Mineração de Dados Educacionais para a Predição de Evasão: Experiência em uma Universidade do Sul do Brasil

Educational Data Mining for Dropout Prediction: an Experience at a University in Southern Brazil

Piero Salaberri
Sandra Dutra Piovesan
Valesca Brasil Irala

RESUMO

A evasão é um problema que assola instituições de ensino superior públicas e privadas em todo o mundo e estratégias para análise dos motivos para o fenômeno abundam em publicações científicas. Muitos trabalhos que objetivam encontrar as técnicas e práticas mais apropriadas e efetivas para a identificação dos indutores de desistência no aluno acabam por se alicerçar no uso de tecnologias para aprimorar a análise dos dados e atingir um maior volume de informação processada. O presente estudo visa identificar boas práticas para o uso de mineração de dados para informações de cunho educacional. Para tanto, investigou-se práticas já existentes na literatura para a estruturação de uma pesquisa com dados de uma universidade pública no interior do estado do Rio Grande do Sul. O estudo conta com testes práticos com os algoritmos Árvore de Decisão C4.5, *Random Forest* e Redes Neurais em diferentes conjuntos de dados. O trabalho demonstra que o algoritmo *Random Forest* conseguiu ter maior precisão na identificação dos alunos em risco de evasão. A partir desta experiência outras instituições poderão basear-se para a definição de suas melhores práticas.

Palavras-chave: Evasão. Universidade. Ensino Superior. Mineração de Dados Educacionais. Algoritmos.

ABSTRACT

Dropout is a problem that plagues public and private higher education institutions around the world and strategies for analyzing the reasons for the phenomenon abound in scientific publications. Many works that aim to find the most appropriate and effective techniques and practices for identifying dropout inducers in students end up being based on the use of technologies to improve data analysis and achieve a greater volume of processed information. The present study aims to identify good practices for the use of data mining for educational information.

Recebido em: 17/08/23
Aprovado em: 25/04/24

Piero Salaberri
pierosalaberri@unipampa.edu.br
Mestre
Universidade Federal do Pampa
Bagé / RS – Brasil

Sandra Dutra Piovesan
sandrapiovesan@unipampa.edu.br
Doutorado
Universidade Federal do Pampa
Bagé / RS – Brasil

Valesca Brasil Irala
valescairala@unipampa.edu.br
Doutorado
Universidade Federal do Pampa
Bagé / RS – Brasil

ABSTRACT

For this purpose, existing practices in the literature were investigated for structuring research with data from a public university in the interior of the state of Rio Grande do Sul. The study includes practical tests with the Decision Tree algorithms C4.5, Random Forest and Neural Networks in different datasets. The work demonstrates that the Random Forest algorithm was able to be more accurate in identifying students at risk of dropping out. From this experience other institutions will be able to base themselves for the definition of their best practices.
Keywords: Dropout. College. Higher Education. Educational Data Mining. Algorithms.

Introdução

A questão da evasão no ensino superior representa um desafio que afeta não apenas as instituições educacionais no Brasil, mas também em todo o mundo. Abundam na literatura científica estudos que se propõem a examinar os fatores subjacentes a esse fenômeno, bem como a desenvolver abordagens viáveis para lidar com ele. É recorrente nas análises presentes nesses trabalhos a constatação de que as abordagens mais eficazes para fomentar a retenção de alunos em risco de evasão baseiam-se na compreensão das dificuldades enfrentadas por esses estudantes. As informações geradas por esse processo precisam refletir, cuidadosamente, a realidade dos alunos. Elas guiarão o planejamento de ações de combate ao abandono do ambiente escolar (Howlett, Ramesh & Perl, 2013). A partir desse entendimento, torna-se possível elaborar as estratégias necessárias para enfrentar esse problema de maneira mais efetiva.

De acordo com os dados apresentados pelo Mapa do Ensino Superior no Brasil em 2023, apenas 26% dos universitários que ingressaram em universidades no ano de 2017 conseguiram concluir suas respectivas graduações. No decorrer de 2021, a taxa de desistência no âmbito universitário alcançou 55%, contrastando com os 26% que alcançaram a graduação e os 18% que mantiveram sua dedicação aos estudos. Essa realidade é motivo de preocupação, visto que compromete a formação acadêmica dos alunos e resulta no não aproveitamento dos recursos alocados para a educação (SEMESP, 2023).

Ainda, sob análise da mesma publicação, é possível identificar que um contingente maior que 50% dos estudantes que adentraram o ensino superior em 2017,

e que deveriam ter seus estudos concluídos, não conseguiram finalizar seus cursos. Durante o período de análise, apenas 18% dos alunos permaneciam matriculados nas universidades (SEMESP, 2023).

Na busca por compreender as causas subjacentes à evasão escolar, diversas abordagens de análise estão disponíveis. Aquelas que se estruturam sobre técnicas computacionais são aprofundadas neste estudo, já que conseguem ampliar o escopo de dados explorados por meio de análises estatísticas, permitindo a identificação de padrões nos quais os dados podem ser agrupados, correlacionados ou classificados, de acordo com o aspecto em análise.

Ao basear-se em dados provenientes dos registros de alunos em uma universidade, a análise de tais dados oferece a oportunidade de criar conhecimento de maneira a antecipar possíveis trajetórias comportamentais dos próprios alunos. Isso pode ser alcançado por meio de experimentos que exploram os dados de forma sistemática, identificando relações entre variáveis que influenciam a evasão. Ao adotar essa abordagem, é possível desenvolver modelos preditivos que contribuem para uma compreensão mais profunda dos fatores que levam à evasão escolar, permitindo assim a implementação de medidas preventivas e de intervenção. A Mineração de Dados constitui um domínio interdisciplinar, que essencialmente incorpora conhecimentos de análise estatística de dados, aprendizado de máquina, identificação de padrões e visualização representativa de informações (Cabena, Hadjinian, Stadler, Verhees & Zanasi, 1998). Isso se fundamenta, em grande parte, na seleção dos algoritmos disponíveis e na utilização dos conjuntos de dados coletados para as análises em questão.

