APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA EM UM CONTEXTO ACADÊMICO COM FOCO NA IDENTIFICAÇÃO DOS ALUNOS EVADIDOS E NÃO EVADIDOS

APPLICATION OF MACHINE LEARNING TECHNIQUES IN AN ACADEMIC CONTEXT WITH A FOCUS ON IDENTIFYING DROPOUT AND DROPOUT STUDENTS

Leandra Cristina Cavina Piovesan Soares [1]
Robson Aparecido Ronzani [2]
Rafael Lima de Carvalho [3]
Alexandre Tadeu Rossini da Silva [4]

Resumo: *A evasão escolar é um dos principais problemas causadores de prejuízos às Instituições de Ensino Superior e a utilização de modelos de predição podem subsidiar decisões para minimização dos prejuízos. Nesse contexto, este trabalho avalia se é possível empregar algoritmos de aprendizado de máquina para gerar modelar o padrão de evasão, a partir de dados de registro acadêmico. Esta hipótese foi validada através de um estudo de caso, usando os dados acadêmicos da Universidade Estadual do Tocantins. Os resultados alcançados pelos experimentos indicaram que a metodologia adotada neste trabalho foi capaz de classificar com elevado grau de confiança os alunos em situação de evasão e de não evasão.*
Palavras-chaves: *Evasão Escolar. Mineração de Dados Educacionais. Aprendizagem de Máquina. Inovação em Gestão Educacional.*

Abstract: *School dropout is one of the main problems that cause losses to Higher Education Institutions and the use of prediction models can subsidize decisions to minimize losses. In this context, this work evaluates whether it is possible to employ machine learning algorithms to generate modeling the evasion pattern, based on academic record data. This hypothesis was validated through a case study, using academic data from the State University of Tocantins. The results achieved by the experiments indicated that the methodology adopted in this work was able to classify students in situations of evasion and non-evasion with a high degree of confidence.*
Keywords: *School Dropout. Educational Data Mining. Learning Machine. Innovation in Educational Management.*

1 Formação: Mestranda em Propriedade Intelectual e Transferência de Tecnologia para a Inovação. Instituição de vinculação: Universidade Federal do Tocantins. Lattes: http://lattes.cnpq.br/0505525976660596. ORCID: https://orcid.org/0000-0003-0347-9160. E-mail: leandra.cavina@gmail.com

2 Especialização em sistemas de apoio a decisão (em andamento). Universidade Federal do Tocantins. Lattes: http://lattes.cnpq.br/1274083409245531. E-mail: ronzani.robson@gmail.com

3 Doutorado em Engenharia de Sistemas e Computação. Universidade Federal do Tocantins. Lattes: http://lattes.cnpq.br/0175648235036864. ORCID: https://orcid.org/0000-0002-5296-8641. E-mail: rafael.lima@uft.edu.br

4 Doutorado em Engenharia de Sistemas e Computação. Universidade Federal do Tocantins. Lattes: http://lattes.cnpq.br/2916003886317695. ORCID: https://orcid.org/0000-0002-6427-4436. E-mail: arossini@uft.edu.br

Introdução

O problema de evasão, ou abandono escolar, tem sido muito discutido no meio acadêmico (GUIMARAES et al, 2019; GOMES et al., 2019; SABBATINI, 2015; SANTOS JUNIOR; REAL, 2019; ADACHI, 2017; CARVALHO, 2017; SOUZA, 2017). Esse fenômeno afeta todos os níveis educacionais, desde os anos iniciais até ao ensino superior, independentemente do tipo de instituição, seja ela pública ou privada, e a sua consequência impacta no desenvolvimento humano (CUNHA e MOSORINI, 2013). De acordo com (Kira, 1998; Gaioso, 2005; BAGGI e colab., 2011), a evasão escolar é definida pela interrupção do ciclo de estudos, em qualquer nível de ensino. Para Kira (1998), a evasão escolar é considerada como fuga ou perda dos alunos antes da conclusão do curso.

Diversos fatores podem levar ao abandono escolar e nem sempre o motivo está relacionado ao mau desempenho acadêmico. Assim, na busca de uma melhor compreensão para esse fenômeno, surgem conceitos e estudos sobre a evasão escolar. De acordo com Souza (2008), a motivação da evasão não está apenas relacionada ao âmbito acadêmico, a causalidade pode advir de aspectos pessoais, sociais e ambientais. Segundo (BAGGI e colab., 2011), o abandono escolar requer uma análise histórica, pois a realidade dos níveis anteriores de ensino podem influenciar no abandono de um curso superior. Nesse contexto, percebe-se o quanto esse fenômeno é complexo uma vez que há vários aspectos envolvidos e requer estudo aprofundado.

