ISSN 2177-8310
DOI: 10.18264/eadf.v16i12881

Estudo de caso

Predição da Evasão em um Curso de Licenciatura em Computação ofertado na Modalidade de Educação a Distância Empregando Redes Neurais Artificiais

Predicting Dropout Rates in a Distance Learning Computer Science Degree Program Using Artificial Neural Networks

Sidnei Renato SILVEIRA[1]*
Vinicius Gadis RIBEIRO[2]
Henrique Carlos de Oliveira CASTRO[3]
Guilherme Bernardino da CUNHA[4]

[1,4] UFSM – Universidade Federal de Santa Maria – Campus Frederico Westphalen. Frederico Westphalen, RS, Brasil.
[2,3] UFRGS – Universidade Federal do Rio Grande do Sul - Porto Alegre, RS, Brasil.

*sidnei.silveira@ufsm.br

Resumo

Este artigo apresenta um experimento acerca dos índices de evasão dos alunos do Curso de Licenciatura em Computação da Universidade Federal de Santa Maria, ofertado na modalidade de Educação a Distância no âmbito da Universidade Aberta do Brasil. O experimento foi conduzido por meio de Redes Neurais Artificiais, com base nos dados dos alunos do referido curso, visando calcular a predição da evasão na turma em andamento. O curso encontra-se na sua terceira oferta, tendo iniciado suas atividades no ano de 2017. Considerando as duas turmas já encerradas (ingressantes de 2017 e 2020/2021), a base de dados foi composta por 201 alunos. A Rede Neural Artificial, após o treinamento, apresentou um índice de 28,89% de predição de evasão para a turma em andamento, que são os ingressantes de 2024. Incluindo os alunos da turma de 2024 que já evadiram (64 alunos, representando 58,72%) e se a predição da rede neural se confirmar até o final desta turma, teremos um percentual de evasão de 70,64%, um percentual alto, mas abaixo do índice de evasão das turmas anteriores. O índice de evasão das turmas de ingressantes de 2017 é de 71,17% e da turma 2020/2021 é de 83,33%. A Rede Neural Artificial implementada poderá ser aplicada para predição da evasão de futuras turmas do curso e, também, em outros cursos ofertados na modalidade de EaD na UFSM.

Palavras-chave: Evasão. Educação a distância. Redes neurais artificiais.

PRÉ

Recebido:10/07/26
Aceito: 04/11/26
Publicado: 12/11/2026
Editores Responsáveis:
Daniel Salvador
Carmelita Portela
Daniela Samira

COMO CITAR ESTE TRABALHO

ABNT: SILVEIRA, S.R. *et al*. Título do Artigo. **EaD em Foco**, 2026;16(1): e2881. 2026. DOI: https://doi.org/10.18264/eadf.v16i1.2881

Predicting Dropout Rates in a Distance Learning Computer Science Degree Program Using Artificial Neural Networks

Abstract

This paper presents an experiment on the dropout rates of students in the Computer Science Bachelor's Degree program at the Federal University of Santa Maria, offered in the Distance Education modality within the scope of the Open University of Brazil. The experiment was conducted using Artificial Neural Networks, based on data from students in the aforementioned course, aiming to calculate the prediction of dropout in the current class. The course is in its third offering, having started its activities in 2017. Considering the two classes that have already ended (entrants from 2017 and 2020/2021), the database consisted of 201 students. After training, the Artificial Neural Networks showed a 28.89% dropout prediction rate for the current class, which consists of students entering in 2024. Including the students from the 2024 class who have already dropped out (64 students, representing 58.72%), and if the Neural Networks's prediction is confirmed by the end of this class, we will have a dropout rate of 70.64%, a high percentage, but below the dropout rate of previous classes. The dropout rate for the 2017 incoming classes is 71.17%, and for the 2020/2021 class it is 83.33%. The Artificial Neural Network implemented can be applied to predict dropout rates in future classes of the course and also in other courses offered in the distance learning modality at UFSM.

Keywords: *Dropout rates. Distance education. Artificial neural networks.*

1. Introdução

A evasão acadêmica no ensino superior constitui um desafio recorrente para instituições educacionais em diversos países, afetando tanto o planejamento institucional quanto as trajetórias formativas dos estudantes. Estudos clássicos sobre o tema indicam que a evasão resulta da interação entre fatores acadêmicos, sociais e institucionais que influenciam o nível de integração do estudante à vida universitária. Nesse contexto, o modelo teórico proposto por Vincent Tinto destaca que a permanência dos estudantes no ensino superior está fortemente relacionada à sua integração acadêmica e social no ambiente institucional (Tinto, 1975).

No caso da Educação a Distância (EaD), o fenômeno da evasão apresenta características próprias e, frequentemente, índices mais elevados do que aqueles observados em cursos presenciais. Fatores como dificuldades de organização do tempo, limitações no acompanhamento pedagógico e desafios relacionados à autonomia dos estudantes podem influenciar a permanência ao longo do curso (Brenner et al., 2021; Pereira et al., 2017). Soso, Kampff e Machado (2023) apontam alguns indicadores, ligados à gestão institucional, que podem contribuir para a permanência dos estudantes em cursos da modalidade de EaD, tais como a integração social, existência de uma equipe multidisciplinar, recursos didáticos e infraestrutura adequada. Diante desse cenário, torna-se relevante o desenvolvimento de estratégias institucionais capazes de identificar precocemente estudantes com maior probabilidade de evasão, possibilitando a implementação de ações voltadas à permanência estudantil.

Neste contexto, o presente trabalho apresenta um estudo sobre os índices de evasão no Curso de Licenciatura em Computação da Universidade Federal de Santa Maria (UFSM), empregando técnicas de Redes Neurais Artificiais (RNAs) para a predição desse fenômeno. O curso é ofertado pela UFSM (Campus Frederico Westphalen/RS) em parceria com a Universidade Aberta do Brasil (UAB) (CAPES, 2026), na modalidade de Educação a Distância.

Atualmente, o curso é ofertado em polos de apoio presencial localizados nas cidades de Balneário Pinhal, Canguçu, Cruz Alta, Hulha Negra, Sapiranga e São Gabriel, todos situados no Estado do Rio Grande do Sul. O objetivo do curso é formar professores para atuar na área de Computação na Educação Básica (Ensino

Fundamental e Médio), bem como na formação profissional em Computação no nível técnico. Dessa forma, a estrutura curricular contempla disciplinas tanto da área de Computação quanto da área de Educação (UFSM, 2026).