Estabelecer objetivos bem definidos é essencial para a obtenção de conhecimento relevante. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), no processo de descoberta de conhecimento, esses objetivos são formulados com base nos propósitos de utilização do sistema, podendo ser categorizados em duas principais abordagens: verificação ou descoberta. O enfoque deste estudo está na busca de identificação dos melhores algoritmos para a descoberta de padrões que possam predizer alunos em risco de evasão. Para tal o estudo se vale da busca de outras experiências, também contextualizadas com dados de alunos de ensino superior.

Trabalhos como o de Souza (2021), indicam que algoritmos do tipo Árvore de Decisão e *Random Forest* desempenham de forma bastante eficiente a análise de

dados de alunos. Todavia existem diversas técnicas que precisam ser exploradas para identificar a melhor opção dado o conjunto de dados e objetivos da pesquisa, O presente trabalho, através da identificação dos algoritmos de maior potencial e de testes práticos daqueles selecionados visa determinar quais opções se mostram mais vantajosas para os conjuntos de dados e atributos específicos analisados durante uma investigação realizada em uma universidade pública localizada no interior do estado do Rio Grande do Sul. Tais achados poderão subsidiar futuros pesquisadores e as próprias instituições a realizarem novas parametrizações ou abordagens a partir das considerações expostas.

Evasão e Mineração de Dados

Dentre as estratégias disponíveis para que se analisem os estudantes e todo o seu contexto, a mineração de dados é uma das técnicas que vem sendo explorada em diversos estudos, posto que pode aglutinar um grande volume de dados e disponibilizar a descoberta de conhecimento realizando associação, classificação ou agrupamento de dados conforme objetivo do pesquisador.

De acordo com Han, Pei e Kamber (2012), o processo de Descoberta de Conhecimento em Bancos de Dados (KDD) é uma sequência de etapas com a finalidade de extrair conhecimento a partir de informações contidas em grandes bases de dados. É crucial ressaltar que o KDD é um processo iterativo, o que implica que a realização das etapas não segue uma linha reta do início ao fim. Faz-se frequentemente necessário retornar a estágios anteriores do processo para, então, prosseguir adiante novamente (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Fayyad, Piatetsky-Shapiro e Smyth (1996), ainda, propõem uma divisão do processo de Descoberta de Conhecimento em Bancos de Dados em seis etapas distintas:

- Preparação dos Dados: Nesse estágio, o foco é incorporar informações relevantes à aplicação e estabelecer as metas a serem alcançadas pelo processo.
- Limpeza dos Dados: Aqui, o objetivo é eliminar dados que possam distorcer a análise. Envolve a aplicação de estratégias para remover ruídos, lidar com

valores ausentes e até mesmo transformar variáveis para reduzir a complexidade, visando aprimorar o desempenho dos algoritmos de análise.

- Seleção de Dados: Nesta etapa, decide-se qual conjunto de dados ou subconjunto será alvo do processo de análise.
- Mineração de Dados: Nesse ponto, é escolhida a tarefa de mineração de dados mais adequada para atingir os objetivos do processo, bem como a seleção da técnica mais apropriada para a tarefa.
- Incorporação do Conhecimento Prévio: Essa etapa consiste em interpretar o modelo descoberto, avaliar sua precisão e buscar melhorias. Isso permite retornar a qualquer fase anterior do processo, eliminando padrões redundantes ou irrelevantes.
- Interpretação dos Resultados: Aqui, os resultados obtidos são integrados ao sistema, permitindo ações baseadas no conhecimento adquirido. Isso pode envolver tomada de decisões informadas ou documentação e apresentação dos resultados às partes interessadas.

Especificamente, o estágio de Mineração de Dados utiliza técnicas de inteligência artificial para identificar relações de similaridade ou discrepância entre os dados. O objetivo é descobrir automaticamente padrões, anomalias e regras, transformando dados aparentemente ocultos em informações valiosas para tomada de decisões ou avaliação de resultados.

Baker, De Carvalho, Da Costa e Neves (2011) destacam que dentre as técnicas disponíveis, as principais utilizadas, em contexto educacional, podem ser categorizadas como:

- Predição: busca prever o valor de um atributo específico seja por classificação (identificar a qual classe um registro pertence ou qual o comportamento do atributo escolhido como classe para cada subconjunto de dados (aluno, por exemplo)) ou regressão (prever um atributo numérico com base em um conjunto de dados);
- Agrupamento: tem como objetivo identificar e agrupar registros semelhantes; e
- Associação: visa descobrir conexões entre variáveis em um conjunto de dados.

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o propósito das tarefas preditivas consiste em antecipar o valor de um atributo específico, também denominado variável, com base nos valores de outros atributos. Em um nível mais conceitual, a predição utiliza uma combinação de variáveis para estimar outras variáveis ou valores desconhecidos. Por exemplo, ao coletar dados sobre alunos matriculados em instituições de ensino superior, é possível examinar todos os dados disponíveis relacionados ao desempenho acadêmico dos alunos e, por meio de técnicas de mineração de dados, identificar as características que estão associadas à evasão. Utilizando o conhecimento adquirido, é viável identificar novos alunos que também possam estar em risco de evasão.

A mineração de dados educacionais tem se mostrado uma ferramenta promissora para enfrentar a evasão universitária. Essa abordagem utiliza técnicas de análise de dados para identificar padrões e tendências nos comportamentos dos estudantes, permitindo uma melhor compreensão das causas da evasão e possibilitando a adoção de medidas preventivas.

MINERAÇÃO DE DADOS EDUCACIONAIS

A Mineração de Dados Educacionais emerge como um campo de pesquisa que aplica técnicas de mineração de dados a conjuntos de dados educacionais. Seu propósito é aprofundar a compreensão de como os estudantes aprendem e interagem com o ambiente educacional, com o intuito de que tal descoberta de conhecimento possa aprimorar os resultados educacionais dos próprios alunos.

