SISTEMA INTELIGENTE PARA PREVISÃO DE EVASÃO ESCOLAR EM UM AMBIENTE UNIVERSITÁRIO

INTELLIGENT SYSTEM FOR PREDICTING SCHOOL EVASION IN A UNIVERSITY ENVIRONMENT

Leandra Cristina Cavina Piovesan Soares [1]
Rafael Lima de Carvalho [2]

Resumo: *O abandono escolar ou evasão escolar é um dos problemas existentes no sistema de ensino brasileiro e que atinge seus mais variados níveis. Especificamente ao Ensino Superior, este problema é considerado de âmbito internacional, tornando-o ainda mais impactante. Em geral, as universidades sabem as taxas gerais de evasão por curso, e ainda que saibam exatamente os alunos evadidos, predizer quais poderão ainda evadir é uma tarefa complexa. Neste sentido, este trabalho apresenta uma metodologia de solução que se utiliza de modelos computacionais inteligentes, especificamente o aprendizado de máquina para aprender os padrões dos alunos evadidos e não-evadidos. Isto significa, que desde o primeiro período a solução permite inferir a probabilidade de o aluno evadir. A pesquisa foi realizada com os dados acadêmicos da Universidade Estadual do Tocantins (Unitins). Os resultados obtidos destacam uma precisão elevada na classificação de alunos propensos à evasão, evidenciando a eficácia da metodologia adotada neste estudo.*

Palavras-chave: *Evasão Escolar Universitária. Mineração de Dados. Aprendizagem de Máquina. Inovação Educacional.*

Abstract: *School dropout is one of the problems that exist in the Brazilian education system and reaches the most varied levels. Specifically in Higher Education, this problem is considered international in scope, making it even more impactful. In general, universities know the general dropout rates per course, and even if they know exactly which students dropped out, predicting which students might still drop out is a complex task. In this sense, this work presents a solution methodology that uses intelligent computational models, specifically machine learning to learn the patterns of dropout and non-dropout students. This means that from the first period the solution allows you to infer the probability of the student dropping out. The research was carried out with academic data from the State University of Tocantins (Unitins). The results achieved reveal high accuracy in classifying students prone to dropout, demonstrating the effectiveness of the methodology adopted in this study.*

Keywords*: University School Dropout. Data Mining. Machine Learning. Educational Innovation.*

1 Mestra em Propriedade Intelectual e Transferência de Tecnologia para a Inovação pela Universidade Federal do Tocantins (UFT). Especialista em Gestão Pública e Qualidade no Serviço pela Universidade Estadual do Tocantins (Unitins), Especialista em Gestão da Tecnologia da Informação pela Faculdade Albert Einsten de Brasília (FALBE). Graduada em Administração com Ênfase em Sistema de Informação pela Universidade Paulista (UNIP). Atualmente é professora na Unitins e coordenadora do Curso de Sistemas de Informação da Unitins. Lattes: http://lattes.cnpq.br/0505525976660596. ORCID: https://orcid.org/0000-0003-0347-9160. E-mail: leandra.cc@unitins.br

2 Doutor em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (UFRJ). Mestre em Sistemas e Computação pelo Instituto Militar de Engenharia (IME). Graduado em Ciência da Computação pela Universidade Federal do Tocantins (UFT). Lattes: http://lattes.cnpq.br/0175648235036864 ORCID: https://orcid.org/0000-0002-5296-8641. E-mail: rafael.lima@uft.edu.br.

Introdução

A temática sobre a evasão escolar faz parte de debates e reflexões entre o meio acadêmico (SABBATINI, 2015; ADACHI, 2017; CARVALHO, 2017; SOUZA, 2017; GUIMARÃES et al, 2019; GOMES et al., 2019; SANTOS JUNIOR; REAL, 2019). Estudos como Tinto (1975) e Brasil (1996a) auxiliam no entendimento dos motivos pelos quais os estudantes deixam o seu curso, reforçando a necessidade de uma maior investigação visando a identificação das causas e motivos associados. Associa-se de forma generalista, que a evasão está relacionada ao curso, mas não necessariamente este seja o problema.

Nota-se que a evasão escolar é composta por diversas variáveis que se interagem e se conflitam em torno da problemática. Estudos como de Souza (2008), Baggi e Lopes (2011), Barlem et al. (2012) relatam as possíveis motivações que podem levar o aluno a abandonar o seu curso. Segundo Souza (2008) a motivação para a evasão não se limita apenas ao âmbito acadêmico, causas podem originar-se de aspectos pessoais, sociais e ambientais.

Para os autores Baggi e Lopes (2011), a análise do abandono escolar exige uma análise das circunstâncias do passado e que pode exercer influência sobre a desistência de um curso superior. No entendimento dos autores Barlem (2012) vários motivos levam o estudante a evadir-se de um curso, sendo ele pelo próprio universo discente, como imaturidade, desconhecimento do curso, dificuldades de adaptação, problemas financeiros e/ou familiares.

Neste contexto, autores como Kira (1998), Gaioso (2005), Baggi e Lopes (2011) definem a Evasão Escolar como a interrupção abrupta dos estudos independentemente do nível escolar. Os autores Schargel e Smink (2002) indicam que a evasão escolar é dividida em categorias de causa: psicológicas, sociológicas, organizacionais, interacionais e as econômicas. Além disso, os autores argumentam que o combate para fenômeno será efetivo, somente quando houver uma abordagem sistêmica, já que o problema não é isolado.

Enquanto que em Soares et al. (2020), mostrou-se a possibilidade de treinar um sistema aprender os padrões de alunos evadidos ou não, o presente trabalho reporta uma metodologia de um modelo por período. Esta metodologia permitiu a geração de um modelo inteligente por período escolar de maneira a tornar o sistema de inferência mais preciso e orientado.

Assim, o objetivo do presente estudo é apresentar os resultados do modelo de inteligência artificial treinado para inferir a probabilidade de o aluno evadir, por período de matrícula. A pesquisa reportada fez uso dos dados da Universidade Estadual do Tocantins (Unitins).