Nos últimos anos, diversas pesquisas têm investigado o uso de técnicas de Aprendizado de Máquina (Machine Learning) e Inteligência Artificial (IA) para a análise de dados educacionais, especialmente com o objetivo de prever a evasão acadêmica em cursos de ensino superior. Esses estudos utilizam diferentes algoritmos de classificação, como regressão logística, árvores de decisão e RNAs, aplicados a dados acadêmicos, demográficos e de interação dos estudantes com Ambientes Virtuais de Aprendizagem (AVAs). Os resultados indicam que modelos preditivos podem contribuir para a identificação precoce de estudantes em risco de evasão, possibilitando o desenvolvimento de estratégias institucionais voltadas à permanência estudantil (Bouihi et al., 2024; Roda-Segarra, De-la-Penã, Mengual-Andrés, 2024; Duro et al., 2026; Vaarma; Li, 2024).

As RNAs são modelos computacionais inspirados no funcionamento do cérebro humano, compostos por neurônios artificiais interconectados que aprendem a partir do ajuste de pesos sinápticos com o objetivo de minimizar erros (Haykin, 2009). Esses modelos pertencem à abordagem conhecida como sistemas conexionistas e foram desenvolvidos com base em analogias com sistemas biológicos, nos quais um grande número de células nervosas opera de forma paralela para processar informações. Em termos computacionais, as RNAs são formadas por diversos elementos de processamento simples, interconectados por meio de pesos ajustáveis. O processo de aprendizado ocorre pela modificação desses pesos, permitindo que o modelo aprenda padrões presentes nos dados (Lorenzi; Silveira, 2011).

Desta forma, o presente estudo investiga a aplicação de RNAs para a predição da evasão acadêmica no Curso de Licenciatura em Computação da UFSM/UAB, utilizando dados institucionais de estudantes ingressantes em diferentes ofertas do curso. O objetivo do trabalho é analisar o potencial dessas técnicas para identificar padrões associados ao abandono do curso e contribuir para o desenvolvimento de estratégias institucionais voltadas à permanência e ao sucesso acadêmico dos estudantes. Para isso, foi realizado um experimento utilizando dados acadêmicos de estudantes ingressantes nas turmas de 2017 e 2020/2021, os quais foram utilizados para o treinamento de uma RNA. Em seguida, o modelo treinado foi aplicado aos dados da turma ingressante em 2024, com o objetivo de realizar a predição de possíveis casos de evasão.

2. Metodologia

O presente trabalho se caracteriza como uma pesquisa de abordagem quantitativa, com finalidade descritiva, baseada na análise de dados acadêmicos de estudantes do Curso de Licenciatura em Computação da UFSM/UAB. De acordo com Sampieri, Collado e Lucio (2013), a pesquisa quantitativa fundamenta-se na coleta e análise de dados numéricos com o objetivo de identificar padrões, relações entre variáveis e tendências observáveis em uma população ou amostra. Nesse sentido, a abordagem quantitativa foi adotada para possibilitar a mensuração e predição das taxas de evasão.

Os dados para a realização do experimento foram obtidos por meio da emissão de relatórios no Portal Acadêmico da UFSM, uma ferramenta disponibilizada para as Coordenações de Cursos, na opção “Consultar Estudante”. Os relatórios foram emitidos por polo de oferta do curso, pois cada polo é armazenado como um curso diferente, com um código de curso próprio. Como o curso tem oferta em 15 polos, os dados foram emitidos em 15 relatórios separados. O experimento baseia-se nos dados das turmas já concluídas (ingressantes de 2017 e 2020/2021) para predizer a evasão dos ingressantes da turma atual (ingressantes de 2024). Sendo assim, a turma dos alunos ingressantes de 2024 (ofertada em 6 polos da UAB) não foi utilizada para a entrada de dados da RNA mas foi utilizada para fazer a predição, após o aprendizado da Rede Neural. Realizou-se, então, a emissão dos relatórios das turmas de 15 polos, sendo gerados 15 relatórios separados.

Após a emissão dos relatórios, os mesmos foram exportados no formato CSV (Comma Separated Values), para que pudessem ser utilizados no experimento empregando RNAs. O portal Acadêmico da UFSM possui esse recurso de exportação no formato CSV. Os relatórios trazem os dados relativos à cota de ingresso,

ano de ingresso, ano da evasão, data de nascimento e percentual de conclusão do curso. A variável ano da evasão não foi empregada no modelo implementado, visando evitar o vazamento de dados (data leakage) (Becker; Recamonde-Mendoza, 2025). Qualquer estudante que não evadiu tem essa variável vazia (zero), fazendo com que a mesma carregue diretamente a resposta da variável de saída ("Evadido: sim ou não"). Desta forma, a RNA poderia facilmente aprender que qualquer valor diferente de zero nesta entrada equivale matematicamente a uma evasão confirmada.

O percentual de conclusão do curso é uma variável importante para identificar casos de evasão precoce, permitindo diferenciar o comportamento de alunos que se matriculam e desistem quase imediatamente (sem concluir nenhuma disciplina ou abandonando o curso com menos de 15% de progresso) daqueles que evadem em etapas mais avançadas. As variáveis relativas à data de nascimento (para calcular a idade dos alunos), gênero e cota de ingresso permitem identificar se existe alguma relação entre a evasão e o perfil socioeconômico e demográfico dos alunos, já que estes fatores atuam de forma combinada com o progresso pedagógico na determinação da persistência acadêmica (Filgueiras, 2023; Oliveira, 2010; Silva Júnior; Silva, 2024).