Os sistemas educacionais abrigam um grande volume de instâncias de informação provenientes de diversos sistemas e registros, abrigando diferentes formatos e dimensões de dados (Romero & Ventura, 2013). Tais técnicas desempenham, também, um papel crucial na investigação comportamental do estudante, auxiliando assim no planejamento, desenvolvimento e operacionalização de soluções tecnológicas mais eficazes para apoiar tanto alunos quanto gestores e educadores.

Quando empregamos tais técnicas com o intuito de previsão, objetiva-se desenvolver modelos que identifiquem elementos vinculados aos dados. Isso é alcançado ao analisar e combinar as características inerentes aos próprios dados examinados, que são referidos como variáveis preditoras. A obtenção de informações sobre os alunos deve buscar extrapolar a investigação dos registros acadêmicos,

explorando opções que reflitam o máximo a vida do aluno, mesmo que seja necessária a aplicação de questionários ou pesquisas de opinião (Baker, De Carvalho, Da Costa, & Neves, 2011). A Mineração de Dados Educacionais pode desempenhar papéis distintos conforme as pesquisas se estruturam, desde a análise de desempenho acadêmico até a predição de alunos em potencial risco, para que a própria instituição possa prover suporte e induzir a permanência dos alunos (Souza, 2021).

Após a etapa de coleta e preparação dos dados, parte-se para a fase dos testes experimentais. Para tal, é necessária a escolha dos algoritmos e da ferramenta que será utilizada para a execução dos mesmos. Dentre as ferramentas disponíveis para o uso se destacam as ferramentas Weka, Rapidminer, KNINE, Orange, R Studio entre outras. Dentre as principais técnicas de classificação compreendem-se diferentes tipos de classificadores, tais como os baseados em árvores de decisão, os baseados em regras, redes neurais, máquinas de vetores de suporte e classificadores naive Bayes. Cada abordagem emprega um algoritmo de aprendizado para discernir um modelo que se adapte de maneira mais precisa à relação entre o conjunto de atributos e as etiquetas de classe dos dados de entrada. O modelo resultante, gerado pelo algoritmo de aprendizado, deve se ajustar de forma eficaz aos dados de entrada e ser capaz de prever com precisão as etiquetas de classe dos registros que ainda não foram observados (Brandão, 2018). Como exemplos de algoritmos baseados na lógica de classificação, tem-se Árvores de Decisão C4.5, CART e *Random Forest*; Redes Neurais Artificiais *Multilayer perceptron;* Naive Bayes; Regressão Linear e Logística e *Support Vector Machines* (SVM), por exemplo.

A próxima seção discorre sobre os testes executados, bem como do processo de escolha das soluções algorítmicas e de ferramentas para execução, indicando métricas de desempenho para cada subconjunto de dados utilizados.

Materias e Métodos

De maneira geral, os algoritmos baseados em árvore de decisão apresentam um desempenho notável, consistentemente figurando entre os líderes ou destacando-se nas comparações, como evidenciado nos estudos conduzidos por Fernandez-Garcia, Gonzalez, Rico e Prieto (2021), Franco, Martínez e Domínguez (2021) e

Lee e Chung (2019). Além disso, foram identificadas investigações que cotejaram diferentes iterações do próprio algoritmo de árvore de decisão. Sunday, Jekayinoluwa, Adedokun, Ajao e Yusuff (2020) e Hamoud, Hashim e Awadh (2018) elegeram o algoritmo C4.5 como a opção superior em termos de desempenho geral quando contrastado com as variantes ID3 e, no caso do primeiro trabalho, com as variações Random Tree e REPTree. NiyoGiSubizo, Nduwimana, Nzobonimpa e Nkurunziza (2022) optaram pela versão Extreme Gradient Boosting.

Dentre todas as análises realizadas, destacou-se como o algoritmo mais eficaz nos estudos conduzidos por Alboaneen, Alenezi, Alrumayh, Almutairi, Alanazi, e Alotaibi (2022), Flores, Heras e Julian (2022), Palacios, Arenas, Jimenez e Villanueva (2021), Perez-Gutierrez (2020), Yağci (2022), Al-Fairouz e Al-Hagery (2020). Notavelmente, em diversas instâncias, como exemplificado por Kabathova e Drlik (2021), o algoritmo *Random Forest* apresentou um desempenho amplamente superior em relação a outros algoritmos avaliados, exibindo diferenças percentuais na métrica de acurácia que chegaram a quase 30%. No contexto do estudo de Fernandez-Garcia, Gonzalez, Rico e Prieto (2021), o algoritmo demonstrou um melhor rendimento para alunos do 3º e 4º semestres, nos quais existe uma disponibilidade considerável de registros acadêmicos.

Outros trabalhos como os de Nabil, Seyam e Abou-Elfetouh (2021); Nuankaew, Namahoot e Phewchean (2020); Siddique, Alam e Marwah (2021) e Miranda e Guzmán (2017), observaram que os algoritmos fundamentados em redes neurais alcançaram o melhor desempenho nos grupos de dados analisados. Por outro lado, no estudo conduzido por Yağci (2022), o desempenho foi notavelmente similar ao do *Random Forest*, que se destacou como a opção de melhor desempenho. Ainda há trabalhos como o de Nascimento, de Carvalho, Lima, Neves e Freitas (2023) que identificou ser o algoritmo Naive Bayes o mais testado em trabalhos de mineração de dados. Todavia, são poucos aqueles nos quais esse algoritmo é o que melhor performa.

Após a avaliação dos estudos mencionados, sob a ótica de um trabalho que visa prever a probabilidade de estudantes abandonarem o ensino superior, utilizando principalmente informações demográficas e acadêmicas, os algoritmos mais promissores revelaram ser: Árvore de Decisão C4.5, *Random Forest* e Redes Neurais. Assim sendo, o estudo utilizará esses algoritmos para determinar qual apre-

senta o melhor desempenho e capacidade de se possa extrair informações mais relevantes nas diversas situações de análise que serão desenvolvidas.