Neste sentido, a solução aqui apresentada permitirá os gestores escolares tomarem decisões mais orientadas aos indivíduos mais propensos a evadir, possibilitando a minimização do problema de evasão. Além do mais, este estudo mostra que as abordagens oferecem *insights* valiosos que, quando aplicados adequadamente, têm o potencial de mitigar significativamente o desafio da evasão escolar.

Desafios da Evasão Escolar no Ensino Superior Brasileiro

A mitigação da taxa de desistência e o prolongamento da permanência dos alunos nas Instituições de Ensino Superior (IES) representam desafios significativos para o cenário educacional brasileiro (INEP, 2019). Percebe-se que estudos anteriores já haviam sidos realizados, como no ano de 1996 com a criação da Comissão Especial de Estudos sobre Evasão nas Universidades Públicas Brasileiras (CEUPB) que contava com a participação das Instituições de Ensino Superior Públicas, Federais e Estaduais.

Dentre os resultados divulgados pela CEUPB, foi reforçado a necessidade de empreendimento de mais esforços sobre a temática, pois considera que este fenômeno é complicado e igualitário para as Instituições de Ensino Superior (IES), necessitando de estudos e análises, já que este fenômeno é provocado por diversas variáveis (BRASIL, 1996a).

Neste aspecto, estudos buscam dar ênfase não apenas pelas razões pelo qual o aluno evadiu, mas também buscam a compreensão por meio de ações preventivas estimulando-os a permanecer no sistema. Assim, para a Administração Pública, a evasão e a continuação por longo tempo dos estudos, são obstáculos que causam reflexos institucionais e sociais, levando-se em

Revista Humanidades e Inovação - ISSN 2358-8322 - Palmas - TO - v.10, n.09

conta os investimentos em recursos humanos e financeiros (PEREIRA; ZAVALA; SANTOS, 2011; BAGGI; LOPES, 2011).

Segundo os autores Rigo et al. (2014) a evasão escolar está presente tanto nas universidades públicas e privadas e o resultado são questões financeiras com a redução de número de alunos formados no ensino superior, causando impacto na cadeia produtiva do nosso país. Lobo (2012) indica que os estudos sobre a evasão, requer um modelo de política de governo orientada a qualidade do ensino.

Corroborando com Lobo (2012) a Associação de Mantenedoras de Ensino Superior (ABMES), reforça também sobre a necessidade da criação de políticas governamentais, com ações orientadas para a qualidade do ensino e uso dos recursos públicos e privados, sendo conduzidos para a promoção de processos e análises direcionadas na realização de atividades (HORTA, 2012).

Nos estudos de Martinho (2014) foi desenvolvido um sistema preditivo inteligente, para prever o risco de evasão dos alunos dos cursos Superiores de Tecnologia (CST) em Automação e Controle da Universidade Federal do Mato Grosso (IFMT). As informações acadêmicas compreendidas para a análise foram dos anos de 2004-2 a 2011-2. As técnicas utilizadas são da Inteligência Artificial, especificamente Redes Neurais Artificiais (*ARTMAP-Fuzzy*) e os resultados indicaram altos índices de acertos do grupo de alunos propensos a evadir, atingindo um percentual de 95% e 100% de acurácia e uma média global de 95% de acertos.

Os dados do CENSO também podem ser fonte de pesquisas para este tipo de fenômeno. Estudos como de (COLPAN, 2019) utilizou os dados do CENSO de 2017 e como resultado descobriu que o Estado do Pará é o maior em índices de evasão no Brasil. Em sua pesquisa, foi diagnosticado que a média de evasão escolar no Estado é de 20%. Do percentual estão um uma escala de 38% a 65% dos alunos com idade acima da recomendada que é de 18 anos, previstas na Lei 9.394/96 (BRASIL, 1996b).

Assim sendo, este fenômeno no ensino superior é prioritário e que precisa ser superado. Os trabalhos que busquem responder a estas indagações são essenciais para o estabelecimento de um campo de discussão, auxiliando as Instituições de Ensino Superior nas políticas educacionais, contribuindo assim, para uma educação superior de qualidade no para o nosso país.

Portanto, propor soluções que busquem meios de identificar de forma proativa e acurada o grupo dos alunos propensos a evadir tornam-se cada vez mais importante para as IES.

Metodologia

A metodologia que foi adotada neste estudo é de natureza quantitativa, abrangendo o processo de coleta, análise e interpretação dos dados acadêmicos com o objetivo de identificar relações entre variáveis e características causais relacionadas à evasão escolar. Como embasamento teórico, foi realizada uma pesquisa bibliográfica, consultando materiais elaborados e publicados em livros e artigos científicos.

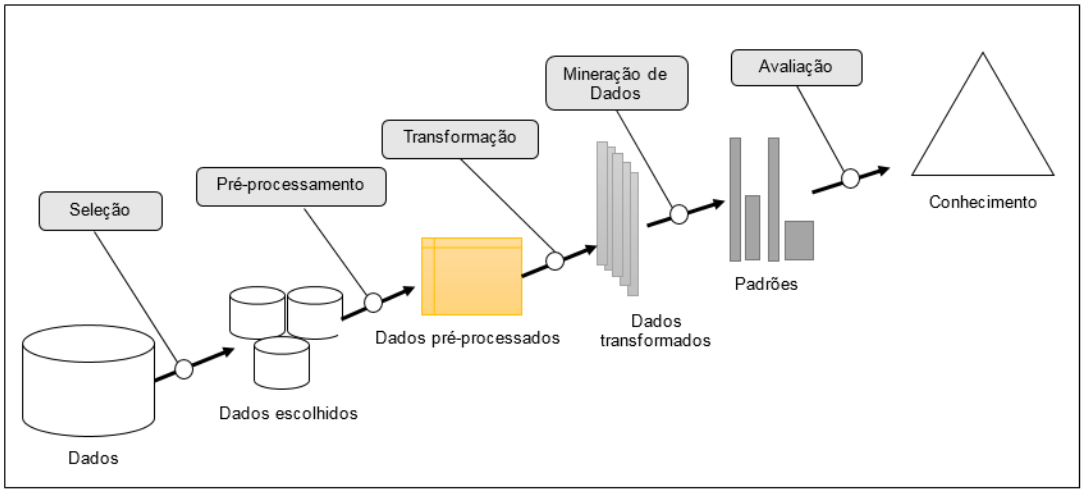
Para a análise dos dados, utilizou-se ferramentas computacionais avançadas, incluindo técnicas de Mineração de Dados Educacionais (MDE) e métodos de Aprendizado de Máquina (AM). Essas abordagens proporcionaram uma compreensão mais aprofundada dos padrões subjacentes nos dados, permitindo identificar indicadores significativos de evasão e suas relações com diversas variáveis.