Estudar a evasão escolar no ensino superior a fim de extrair informações/conhecimentos que possam ser utilizados para minimizá-lo é o objetivo deste trabalho. Assim, assume-se como hipótese a ser verificada que técnicas de Aprendizado de Máquina podem ser aplicadas de maneira satisfatória em mineração de dados educacionais. Para verificar a hipótese, foram utilizados, como estudo de caso, os dados acadêmicos da Universidade Estadual do Tocantins (Unitins). Por meio de técnicas computacionais de Aprendizado de Máquina, os dados foram investigados e analisados a fim de prever a evasão antes de um aluno abandonar o curso.

Aprendizado de Máquina (AM) é um ramo da Inteligência Artificial que compreende o desenvolvimento de técnicas e sistemas capazes de adquirirem conhecimento de maneira automatizada. A construção de um sistema de AM é baseada em experiências acumuladas através da solução de problemas anteriores (MONARD e BARANAUSKAS, 2003). Deste modo, a partir de dados acadêmicos da Unitins, esta pesquisa utilizou AM para produzir um classificador preditivo de alunos evadidos e não evadidos.

Para melhor compreensão deste trabalho, será apresentado (a): o contexto da evasão escolar no ensino superior, em especial da Unitins; a relação de procedimentos metodológicos utilizados nesta pesquisa; o conjunto de testes e seus resultados, seguidos de uma discussão sobre eles.

O contexto da evasão escolar no ensino superior

O fenômeno da evasão escolar e o estímulo à permanência por longo tempo dos alunos nas Instituições de Ensino Superior (IES) é um dos grandes desafios para a educação brasileira (INEP, 2019b). Para a Comissão Especial de Estudos sobre Evasão nas Universidades Públicas Brasileiras (CEUB), este estudo é complexo e comum as IES do mundo contemporâneo e que vem sendo influenciado por diversas variáveis, nos quais provocam a necessidade de estudos e análises sobre tal tema (BRASIL, 1996b).

Para (FILHO e colab., 2007) a evasão escolar é um problema de ordem internacional que afeta os sistemas educacionais. No que tange ao setor público, os recursos investidos tem seu retorno comprometido por causa da evasão escolar, uma vez que gera ociosidade de vagas, equipamentos e espaços físicos; para o setor privado há perda de receitas. Além disso, a evasão escolar causa desperdícios sociais, acadêmicos e econômicos. Segundo a Comissão Especial de Estudos sobre Evasão nas Universidades Públicas Brasileiras (CEUB), a definição desse fenômeno é como a saída definitiva do aluno de seu curso de origem, sem concluí-lo (BRASIL, 1996b). Para melhor entendimento sobre a distinção de formas de evasão escolar, a CEUB apresenta, por meio do quadro 1(um) os tipos de evasão.

Quadro 1 - Distinção sobre conceitos de Evasão Escolar

Tipo	Descrição
Evasão do Curso	É quando o estudante é desligado do curso superior em situações diversas tais como: abandono (deixa de matricular-se), desistência (oficial), transferência ou reopção (mudança de curso), exclusão por norma institucional.
Evasão da Matrícula	É quando o estudante desliga-se da instituição na qual está matriculado.
Evasão do Sistema	Quanto o estudante abandona de forma definitiva ou temporária o ensino superior.

Fonte: Elaborado pelos autores com base em BRASIL (1996b).

Os conceitos apresentados no quadro 1 (um) são importantes para compreender os trabalhos que mapearam as causas que levam os alunos a abandonarem seus cursos. Um estudo realizado por (COLPANI, 2019) identificou que o Pará foi o Estado com maior índice de evasão escolar no Brasil em 2017. O autor utilizou os dados públicos do Censo no seu estudo e os resultados apontaram que a média de evasão nas escolas é de 20%. Deste percentual, entre 38% a 65% são dos alunos estão com idade acima recomendada pela Lei 9.394/96, que é de 18 anos (BRASIL, 1996a). Assim, o indicador da taxa de distorção de série/idade foi a variável associada com a evasão escolar.

Souza (2008) realizou um estudo nos cursos de Engenharias da Universidade Federal Fluminense (UFF) e identificou que 32% das evasões estão relacionadas aos cursos de Engenharia Metalúrgica e as disciplinas obrigatórias são as que mais causam reprovações entre os alunos. O quadro 2 (dois) apresenta o ranqueamento das disciplinas que possuem maiores reprovações.

Quadro 2 - Ranqueamento das disciplinas com maiores reprovações

Código disciplina	**Nome disciplina**	**Casos**
GMA04043	Cálculo diferencial e integral Aplicado I	761 (14%)
GFI05100	Física geral e experimental XIII	674 (13%)
GAN06118	Álgebra linear aplicada	515 (10%)
GGM02055	Introdução à geometria descritiva	418 (8%)
TCC03060	Introdução à informática	350 (7%)
TCC03063	Programação de computadores III	234 (4%)
GMA06071	Equações diferenciais aplicadas	206 (4%)
GMA06074	Cálculo diferencial e integral aplicado II	171 (3%)
GMA04004	Cálculo diferencial e integral IV	156 (3%)
GFI05102	Física geral e experimental XX	155 (3%)

GFI05101	Física geral e experimental XIX	151 (3%)

Fonte: Elaborado pelos autores com base em SOUZA (2008).

