérios de inclusão, exclusão e qualidade dos artigos encontrados. Os artigos científicos considerados foram limitados a uma janela temporal de até 5 anos de publicação.

3.4 Critérios de Inclusão, Exclusão e Qualidade

Um passo crucial na condução de uma RSL é estabelecer critérios que ajudem na seleção adequada do material bibliográfico. Esses critérios geralmente são organizados em três categorias: Critérios de Inclusão, Critérios de Exclusão e Critérios de Qualidade. Os artigos foram obtidos por meio da string de busca aplicada nas bases de dados, utilizando as questões formuladas para a pesquisa. Em seguida, eles foram selecionados conforme os critérios estabelecidos, descritos na Tabela 4.

Tabela 4 – Critérios

Critérios de Inclusão (I)
(I.1) Artigos de periódicos, dissertações e teses de repositórios institucionais;
(I.2) Publicações que estejam no tema proposto, independente de idioma;
(I.3) Documentos disponíveis publicamente para

leitura (*Open Source*);
(I.4) Publicações correlacionadas à ciência da computação ou aplicados na área de educação;
(I.5) Publicações dos últimos 5 anos.
Critérios de Exclusão (E)
(E.1) Publicações que fujam do tema abordado e que possuam mais de 5 anos de publicação;
(E.2) Trabalhos que não sejam artigos científicos, dissertações, teses ou qualquer tipo de trabalho acadêmico;
(E.3) Documentos duplicados;
(E.4) Publicações que tiveram menos de 10 citações.
Critérios de Qualidade (Q)
(Q.1) Definição clara dos objetivos da pesquisa;
(Q.2) Contexto adequado onde foi realizada a pesquisa;
(Q.3) Estratégia de extração de dados adequada aos objetivos da pesquisa;
(Q.4) Descrição dos resultados encontrados;
(Q.5) Evidência da contribuição para o aluno ou professor.

Fonte: Adaptado de Kitchenham (2012).

4. Análises e Resultados

Este trabalho constitui uma pesquisa bibliográfica sobre MDE, empregando uma abordagem de revisão sistemática da literatura para uma análise do estado da arte. No total, foram utilizados 5 artigos para a realização das avaliações necessárias. A Figura 1 descreve as etapas do processo de pesquisa e seleção dos artigos de forma detalhada.

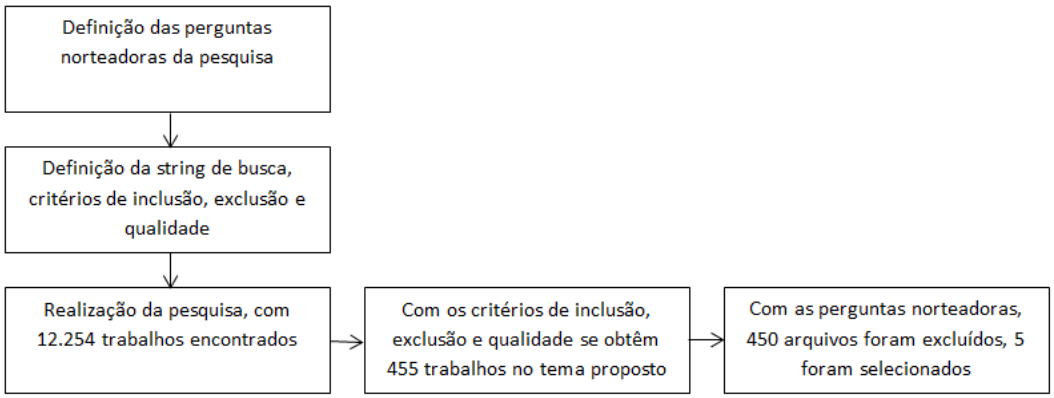

Figura 1 – Processo de pesquisa e seleção de artigos.

Fonte: Adaptado de Kitchenham (2012).

Pode-se observar que na Tabela 5 é apresentado o total de artigos encontrados através dos critérios de inclusão, exclusão e qualidade, por cada site, tendo sido encontrados 6.134 no Periódicos CAPES e 6.120 no Scholar Google.

Tabela 5 – Resultado das pesquisas realizadas

Base	Fase 1	Fase 2	Fase 3
Periódicos CAPES	6.134	234	1
Scholar Google	6.120	221	4
Total	**12.254**	**455**	**5**

Fonte: Elaborado pelo autor.