Segundo os autores Rigo et al. (2014) com o aumento substancial do uso das tecnologias de informação e comunicação (TIC) fez com que aumentasse substancialmente as bases de dados, fazendo com que a capacidade de geração dos dados fosse maior do que a conhecimento dos pesquisadores e analistas. Este cenário também está presente no ambiente acadêmico, onde as instituições de ensinos possuem uma gama de sistemas informatizados, como ambiente de educação a distância, sistemas acadêmicos e comunidades virtuais.

Para a pesquisa utilizou-se de ferramentas tecnológicas como: *Spyder*, linguagem de programação *Python*, f*rameworks Scikit-learn* e Pandas. Para a extração do conhecimento na base de dados ou *Knowledge Discovery in Databases* (KDD), aplicou-se algoritmos de análises e descobertas de dados sob controles de eficiência computacional aceitável. As fases desenvolvidas para a descoberta do conhecimento em banco de dados é: a seleção, o pré-processamento, a

transformação, a mineração de dados e avaliação (FAYYAD et al.,1996). A figura 1 apresenta o detalhamento deste processo do KDD.

Figura 1. Estágio do processo de KDD

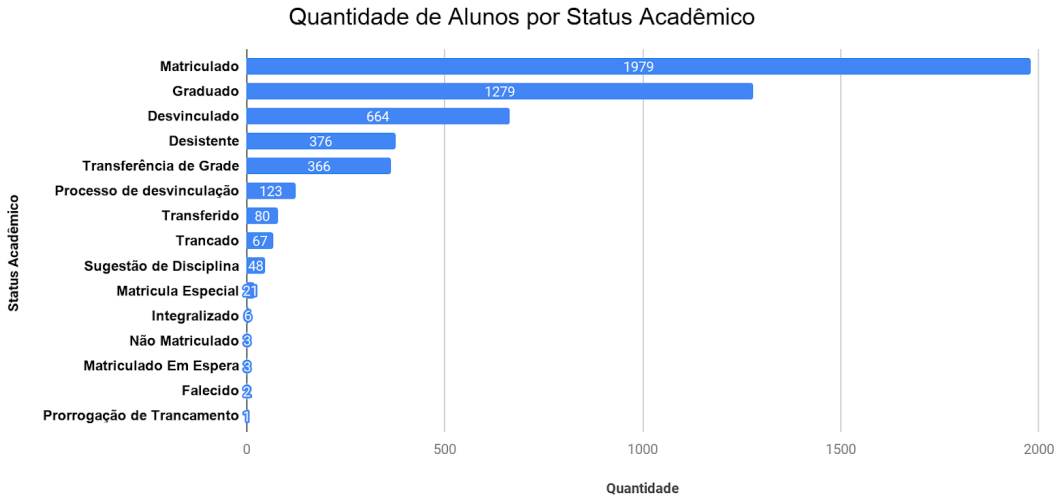


Fonte: Elaborado pelos autores com base em Fayyad et. al. (1996).

Para a aferição e treinamento dos classificadores, foram utilizados 12 (doze) cursos presenciais, referente aos câmpus de Araguatins, Augustinópolis, Dianópolis e Palmas, onde o período de análise foi entre os períodos letivos de 2010-2 até 2019-2. As tabelas utilizadas foram: Cadastro de alunos, Matrizes curriculares e Histórico acadêmico. Após análise das tabelas, foram selecionadas 18 (dezoito) variáveis, sendo: RA, idade, sexo, estado civil, cor/raça, estado natal, naturalidade, UF, cidade, tipo de instituição, ano de formação do 2º grau, semestre de ingresso, tipo do ingresso, código da matriz curricular, último semestre cursado, câmpus, curso e turno.

A partir da extração dos dados da tabela "Cadastro de Alunos", chegou-se nos quantitativos de alunos por status acadêmico. O status acadêmico é um dado importante, pois é através dele que serão identificados os alunos que farão parte dos testes e treinamentos dos classificadores. A figura 2 (dois) apresenta os quantitativos de alunos por status acadêmico. A seguir serão apresentados os status acadêmicos mais relevantes para a pesquisa.

Figura 2. Alunos por tipo de status acadêmico



Fonte: Elaborado pelos autores (2020).

Posteriormente, realizou-se a seleção de dados com base nos status acadêmicos mais expressivos para a pesquisa, incluindo Graduado, Desvinculado e Desistente. Foi aplicado um filtro ao conjunto de 5.018 (cinco mil e dezoito) registros, resultando na escolha de 2.319 (dois mil e trezentos e dezenove) alunos que atendiam aos critérios estabelecidos.

Com o intuito de aprimorar a clareza dos dados, introduziu-se um novo status denominado "Evadido", resultante da combinação dos status Desistente e Desvinculado. Adicionalmente, o status anteriormente denominado Graduado foi renomeado para "Formado" a fim de evitar possíveis ambiguidades com os dados originais.

A etapa subsequente envolveu o pré-processamento dos dados, que incluiu o preenchimento dos valores ausentes nos atributos. Para o atributo idade, os dados não informados foram substituídos pela média da idade. Quanto aos demais atributos nulos ou vazios, foram categorizados como "não informado". Adicionalmente, os dados não explícitos foram extraídos do histórico acadêmico e da matriz curricular. Essa abordagem visou identificar totalizadores como disciplinas da matriz, disciplinas cursadas, disciplinas aprovadas e disciplinas reprovadas, levando em consideração médias e faltas. Para melhor exemplificar essa etapa, a tabela 1 apresenta os quantitativos de dados por atributos de classe.