O INEP (2019a) traçou perfil dos ingressantes do ano de 2018 nas Redes Federais que mudaram de Unidade Federativa para estudar fora de seu local de residência. Do total de 309.266 ingressantes, 10,4% haviam desistido no primeiro ano de curso e 4,1% estavam com a matrícula trancada. Já no que tange ao contexto do Estado do Tocantins, que é objeto de estudo deste trabalho, 830 alunos foram oriundos de outros Estados (INEP, 2019a). Nota-se a necessidade de maior atenção para este grupo específico de alunos, pois a evasão escolar destes alunos pode estar relacionada a vários aspectos, como condições emocionais, educacionais ou ambientais. Nesse sentido, Souza (2008) assevera que a evasão nem sempre está relacionada ao baixo desempenho acadêmico e, por isso, é necessário investigar outras aspectos. Em suma, as IES, a fim de encontrar soluções para o problema de evasão escolar, devem analisar dados de seus acadêmicos, identificar situações eminentes de abandono e propor ações para minimizar a evasão escolar.

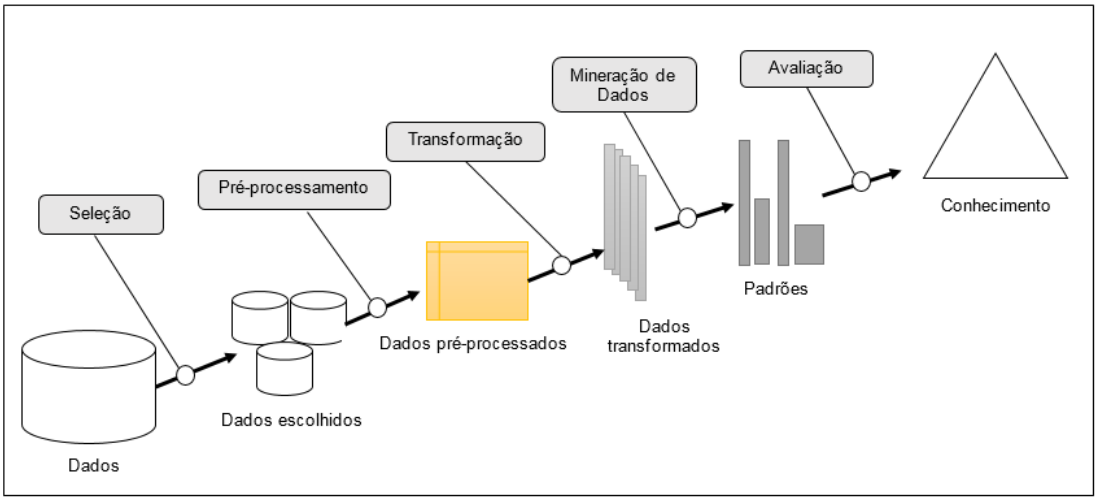
Metodologia

Pela própria natureza dos dados acadêmicos da Unitins, a natureza da pesquisa é quantitativa e envolveu processo de coleta, análise e interpretação dos dados no intuito de descobrir relações entre as variáveis e características de causalidade ao fenômeno de evasão escolar. Como fundamentação teórica, foi utilizada a pesquisa bibliográfica a partir de materiais já elaborados e previamente publicados em livros e artigos científicos.

Para o domínio dos dados foi aplicada a técnica de Mineração de Dados Educacionais (MDE) com a utilização de métodos de AM. Para isso foram utilizadas as ferramentas *Spyder,* linguagem de programação *Python* e os *frameworks Scikit-learn* e Pandas. Segundo (RAMESH e colab., 2013), a MDE auxilia na identificação de padrões para a tomada de decisão, cujo o processo é feito por meio da coleta dos dados, análise da informação e a geração do conhecimento. Já o AM fornece a base técnica para a MDE, que transforma dados brutos em informações de mais fácil compreensão, como previsões, correlações e relações de causalidade, o que, no processo de análise, auxilia na compreensão e explicação de fenômenos.

Fayyad et al (1996) propuseram um processo denominado Descoberta de Conhecimento em Banco de Dados, do inglês *Knowledge Discovery in Databases* (KDD), com as seguintes etapas: seleção, pré-processamento, transformação, mineração de dados e avaliação. As três primeiras etapas (seleção, pré-processamento e transformação) visam escolher, tratar, corrigir, normalizar e enriquecer os dados que serão processados para gerar conhecimento. Com os dados prontos para serem processados, a etapa de mineração de dados faz uso de algoritmos que extraem padrões dos dados e os padrões descobertos devem ser interpretados em uma fase de avaliação. A figura 1 (um) apresenta as principais etapas do KDD.