As cotas de ingresso na UFSM compreendem (UFSM, 2024): AC: Ampla Concorrência, para pessoas que não se encaixam nas cotas ou que não desejam participar do SiSU por meio delas; LB-PPI: Para candidatos autodeclarados pretos, pardos ou indígenas, com renda familiar bruta per capita igual ou inferior a 1 salário mínimo e que tenham cursado integralmente o Ensino Médio em Escolas Públicas; LB-Q: Para candidatos autodeclarados quilombolas, com renda familiar bruta per capita igual ou inferior a 1 salário mínimo e que tenham cursado integralmente o Ensino Médio em Escolas Públicas; LB-PCD: Para candidatos com deficiência, que tenham renda familiar bruta per capita igual ou inferior a 1 salário mínimo e que tenham cursado integralmente o Ensino Médio em Escolas Públicas; LB-EP: Para candidatos com renda familiar bruta per capita igual ou inferior a 1 salário mínimo que tenham cursado integralmente o Ensino Médio em Escolas Públicas; LI-PPI: Para candidatos autodeclarados pretos, pardos ou indígenas, independentemente da renda, que tenham cursado integralmente o Ensino Médio em Escolas Públicas; LI-Q: Para candidatos autodeclarados quilombolas, independentemente da renda, que tenham cursado integralmente o Ensino Médio em Escolas Públicas; LI-PCD: Para candidatos com deficiência, independentemente da renda, que tenham cursado integralmente o Ensino Médio em Escolas Públicas; LI-EP: Para candidatos que, independentemente da renda, tenham cursado integralmente o Ensino Médio em Escolas Públicas.

Em um segundo momento foi necessário unir todos os arquivos CSV, gerados por curso, em dois arquivos: um com os dados das turmas de 2017 e 2020/2021 (para aprendizado da RNA) e um segundo arquivo com os dados da turma de 2024 (para realizar a predição). Os arquivos foram reunidos em planilhas, utilizando-se o Microsoft Excel.

A partir dos dados reunidos, iniciou-se o pré-processamento dos dados, onde se identificou que os alunos concluintes (formados), das turmas de ingressantes de 2017 e 2020/2021 constavam como evadidos. Sendo assim, adicionou-se uma coluna na planilha contendo a informação “Evadido (sim ou não)”, tomando-se o cuidado para não incluir os alunos formados como evadidos. Outro item relevante que não constava dos dados era o gênero dos alunos. Incluiu-se, então, mais uma coluna, contendo o gênero (masculino ou feminino). A data de nascimento foi empregada para calcular a idade dos alunos, visando utilizar essa informação nas análises posteriores. As cotas de ingresso tinham nomenclaturas diferentes, pois os ingressantes eram dos anos de 2017, 2020/2021 e 2024 e houve uma alteração nos códigos das cotas (UFSM, 2024). Sendo assim, todas as cotas foram revisadas para que seguissem a nomenclatura atual. Após o pré-processamento dos dados, as planilhas criadas foram exportadas em formato CSV, para que pudessem ser utilizadas pela RNA.

Para a criação das RNAs foi preciso que todos os dados fossem representados de forma numérica. Os dados para uma RNA precisam ser numéricos (Luger, 2004), já que as RNAs são modelos matemáticos baseados em operações algébricas (multiplicações e somas de pesos e entradas). Portanto, qualquer informação, seja

texto, imagem ou som, deve ser convertida em números (vetores ou tensores) para que a rede possa processar e aprender.

Desta forma, os dados foram assim formatados: as cotas receberam um código de 1 a 7, sendo assim numeradas: AC (1), LB-EP (2), LI-EP (3), LB-PPI (4), LI-PPI (5) e LI-PCD (6). Não havia nenhum aluno ingressante nas demais cotas (LB-Q, LI-Q e LB-PDC). O gênero foi convertido para 0 (zero) correspondente ao gênero masculino e 1 (um) para o gênero feminino. Além disso, a coluna ano de nascimento foi excluída, já que foi utilizada para calcular a coluna idade.

Visando assegurar a privacidade dos alunos, todos os dados foram anonimizados. Informações tais como os números de matrícula e nomes dos alunos foram omitidos na geração dos arquivos CSV transformados. Com relação à utilização de dados educacionais, o Referencial de IA na Educação destaca a importância da privacidade e da proteção de dados pessoais:

> A incorporação de sistemas de inteligência artificial nos contextos educacionais requer a observância rigorosa de princípios relacionados à privacidade, à proteção de dados pessoais e à segurança da informação (BRASIL, 2025, p. 49).

Os resultados obtidos por meio do experimento, bem como a discussão dos mesmos, são apresentados na próxima seção.

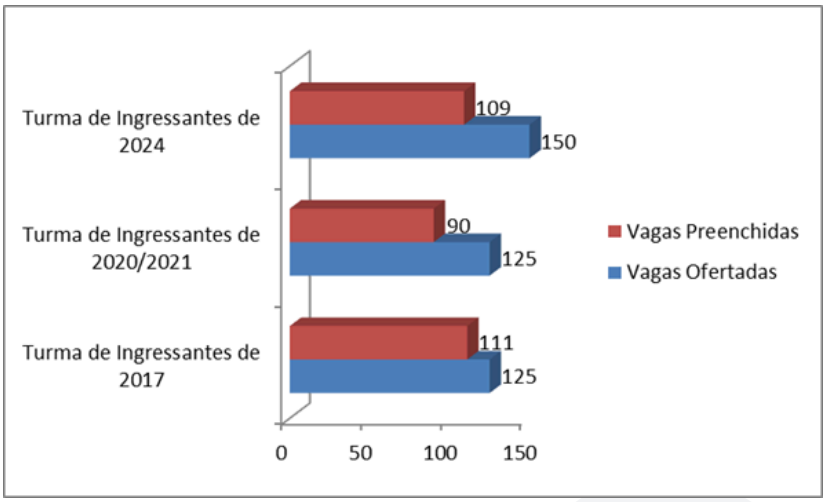
3. Resultados e Discussão

O conjunto de dados empregado no experimento é composto pelos dados de 310 alunos do Curso de Licenciatura em Computação da UFSM/UAB, sendo 111 ingressantes na turma de 2017, 90 ingressantes da turma de 2020/2021 e 109 ingressantes em 2024. Na turma 2020/2021, houve dois ingressos - um em novembro de 2020 e outro no início de 2021. Essa exceção deveu-se à Pandemia de COVID-19, em que o isolamento social dificultou a realização do processo seletivo para o ingresso dos estudantes.