A técnica de validação cruzada é amplamente empregada para garantir a utilização abrangente do conjunto de dados, dividindo-o em várias partes, também conhecidas como pastas. A validação cruzada com k = 10 pastas é frequentemente citada na literatura como o melhor valor a ser utilizado (Castro & Ferrari, 2006). Dessa forma, todas as abordagens de modelo serão submetidas à validação cruzada de 10 pastas, com o propósito de avaliar se essa abordagem afeta o desempenho, considerando que essa técnica demonstrou ser eficaz na maioria substancial dos estudos examinados.

Os modelos foram desenvolvidos utilizando a plataforma Weka (Waikato, 1999), que oferece uma interface para a criação, ajuste e avaliação dos modelos, bem como para os dados submetidos a teste. Alturki, Hulpuș e Stuckenschmidt (2020), em sua revisão, ressaltaram que essa ferramenta foi a mais comumente utilizada em pesquisas dedicadas à previsão de evasão no ensino superior ao longo da última década. Autores como Nascimento*,* de Carvalho, Lima, Neves e Freitas (2023) corroboram tal achado, indicando que a ferramenta WEKA está entre as mais frequentemente empregadas nos estudos voltados a investigar as melhores abordagens para a exploração de dados institucionais. Martins (2017) chegou à conclusão de que o WEKA oferece uma interface mais simples, resultando em uma ferramenta mais prática, de melhor usabilidade e com curva de aprendizagem menor. Além disso, observou, também, que conta com uma quantidade satisfatória de documentação.

Como fonte de dados para a condução do experimento de pesquisa, o autor requisitou acesso à base de dados da instituição pesquisada, visando obter informações relacionadas aos alunos. A base de dados da instituição tem em torno de 40.000 vínculos entre alunos regulares, com vínculo concluído e evadidos. Após a obtenção de autorização, o autor obteve acesso integral aos registros armazenados nos bancos de dados mantidos pela equipe técnica responsável pela gestão dos dados informatizados.

Nesse contexto, uma variedade de informações estava disponível para acesso e análise, abrangendo dados demográficos (dados relacionados à cidade de origem, idade, sexo e estado civil); de ingresso (dados sobre tipo de acesso e média utilizada para admissão); relativos ao curso (tipo de curso EaD ou presencial,

formação bacharelado/licenciatura/tecnológico e turno); dados de ordem Pessoal/Familiar (etnia, se possui deficiência e qual, escola de origem pública ou privada, número de membros da família e dados de renda); de vínculo acadêmico (número total de disciplinas vencidas e reprovadas, número de disciplinas vencidas e reprovadas por semestre do 1 ao 10, carga horária exigida e vencida para integralização do curso, e carga horárias de atividades complementares exigidas e vencidas); assistência estudantil (solicitações de benefício enviadas e deferidas, e tipos de benefícios recebidos com, número de competências recebidas e valor médio); do uso da biblioteca (dados sobre empréstimos de livros físicos e digitais); e Restaurante Universitário (número de refeições, valor pago pelo aluno e valor subsidiado pela instituição). Esse conjunto abrangente de informações possibilitou uma investigação das possíveis implicações de cada um desses atributos no contexto da evasão de alunos naquela universidade.

Os estudos realizados por Lanot (2012) e Borin (2014), ambientados na maior unidade acadêmica da universidade, ressaltam a importância de considerar o contexto temporal das aprovações e reprovações em relação à estrutura curricular dos cursos. É fundamental entender em que fase do currículo esses eventos ocorrem. No entanto, analisar essa dinâmica para cada curso na instituição se tornaria impraticável devido ao tempo necessário para execução. Para contornar essa limitação, os dados coletados foram categorizados por semestre em que a disciplina está integrada ao curso. Dessa forma, foi possível examinar uma abordagem que também proporciona uma visão consistente em comparação com os modelos gerados individualmente para cada curso.

Para a compreensão das métricas utilizadas na próxima seção do artigo, intitulada de Análise dos Dados, faz-se necessário expressar a definição de cada indicador presente. Todas as métricas partem da matriz de confusão, que é construída pelo próprio algoritmo após sua execução. De acordo com Castro e Ferrari (2006), a tabela em questão representa os acertos e erros do modelo, ao ser comparado com o resultado desejado. Os acertos do modelo formam a diagonal principal, enquanto os demais valores correspondem aos erros ocorridos. Quando se foca em uma classe específica, que possui dois possíveis valores (por exemplo, “aluno evadiu: SIM ou NÃO”), isso é chamado de cenário binário. Nesse contexto, a representação mais comum é exemplificada pela matriz exibida na Figura 1.

Figura 1. Matriz de confusão.

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: (Castro & Ferrari, 2006).

Os campos da matriz são preenchidos com os seguintes campos:

- Verdadeiro Positivo (**VP**): indica a correta classificação da classe positiva.;
- Falso Negativo (**FN**): ocorre quando o modelo prevê a classe negativa, mas o valor real é positivo;
- Falso Positivo (**FP**): ocorre quando o modelo prevê a classe positiva, mas o valor real é negativo; e
- Verdadeiro Negativo (**VN**): representa a correta classificação da classe negativa.

Com base nos valores dos campos acima descritos pode-se calcular métricas de desempenho, como segue:

- **Acurácia**: é o número de classificações corretas dividido pelo número total de classificações. Pode ser traduzida pela fórmula (VP+VN)/(VP+FP+-VN+FN);
- **Precisão**: mede a exatidão do algoritmo, ou seja, dentre todas as classificações de classe positiva que o modelo fez, quantas estão corretas. Quando os Falsos Positivos têm um impacto maior do que os Falsos Negativos, a precisão pode ser uma métrica relevante para análise. Pode ser expressa como VP/(FP+VP);
- ***Recall***: avalia quantas das situações esperadas como positivas foram classificadas corretamente. É uma métrica interessante para quando Falsos Negativos são considerados mais prejudiciais que os Falsos Positivos. Pode ser expressa como VP/(FN+VP); e

- **F-*Score***: média harmônica entre precisão e recall, ou seja, quando tem-se um F1-Score baixo, é um indicativo de que ou a precisão ou o recall está baixo. Pode ser expressa como (2 x Precisão X *Recall*) / (Precisão + *Recall*).