Figura 1- Principais etapas do processo de KDD.

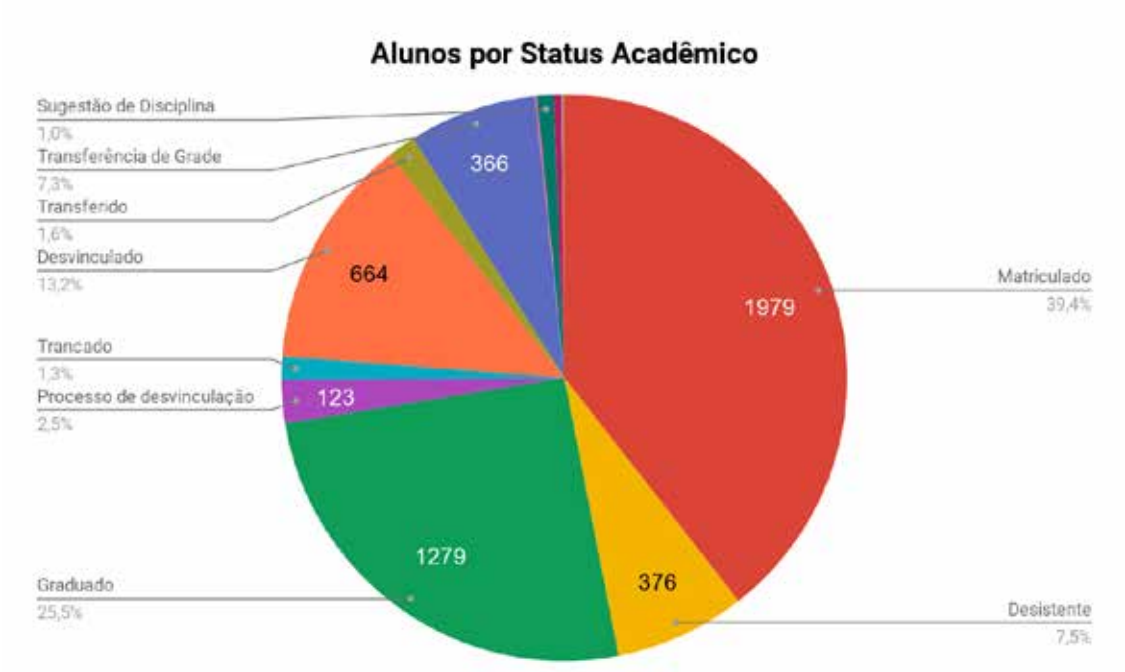


Fonte: Elaborado pelos autores com base em Fayyad et. al. (1996).

Compreensão dos dados

Para obtenção do acesso aos dados, esta pesquisa foi submetida e aprovada pelo Comitê de Ética em Pesquisa com Seres Humanos (CEP), sob o número 29598820.3.0000.8023. Para a medição e treinamento dos classificadores foi utilizada a base de dados acadêmica dos 12 (doze) cursos presenciais da Unitins. Os dados estavam distribuídos em 18 (dezoito) variáveis de três tabelas ("cadastro de alunos", "matrizes curriculares" e "histórico acadêmico"): 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 figura 2 (dois) apresenta o quantitativo de alunos cadastrados na tabela de "cadastro de aluno". Os acadêmicos estão classificados por *status*, que é a situação de vínculo em que eles se encontram.

Figura 2 - Quantitativo de alunos por tipo de status acadêmico



Fonte: Elaborado pelos autores com base em UNITINS (2020).

Posteriormente, foi feita uma seleção de dados a partir dos *status* acadêmicos mais relevantes para esta pesquisa: Graduado, Desvinculado e Desistente. Assim, do total de 5.018 (cinco mil e dezoito), foram selecionados 2.319 (dois mil e trezentos e dezenove) alunos. Um novo *status*, denominado Evadido, foi criado a partir da junção dos *status* Desistente e Desvinculado e o *status* graduado passou a se chamar Formado, para não gerar quaisquer ambiguidades com o status dos dados originais.

Em seguida, na etapa de pré-processamento foi realizado preenchimento dos valores para os atributos que não estavam preenchidos. Para o atributo idade foi definida a média de idade da base de dados, os demais atributos foram registrados como "não informado". Os dados não implícitos foram extraídos a partir do histórico acadêmico e da matriz curricular, na busca da identificação dos totalizadores, como: disciplinas da matriz, disciplinas cursadas, disciplinas aprovadas e disciplinas reprovadas por médias e faltas. A etapa de mineração de dados será abordada com detalhes na próxima seção.