Tabela 1. Atributos com valores nulos

Qtde de atributos com valores nulos	**Etapa de pré-processamento dos dados**					
	Qtde de registros da Classe "Evadido"	**% Classe Evadido**	**Qtde de registros da Classe "Graduado"**	**% Classe Graduado**	**Total das classes**	**% Total das classes**
Nenhum	495	47,60%	730	57,08%	1225	52,85%
1	455	43,75%	407	31,82%	862	37,19%
2	72	6,92%	37	2,89%	109	4,70%
3	16	1,54%	103	8,05%	119	5,13%
4	0	0,00%	2	0,16%	2	0,09%
5	1	0,10%	0	0,00%	1	0,04%
6	1	0,10%	0	0,00%	1	0,04%
Total de registros	**1040**		1279		2319	

Fonte: Elaborado pelos autores (2020).

Nota-se na tabela 1 (um), que o maior percentual equivale a coluna "nenhum". Percebe-se que 52,85% dos atributos pertencentes a esta classe, estão com dados preenchidos, não sendo necessário o manuseio e/ou preenchimento de dados ausentes. A partir da tabela 1 (um), foi construída a tabela 2 (dois) que apresenta os atributos que possuem maiores quantitativos nulos.

Tabela 2. Atributos com maiores quantidades nulas

Atributo	Descrição	Qtde de registros da Classe "Evadido"	Qtde de registros da Classe "Graduado"	Total das classes
CORRACA	Cor e raça	520	463	983
ANOULTIMAINST	Último ano cursado no ensino médio	56	105	161
ULTIMO_SEMESTRE	Último semestre cursado	0	103	103
ESTADOCIVIL	Estado civil	67	24	91
TOTAL		643	695	1338

Fonte: Elaborado pelos autores (2020).

Observando-se os números do atributo CORRACA é o que possui um maior quantitativo de alunos, que estão sem esse dado preenchido. Esse reflexo é decorrente da não obrigatoriedade desse dado, já que somente a partir de 2017 foi homologado o parecer de número CNE/CEB Nº: 4/2017, que instituiu as Instituições de Ensino Básica e Superior a inclusão da raça/cor para o uso nos Censos Educacionais (MEC, 2017). A partir de então, a Unitins incluiu este atributo como obrigatório nos processos de matrícula para os alunos novatos e de rematrícula para os veteranos. Os demais atributos mesmo sendo importantes, observa-se que a maioria se encontra na classe graduada.

Solução Proposta

O Aprendizado de Máquina (AM) é um sistema capaz de adquirir conhecimentos de forma automática (MITCHELL, 1997). Neste estudo utilizou-se o algoritmo de Floresta Aleatória (*Random Forests*) o AM como método na fase da mineração de dados. O treinamento do algoritmo a base de dados foi dividida em dois conjuntos aleatórios, ficando 70% dos dados para treinamento, e 30% foram separados para validação do treinamento. Para o tipo de aprendizado indutivo o escolhido foi o aprendizado supervisionado.

A fase de validação cruzada, adotou-se o algoritmo *Grid Search*, que cria uma grade multidimensional com diversas combinações de parâmetros a serem avaliados. No contexto do algoritmo de Florestas Aleatórias, foram testados dois parâmetros: o critério de entropia e a quantidade de estimadores (árvores na floresta). O critério de entropia padrão foi a entropia (*entropy*), comparado com o critério de Gini. A variação dos estimadores foi testada no intervalo de 10 (dez) a 120 (cento e vinte), valores determinados empiricamente e dependentes do problema específico.

Durante a etapa de validação cruzada, buscou-se a combinação de parâmetros que maximizasse a acurácia do modelo. As métricas utilizadas incluíram a Matriz de Confusão, que revela a distribuição das classes no modelo avaliado; a *Receiver Operating Characteristic* (ROC) e a Área sob a Curva ROC (ROC-AUC), que avaliam a qualidade geral do modelo.

No contexto desta pesquisa, o Verdadeiro Positivo (VP) representa a quantidade de alunos que evadiram e foram corretamente identificados pelo modelo como evadidos. O Verdadeiro Negativo (VN) refere-se aos alunos que não evadiram e foram corretamente classificados como não evadidos. Por outro lado, o Falso Positivo (FP) corresponde aos alunos erroneamente classificados como evadidos, embora não tenham evadido. O Falso Negativo (FN) representa os alunos que evadiram, mas foram incorretamente classificados como não evadidos.

Nesse contexto, minimizar o FN é crucial, uma vez que a não identificação de alunos que evadiram pode resultar na ausência de ações preventivas. Com base nos valores de VP, FP, VN e FN, constrói-se a matriz de confusão, apresentada na Figura 3.

Figura 3. Matriz de confusão das classes (Evadido e Não Evadido)

Condição Atual / Predição	**Evadido (P)**	**Não Evadido (N)**
Aluno Evadido (P)	VP (Verdadeiro Positivo)	FP (não evadiu, mas não é extraído)
Aluno não Evadido (N)	FN (evadido, mas não é extraído)	VN (Verdadeiro Negativo)

Fonte: Elaborado pelos autores com base em (RASCHKA, 2015).