Na primeira turma do curso, no ano de 2017, foram ofertadas 125 vagas, sendo 25 em cada um dos polos da UAB selecionados à época, todos no Estado do Rio Grande do Sul: Cruz Alta, Palmeira das Missões, Sarandi, Seberi e Três Passos. Das 125 vagas foram preenchidas 111 (88,8%). Na segunda turma, ofertada no final do ano de 2020 e início de 2021, novamente foram ofertadas 125 vagas, sendo 25 vagas para cada polo. Os polos de oferta da segunda turma foram nas cidades de Cachoeira do Sul, Constantina, Jacuizinho, Tio Hugo e Três de Maio. Na segunda turma, das 125 vagas ofertadas, 90 vagas foram preenchidas (72%). Infere-se que a redução no número de vagas preenchidas (matrículas) tenha sido em razão da Pandemia de COVID-19. Devido à pandemia e ao reduzido número de ingressantes na primeira chamada (novembro de 2020), houve um segundo ingresso na mesma turma (no início do ano de 2021). A terceira turma (que está em andamento), com ingresso em 2024, teve 150 vagas ofertadas em 6 polos do RS: Balneário Pinhal, Canguçu, Cruz Alta, Hulha Negra, São Gabriel e Sapiranga, tendo sido preenchidas 109 vagas (72,67%). A Figura 1 apresenta, de forma gráfica, o número de ingressantes e de vagas em cada uma das ofertas.

Para o desenvolvimento da primeira etapa do processo, que compreende a aprendizagem supervisionada da RNA, foram utilizados os dados dos alunos ingressantes nos anos de 2017 e 2020/2021, por meio de um arquivo CSV gerado conforme definido na seção anterior (Metodologia). O arquivo CSV empregado possui os dados de 201 alunos.

Figura 1 – Gráfico das Vagas Ofertadas e Preenchidas em cada Oferta



Fonte: Os autores (2026)

Para a criação da RNA o arquivo CSV foi dividido em dois: um arquivo com as informações de entrada (cota, ano de ingresso, percentual vencido do curso, gênero e idade), Figura 2 (a), e um arquivo contendo apenas o resultado da classificação da RNA, indicando se o aluno evadiu ou não, Figura 2 (b).

Figura 2: Recorte dos Arquivos gerado para uso na RNA (a) – Entradas (b) Saídas

	A	B	C	D	E	F	G
1	Cota	Ano Ingresso	Evadido	% Vencido	Ano Nascimento	Gênero	Idade
2	1	2017	0	100.00	1981	0	45
3	1	2017	1	30.27	1993	1	33
4	2	2017	1	32.53	1990	1	36
5	1	2017	0	100.00	1999	0	27
6	1	2017	1	32.08	1977	1	49
7	1	2017	0	100.00	1985	0	41
8	1	2017	1	37.50	1988	1	38
9	1	2017	1	0.00	1997	0	29
10	1	2017	1	0.00	1962	0	64
11	3	2017	0	100.00	1997	0	29
12	1	2017	1	0.00	1983	0	43
13	1	2017	1	10.84	1999	0	27
14	1	2017	0	100.00	1983	0	43
15	1	2017	1	10.84	1994	1	32

(a)

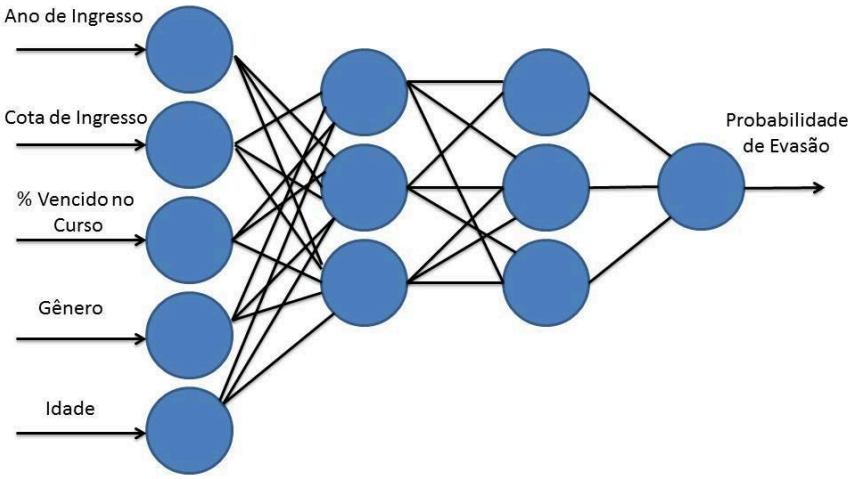
	A	B
1	Evadido	
2	0	
3	1	
4	1	
5	0	
6	1	
7	0	
8	1	
9	1	
10	1	
11	0	

(b)

Fonte: Os autores (2026)

Modelou-se uma RNA multicamadas (MLP MultiLayer Perceptron) (Haykin, 2009), com 5 neurônios na camada de entrada (representando as informações de entrada), duas camadas intermediárias (ou camadas ocultas), cada uma com 3 neurônios e uma camada de saída com apenas um neurônio (representando a probabilidade de evasão). A arquitetura da rede foi definida considerando o número de variáveis disponíveis na base de dados e a natureza do problema de classificação binária (evasão ou não evasão). Optou-se por uma estrutura relativamente simples, com duas camadas ocultas e três neurônios em cada camada, buscando equilibrar a capacidade de aprendizagem do modelo com a dimensão limitada do conjunto de dados utilizado. A Figura 3 apresenta a estrutura da RNA.

Figura 3: Modelo da RNA implementada



Fonte: Os autores (2026)

A camada de entrada conta com os dados utilizados para identificar a possibilidade de evasão dos alunos, que são os dados contidos na base de dados da UFSM: cota de ingresso, ano de ingresso, percentual vencido no curso, gênero e idade do aluno. A saída é a probabilidade de evasão. Nesse contexto, a RNA tem, por objetivo, o de realizar a predição da evasão, com base nos dados dos alunos. Segundo Luger (2004), tarefas relacionadas à predição estão entre as possíveis aplicações de uma RNA.

A RNA foi construída com tensores, que são estruturas de dados fundamentais na IA e aprendizado profundo (Deep Learning), funcionando como arrays multidimensionais (Busson et al., 2018). Tensores são as estruturas de dados utilizadas para representar e manipular dados (como imagens, textos e sons) em modelos de RNAs. Os dados dos arquivos no formato CSV foram convertido para o formato de tensores, por meio da codificação da RNA, implementada na linguagem de programação Python.

O Aprendizado Profundo (Deep Learning) utiliza RNAs com múltiplas camadas, permitindo o aprendizado automático de características complexas, a partir de grandes volumes de dados. É uma evolução do aprendizado de máquina (GoodFellow; Bengio; Courville, 2016). Deep Learning é uma subárea da IA que utiliza RNAs com múltiplas camadas para aprender padrões complexos.