Análise dos Dados

Foram conduzidos testes iniciais utilizando um conjunto de 86 atributos. Os resultados destes testes estão apresentados no Quadro 1. Inicialmente, foram empregados os algoritmos *Random Forest* (RF), Árvore de Decisão C4.5 e Redes Neurais com as configurações padrão do software WEKA. Observou-se que os algoritmos C4.5 e RF apresentaram os melhores desempenhos, alcançando taxas de acerto de 61,25% e 60,07%, respectivamente. Entretanto, ao analisar a Matriz de Confusão, identificou-se que o algoritmo RF obteve um maior número de acertos nos casos de alunos que efetivamente evadiram (3428 acertos contra 3286 do algoritmo C4.5). Dado um contexto em que é crucial fornecer suporte aos alunos em risco, a decisão foi tomada em favor do algoritmo com maior precisão na detecção desses estudantes.

Quadro 1. Métricas dos algoritmos testados com 86 atributos.

Algoritmo/Base	Métrica	Valor	Matriz de confusão (a= SIM, b= NAO)
Random Forest com base original (86 atributos)	Acurácia Precisão *Recall* F-*Score*	60,0658% 0,596 0,601 0,596	a b 3428 3692 2743 6251
C4.5 com base original (86 atributos)	Acurácia Precisão *Recall* F-*Score*	61,2511% 0,608 0,613 0,605	a b 3286 3834 2410 6584
Redes Neurais *Multilayer Perceptron* com base original (86 atributos)	Acurácia Precisão *Recall* F-*Score*	59,3583% 0,588 0,594 0,586	a b 3164 3956 2593 6401

Fonte: (Autores, 2023).

Ao examinarmos as matrizes de confusão dos algoritmos avaliados com o conjunto de 86 atributos, constata-se uma acurácia globalmente baixa, variando entre 59,36% e 61,25%. O algoritmo de Redes Neurais ficou em último lugar nessa avaliação, demandando maior capacidade computacional e, consequentemente, mais tempo de execução. Devido à sua menor efetividade, optou-se por excluí-lo das próximas etapas de testes. Ainda, um número maior de falsos positivos podem indicar mais alunos com perfil similar aqueles identificados como em risco de evasão, podendo servir de alerta para quais alunos a universidade pode direcionar sua atenção. Reforçando *Random Forest* como a melhor opção dentre o grupo testado.

O campo da aprendizagem de máquina não apenas possibilita avaliar a eficácia e precisão dos algoritmos, mas também permite investigar a contribuição de cada atributo ao longo do processo. Essa análise oferece insights sobre o grau de relevância das decisões tomadas pelo algoritmo com base nas informações disponíveis na base de dados.

Nos algoritmos de árvore de decisão, o cálculo da importância ou relevância dos atributos durante o treinamento de um conjunto de dados ocorre ao escolher as variáveis ou características em cada nó que maximizam a redução de erros no processo geral de previsão. Nesse contexto, os atributos mais significativos em uma árvore de decisão são classificados com base na sua capacidade de reduzir o erro quando são considerados, ponderando essa redução de acordo com o número de observações relacionadas ao nó. No caso do algoritmo *Random Forest*, que consiste em dividir a base original, aleatoriamente, criando várias árvores menores, esse procedimento é realizado individualmente para cada uma das árvores criadas. Posteriormente, essas importâncias individuais são agregadas por meio de médias para determinar a importância de uma característica específica no contexto geral (Thorn, 2020).

A etapa inicial de testes revelou um considerável número de atributos com pouca relevância no processo de mineração de dados. A relevância dos atributos do conjunto de dados analisado é demonstrada nas Figuras 2 e 3. Na Figura 2, constam os 15 atributos de maior relevância, com valores de índice variando entre 0,31 e 0,28. A partir desta seleção, foi notável que as disciplinas concluídas pelos alunos surgiram como o atributo de maior significância, indicando a sua relevância predominante. Os atributos relacionados a disciplinas vencidas (DISC_VENC, DISC_VENC_

SEM_2, DISC_VENC_SEM_3, DISC_VENC_SEM_1, DISC_VENC_SEM_4, CH_VENCIDA, DISC_VENC_SEM_7) sugerem que o desempenho acadêmico desempenha um papel fundamental no processo de tomada de decisão em relação à evasão, uma vez que alunos que concluem com êxito um maior número de disciplinas e apresentam menor quantidade de reprovações geralmente apresentam um risco reduzido de abandonar os estudos. Essa observação é consistentemente alinhada com as publicações destinadas à análise da evasão, que apontam que os dados acadêmicos, especialmente os relacionados à quantidade de disciplinas aprovadas e reprovadas, representam os indicadores mais sólidos para prever a evasão.