A métrica de acurácia ou exatidão (AC) quantifica o percentual de previsões corretas em relação ao desempenho global do modelo. A acurácia é calculada pela soma das previsões corretas, dividida pelo número total de previsões, como expresso na seguinte equação:

$$AC = \frac{VP+VN}{FP+FN+VP+VN} \qquad (1)$$

Devido ao desbalanceamento das bases, observam-se disparidades significativas entre o número de amostras em cada classe. Nesse cenário, é comum recorrer a métricas adicionais, tais como Precisão, Revocação, F1-score e a curva ROC. Para melhor interpretar essas métricas, é essencial estabelecer a Taxa de Falsos Positivos (FPR) e a Taxa de Verdadeiro Positivo (TPR), definidas pelas equações (2) e (3), respectivamente. Essas taxas fornecem insights cruciais sobre as amostras positivas que foram identificadas corretamente em relação ao conjunto total de amostras positivas.

$$FPR = \frac{FP}{N} = \frac{FP}{FP+VN} \qquad (2)$$

$$TRP = \frac{VP}{P} = \frac{VP}{FN+VP} \qquad (3)$$

Assim, a Precisão (PRE) é calculada como a razão entre o número de Verdadeiros Positivos (VP) e a soma de Verdadeiros Positivos (VP) e Falsos Positivos (FP), conforme definido na equação (4). A métrica PRE avalia se aqueles classificados como positivos são, de fato, corretos. Por outro lado, a Revocação ou Recall (REC), expressa na equação (5), mensura a completude dos resultados, sendo, na verdade, sinônimo de Taxa de Verdadeiros Positivos (TPR). O Recall é especialmente útil em situações em que os Falsos Negativos (FN) são considerados mais prejudiciais do que os Falsos Positivos (FP).

$$PRE = \frac{VP}{VP+FP} \qquad (4)$$

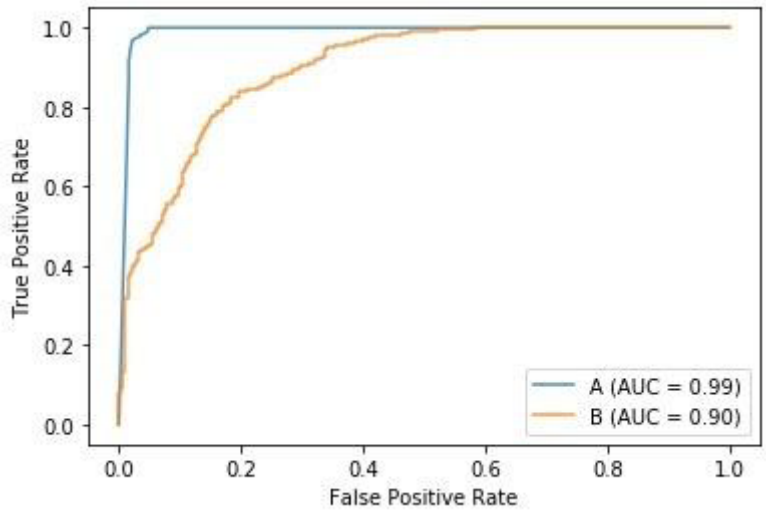
$$REC = TRP = \frac{VP}{P} = \frac{VP}{FN+VP} \qquad (5)$$

Para consolidar as informações de Precisão (PRE) e Revocação (REC) em uma única medida, recorremos à métrica F1-Score, expressa na equação (7). O F1-Score representa uma média harmônica entre as duas métricas, possibilitando a avaliação abrangente da qualidade do modelo por meio de um único valor. Desse modo, valores elevados de F1 indicam uma maior qualidade no desempenho do modelo avaliado.

A avaliação da qualidade de um processo de classificação em termos de sensibilidade e especificidade frequentemente utiliza a curva ROC como uma ferramenta estatística (MARTÍNEZ-CAMBLOR; PARDO-FERNÁNDEZ, 2019). Essa curva representa graficamente a relação entre a Taxa de Verdadeiros Positivos (TPR) e a Taxa de Falsos Positivos (FPR). A Figura 4 ilustra a curva ROC gerada para dois classificadores, A e B.

No gráfico da Figura 4, o classificador A é interpretado como superior ao B, conforme indicado pela curva ROC. Uma simplificação dessa comparação é obtida através do cálculo da área sob a curva, conhecida como ROC-AUC, que naturalmente varia de 0 (zero) a 1 (um) (RASCHKA, 2015).

Figura 4. Curva AUC (A e B)



Fonte: Soares et al. (2020).

Portanto, a avaliação do modelo de classificação pode ser conduzida por meio de diversas métricas, adaptadas às características específicas dos dados. É essencial considerar, durante a avaliação de modelos de classificação, fatores como a proporção de dados em cada classe e o objetivo específico da precisão almejada.

Resultados

Baseado nos treinamentos dos classificadores para os métodos aprendidos, apresenta-se os resultados obtidos por meio do Sistema Inteligente. As informações geradas pela solução inteligente, permitiu a identificação da predisposição da evasão escolar relacionada aos câmpus e cursos. As métricas de avaliação empregadas para o modelo adotado foram: Acurácia, Precisão, Revogação, F1-*Score* e ROC-AUC.

Os resultados apresentados nas tabelas 3 e 4, foram gerados a partir de um modelo aprendido dos alunos evadidos e não evadidos, permitindo assim, a inferência da probabilidade do aluno evadir deste o primeiro período do curso. Para os resultados o modelo considerou apenas os alunos matriculados no segundo semestre de 2019, ou seja, um total de 1979 alunos ativos que estavam cursando disciplinas no semestre letivo de 2019-2.

Tabela 3. Propensão de evasão escolar por câmpus

Câmpus	Qtde de Alunos Propensos a Evasão	Qtde de Alunos Predispostos a Formar	Total Alunos Matriculados	Propensão de Evasão
Palmas	682	156	838	81,38%
Araguatins	147	94	241	60,69%
Augustinópolis	314	220	534	58,80%
Dianópolis	168	198	366	45,90%

Fonte: Elaborado pelos autores (2020)

Tabela 4. Cursos com maiores índices de propensão de evasão

Câmpus	Curso	Qtde de Alunos Propensos a Evasão	Qtde de Alunos Predispostos a Formar	Total Alunos Matriculados	Propensão de Evasão
Palmas	Sistemas de Informação	127	10	137	92, 7%
	Serviço Social	135	29	164	82,31%
	Direito	230	62	292	78,76%
	Engenharia Agronômica	190	55	245	77,55%
Augustinópolis	Enfermagem	113	54	167	67,66%

Fonte: Elaborado pelos autores (2020).