O Aprendizado Profundo tenta simular, de forma simplificada, o funcionamento do cérebro humano usando neurônios artificiais, camadas ocultas, ajuste automático de pesos e grandes volumes de dados. Quanto mais camadas a Rede Neural possui, mais profundo é o modelo.

O processo de treinamento é uma das partes mais importantes das RNAs, cuja função é modificar os pesos sinápticos da RNA de forma ordenada para alcançar o objetivo de processo desejado. A modificação dos pesos pode ser realizada no modo padrão ou no modo batch. No modo padrão a correção dos pesos é feita a cada apresentação de um exemplo do conjunto à RNA. No modo batch a apresentação é feita por ciclo, ou seja, a correção é feita quando todos os exemplos do conjunto são apresentados à rede (Russell; Norvig, 2021). A implementação empregada neste experimento utiliza o modo batch, com grupos de 10 registros a cada ciclo.

A RNA deste experimento implementa o algoritmo de Backpropagation (retropropagação). Este algoritmo caracteriza-se por realizar a correção dos pesos, da camada de saída para a camada de entrada da RNA. Segundo Luger (2004), a retropropagação fornece um algoritmo para atribuir aos neurônios a sua "parcela de culpa" pelo erro da RNA e ajustar os pesos correspondentemente. O algoritmo inicia pela camada de saída e propaga o erro retroativamente através das camadas ocultas, ajustando os pesos. O processo de aprendizagem busca, então, o melhor conjunto de pesos da RNA, por meio de uma adaptação por correção de erros, buscando minimizar a diferença entre a soma ponderada das entradas pelos pesos (saída calculada pela RNA) e a saída desejada, ou seja, o erro da resposta atual da rede (Haykin, 2009; Russell; Norvig, 2021). As funções de ativação utilizadas foram as funções ReLU e Sigmóide que serão explicadas posteriormente.

A RNA foi implementada utilizando-se a linguagem de programação Python, com diferentes bibliotecas: Pandas: biblioteca Python de código aberto, rápida e flexível, essencial para manipulação e análise de dados, oferecendo o acesso a estruturas de dados tais como DataFrame (tabelas) e arrays, facilitando a limpeza, transformação e exploração de grandes conjuntos de dados e a leitura de diversos formatos de arquivos como, por exemplo, o formato CSV empregado neste experimento; Scikit-learn: principal biblioteca de aprendizado de máquina (Machine Learning) de código aberto na linguagem Python; Seaborn: focada na criação de visualizações de dados estatísticos, tais como gráficos e mapas de calor, sendo empregada para visualização da matriz de confusão; Pytorch (torch): uma das bibliotecas mais populares para aprendizado de máquina e Deep Learning.

A partir do arquivo CSV com os dados dos ingressantes de 2017 e 2020/2021, totalizando 201 alunos, a RNA foi treinada por meio da aprendizagem supervisionada. Na aprendizagem supervisionada os valores de entrada e de saída desejados para a RNA são fornecidos por um supervisor/professor externo. Segundo Luger (2004), a aprendizagem de uma RNA surge por meio de um processo de aprendizado (ou adaptação) pelo qual as conexões entre os neurônios são ajustadas, por meio de diferentes camadas.

Para realizar o treinamento utilizou-se a validação cruzada. O conjunto de dados foi dividido em 5 partes (ou partições), garantindo que cada parte mantivesse a mesma proporção de classes do total. Optou-se pela validação cruzada tendo-se em vista que a base de dados é pequena (201 alunos) e, uma única divisão aleatória (base de testes e validação) poderia concentrar padrões muito fáceis de classificar no conjunto de validação, inflando artificialmente a acurácia. Cada uma das 5 partições preservou a proporção estatística de 76,6% de alunos evadidos e 23,4% de alunos ativos/formados. Esta consistência proporcional impede que a RNA sofra de distorção de aprendizado induzida pelo desbalanceamento de classes, garantindo uma validação cientificamente correta e livre de vazamentos (data leakage). As 5 partições foram assim constituídas: partições 1, 2, 3 e 4 com 40 alunos cada, sendo 31 evadidos e 9 não-evadidos, cada uma destas quatro partições representando 19,9% da base e a 5ª partição com 41 alunos, sendo 30 evadidos e 11 não-evadidos, representando 20,4% da base. A 5ª partição tem 1 aluno a mais do que as outras, devido ao número total de alunos da base ser ímpar (201 alunos).

Para cada uma das 5 partições, treinou-se uma nova instância da RNA, a partir do subconjunto de treinamento correspondente. Ao término de cada iteração, os parâmetros ajustados foram descarregados, impedindo qualquer transferência de aprendizado (knowledge transfer) ou retenção de memória entre as partições (GoodFellow; Bengio; Courville, 2016). Dessa forma, as métricas consolidadas do experimento (acurácia média de 89,55% e sensibilidade média de 93,57%) refletem a média do desempenho de 5 RNAs totalmente autônomas, atestando a capacidade real de generalização do modelo e a ausência de viés metodológico.

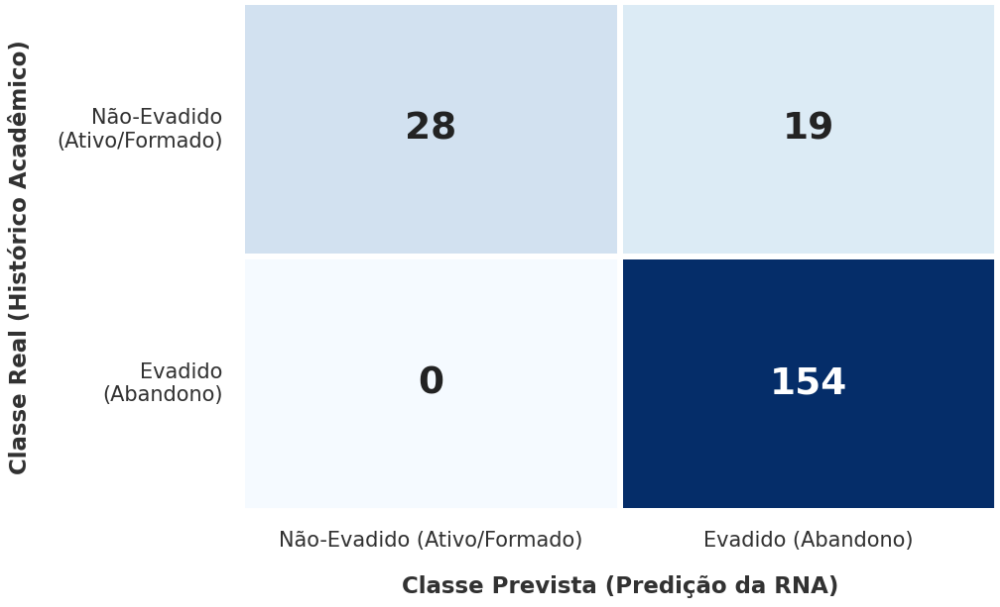
A avaliação da RNA foi realizada por meio de métricas consolidadas, tais como matriz de confusão, acurácia, precisão, recall e F1-Score. Essas métricas são amplamente reconhecidas como fundamentais para validar a eficácia de modelos de classificação. A acurácia em uma RNA é uma métrica de avaliação utilizada em tarefas de classificação para medir a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões realizadas, representando a taxa de acerto do modelo, indicando quão frequentemente a rede neural acerta a classe correta (evadido ou não no caso deste experimento). A fórmula da acurácia é calculada com base na divisão do número de previsões corretas da RNA pelo total de previsões.