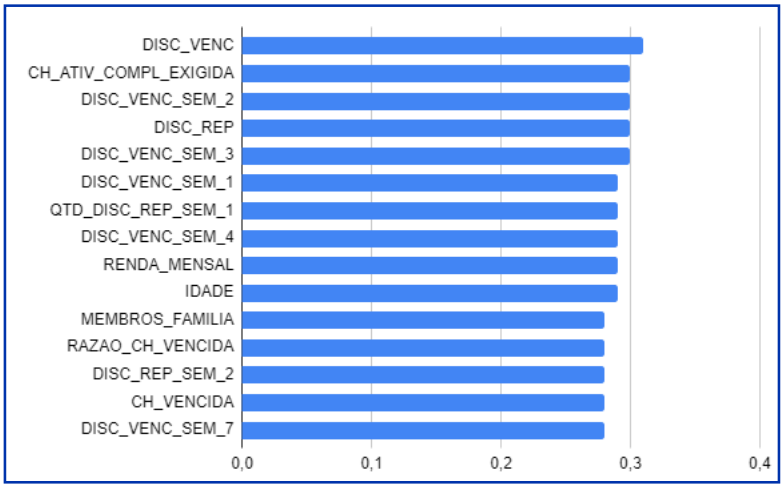
Em segundo lugar de importância, encontra-se a quantidade de disciplinas reprovadas. Além disso, chama a atenção o fato de que as disciplinas vencidas nos dois primeiros anos do percurso formativo possuem alta significância, enquanto a quantidade de disciplinas reprovadas no primeiro semestre também é considerada relevante. Sendo estes os atributos mais significativos, juntamente ao cumprimento de atividades complementares, pode ser considerado um indicativo de que o desempenho acadêmico e o envolvimento com a instituição podem desempenhar um papel crucial na prevenção da evasão. Esses atributos também trazem claro alerta que os alunos ingressantes têm maior probabilidade de estar em risco de abandono. Portanto, qualquer ação institucional deve sempre levar em conta essa diretriz.

A carga horária das atividades complementares exigidas (CH_ATIV_COMPL_EXIGIDA) também desempenha um papel crucial. Alunos que satisfatoriamente completam essas atividades parecem demonstrar um maior nível de envolvimento e possivelmente um comprometimento mais sólido com o curso. Além disso, é válido ponderar sobre o impacto de fatores socioeconômicos, como renda mensal e número de membros na família, e até que ponto eles podem influenciar a probabilidade de evasão. Baixa renda ou uma maior dependência financeira podem ser fatores de risco. A idade dos alunos também pode ser relevante, indicando diferentes estágios de maturidade e engajamento com os estudos. Entretanto, não foi possível analisar se alunos mais velhos refletem tal observação e podem ter um risco menor de evasão, dado que eles frequentemente enfrentam a necessidade de equilibrar a jornada de trabalho com os estudos.

A relação entre a carga horária cumprida e a carga horária total (RAZAO_CH_VENCIDA) pode servir como um indicador do progresso acadêmico do aluno.

Aqueles com uma proporção elevada apresentam um avanço consistente em relação ao currículo, o que, por conseguinte, também pode contribuir para a mitigação do risco de evasão.

Figura 2. Principais atributos segundo algoritmo *Random Forest.*



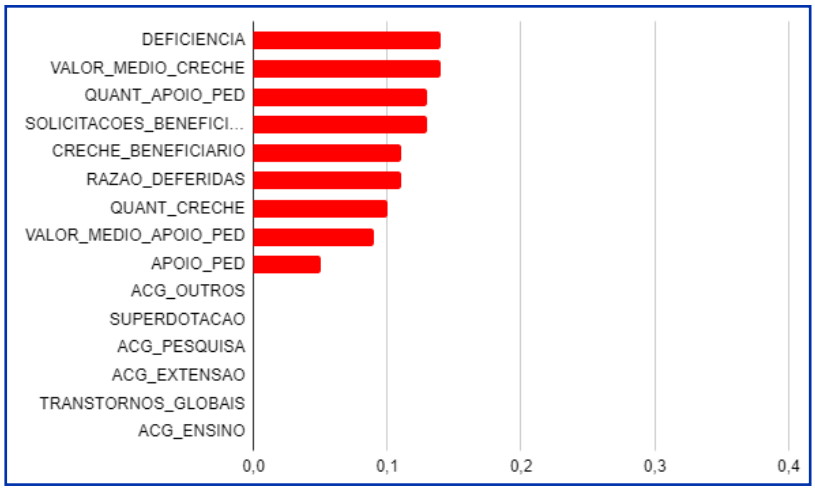
Fonte: (Autores, 2023).

Na Figura 3, podemos observar os atributos de menor significância, ou seja, aqueles que obtiveram pontuações abaixo de 0,14. Os atributos associados a atividades extracurriculares (ACG_OUTROS, ACG_PESQUISA, ACG_EXTENSAO, ACG_ENSINO) não receberam pontuações consideráveis, o que pode indicar, principalmente, uma limitada coleta de dados, possivelmente devido ao fato de que os alunos fornecem essas informações ao final do curso. Mesmo alunos que eventualmente abandonaram o curso poderiam ter participado de atividades desse tipo, mas talvez não tenham havido registros dessas atividades. Estimular o registro ao longo de toda a graduação do estudante provavelmente aumentaria a relevância desses atributos. Por outro lado, atributos relacionados a deficiências também mostraram baixa relevância. Isso também pode refletir uma quantidade reduzida de casos na amostra pesquisada, em contraste com o cenário anterior. No caso anterior, supõe-se que o problema esteja relacionado ao momento em que os registros são feitos. Já no que diz respeito a alunos com deficiência, é plausível especular que o número

de matrículas seja pequeno, o que poderia tornar esse atributo menos relevante em comparação com os outros.

Os atributos remanescentes deste recorte estão associados à assistência estudantil, embora essa seja uma área amplamente considerada ao se planejarem estratégias para manter os alunos no ambiente acadêmico. Benefícios principalmente ligados ao suporte pedagógico e serviços de creche podem não apresentar uma relação direta com a evasão, mas podem servir como indicadores de outros fatores socioeconômicos. Seria necessário investigar o número de auxílios concedidos aos alunos beneficiados e avaliar o impacto deles na continuidade desses estudantes.

Figura 3. Atributos menos relevantes segundo algoritmo *Random Forest.*



Fonte: (Autores, 2023).

A respeito da análise abrangente do conjunto de dados, foi identificado que 32 atributos na amostra exibiram uma significância inferior a 0,2. Portanto, foram realizados testes sem a inclusão desses dados nas amostras, com o objetivo de avaliar se tal exclusão teve algum impacto no desempenho global dos algoritmos. Os resultados desses testes estão documentados no Quadro 2, assim como os testes conduzidos com os 37 atributos (significância superior a 0,25) e 28 atributos (significância superior a 0,26). Os testes foram conduzidos empregando os algoritmos *Random Forest* e C4.5, e os resultados foram altamente satisfatórios, superando os

testes realizados com a amostra completa. Isso resultou em um aumento de mais de 20% na acurácia para ambos os algoritmos.