Nota-se que o câmpus de Palmas é o maior em índices de chances de evasão, em seguida o câmpus de Araguatins. Além disso, percebe-se que dos 12 cursos presenciais estudados, os cursos de Palmas são os que estão no *ranking* de evasão. Chama-se a atenção para o curso de Sistemas de Informação com a baixa probabilidade de graduação dos alunos matriculados neste curso.

Como forma de avaliar os resultados obtidos pelo Sistema Inteligente, apresenta-se a tabela 6 que expõe os resultados obtidos a partir do conjunto de treinamento que utilizou 70% da base de dados. Para tal, foi utilizado o algoritmo de Floresta Aleatória pela técnica de validação. Observa-se que a média da acurácia entre os períodos é de 95,21% para os subconjuntos gerados pelo processo de validação cruzada, com subdivisão balanceada de 10 conjuntos.

Isto significa que a parte de 70% foi dividida em 10 subconjuntos e, portanto, 10 rodadas foram feitas. Em cada rodada, 9 partes foram usadas para treinamento e 1 para testes, gerando-se,

portanto, 10 classificadores de floresta aleatória. A Tabela 5 exibe as médias das métricas utilizadas nos 10 classificadores.

Tabela 5. Resultados dos treinamentos a partir do 1º período dos cursos presenciais

Métricas de Aprendizado de Máquina	Resultados por período							
	1º Período	2º Período	3º Período	4º Período	5º Período	6º Período	7º Período	8º Período
Acurácia	93,9	94,3	94,1	94,7	94,8	96,2	96,5	97,2
Precisão	93,1	92,7	92,3	93,4	93,4	94,8	94,8	95,7
Revogação	96,5	97,8	97,8	97,6	97,8	98,7	99,3	99,6
F1	94,7	95,1	94,9	95,4	95,5	96,7	97	97,6
ROC-AUC	93,5	93,7	93,5	94,2	94,4	95,8	96,2	96,9

Fonte: Elaborado pelos autores (2020).

Para avaliação dos resultados foi utilizada a matriz de confusão, que está apresentada na tabela 6. A tabela 6 foi gerada por meio da validação cruzada e apresenta o quanto a metodologia utilizada nesta pesquisa foi eficaz em classificar corretamente os alunos evadidos e não evadidos. Os resultados evidenciam-se que a média dos períodos entre os Verdadeiro Negativo (VN) foi corretamente classificado em 78,1%. Isto significa que o classificador identificou corretamente os alunos não evadidos com uma precisão média de 78,1%.

Já para os Verdadeiro Positivo (VP), chamado de aluno evadido a média de acertos foi de 53,1%. É importante ressaltar que esta etapa serviu apenas para realizar o ajuste otimizado dos parâmetros (via um mecanismo chamado *Grid Search*) a serem utilizados no classificador final. O processo de *Grid Search* consiste em estipular um conjunto de parâmetros onde a combinação destes forma uma grid multidimensional. Cada ponto é uma combinação possível dos valores a serem explorados para otimização.

Tabela 6. Matriz de Confusão gerada a partir do treinamento dos períodos dos cursos

Matriz de Confusão - Validação Cruzada por Período	Resultados por período			
	VP	FP	FN	VN
1º Período	52,6	5,6	2,8	76,8
2º Período	52,1	6,1	1,7	77,9
3º Período	51,8	6,4	1,7	77,9
4º Período	52,8	5,4	1,9	77,7
5º Período	52,8	5,4	1,7	77,9
6º Período	54	4,2	1	78,6
7º Período	54	4,2	0,5	79,1

8º Período	54,7	3,5	0,3	79,3

Fonte: Elaborado pelos autores (2020).

A tabela 7 apresenta os resultados dos classificadores treinados por cada período, utilizando-se 70% de toda a base respectiva ao seu período. Os demais 30% foram utilizados para testar os classificadores finais, respeitando-se a informação do período. O algoritmo utilizado foi de Floresta Aleatória, com seus respectivos parâmetros otimizados pelo processo de *Grid Search.* O treinamento foi realizado utilizando-se os 70% da base de dados. Para a extração das medidas de qualidade, para cada período foi testado com 30% da base. É importante ressaltar que os 30% de teste são desconhecidos do processo de treinamento e, portanto, são dados não vistos pelo classificador.

Tabela 7. Resultado dos classificadores dos grupos de 70% de treinamento e 30% para teste

Métricas de Aprendizado de Máquina	Resultados por período							
	1º Período	2º Período	3º Período	4º Período	5º Período	6º Período	7º Período	8º Período
Acurácia	92,7	92,7	93,5	93,4	94,5	94,7	96,1	95,7
Precisão	91,4	90,2	92,2	91,7	92,5	92,3	93,9	93,9
Revogação	96,4	97,9	97	97,3	98,5	99,1	99,7	99,1
F1	93,8	93,9	94,5	94,4	95,4	95,6	96,7	96,4
ROC-AUC	92	91,7	92,9	92,6	93,8	93,8	95,4	95,1

Fonte: Elaborado pelos autores (2020).

O resultado gerado pela média entre os períodos para a acurácia é de 94,16%. Nota-se que os resultados obtidos na validação de 30% da base de dados, obtêm-se os resultados semelhantes com os alcançados durante o treinamento dos 70% dos dados. Pelos resultados, podemos concluir que o preditor se ajustou de tal forma, que conseguiu extrair o padrão representado por alunos evadidos e não-evadidos com uma precisão bastante elevada.

Além disso, a Tabela 8 exibe a matriz de confusão, evidenciando os resultados do grupo de treinamento de 70% e do conjunto de testes de 30%. Esta matriz reflete a precisão na classificação da situação do acadêmico, indicando se evadiu ou não. Nessa análise, destaca-se o desempenho positivo dos métodos empregados nesta pesquisa ao alcançar valores significativos de Verdadeiros Positivos (VP) para alunos evadidos e Verdadeiros Negativos (VN) para alunos não evadidos.