A matriz de confusão (apresentada na Figura 4) é uma tabela que avalia o desempenho de modelos de classificação em Machine Learning, comparando os valores reais com as previsões feitas pelo algoritmo. A matriz de confusão possui quatro resultados: VP (Verdadeiro Positivo), quando o conteúdo avaliado é verdadeiro e o modelo trouxe como resultado um valor positivo, indicando que o modelo acertou – no caso do modelo implementado o resultado compreende a identificação correta do aluno evadido ou não; FN (Falso Negativo), quando houve erro do modelo (previu a classe negativa quando o valor real era positivo); VN (Verdadeiro Negativo), quando o conteúdo avaliado é negativo e o modelo trouxe como resultado um valor negativo, indicando acerto do modelo e FP (Falso Positivo), quando o resultado da avaliação da RNA é positivo e

o modelo trouxe como resultado um valor negativo, indicando erro do modelo (Pinho; Gaspar; Sassi, 2024; Ramos et al., 2018). De acordo com a Figura 4, temos VP igual a 154, VN igual a 28, FP são 19 e FN 0 (nenhum aluno evadido passou despercebido).

Os valores da soma da diagonal principal da matriz de confusão (28 + 154 na Figura 4) representam os acertos da RNA: quando o valor da saída era 0 (zero) e a RNA indicou 0 e quando era 1 e a RNA também indicou corretamente. Com base na matriz de confusão, a acurácia da RNA foi de 90,55%. Cabe destacar que este resultado é diferente da acurácia média da avaliação cruzada apresentada anteriormente (89,55%). Essa diferença ocorre devido ao método estatístico de agregação dos resultados em cada um dos dois cálculos, já que medem o desempenho da RNA sob perspectivas diferentes: a acurácia global (matriz de confusão) é obtida somando diretamente todos os acertos da matriz de confusão divididos pelo total absoluto de alunos testados na base, ou seja, trata todas as 201 previsões realizadas ao longo das 5 rodadas como se fossem um único grande teste consolidado. A acurácia média das partições (89,55%) é a média aritmética simples das taxas de acurácia obtidas individualmente em cada um dos 5 subconjuntos de teste, avaliando o desempenho médio do algoritmo atribuindo peso igual a cada uma das 5 rodadas de teste. A diferença de um ponto percentual entre a acurácia da matriz de confusão e da validação cruzada deve-se ao fato de que a base tem um número total ímpar (201 alunos) e uma das partições ficou com 41 alunos (no cálculo global da matriz essa parte tem um peso ligeiramente maior).

Figura 4: Matriz de Confusão



Fonte: Os autores (2026)

Além da acurácia, para validar a RNA, utilizaram-se as métricas de precisão, recall e F1-score. A precisão é a métrica que mede a proporção de previsões positivas corretas em relação ao total de previsões positivas feitas pelo modelo. O cálculo da precisão é realizado por meio da divisão dos valores de VP pela soma de VP + FP, apresentados na matriz de confusão (Figura 4). De acordo com os dados da matriz de confusão, a precisão da RNA foi de 89,02%.

O recall (ou sensibilidade) mede a proporção de casos positivos reais que o modelo conseguiu identificar corretamente. O recall é calculado pelo valor obtido em VP dividido pela soma de VP + FN. De acordo com os resultados apresentados na matriz de confusão (Figura 4), a RNA obteve um recall de 100%.

O F1-score é uma métrica de avaliação resultante da combinação da precisão e do recall em um único valor, por meio da média harmônica. Esta métrica é indicada para não ocorrerem erros provenientes de dados desbalanceados. Por exemplo, se uma base de dados tem 95% de exemplos de uma classe e 5% de outra, a

métrica da acurácia pode resultar em valores artificialmente inflados. Sendo assim, a métrica F1-score mostra se a RNA realmente aprendeu a encontrar a classe minoritária. A RNA atingiu 94,19% nesta métrica, aplicando-se a fórmula 2 x (precisão x recall) / (precisão + recall).

Os valores apresentados anteriormente foram calculados diretamente a partir dos resultados apresentados na matriz de confusão (Figura 4). Considerando-se a avaliação cruzada realizada em 5 partes, tem-se a média aritmética das métricas calculadas individualmente em cada uma das partes, com os seguintes resultados: acurácia média (89,55%), precisão média (91,73%), recall médio (93,57%) e F1-score médio (92,56%).

Apesar dos resultados obtidos indicarem elevada capacidade de classificação, estudos futuros poderão ampliar o conjunto de dados e realizar novas validações do modelo, a fim de avaliar sua capacidade de generalização em diferentes contextos. Pode-se inferir que os percentuais alcançados nas métricas empregadas devam-se ao tamanho reduzido da base de dados empregada, já que o experimento foi realizado apenas com os alunos de um dos cursos da universidade e que o referido curso ofertou, até o momento, apenas 3 turmas desde a sua criação.