O objetivo principal dos trabalhos de Taeza-Cruz e Capili-Kummer (2023), Guleria e Sood (2023), Cruz e Encarnacion (2021) e Repaso e Capariño (2020) foi analisar o desempenho dos alunos por meio da aplicação de técnicas de mineração de dados, seja utilizando ferramentas sintáticas ou desenvolvendo software com a mesma finalidade. Já o trabalho de Masci et al. (2018), o objetivo principal foi descobrir quais características dos estudantes estão associadas às pontuações dos testes do PISA. A Tabela 6 apresenta os artigos selecionados.

Tabela 6 – Artigos selecionados

Autores	Ano de Publicação	Tema
TAEZA-CRUZ, Maria Elisa Linda; CAPILI-KUMMER, Marifel Grace	2023	Decision Support System to Enhance Students' Employability using Data Mining Techniques for Higher Education Institutions
GULERIA, Pratiyush; SOOD, Manu	2023	Explainable AI and machine learning: performance evaluation and explainability of classifers on educational data mining inspired career counseling
CRUZ, Maria Elisa Linda Taeza; ENCARNACION, Riah Elcullada	2021	Analysis and Prediction of Students' Academic Performance and Employability Using Data Mining Techniques: A Research Travelogue
REPASO, Jennifer Anne A.; CAPARIÑO, Elenita T.	2020	Analyzing and Predicting Career Specialization using Classification Techniques
MASCI et al.	2018	Student and school performance across countries: A machine learning approach

Fonte: Elaborado pelo autor.

Os artigos escolhidos passaram por uma leitura minuciosa e, em seguida, os dados relevantes foram extraídos. A discussão dos resultados foi conduzida por meio da organização de perguntas direcionadas à pesquisa. Logo após, os dados foram sintetizados e analisados para responder às questões de pesquisa.

4.1 Questões de Pesquisa

Q1: É passível de automação um sistema de avaliação e predição de desempenho de estudantes deterministicamente com previsibilidade de desempenho profissional?

Taeza-Cruz e Capili-Kummer (2023) desenvolveram um sistema para o ensino superior com foco na empregabilidade dos estudantes e na oferta de uma educação de qualidade que produza graduados qualificados para oportunidades de emprego. Este sistema foi considerado útil e necessário, dado o compromisso da Universidade de Nizwa com a empregabilidade dos estudantes. O uso do sistema desenvolvido simplificaria a identificação de estudantes em risco por parte dos orientadores acadêmicos e permitiria a eles fornecer intervenções rápidas e apropriadas para ajudar seus orientados a melhorar seu desempenho acadêmico. Este estudo mostrou que modelos determinísticos, como o modelo Árvore de Decisão J48, é uma das técnicas mais apropriadas de mineração de dados para prever a empregabilidade de estudantes, onde este obteve uma precisão de 96,6135%. Os resultados sugerem que a situação acadêmica, a nota do estágio e a carga horária são as variáveis mais significativas da empregabilidade dos alunos, conforme a Tabela 7.

Tabela 7 – Variáveis significativas da empregabilidade

Variável	Descrição	Valores
Situação Acadêmica	Classificação do CGPA para o grau de bacharel.	Excelente com Honras, Excelente, Muito Bom, Bom, Satisfatório e Em Período de Avaliação.
Nota do Estágio	A nota final recebida pelo estudante durante o estágio.	A (4.0); A- (3.7); B+ (3.3); B (3.0); B- (2.7); C+ (2.3); C (2.0); C- (1.7); D+ (1.3); D (1.0); F (0).
Carga horária	Horas de crédito completadas pelo estudante.	9 – 160 horas.

Fonte: Adaptada de Taeza-Cruz e Capili-Kummer, (2023).

Guleria e Sood (2023) desenvolveram um framework onde este utiliza técnicas de aprendizado de máquina para realizar predições de fatores acadêmicos e de oportunidades de emprego como parte do modelo de orientação profissional proposto.

Os modelos de aprendizado de máquina precisam ser treinados com os conjuntos de dados educacionais coletados em várias instituições educacionais, com um maior número de atributos para avaliação dos estudantes, levando em consideração fatores como oportunidades de emprego com o objetivo de aprimorar a empregabilidade e o desenvolvimento de habilidades. No entanto, o escore de Recall e F-Measure alcançado pelo Naive Bayes para realizar predições é de 91,2% e 90,7%, que é o melhor em comparação com os escores dos modelos de Regressão Logística, Árvore de Decisão, SVM, KNN e Ensemble considerados no estudo. Algumas variáveis preditoras foram utilizadas, como notas do ensino médio, notas do ensino superior, experiência de trabalho, nota em teste eletrônico de empregabilidade e notas de curso de especialização (se houver), conforme Tabela 8.