Tabela 8. Matriz de Confusão do grupo 70% de treinamento de 30% para teste

Matriz de Confusão - Validação Cruzada por Período	Resultados por período Grupos de 70 e 30%			
	VP	FP	FN	VN
1º Período	218	31	12	330
2º Período	213	36	7	335

3º Período	221	28	10	332
4º Período	219	30	9	333
5º Período	222	27	5	337
6º Período	221	28	3	339
7º Período	227	22	1	341
8º Período	227	22	3	339

Fonte: Elaborado pelos autores (2020).

Ao analisar os resultados apresentados na Tabela 8, destaca-se que nos 30% aleatoriamente selecionados, os quais não foram utilizados durante a fase de treinamento, a quantidade de Falsos Negativos (FN) é nula. Ao observar a matriz de confusão na Tabela 7, referente aos dados de treinamento, percebe-se que esse valor é inferior a 1%. Em outras palavras, nos 70% dos dados, nos quais a validação cruzada foi conduzida em k=10 grupos com o algoritmo treinado repetidamente em 10 iterações, utilizando 9 subgrupos para treinamento e 1 para testes, a média dos FN foi também próxima de zero.

Vale ressaltar que o Falso Negativo representa o cenário em que o aluno evadiu, mas foi erroneamente classificado como não evadido. Esse caso é crucial para a aplicação em consideração. Se o sistema rotular um aluno propenso à evasão como não evadido, os gestores educacionais podem perder a oportunidade de intervir antes que o aluno tome a decisão de evadir. Ao observar os Falsos Positivos, equívocos nesse aspecto indicariam um erro por excesso, pois esses casos também seriam considerados na formulação de políticas de enfrentamento ao problema. Ao analisar a Tabela 4, percebe-se que apenas 17 amostras se enquadram nessa situação, representando uma proporção de apenas 0,02% em relação ao total de amostras na base de dados de validação. Esse excedente é mínimo e, portanto, aceitável dada a relevância informativa do experimento.

Considerações Finais

A evasão universitária é um desafio abrangente que permeia todos os níveis de ensino, caracterizado por uma complexidade de fatores que nem sempre se restringem ao ambiente acadêmico. A aplicação de ferramentas de inteligência computacional, como as técnicas de Aprendizado de Máquina, tem desempenhado um papel significativo em estudos sobre evasão escolar, ao extrair informações e conhecimentos de bases de dados. Essa abordagem, por sua vez, pode subsidiar de maneira valiosa o processo de tomada de decisão das Instituições de Ensino Superior (IES).

Este estudo validou, por meio de um estudo de caso, a hipótese de que as técnicas de Aprendizado de Máquina podem ser aplicadas de maneira satisfatória na Mineração de Dados Educacionais. Os experimentos conduzidos utilizando a base de dados acadêmica da Universidade Estadual do Tocantins (Unitins) demonstraram que essas técnicas são eficazes. Os resultados obtidos indicam que métodos de Mineração de Dados Educacionais podem contribuir para o desenvolvimento de um classificador de alta confiança, automatizando a tarefa de identificar alunos propensos à evasão.

Ao segmentar os alunos em categorias específicas, é possível conduzir estudos mais focalizados, buscando identificar padrões que levam à evasão. Isso é particularmente relevante, considerando que muitas vezes um grupo reduzido de profissionais nos setores acadêmicos é responsável por entender as razões por trás da evasão e propor soluções.

Os resultados do experimento computacional apresentado neste estudo sugerem que é viável construir um modelo preditivo de evasão com base nos registros acadêmicos. Essa abordagem, ao incorporar tecnologias avançadas, lança luz sobre a possibilidade de orientar políticas públicas

para combater os prejuízos causados pela evasão no ensino superior. Conclui-se que os modelos de Inteligência Artificial podem formar uma base sólida de informações precisas, permitindo uma gestão mais inteligente, focada e eficaz no combate à evasão escolar no contexto universitário.

Referências

ADACHI, Ana Amélia Chaves Teixeira (2017). **Evasão de estudantes de cursos de graduação da USP – Ingressantes nos anos de 2002, 2003 e 2004. 2017.** 294p. Tese. (Doutorado em Educação) – Faculdade de Educação, Universidade de São Paulo, São Paulo-SP, 2017.

BAGGI, Cristiane Aparecida Dos Santos; LOPES, Doraci Alves. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. Avaliação: **Revista da Avaliação da Educação Superior.** Campinas, [s.n.], 2011, Disponível em: http://dx.doi.org/10.1590/s1414-40772011000200007. Acesso em: 26 mai 2020.

BARLEM, Jamila Geri Tomaschewski et al. (2012). Opção e evasão de um curso de graduação em enfermagem: percepção de estudantes evadidos. **Revista Gaúcha de Enfermagem.** [S.l: s.n.], 2012. Disponível em: http://dx.doi.org/10.1590/s1983-14472012000200019. Acesso em: 26 mai 2020.

BRASIL. Ministério da Educação. Secretaria de Educação Superior. Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras. **Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas**. 1996a. Disponível em: http://www.andifes.org.br/wp-content/files_flutter/Diplomacao_Retencao_Evasao_Graduacao_em_IES_Publicas-1996.pdf.Acesso em: 11 abr. 2020.

BRASIL, Lei nº 9.394, de 20 de dezembro de 1996. **Estabelece as diretrizes e bases da educação nacional.** Brasília, DF, dez 1996b

CARVALHO, Alessandro Pires (2017). **Fatores institucionais associados à evasão na educação superior.** 2017. 90 f. Dissertação. (Mestrado em Administração) Programa de Pós-Graduação em Administração, Universidade Federal de Goiás, Goiânia.

COLPANI, Rogério. Mineração de Dados Educacionais: um estudo da evasão no ensino médio com base nos indicadores do Censo Escolar. **Informática na educação:** teoria & prática. [S.l: s.n.], 2019. Disponível em: <http://dx.doi.org/10.22456/1982-1654.87880> Acesso em: 11 abr. 2020.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; **From Data Mining to Knowledge Discovery in Databases.** 1996. Disponível em: https://wvvw.aaai.org/ojs/index.php/aimagazine/article/view/1230. Acesso em: 28 fev 2020.