Métodos de Aprendizado de Máquina

O algoritmo utilizado como método de AM na etapa de mineração de dados foi da Floresta Aleatória *(Random Forests).* Segundo BREIMAN (2001), as melhorias significativas na precisão da classificação resultaram o crescimento de um conjunto de árvores, que votam na classe mais popular, ficando conhecida como Florestas Aleatórias. De acordo com Montaño (2016), esse método é mais adequado para bases de dados que possuem diversas variações e requer poucos ajustes de parâmetros, caso da base de dados utilizada neste trabalho. Para o treinamento do algoritmo, a base de dados foi particionada em dois conjuntos aleatórios, um com 70% dos dados a serem usados no treinamento, proporcional de acordo com as classes, e os demais 30% foram reservados para validação do treinamento. -

A validação cruzada (*cross validation*), que é uma técnica para avaliar a capacidade de generalização de um modelo a partir de um conjunto de dados, foi utilizada nas etapas de treinamento e avaliação. Na validação cruzada é feito particionamento do conjunto de dados em subconjuntos mutuamente exclusivos, um elemento do conjunto é separado para testes (não participa do treinamento) e os demais elementos são utilizados no treinamento. Esse processo foi repetido *k* vezes alternando os elementos que foram utilizados para o grupo treinamento e para teste. A métrica de avaliação utilizada foi resultante das médias aritméticas dos *k* testes realizados (SCHAFFER, 1993).

Os algoritmos de aprendizado de máquina, em geral, possuem parâmetros que podem ser ajustados para um melhor desempenho, segundo alguma métrica. Na etapa de validação cruzada, foi utilizado o algoritmo chamado *Grid Search* que cria uma grade multidimensional com a combinação dos parâmetros a serem avaliados. Para o algoritmo de Florestas Aleatórias, dois parâmetros foram testados: o critério de entropia e a quantidade de estimadores (árvores da floresta). Os critérios de entropia foram o padrão (*entropy*) e o critério de Gini. Enquanto que os valores de estimadores foram testados num intervalo entre 10 (dez) e 120 (cento e vinte). Estes valores do intervalo foram obtidos de forma empírica e em geral são dependentes do problema avaliado. Durante a etapa de validação cruzada, buscou-se obter a combinação de valores que maximizou o valor da acurácia, métrica que será explicada a seguir.

Métricas de avaliação da solução proposta

Em geral, sistemas de aprendizado de máquina são avaliados por meio de métricas pontuais, aplicadas tanto na base de dados de treinamento quanto de teste. De forma a subsidiar a discussão das métricas a serem adotadas para a solução proposta, a seguir, são explicadas duas métricas de modelos de predição considerando um problema com apenas duas classes (identificadas neste trabalho como positivo e negativo), que poderão ser utilizadas para avaliar e analisar os resultados gerados através da etapa de Mineração de Dados. Essas métricas são: Matriz de Confusão, que permite identificar a frequência das classes no modelo avaliado; A *Receiver Operating Characteristic* (ROC) e Área sob a Curva ROC ou *Area Under the Curve* (ROC-AUC), que serve para avaliar a qualidade do modelo.

A primeira maneira de se observar o desempenho de um classificador binário e através de uma Matriz de Confusão. Esta consiste de uma matriz quadrada onde são dispostas as previsões (linhas) e os valores verdadeiros (colunas). Ela é utilizada para mostrar a quantidade de acertos e erros de maneira que seja possível verificar a quantidade de amostras confundidas pelo sistema. Para tanto, é preciso definir o significado dos seguintes termos: Verdadeiro Positivo (VP), Falso Positivo (FP), Verdadeiro Negativo (VN) e Falso Negativo (FN) [RASCHKA, 2015].

Para o problema proposto neste trabalho, Verdadeiro Positivo (VP) é a quantidade de alunos que evadiram e o modelo identificou como evadido. O Verdadeiro Negativo (VN) é a quantidade de alunos que não evadiram e o modelo reportou como não evadido. Falso Positivo (FP) é a quantidade de alunos que não evadiram, mas foram erroneamente classificados como tal. Falso Negativo (FN) é a quantidade de alunos que evadiram, mas foram classificados como não evadido. Neste sentido, minimizar o FN é importante uma vez que, ao não identificar o aluno que evadiu, ações que poderiam evitar a evasão nestes casos podem deixar de ser tomadas.

Com os valores definidos de VP, FP, VN e FN é construída a matriz de confusão, a qual é exibida na Figura 3.