Após o treinamento e testes de validação, os dados dos alunos ingressantes de 2024 (que ainda não evadiram) foram submetidos à avaliação pela RNA. Para realizar a avaliação foram excluídos dos 109 alunos os que já evadiram, pois existem muitos casos de evasão precoce (alunos que se matriculam e não concluem nenhuma disciplina ou que desistem do curso com menos de 15% do curso concluído). A predição foi realizada, então, com o conjunto dos dados de 45 alunos, que estão, atualmente, cursando as disciplinas do quinto semestre do curso.

A variável “percentual de conclusão do curso” refere-se ao acúmulo de créditos integralizados pelo discente até o encerramento do período letivo imediatamente anterior à consulta. No caso da turma ativa de 2024, a extração dos dados via Portal Acadêmico ocorreu após a conclusão do 4º semestre do curso (os dados foram extraídos no início do ano de 2026 e os alunos da turma de 2024 integralizaram as disciplinas do 4º semestre do curso no final de 2025), garantindo que os percentuais de integralização fossem resultantes do histórico acadêmico já consolidado (1º ao 4º semestre). Como a informação é extraída diretamente dos registros oficiais em tempo real, sem depender de desempenhos futuros do semestre em andamento, fica assegurada a total disponibilidade temporal do dado no momento da realização da predição.

Ao realizar a entrada destes dados na RNA, a mesma indicou a previsão de 13 evasões neste grupo, representando 28,89%. Se incluirmos os alunos da turma de 2024 que já evadiram (64 alunos, representando 58,72% dos alunos ingressantes em 2024) e se a predição da RNA se confirmar até o final desta turma, teremos um percentual de evasão de 70,64%, um percentual alto, mas abaixo do índice de evasão das turmas anteriores. O índice de evasão das turmas de ingressantes de 2017 é de 71,17% e de 2020/2021 é de 83,33%.

4. Conclusões

A RNA implementada, com base nos dados empregados, atingiu 89,55% de acurácia nos testes realizados considerando-se a validação cruzada, além dos índices elevados nas demais métricas empregadas (precisão, recall e F1-score). Apesar dos resultados obtidos indicarem elevada capacidade de classificação, estudos futuros poderão ampliar o conjunto de dados e realizar novas validações do modelo, a fim de avaliar sua capacidade de generalização em diferentes contextos. Acredita-se que o modelo possa ser empregado, então, na predição da evasão de turmas futuras da Licenciatura em Computação EaD da UFSM. Podem ser incluídas novas características na RNA, buscando-se mais dados disponíveis nas bases da universidade. Desta forma, pode-se criar novos modelos de RNAs com mais características (neurônios de entrada), buscando refinar a predição da evasão.

O presente trabalho não abordou as causas e/ou motivos da evasão, o que pode ser realizado em trabalhos futuros empregando ou não RNAs. Dados do Censo da Educação Superior (INEP, 2024) indicam um

percentual de evasão de 41,6% nos cursos ofertados na modalidade de EaD. No caso específico dos cursos da área de Computação, estudos revelam que as dificuldades relacionadas aos processos de aprendizagem na área de programação de computadores fazem com que os índices de evasão sejam maiores. O ensino da programação de computadores não é fácil e, devido a isso, muitas universidades têm discutido, com frequência, os currículos de seus cursos da área de Computação, em busca de alternativas para diminuir o índice de evasões dos cursos desta área (Castro et al., 2003; Garlet et al., 2018; Hoed, 2017; Souza et al., 2018). Além das dificuldades relacionadas à área de Computação, na modalidade a distância, os processos de construção do conhecimento se dão, em muitos casos, de forma autônoma (Brenner et al., 2021), exigindo uma maior maturidade e resiliência por parte dos alunos. A falta de autonomia é um dos motivos da evasão em cursos ofertados na modalidade de EaD (Pereira et al., 2017).

Com relação ao emprego de IA, como a RNA criada e empregada neste estudo, destaca-se que a IA pode analisar bases de dados educacionais para prever a evasão. Dessa forma, os resultados obtidos neste estudo indicam que técnicas de IA, como as RNAs, podem contribuir para a análise de dados educacionais e para a identificação de padrões associados à evasão acadêmica. A utilização de modelos preditivos baseados em dados institucionais pode apoiar gestores e equipes pedagógicas no desenvolvimento de estratégias de acompanhamento e intervenção voltadas à permanência estudantil. Segundo Filgueiras (2023), a aplicação de IA em políticas públicas educacionais pode otimizar processos, personalizar o aprendizado, identificar padrões e fornecer insights relevantes para a tomada de decisões. Nesse sentido, espera-se que os resultados apresentados possam contribuir para futuras pesquisas sobre o uso de técnicas de IA aplicadas à análise de dados educacionais e à melhoria dos processos de permanência e sucesso acadêmico no ensino superior.

Referências Bibliográficas

BECKER, A. E.; RECAMONDE-MENDOZA, M. Mind the Gap: Investigating the Impact of Data Leakage on Machine Learning Predictive Models. **Brazilian Conference on Intelligent Systems – BRACIS**, Anais. Porto Alegre: Sociedade Brasileira de Computação, 2025. Disponível em: https://sol.sbc.org.br/index.php/bracis/article/view/40816. Acesso em: 14 ago. 2026.

BOUIHI, B. *et al.* Prediction of Higher Education Student Dropout based on Regularized Regression Models. **Engineering, Technology & Applied Science Research**, v. 14, n. 6, p. 17811-17815, 2024. Disponível em: https://etasr.com/index.php/ETASR/article/view/8644. Acesso em: 21 abr. 2026.

BRASIL. Ministério da Educação. **Referencial de Inteligência Artificial na Educação**. Brasília: MEC, 2025. Disponível em: https://www.gov.br/mec/pt-br/referencial-de-ia-na-educacao. Acesso em: 23 mar. 2026.

BRENNER, C. E. B. et al. **Educação a Distância na UFSM**: histórias e perspectivas. Santa Maria: UAB/UFSM, 2021. Disponível em: https://repositorio.ufsm.br/handle/1/25833. Acesso em: 28 jun. 2025.

BUSSON, A. *et al.* Desenvolvendo Modelos de Deep Learning para Aplicações Multimídia no Tensorflow. **Anais do XXIV Simpósio Brasileiro de Sistemas Multimídia e Web**: Minicursos. p.67-116, 2018. Disponível em: https://www.researchgate.net/publication/328341764_Desenvolvendo_Modelos_de_Deep_Learning_para_Aplicacoes_Multimidia_no_Tensorflow. Acesso em: 20 abr. 2026.