Observa-se que o algoritmo *Random Forest*, que se valeu de uma base de testes com 52 atributos, demonstrou o melhor desempenho nessa experiência, obtendo a maior taxa de acerto tanto para os casos de evasão ("SIM") quanto para os casos de não evasão ("NÃO") (linha destacada no Quadro 2).

Pode ser observado que à medida que o número de atributos diminuiu nos testes subsequentes, houve uma correspondente redução na eficácia dos algoritmos de aprendizado de máquina.

Quadro 2. Métricas dos algoritmos testados com 52, 37 e 28 atributos.

Algoritmo/Base	Métrica	Valor	Matriz de confusão (a= SIM, b= NAO)
***Random Forest* com atributos acima de 0,2 (52 atributos)**	**Acurácia** **Precisão** ***Recall*** **F-*Score***	**84,4424%** **0,842** **0,844** **0,843**	a b 4653 2281 1960 18366
C4.5 com atributos acima de 0,2 (52 atributos)	Acurácia Precisão *Recall* F-*Score*	82,6082% 0,826 0,826 0,826	a b 4528 2406 2335 17991
Random Forest com atributos acima de 0,25 (37 atributos)	Acurácia Precisão *Recall* F-*Score*	84,4855% 0,843 0,845 0,844	a b 4476 2145 1935 17742
C4.5 com atributos acima de 0,25 (37 atributos)	Acurácia Precisão *Recall* F-*Score*	82,7059% 0,828 0,827 0,828	a b 4410 2211 2337 17340
Random Forest com atributos acima de 0,26 (28 atributos)	Acurácia Precisão *Recall* F-*Score*	82,9808% 0,828 0,830 0,829	a b 4067 2286 2051 17079

Fonte: (Autores, 2023).

Considerações Finais

A mineração de dados está sendo usada no ensino superior para analisar as taxas de abandono escolar e identificar padrões que podem ajudar os administradores ou gestores acadêmicos a tomar decisões mais embasadas. Ao analisar-se os dados relacionados às informações dos estudantes, técnicas de mineração de dados podem ser usadas para identificar estudantes em risco de abandono escolar e planejar intervenções apropriadas. Além disso, a mineração de dados pode ser usada para melhorar os processos de avaliação e tomada de decisão no ensino superior, utilizando o conhecimento gerado como subsídio para a construção de cenários que caracterizam a universidade como uma rede de amparo e potencializadora de todos os alunos.

Neste estudo, o objetivo proposto foi cumprido ao demonstrar com clareza que tanto o conjunto de dados quanto a seleção dos algoritmos e técnicas de mineração de dados desempenham um papel crucial na precisão dos resultados obtidos. Ao analisar-se os atributos destacados, é evidente que buscar uma profusão máxima de informações nem sempre constitui a estratégia mais vantajosa, pois nem todas as informações se mostram relevantes no processo de descoberta de conhecimento. Além disso, atributos com dados limitados ou com um número reduzido de instâncias acabam perdendo sua relevância no contexto da aprendizagem de máquina. Logo, não se pode generalizar que os atributos considerados menos significativos neste estudo também seriam irrelevantes em outros cenários de avaliação.

Uma conclusão relevante que emerge é a importância de a universidade manter uma base de dados sólida e constantemente atualizada. É imperativo criar uma cultura organizacional que promova a qualidade dos registros dos alunos e a rapidez na armazenagem das informações, pois os dados sobre o percurso acadêmico dos alunos só serão úteis para estudos que buscam identificar riscos de evasão se forem registrados de maneira oportuna. Registrar elementos como componentes curriculares, atividades extracurriculares ou disciplinas complementares apenas no final da trajetória acadêmica inviabiliza a condução de estudos preditivos de evasão.

Outros dados provenientes da relação entre professores e discentes podem ser empregados como atributos para essa previsão, como a frequência dos alunos. Esse atributo é um preditor comum em trabalhos similares, porém, para seu uso

eficaz, é essencial garantir que todos os professores registrem de forma regular as presenças e ausências dos alunos. Ainda que a omissão ocorra apenas em uma pequena parcela dos docentes, ela pode causar um desequilíbrio entre os alunos, impactando o erro dos algoritmos.

Ao analisar-se a eficácia dos algoritmos, torna-se evidente que, apesar da similaridade das métricas apresentadas para uma mesma base de dados, há variações na precisão dos algoritmos. Portanto, a análise da matriz de confusão torna-se essencial para guiar a seleção do algoritmo mais apropriado. Algoritmos que demonstrem maior precisão e, principalmente, apresentem valores mais elevados na diagonal principal devem ser escolhidos.

O algoritmo *Random Forest* demonstrou ser a melhor alternativa dentre as soluções elencadas nesse estudo, sendo, consequentemente, selecionado como a melhor escolha para previsões que buscam identificar alunos em risco de evasão. Tal conclusão está em sintonia com uma vasta gama de trabalhos que também identificaram nele este alto potencial preditivo. Resta a próximos pesquisadores, explorar estratégias de parametrização do algoritmo para que seja possível analisar alguma melhoria em seu desempenho. Estratégias relacionadas à base de dados também podem ser analisadas, principalmente, relacionadas ao pré-processamento de dados e aos dados faltantes. Pode-se analisar a retirada de atributos com muitos dados omissos e a uniformização de escala, dimensões ou natureza dos dados.