Tabela 8 – Variáveis preditoras

Variável	Tipo de Dados	Exemplo
Notas (ensino médio, ensino superior, teste de empregabilidade, especialização (se houver))	Float	78.33, 73.6, 59.43
Experiência de Trabalho	Boolean	Yes/No

Fonte: Adaptada de Guleria e Sood (2023).

Cruz e Encarnacion (2021) realizaram uma revisão da literatura, onde vários estudos mostraram resultados notáveis na predição da empregabilidade dos estudantes. As técnicas de classificação foram predominantemente usadas para tarefas de predição e análise, especificamente o uso dos algoritmos J48 (C4.5), Naïve Bayes e algoritmos de Árvore de Decisão CHAID, devido aos seus altos resultados de precisão. O estudo revelou algumas variáveis comuns que foram usadas para classificação, como média acumulada de notas (CGPA), gênero, idade, grupo de estudo e aulas particulares, conforme Tabela 9. Além disso, a revisão sistemática da literatura e estudos relacionados demonstrou que o desempenho acadêmico influencia significativamente a empregabilidade dos estudantes.

Tabela 9 – Variáveis relacionadas ao estudante

Variável	Descrição	Valores
Gênero	Gênero do estudante	Masculino, Feminino.
Comunidade	Status social, conforme classes sociais da Índia.	SC (Castas Programadas), ST (Tribos Programadas), MBC (Classes Mais Atrasadas), BC (Classes Atrasadas), OC (Outras).
Educação dos pais	Indica se os pais	Ambos educados, mãe

	foram alfabetizados.	educada, pai educado, ambos não educados.
Localidade de Moradia	Localidade onde mora o estudante.	Empresarial, Município, Zona Rural.
Transporte	Modo de transporte para escola.	A pé, bicicleta, ônibus coletivo, ônibus escolar, automóvel próprio.
Tipo de Escola	Determina o tipo de escola que o aluno estudou o nível superior.	Coeducação, Meninos, Meninas.
Aulas Particulares	A maioria dos pais envia seus filhos para aulas particulares.	Sim, Não.
Grupo de Estudo	Seis tipos de grupos de estudo, baseados em disciplinas básicas, são oferecidos no nível superior.	Primeiro, segundo, terceiro, quarto, quinto, sexto.
Localidade da Escola	Localidade da escola do estudante.	Empresarial, Município, Zona Rural.
Acesso à celular	Indica se o estudante possui celular.	Sim, Não.
Acesso a esportes	Indica se o aluno possui acesso a esportes e lazer.	Sim, Não.
Acesso a computador	Indica se o aluno possui acesso a computador.	Sim, Não.
Acesso à internet	Indica se o aluno possui acesso à internet.	Sim, Não.
Profissão do pai	Indica a ocupação do pai do estudante.	Trabalhador braçal, Agricultor, Tecelão, Emprego no setor privado, Emprego no setor público, Empresário, Não aplicável.
Profissão da mãe	Indica a ocupação da mãe do estudante.	Dona de Casa, Trabalhadora braçal, Agricultora, Tecelã, Emprego no setor privado, Emprego no setor público, Não aplicável.

Salário dos pais	Indica a soma dos salários dos pais.	0..0.9K, 1K..2.9K, 3K...4.9K, 5K..9K, 10K..20K, acima de 20K, Não aplicável.
Média acumulada de notas (CGPA)	Notas obtidas no nível superior. É dividido em sete classes.	O – 90% a 100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%, F - < 40%.

Fonte: Adaptada de Cruz e Encarnacion (2021).

Repaso e Capariño (2020) revelaram que algoritmos de classificação, como Naïve Bayes, podem ser apropriados para prever a especialização de carreira de um graduado. Cada disciplina cursada habilita o aluno a uma competência técnica e também a uma habilidade interpessoal. O estudo apresenta algumas habilidades interpessoais como: resolução de problemas, criatividade, trabalho em equipe, utilização estratégica de tecnologia da informação e planejamento, conforme Tabela 10. Os estudantes do primeiro ano demonstraram um desempenho razoável em sua primeira disciplina de Programação e Redes, tiveram resultado satisfatório em disciplinas, como Análise e Projeto de Sistemas, Projeto de Conclusão de Curso e Estágio. Nesse sentido, o departamento de tecnologia da informação do campus e os professores puderam implementar programas para melhorar o desempenho dos estudantes, especialmente no primeiro ano, especificamente em Programação e em algumas disciplinas do segundo e terceiro ano, como Redes e Engenharia de Software. Além disso, as boas práticas do departamento foram mantidas para sustentar o desempenho dos estudantes, especialmente em disciplinas de Comunicação e Ética.