Figura 3 - Matriz de confusão com as duas 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 acurácia ou exatidão (AC) diz o quanto o modelo acertou dentro de seu desempenho geral. A acurácia é obtida tomando-se a soma das previsões corretas, e dividindo-a pelo número total de previsões, respectivamente, conforme mostra a equação (1):

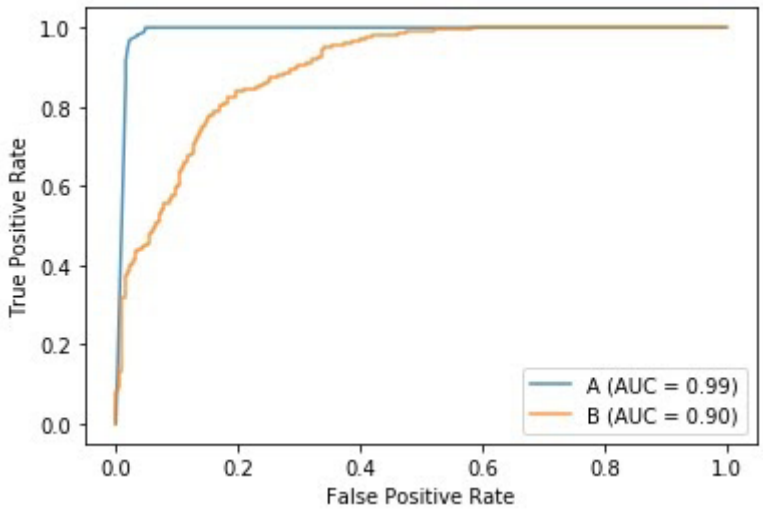
Devido à bases de dados desbalanceadas, onde existem diferenças muito grandes entre o número de amostras de cada classe, em geral também utiliza-se outras métricas tais como Precisão e Revogação, F1-*score* curva ROC. Para dar suporte a estas métricas, é necessário estabelecer a Taxa de Falsos Positivos (FPR) e Taxa de Verdadeiro Positivo (TPR), que são definidas pelas equações (2) e (3), respectivamente. Estas taxas fornecem informações importantes, para as amostras positivas que foram identificadas corretamente no conjunto de total das amostras positivas.

Posto isto, define-se a Precisão (PRE) como a razão entre o número de verdadeiros positivos (VP) e a soma de Verdadeiro Positivos (VP) e Falso Positivos (FP), definida na equação (4). A métrica PRE serve para medir se daqueles que foram classificados como positivos, efetivamente estão corretos. Já a revogação ou *recall* (REC), apresentada na equação (5), é quão completo os resultados estão, onde o REC de fato é o sinônimo de TPR. O REC pode ser realizado em situações que os FN são considerados mais prejudiciais do que os FP.

Para combinar as duas informações PRE e REC e mua medida só, há uma métrica chamada F1-*Score*, definida na equação (7). A F1-*Score* é considerada uma média harmônica entre duas medidas PRE e REC, permitindo que seja possível a identificação da qualidade geral do modelo, através de um único valor. Portanto, valores altos de F1 implicam em uma qualidade maior do modelo avaliado.

De acordo com (MARTÍNEZ-CAMBLOR e PARDO-FERNÁNDEZ, 2019), a curva de ROC é uma ferramenta utilizada na estatística para estudar a qualidade de um processo de classificação em termos de sensibilidade e especificidade. Quando a curva padrão estiver com os valores maiores no marcador de diagnóstico, significa que existe maior probabilidade de se manter as características estudadas. Um gráfico ROC é a relação entre TPR e FPR. A figura 4 apresenta a curva ROC gerada para dois classificadores A e B. No gráfico da figura 4, o classificador A é interpretado como sendo melhor do que B, segundo a curva ROC. Esta medida pode ser simplificada através do cálculo da área sob a curva, chamado de ROC-AUC, que naturalmente varia de 0 (zero) a 1 (um) (RASCHKA, 2015).

Figura 4 - Curva AUC (A e B)



Fonte: Elaborado pelos autores (2020).

Portanto, a avaliação do modelo de classificação pode ser feita por meio de diferentes métricas de acordo com as características dos dados. É importante que se leve em consideração, durante a avaliação de modelos de classificação, fatores como proporção de dados de cada classe e o objetivo da precisão.

Resultados

Com base nas métricas de Aprendizado de Máquina, apresenta-se os resultados que foram obtidos para a avaliação do modelo. As métricas utilizadas para este estudo foram: Acurácia, Precisão, Revogação, F1-*Score* e ROC-AUC.

A tabela 2 (dois) expõe os resultados obtidos no conjunto treinamento, que utilizou 70% da base de dados, com o algoritmo Floresta Aleatória pela técnica de validação cruzada. Como resultado, destaca-se acurácia média de 97,4%, entre os subconjuntos gerados pelo processo de validação cruzada.

Tabela 2 - Resultados dos treinamentos

Métricas de Aprendizado de Máquina	Resultado
Acurácia	97,4
Precisão	96,2
Revogação	99,2

F1	97,6
ROC-AUC	97,1

Fonte: Elaborado pelos autores (2020).

Como métrica de avaliação dos resultados, a matriz de confusão foi calculada e está apresentada na tabela 3 (três). Os dados contidos na tabela 3 (três) representam a média gerada pela validação cruzada e apresenta a porcentagem de acertos da metodologia adotada neste trabalho em relação à sua capacidade de classificar alunos evadidos e não evadidos. Destaca-se que 88,8% dos alunos que não evadiram, chamado de Verdadeiro Negativo (VN), e que 69,3% dos acadêmicos evadidos, chamado de Verdadeiro positivo (VP), foram classificados corretamente.