GAIOSO, N. P. L (2005). **O fenômeno da evasão escolar na educação superior no Brasil.** 2005. 75 f. Dissertação (Mestrado em Educação) Programa de Pós-Graduação em Educação da Universidade Católica de Brasília, Brasília, DF, 2005.

GOMES, E. C.; SOARES, D. B.; DESIDÉRIO, S. N.; ROCHA, A. S. DA R. S. DA. (2019). Evasão no curso de Licenciatura em Física da Universidade Federal Do Tocantins: diagnóstico e primeiros resultados de um projeto de intervenção. **Revista Observatório**, v. 5, n. 5, p. 482-508, 1 ago.

GUIMARÃES, Orlineya Maciel; MARTINS, Eliana Canteiro Bolorino; LIMA, Maria Jose de Oliveira (2019). A Evasão no Ensino Superior: a Unesp Câmpus de Franca - Período de 2013-2018. **CAMINE:** Caminhos da Educação. Franca, v. 11, n. 2, p. 136-161, mar. 2020. ISSN 2175-4217. Disponível em: https://periodicos.franca.unesp.br/index.php/caminhos/article/view/3013. Acesso em: 15 abr

2020.

HORTA, Cecília Eugenia Rocha. **Associação Brasileira de Mantenedoras de Ensino Superior Evasão no ensino superior brasileiro.** (2012). Disponível em: https://abmes.org.br/arquivos/publicacoes/Cadernos25.pdf. Acesso em: 27 abr 2020.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2019). Sem desistências, número de graduados poderia dobrar no Brasil. http://inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/sem-desistencias-numero-de-graduados-poderia-dobrar-no-brasil/21206. Acesso em 08 out 2019.

KIRA, L. P. (1998) **A evasão no ensino superior:** o caso do curso de pedagogia da Universidade Estadual de Maringá (1992-1996). Dissertação (Mestrado em Educação), Universidade Metodista de Piracicaba, 106 p.

LOBO, M. B. de C. M. (2012). **Panorama da evasão no ensino superior brasileiro:** aspectos gerais das causas e soluções. ABMES Cadernos. Brasília, set./dez. 2012.

MARTÍNEZ-CAMBLOR, Pablo e PARDO-FERNÁNDEZ, Juan C. (2019). **Parametric estimates for the receiver operating characteristic curve generalization for non-monotone relationships. Statistical methods in medical research,** v. 28, n. 7, p. 2032–2048, Jul 2019.

MARTINHO, Valquiria Ribeiro de Carvalho (2014). **Sistema Inteligente para a predição de grupo de evasão discente.** Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/100340/000751146.pdf?sequence=1&isAllowed=y. Acesso em: 26 mai 2020.

MEC. Ministério da Educação (2017). **PARECER HOMOLOGADO** Despacho do Ministro, publicado no D.O.U. de 12/1/2018, Seção 1, Pág. 12. Disponível em: http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=67801-pceb004-17-pdf&category_slug=julho-2017-pdf&Itemid=30192. Acesso em: 08 jun 2020.

MITCHELL, Tom M. (1997). **Machine learning**. 1997. Burr Ridge, IL: McGraw Hill, v. 45, 1997.

PEREIRA, R. S.; ZAVALA, A. A.; SANTOS, A. C. (2011). **Evasão na Universidade Federal de Mato Grosso.** Revista de Estudos Sociais, v. 13, n. 26, p. 74-86, 2011.

RASCHKA, Sebastian (2015). **Python Machine Learning.** [S.l.]: Packt Publishing Ltd, 2015.

RIGO, S. J. et al.(2014) . **Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar:** oportunidades e desafios. Revista Brasileira de Informática na Educação. Disponível em: https://www.br-ie.org/pub/index.php/rbie/article/view/2423. Acesso em: 12 mai 2020

SABBATINI, M. (2015). **Concepções e estratégias da aprendizagem participativa na educação a distância (EAD):** contribuição das práticas dialógicas e comunicacionais para a autonomia discente. Revista Observatório, v. 1, n. 3, p. 80-99, 26 dez.

SANTOS JUNIOR, J. DA S.; REAL, G. C. M (2019). Fator institucional para a evasão na educação superior. **Revista Internacional de Educação Superior**, v. 6, p. e020037, 27 dez.

SCHARGEL, F. P.; SMINK, J. **Estratégias para auxiliar o problema de evasão escolar.** Tradução de Luiz Frazão Filho. Rio de Janeiro: Dunya, 2002

SOARES, L. C. C. P. et al. (2020). **APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA EM**

UM CONTEXTO ACADÊMICO COM FOCO NA IDENTIFICAÇÃO DOS ALUNOS EVADIDOS E NÃO EVADIDOS. Disponível em: https://revista.unitins.br/index.php/humanidadeseinovacao/article/view/3293. Acesso em: 08 jun 2020.

SOUZA, Solange Lima de (2008). **Evasão no Ensino Superior:** Um estudo utilizando a mineração de dados como ferramenta de gestão do conhecimento em um banco de dados referente à graduação de Engenharia. Disponível em: http://livros01.livrosgratis.com.br/cp064905.pdf. Acesso: 11 abr 2020.

SOUZA, Thays Santos (2017). **Estudo sobre a evasão em cursos de graduação presenciais na Universidade Federal de Goiás** – UFG. 2017. 214 f. Dissertação. (Mestrado Profissional em Gestão Organizacional)-Programa de Pós-Graduação em Gestão Organizacional, Universidade Federal de Goiás, Catalão.

TINTO, V. (1975). **Dropout from Higher Education:** A theoretical synthesis of recent research. Review of Educational Research Winter, v. 45, n. 1, p. 89-125, 1975.

UNITINS (2020). Universidade Estadual do Tocantins. **Graduação.** Disponível em: https://www.unitins.br/nportal/graduacao. Acesso em: 11 abr 2020.

Recebido em 12 de abril de 2022.
Aceito em 16 de maio de 2023.