Tabela 10 – Atributos e habilidades correspondentes

Disciplina/Atributo	Habilidades Interpessoais	Valores
Programação	Resolução de problemas, criatividade.	VG (1.0, 1.25, 1.5), G (1.75, 2.0, 2.25), F (2.5, 2.75, 3.0)
Banco de Dados	Resolução de problemas, criatividade.	
Desenvolvimento Web	Resolução de problemas, criatividade.	
Redes de Computadores	Resolução de problemas, criatividade.	
Análise e Projeto de Sistemas	Resolução de problemas, criatividade, trabalho em equipe, utilização estratégica da tecnologia.	
Engenharia de Software	Resolução de problemas, criatividade.	
Sistemas Operacionais	Resolução de problemas, criatividade.	

Comunicação	Habilidade de comunicação.	
Ética Profissional	Trabalho em equipe, ética de trabalho.	
Gerenciamento de Projetos	Resolução de problemas, criatividade, trabalho em equipe.	
Estágio	Trabalho em equipe.	

Fonte: Adaptada de Repaso e Capariño (2020).

Q2: Quais tecnologias atuais poderiam ser usadas em um protótipo nesse sentido?

Taeza-Cruz e Capili-Kummer (2023) empregaram métodos descritivos e de desenvolvimento, incluindo métodos ágeis Scrum e a linguagem de programação Hypertext Preprocessor (PHP) para criar o site.

Guleria e Sood (2023) realizaram a experimentação usando a caixa de ferramentas de aprendizado de máquina do MATLAB, como por exemplo, na avaliação de desempenho dos modelos de aprendizado de máquina caixa branca e caixa preta.

Nos estudos de Taeza-Cruz e Capili-Kummer (2023), Guleria e Sood (2023) e Repaso e Capariño (2020) foi predominante o uso do framework Weka. Cruz e Encarnacion (2021) em sua revisão da literatura também encontraram trabalhos que utilizaram esse software para criar modelos de predição. A Tabela 11 apresenta os artigos selecionados e um resumo das respostas de forma sintética para as questões de pesquisa 1 e 2:

Tabela 11 – Resultados dos artigos selecionados

Estudo	**TAEZA-CRUZ; CAPILI-KUMMER, 2023**	**GULERIA; SOOD, 2023**	**CRUZ; ENCARNACI-ON, 2021**	**REPASO; CAPARIÑO, 2020**
Sistema	Predição de empregabilidade dos estudantes	Aconselhamen-to de carreira de estudantes	Revisão da literatura	Predição de especialização de graduados
Nível de escolaridade	graduação	graduação	graduação	graduação
Origem	Omã	Tailândia	Diversas	Filipinas
Framework	Weka	Weka	Diversos	Weka
Técnica de melhor acurácia	Árvore de Decisão J48	Naive Bayes	Árvore de Decisão J48	Naïve Bayes e Random Forest

Percentual	96,6135%	89,2%	-	65%

Fonte: Elaborado pelo autor.

Q3: Ranks (como TOEFL, GRE, ENEM, PISA) são eficazes na elaboração de diretrizes e guias para as instituições de ensino terem melhores egressos?