Tabela 3 - Matriz de Confusão gerada a partir do treinamento

Matriz de Confusão - Validação Cruzada	
VP (69,3%)	FP (3,5%)
FN (0,7%)	VN (88,8%)

Fonte: Elaborado pelos autores (2020).

Em seguida, o algoritmo de Floresta Aleatória, com seus parâmetros otimizados pelo processo de *Grid Search*, foi treinado nos 70% da base de dados e testado utilizando os demais 30%, que não foram apresentados ao classificador, em qualquer momento, durante o treinamento. A Tabela 4 (quatro) traz os resultados alcançados.

Tabela 4 - Resultado dos classificadores dos grupos de 70% de treinamento e 30% para teste

Métricas de Aprendizado de Máquina	Resultado
Acurácia	97,5
Precisão	95,7
Revogação	100
F1	97,8
ROC-AUC	97,2

Fonte: Elaborado pelos autores (2020).

Percebe-se que os resultados dos dados de validação (30% da base de dados), obtêm-se os resultados semelhantes com os alcançados durante o treinamento (70% dos dados). Este fato é um indicador de que o modelo se adaptou (aprendeu) de tal forma que conseguiu generalizar o conhecimento para novos dados (objetivo principal de qualquer preditor).

A matriz de confusão da tabela 5 (cinco) apresenta a média gerada pela validação cruzada e expressa a quantidade de acertos em classificar a situação do acadêmico: evadido ou não evadido. Nesse sentido, observa-se o bom desempenho dos métodos utilizados neste trabalho na obtenção de VP (alunos evadidos) e de VN (alunos não evadidos).

Tabela 5 - Resultado da Matriz de Confusão do grupo 70% de treinamento de 30% para teste

Matriz de Confusão - Grupos de 70 e 30%	
VP (295)	FP (17)
FN (0)	VN (384)

Fonte: Elaborado pelos autores (2020).

Observando os resultados apresentados na Tabela 5 (cinco), percebe-se que nos 30% aleatoriamente selecionados, e consequentemente não utilizados durante a fase de treinamento, a quantidade de Falsos Negativos (FN) é zero. Observando a matriz de confusão exibida na Tabela 3 (três), que remete aos dados de treinamento, percebe-se que este valor é menor que 1%. Ou seja, nos 70% dos dados, a validação cruzada separou em k=10 grupos, e o algoritmo foi treinado por 10 (dez) vezes seguidas, sempre utilizando-se 9 (nove) subgrupos para treinamento e 1 (um) para testes. Ainda neste cenário, o valor médio dos FN foi menor também próxima de zero.

É interessante ressaltar que o Falso Negativo representa o caso em que o aluno evadiu, mas foi classificado como não evadido. Este caso é o mais sensível para a aplicação considerada. Perceba que, se o sistema classificar um aluno provável de evadir como se não evadisse, os gestores do ensino não poderiam também considerar a situação deste aluno para criar elementos para tentar alcançá-lo, antes que este decida por evadir. Observando os Falsos Positivos, errar neste aspecto significaria errar por excesso, pois eles também seriam considerados para elaboração de políticas para lidar com o problema. Ainda analisando a Tabela 5 (cinco), percebe-se que 17 (dezessete) amostras se encaixam nesta situação. Isto é apenas uma proporção de 0,02 (dois centésimos), quando comparado com o total de amostras considerada na base de dados de validação. O excesso é mínimo e, portanto, suportável pelo poder informativo do experimento.

Considerações Finais

A evasão escolar é um problema que afeta todos os níveis de ensino. Os aspectos relacionados a esse fenômeno são diversos e complexos e nem sempre estão associados ao âmbito acadêmico. O uso de ferramentas de inteligência computacional tais como técnicas de AM, tem contribuído bastante para estudos sobre Evasão Escolar ao extrair, de bases de dados, informações e conhecimentos que podem subsidiar o processo de tomada de decisão das IES.

Este trabalho validou, por meio de um estudo de caso, a hipótese de que as técnicas AM podem ser usadas de maneira satisfatória em MDE, o que foi verificado por experimentos utilizando a base de dados acadêmica da Unitins. Os resultados alcançados neste trabalho indicam que os métodos de MDE podem contribuir para o desenvolvimento de um classificador de confiança elevada, ou seja, capaz de automatizar a tarefa de classificar alunos em estado de evasão. Uma vez feita esta separação, é possível conduzir um estudo mais focado no grupo identificado e tentar encontrar padrões que levaram à evasão.