CAPES. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. **Universidade Aberta do Brasil (UAB).** Brasília: CAPES, 2026. Disponível em: https://www.gov.br/capes. Acesso em: 15 abr. 2026.

CASTRO, C. T. et al. Utilizando Programação Funcional em Disciplinas Introdutórias de Computação, In**: Anais do XI Workshop de Educação em Computação** – WEI, Campinas/SP, 2003.

DURO, B. *et al.* Machine learning and deep learning for dropout prediction in higher education: a review. **Computers**, v. 15, n. 3, p. 164, 2026. Disponível em: https://www.mdpi.com/2073-431X/15/3/164. Acesso em: 21 abr. 2026.

FILGUEIRAS, F. Artificial Intelligence and Education Governance. **Sage Journals**, v.19, n.3, 2023. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/17461979231160674. Acesso em: 20 ago. 2025.

GARLET, D.; BIGOLIN, N. M.; SILVEIRA, S. R. Ensino de Programação de Computadores na Educação Básica: um estudo de caso. **Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica**, v.9, n.2, 2018. Disponível em: http://periodicos.unifacef.com.br/index.php/resiget/article/view/1604/1144. Acesso em: 10 mai., 2024.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge: MIT Press, 2016.

HAYKIN, S. **Neural networks and learning machines**. 3. ed. Prentice Hall, 2009.

HOED, R. M. **Análise da Evasão em Cursos Superiores**: o caso da evasão em cursos superiores da área de computação. Brasília: UnB – Programa de Pós-graduação em Computação Aplicada. (Dissertação de Mestrado), 2017. Disponível em: http://repositorio.unb.br/handle/10482/22575. Acesso em: 15 abr. 2026.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Censo da Educação Superior**. Disponível em: https://www.gov.br/inep/pt-br/centrais-de-conteudo/noticias/censo-da-educacao-superior/inep-divulga-resultado-do-censo-superior-2024. Acesso em: 15 abr. 2026.

LORENZI, F.; SILVEIRA, S. **Desenvolvimento de Sistemas de Informação Inteligentes**. Porto Alegre: Editora do UniRitter, 2011.

LUGER, G. F. **Artificial intelligence**: structures and strategies for complex problem solving. 5. ed. Boston: Pearson, 2004.

OLIVEIRA, D. A. **Políticas Educacionais no Brasil**. Petrópolis: Vozes, 2010.

PEREIRA, A. S. et al. **Metodologia da Aprendizagem em EaD**. Santa Maria/RS: UAB/UFSM, 2017.

PINHO, C. M. A.; GASPAR, M. A.; SASSI, R. J. Aplicação de Técnicas de Inteligência Artificial para Classificação de Fuga ao Tema em Redações. **Educação em Revista UFMG**, v. 40, 2024. Disponível em: https://www.scielo.br/j/edur/a/YPPVRRKGdLdFbXTQXMmZgCc/?lang=pt. Acesso em: 13 ago. 2026.

RAMOS, J. L. C. *et al.* Um Estudo Comparativo de Classificadores na Previsão da Evasão de Alunos em EaD. **Brazilian Symposium on Computers in Education**, 2019. Anais. São Paulo: Sociedade Brasileira de Computação, 2018. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/41553. Acesso em: 13 ago. 2026.

RODA-SEGARRA, J.; DE-LA-PEÑA, C.; MENGUAL-ANDRÉS, S. Effectiveness of artificial intelligence models for school dropout: a meta-analysis. **REMIE – Multidisciplinary Journal of Educational Research**, v. 14, n. 3, 2024 (October). Disponível em: https://hipatiapress.com/hpjournals/index.php/remie/en/article/view/13342. Acesso em: 21 abr. 2026.

RUSSELL, S.; NORVIG, P. **Artificial intelligence**: a modern approach. 4. ed. Pearson, 2021.

SAMPIERI, R. H.; COLLADO, C. F.; LUCIO, M. P. B. **Metodologia de pesquisa**. 5. ed. Porto Alegre: Penso, 2013.

SILVA JÚNIOR, T.M.; SILVA J. E. Aplicação de um modelo de machine learning para prever a evasão escolar. **Revista E & S Estratégias e Soluções**. 2024. Disponível em: https://revistaes.com.br/artigos/aplicacao-de-um-modelo-de-machine-learning-para-prever-a-evasao-escolar. Acesso em: 14 ago. 2026.

SOSO, F. S.; KAMPFF, A. J. C.; MACHADO, K. G. W. Evasão e Permanência na Educação Superior a Distância na América Latina. **Eventos Pedagógicos**, 14(1); 124-143, 2023. Disponível em: https://www.researchgate.net/publication/371159491_EVASAO_E_PERMANENCIA_NA_EDUCACAO_SUPERIOR_A_DISTANCIA_NA_AMERICA_LATINA. Acesso em: 21 abr. 2026.

SOUZA, N. G.; SILVEIRA, S. R.; PARREIRA, F. J. Proposta de uma Metodologia para apoiar os Processos de Ensino e de Aprendizagem de Lógica de Programação na Modalidade de Educação a Distância. **ECCOM – Educação,**

Cultura e Comunicação, v. 9, 2018. Disponível em: http://unifatea.com.br/seer3/index.php/ECCOM/article/view/851. Acesso em maio, 2020.

TINTO, V. Dropout from higher education: a theoretical synthesis of recent research. **Review of Educational Research**, v. 45, n. 1, p. 89–125, 1975. Disponível em: https://www.jstor.org/stable/1170024. Acesso em: 20 abr. 2026.

UFSM. Universidade Federal de Santa Maria. **Curso de Licenciatura em Computação** – informações institucionais. Santa Maria: UFSM, 2026. Disponível em: https://www.ufsm.br/frederico. Acesso em: 2026.

UFSM. Universidade Federal de Santa Maria. **Política de ações afirmativas e sistema de cotas**. Santa Maria: UFSM, 2024. Disponível em: https://www.ufsm.br/pro-reitorias/prograd/sisu/documentos-cotas. Acesso em: 15 abr. 2026.

VAARMA, M.; LI, H. Predicting student dropouts with machine learning: an empirical study in finnish higher education. **Technology in Society**, v. 76, 2024 (March). Disponível em: https://www.sciencedirect.com/science/article/pii/S0160791X24000228. Acesso em: 20 abr. 2026.

PRLO