Masci et al. (2018) aplicaram novos métodos de análise, retirados da literatura de aprendizado de máquina, para examinar os determinantes das notas dos alunos no PISA e do valor agregado das escolas. Os resultados confirmaram muitos dos relacionamentos que já eram até então conhecidos a partir da análise estatística, mas fornecem uma compreensão nova e enriquecida de como as não-linearidades entre os fatores quanto as interações entre eles determinam o desempenho educacional. Essas percepções derivam do reconhecimento de que o processo educacional é complexo, desconhecido em seus mecanismos específicos e heterogêneo entre os países participantes da pesquisa. Uma parte do estudo mostra que as variáveis relacionadas aos alunos conseguem explicar parte da variabilidade de seu desempenho: Índice socioeconômico, ansiedade, motivação, gênero e educação dos pais são algumas das variáveis mais influentes. Sua associação com as notas do PISA e sua capacidade de explicar a variabilidade no sucesso acadêmico dos alunos diferem substancialmente entre os países pesquisados. A porcentagem da variabilidade do sucesso acadêmico dos alunos explicada no valor agregado das escolas também varia entre os países. Os países nos quais a variância estimada do valor agregado das escolas é alta são caracterizados pela heterogeneidade no nível da escola. Pelo contrário, os países nos quais a variância do valor agregado das escolas é limitada em magnitude oferecem uma experiência mais homogênea entre as escolas. Existem claras implicações de política ao notar, por exemplo, que a relação entre alunos e professores tem uma influência relativamente alta no Canadá, Japão e Espanha, mas não em outros lugares. Em muitos países, as ações que podem melhorar de forma mais eficaz os resultados educacionais não são políticas educacionais em si, mas sim políticas sociais.

5. Considerações Finais

Durante a RSL realizada, ficou evidente que a MDE é uma área vasta, mas que ainda necessita de maior estruturação e estudo, devido à diversidade e ao volume significativo de dados na área educacional, que é bastante ampla. Contudo, através da pesquisa observa-se que a utilização de dados educacionais traz benefícios significativos para a gestão escolar.

Os artigos selecionados através dos critérios de inclusão, exclusão e qualidade demonstraram conter conteúdo teórico relevante, permitindo uma compreensão aprofundada sobre o tema em questão, sua importância e como ele pode impactar a gestão educacional. No entanto, uma questão de pesquisa (Q3) ainda não foi respondida de maneira abrangente com esta revisão, indicando a necessidade de estender a pesquisa para outras bases de dados.

Foi possível compreender que os algoritmos determinísticos, como Árvore de Decisão J48 e Naive Bayes, se destacam na literatura como modelos mais eficazes na predição de desempenho profissional. Observou-se também que é passível de automação sistemas com esse tipo de previsibilidade. Além disso, observou-se o uso de vários atributos/variáveis, como idade, gênero, experiência de estágio, notas, instituição, coeficiente de rendimento, habilidades técnicas, comunicação, trabalho em equipe, dentre outras.

A experiência vivenciada nesse trabalho possui características próprias que não se pode deixar de registrá-las como prática de lições aprendidas e para aproveitamento em um trabalho futuro. Pode-se destacar a RSL, que serviu para reunir, avaliar e sintetizar todas as evidências disponíveis sobre as questões de pesquisa, oferecendo uma visão completa e imparcial do estado do conhecimento atual. Isso é fundamental para a progressão do conhecimento e para a tomada de decisões sobre o tema abordado.

Ao longo do desenvolvimento deste trabalho, puderam ser identificadas algumas possibilidades de melhoria e de continuação a partir de futuras pesquisas, as quais incluem a implementação do sistema de predição de desempenho acadêmico dos estudantes, com a possibilidade de realizar a integração com os modelos de classificação através do framework Weka.

Os resultados gerados por um sistema de predição de desempenho podem proporcionar uma experiência satisfatória para os usuários, permitindo que os estudantes tomem medidas para aprimorar seus resultados, que são exibidos e, posteriormente, trabalhem para melhorar seu desempenho. Além disso, os professores podem desempenhar um papel fundamental ao orientar os alunos adequadamente e inspirá-los a alcançarem melhores resultados ao longo do curso.

Referências

AHMED, Saja Taha et al. Developed third iterative dichotomizer based on feature decisive values for educational data mining. Indonesian Journal of Electrical Engineering and Computer Science, 18(1):209–217, 2020.

ARAUJO, Haroldo Alexandre de et al. Algoritmo simulated annealing: uma nova abordagem. 2001.

CRUZ, Maria Elisa Linda Taeza; ENCARNACION, Riah Elcullada. Analysis and prediction of students' academic performance and employability using data mining techniques: A research travelogue. The Eurasia Proceedings of Science Technology Engineering and Mathematics, v. 16, p. 117-131, 2021.

GULERIA, Pratiyush; SOOD, Manu. Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling. Education and Information Technologies, v. 28, n. 1, p. 1081-1116, 2023.

HU, Qian; RANGWALA, Huzefa. Towards fair educational data mining: A case study on detecting atrisk students. In Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020), pages 431–437, 2020.

HUNG, Hui-Chun et al. Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education. Symmetry, 12(2):213, 2020.