Sabe-se que os setores acadêmicos que lidam com diversos processos e, em geral, um grupo pequeno de trabalho é alocado na tarefa de tentar entender as razões da evasão nas instituições e propor soluções. O experimento computacional reportado nesta pesquisa dá fortes indicativos de que é possível aproximar um modelo de evasão a partir dos registros acadêmicos.

Como trabalhos futuros, pretende-se conduzir uma investigação sobre a capacidade de um sistema de aprendizado de máquina de agir como preditor em dados parciais. Visto que, ao final de cada semestre letivo, o sistema poderia propor uma predição de evasão e reportar este percentual aos coordenadores de curso ou para a gestão acadêmica. Uma vez de posse destes dados, poderiam tomar decisões fundamentadas e mais acuradas para alcançar o aluno provável de evadir, antes do acontecimento em si. Por fim, é interessante ressaltar que o fator humano será cada vez mais necessário, uma vez que a máquina será capaz de indicar a porcentagem de previsão

de evasão, mas o entendimento do que poderia levar o aluno a evadir terá que ser interpretado por outro ser humano.

Referências

ADACHI, Ana Amélia Chaves Teixeira. 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 e DOS SANTOS BAGGI; Cristiane Aparecida e LOPES, Doraci Alves. **Evasão e avaliação institucional no ensino superior:** uma discussão bibliográfica. Avaliação: Revista da Avaliação da Educação Superior (Campinas). [S.l: s.n.]. Disponível em: <http://dx.doi.org/10.1590/s1414-40772011000200007>. 2011.

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

BRASIL. Ministério da Educação. Secretaria de Educação Superior (1996b). **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. Acesso em: 11/04/2020. Disponível em: http://www.andifes.org.br/wp-content/files_flutter/Diplomacao_Retencao_Evasao_Graduacao_em_IES_Publicas-1996.pdf.
BREIMAN.L. Random forests. Machine Learning, 45:5–32, 2001.

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: t**eoria & prática. [S.l: s.n.]. Disponível em: <http://dx.doi.org/10.22456/1982-1654.87880>. 2019.

CUNHA, E.R, MOROSINI, M.C.(2013). **Evasão na Educação Superior:** Uma temática em Discussão. Disponível em:https://paginas.uepa.br/seer/index.php/cocar/article/view/283. Acesso em: 14/10/2019.

FAYYAD, U. M., Piatetsky Shapiro, G., Smyth, P. & Uthurusamy, R. "**Advances in Knowledge Discovery and Data Mining"** 1996, AAAIPress, The Mit Press.

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

FILHO, Roberto Leal Lobo e. Silva e colab. **A evasão no ensino superior brasileiro**. Cadernos de Pesquisa. [S.l: s.n.]. Disponível em: <http://dx.doi.org/10.1590/s0100-15742007000300007>. 2007.

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

GUIMARÃES, Orlineya Maciel; MARTINS, Eliana Canteiro Bolorino; LIMA, Maria Jose de Oliveira (2020). A Evasão no Ensino Superior: A Unesp Câmpus de Franca - Período DE 2013-2018. **CAMINE: Caminhos da Educação = Camine: Ways of Education**, 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.

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.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2019a). **Censo da Educação Superior 2018.** Divulgação dos Resultados. Brasilia- DF- 19 de Setembro de 2019.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2019b). **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 08/10/2019.

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.

MARTÍNEZ-CAMBLOR, Pablo e PARDO-FERNÁNDEZ, Juan C. 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.

MELO, Francisco. Area under the ROC Curve. **Encyclopedia of Systems Biology**. [S.l.]: Springer, New York, NY, 2013. p. 38–39. . Acesso em: 31 out 2019.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto (2003). **Conceitos sobre Aprendizado de Máquina**. Disponível em: <http://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf>. Acesso em: 9 dez 2019.

MONTAÑO. RAZER. **Aplicação de Técnicas de Aprendizado de Máquina na Mensuração Florestal.** Disponível em: https://www.acervodigital.ufpr.br/bitstream/handle/1884/45346/R%20-%20T%20-%20RAZER%20ANTHOM%20NIZER%20ROJAS%20MONTANO.pdf?sequence=1&isAllowed=y. Acesso em: 16 de Mar. 2020.

RAMESH, V. e PARKAVI, P. e RAMAR, K. **Predicting Student Performance:** A Statistical and Data Mining Approach. International Journal of Computer Applications. [S.l: s.n.]. Disponível em: <http://dx.doi.org/10.5120/10489-5242>. 2013.

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

SCHAFFER. C. **Selecting a classification method by cross-validation**. Disponivel em: https://link.springer.com/content/pdf/10.1007/BF00993106.pdf. Acesso em: 17 de Mar. 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.

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: 11/04/2020.

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.

UNITINS (2020). Universidade Estadual do Tocantins. **Graduação**. Disponível em: https://www.unitins.br/nportal/graduacao. Acesso em: 11/04/2020.

Recebido em 7 de maio de 2020.
Aceito em 8 de maio de 2020.