Referências

Alboaneen, D., Alenezi, M., Alrumayh, A., Almutairi, A., Alanazi, S., & Alotaibi, M. (2022). Development of a web-based prediction system for students' academic performance. *Data, 7*(2), 21. https://doi.org/10.3390/data7020021

Al-Fairouz, E., & Al-Hagery, M. (2020). The most efficient classifiers for the student's academic dataset. *International Journal of Advanced Computer Science and Applications, 11*(9), 501–506. https://doi.org/10.14569/IJACSA.2020.0110960

Alturki, S., Hulpuș, I., & Stuckenschmidt, H. (2020). Predicting academic outcomes: A survey from 2007 till 2018. *Technology, Knowledge and Learning, 27*(1), 275-307. https://doi.org/10.1007/s10758-020-09476-0

Baker, R., De Carvalho, A., Da Costa, E., & Neves, P. (2011). Mineração de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação, 19*(2), 3-13. https://doi.org/10.5753/rbie.2011.19.02.03

Borin, J. (2014). *Desenvolvimento de um software para análise de evasão na Unipampa Campus Bagé utilizando técnicas de mineração de dados* (Unpublished doctoral dissertation). Universidade Federal do Pampa, Bagé.

Brandão, I. (2018). *Framework de mineração de dados educacionais em ambiente de cursos a distância governamental* (Unpublished undergraduate thesis). Universidade de Brasília, Brasília.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: From concept to implementation*. Prentice-Hall, Inc.

Castro, L. N., & Ferrari, D. G. (2016). *Introdução* à mineração de dados: Conceitos básicos, algoritmos e aplicações. Saraiva.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Advances in knowledge discovery and data mining. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *From data mining to knowledge discovery: An overview* (pp. 1-34). American Association for Artificial Intelligence.

Fernandez-Garcia, A., Gonzalez, S., Rico, M., & Prieto, E. (2021). A real-life machine learning experience for predicting university dropout at different stages using academic data. *IEEE Access, 9*, 133076–133090. https://doi.org/10.1109/ACCESS.2021.3119596

Flores, V., Heras, S., & Julian, V. (2022). Comparison of predictive models with balanced classes using the SMOTE method for the forecast of student dropout in higher education. *Electronics, 11*(3), 457. https://doi.org/10.3390/electronics11030457

Franco, E., Martínez, R., & Domínguez, V. (2021). Predictive models of academic risk in computing careers with educational data mining. *Revista de Educación a Distancia, 21*(66). https://doi.org/10.6018/red.425981

Hamoud, A., Hashim, A., & Awadh, W. (2018). Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence, 5*(2), 26. https://doi.org/10.9781/ijimai.2018.02.004

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*. Elsevier/Morgan Kaufmann. https://www.sciencedirect.com/science/book/9780123814791

Howlett, M., Ramesh, M., & Perl, A. (2013). *Política pública: Seus ciclos e subsistemas – uma abordagem integral*. Campus.

Kabathova, J., & Drlik, M. (2021). Towards predicting student dropout in university courses using different machine learning techniques. *Applied Sciences, 11*(7), 3130. https://doi.org/10.3390/app11073130

Lanot, A. (2012). *Mineração de dados aplicada na identificação da propensão à evasão na universidade* (Unpublished undergraduate thesis). Universidade Federal do Pampa, Bagé.

Lee, S., & Chung, J. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences, 9*(15), 3093. https://doi.org/10.3390/app9153093

Martins, B. (2017). *Uma discussão sobre diferentes ambientes de software para mineração de dados* (Unpublished undergraduate thesis). Universidade Federal do Maranhão, São Luís.

Miranda, M., & Guzmán, J. (2017). Análisis de la deserción de estudiantes universitarios usando técnicas de minería de datos. *Formación Universitaria, 10*(3), 61-68. https://doi.org/10.4067/s0718-50062017000300007

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access, 9*, 140731–140746. https://doi.org/10.1109/ACCESS.2021.3119596

Nascimento, F., de Carvalho, A. A., Lima, C. F., Neves, P. V., & Freitas, C. (2023). Mineração de dados educacionais: Uma revisão sistemática da literatura. *Zenodo, 27*(120), 1-21. https://doi.org/10.5281/zenodo.7763620

Niyogisubizo, J., Nduwimana, A., Nzobonimpa, L., & Nkurunziza, D. (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*, 100067. https://doi.org/10.1016/j.caeai.2022.100067

Nuankaew, P., Namahoot, C., & Phewchean, N. (2020). Prediction model of student achievement in business computer disciplines. *International Journal of Emerging Technologies in Learning (iJET), 15*(20), 160. https://doi.org/10.3991/ijet.v15i20.15273

Palacios, C., Arenas, M., Jimenez, E., & Villanueva, P. (2021). Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. *Entropy, 23*(4), 485. https://doi.org/10.3390/e23040485

Pérez-Gutiérrez, B. (2020). Comparación de técnicas de minería 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

Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008). Data mining algorithms to classify students. In *Proceedings of the International Conference on Educational Data Mining* (pp. 8-17).

SEMPESP. (2023). Mapa do ensino superior 2023. https://www.semesp.org.br/wp-content/uploads/2023/06/mapa-do-ensino-superior-no-brasil-2023.pdf

Siddique, A., Alam, S., & Marwah, A. (2021). Predicting academic performance using an efficient model based on fusion of classifiers. *Applied Sciences, 11*(24), 11845. https://doi.org/10.3390/app112411845

Souza, V. F. (2021). Mineração de dados educacionais com aprendizagem de máquina. *Revista Educar Mais, 5*(4), 766-787. https://doi.org/10.15536/reducarmais.5.2021.2417

Sunday, K., Jekayinoluwa, J., Adedokun, A., Ajao, T., & Yusuff, A. (2020). Analyzing student performance in programming education using classification techniques. *International Journal of Emerging Technologies in Learning (iJET), 15*(2), 127–144. https://doi.org/10.3991/ijet.v15i02.11527

Waikato, Departments of Computer Science and Software Engineering. (1999). *Weka 3: Machine learning software in Java*. https://www.cs.waikato.ac.nz/ml/weka/

YAğCi, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1), 1-19. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s40561-022-00192-z