HUSSAIN, Sadiq et al. Educational data mining and analysis of students' academic performance using weka. Indonesian Journal of Electrical Engineering and Computer Science, 9(2):447–459, 2018.

JASSIM, Mustafa Abdalrassual. Analysis of the performance of the main algorithms for educational data mining: a review. In: IOP conference series: materials science and engineering. IOP Publishing, 2021. p. 012084.

KAUR, Amritpreet et al. Classification and prediction based enhanced j48 and reptree algorithms to predict corona virus pandemic. no. July, 2020.

KEELE, Staffs et al. Guidelines for performing systematic literature reviews in software engineering. 2007.

KITCHENHAM, Barbara A. Systematic review in software engineering: where we are and where we should be going. In: Proceedings of the 2nd international workshop on Evidential assessment of software technologies. 2012. p. 1-2.

KUH, George D. et al. What matters to student success: A review of the literature. Washington, DC: National Postsecondary Education Cooperative, 2006.

MARQUES, L.T., MARQUES, B.T., SILVA, C.A.M. . A Descoberta das Causas da Retenção Acadêmica Utilizando Mineração de Dados: Uma Revisão Sistemática da Literatura. RENOTE, Porto Alegre, v. 20, n. 1, p. 263–272, 2022.

LEE, L. E. et al. Evaluation of prediction algorithms in the student dropout problem. Journal of Computer and Communications, 8(03):20, 2020.

MAHAJAN, Ginika; SAINI, Bhavna. Educational data mining: A state-of-the-art survey on tools and techniques used in edm. International Journal of Computer Applications & Information Technology, 12(1):310–316, 2020.

MAHESHWARI, Ekansh et al. Prediction of factors associated with the dropout rates of primary to high school students in india using data mining tools. In Frontiers in Intelligent Computing: Theory and Applications, pages 242– 251. Springer, 2020.

MASCI, Chiara; JOHNES, Geraint; AGASISTI, Tommaso. Student and school performance across countries: A machine learning approach. European Journal of Operational Research, v. 269, n. 3, p. 1072-1085, 2018.

NASCIMENTO, F.A. et al. Mineração de Dados Educacionais: uma Revisão Sistemática da Literatura. Rio de Janeiro: Editora Oston Ltda. 2023. DOI: 10.5281/zenodo.7763620

QIU, Jiezhong et al. Gcc: Graph contrastive coding for graph neural network pre-training. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1150–1160, 2020.

REPASO, Jennifer Anne A.; CAPARIÑO, Elenita T. Analyzing and predicting career specialization using classification techniques. International Journal, v. 9, n. 1.3, 2020.

REZENDE, Solange Oliveira. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.

SOUZA, V.F., SANTOS, T.C.B. Processo de Mineração de Dados Educacionais aplicado na Predição do Desempenho de Alunos: Uma comparação entre as Técnicas de Aprendizagem de Máquina e Aprendizagem Profunda. Revista Brasileira de Informática na Educação – RBIE, 2021.

TAEZA-CRUZ, Maria Elisa Linda; CAPILI-KUMMER, Marifel Grace. Decision Support System to Enhance Students' Employability using Data Mining Techniques for Higher Education Institutions. International Journal of Computing and Digital Systems, 2023.

VIHERÄKOSKI, Johanna. Strengths in career development: Modeling strength-based career counseling through reflecting customer experience. 2020.

VIVIAN, R. L. et al. Mineração de Dados Educacionais e Análise de Sentimentos em Ambientes Virtuais de Aprendizagem: um Mapeamento Sistemático. EaD em Foco, v. 12, n. 2, e1786, 2022.

WAHEED, Hajra et al. Predicting academic performance of students from VLE big data using deep learning models. Computers in Human behavior, v. 104, p. 106189, 2020.

YANG, Cuibi; HUAN, Shuliang; YANG, Yong. A practical teaching mode for colleges supported by artificial intelligence. International Journal of Emerging Technologies in Learning (IJET), v. 15, n. 17, p. 195-206, 2020.

YORK, Travis T.; GIBSON, Charles; RANKIN, Susan. Defining and measuring academic success. Practical assessment, research, and evaluation, v. 20, n. 1, p. 5, 2015.

YULIANTO, Lili Dwi; TRIAYUDI, Agung; SHOLIHATI, Ira Diana. Implementation Educational Data Mining For Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4. 5: Implementation Educational Data Mining For Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4. 5. Jurnal Mantik, v. 4, n. 1, p. 441-451, 2020.