ARTIGO

Pode a inteligência artificial apoiar ações contra evasão escolar universitária?

Wanderci Alves Bitencourt [a]
Diego Mello Silva [b]
Gláucia do Carmo Xavier [c]

Resumo

A evasão escolar é uma preocupação mundial devido às consequências negativas para toda a sociedade, sendo preciso investigá-la para compreendê-la e atuar de forma antecipada, mitigando seu risco de ocorrência. Esse trabalho propõe o emprego de Mineração de Dados Educacionais com técnicas de Aprendizado de Máquina para identificar as variáveis que são importantes para a caracterização do perfil do estudante em risco de evasão. As técnicas Máquina de Vetores de Suporte, *Gradient Boosting Machine*, Floresta Aleatória e comitê de máquina foram aplicadas a 1.429 registros de estudantes dos cursos superiores de um dos *campi* do IFMG, entre 2013 e 2019. Os resultados obtidos sugerem superioridade de desempenho do comitê de máquina, por meio do qual se obteve a importância das variáveis sobre o fenômeno em estudo, o que permitiu traçar o perfil do estudante evasor, por período. Tais resultados viabilizaram a proposição de um processo de detecção e acompanhamento desses estudantes.

Palavras-chave: Evasão. Aprendizado de Máquina. Estudantes Universitários.

1 Introdução

O abandono escolar, denominado também de evasão escolar, provoca graves consequências sociais, acadêmicas e econômicas (BAGGI; LOPES, 2011). Caracteriza-se como um fenômeno complexo, associado à não concretização de expectativas e que pode ter como causas múltiplas razões (FRITSCH; ROCHA; VITELLI, 2015). Evidências empíricas sugerem que existem fatores internos e

[a] Instituto Federal de Minas Gerais, Formiga, MG, Brasil.
[b] Instituto Federal de Minas Gerais, Formiga, MG, Brasil.
[c] Instituto Federal de Minas Gerais, Ribeirão das Neves, MG, Brasil.

Recebido em: 25 abr. 2020
Aceito em: 16 jun. 2021

externos que podem atuar como possíveis influenciadores da decisão do estudante em permanecer ou evadir do curso, tais como atributos demográficos, acadêmicos, pessoais e familiares (FRITSCH; ROCHA; VITELLI, 2015). Adicionalmente, a análise do risco de evasão pode ser realizada sobre a perspectiva do indivíduo, da escola ou do sistema de ensino, o que pode ocasionar resultados distintos à investigação, uma vez que diferentes atores atribuem diferentes significados às experiências (FIGUEIREDO; SALLES, 2017).

Um modelo que busca explicar a evasão no Ensino Superior é o proposto por Tinto (1997). Nesse modelo, a evasão é associada a fatores anteriores ao ingresso do estudante na instituição (chamados *ex-ante*) e a fatores posteriores ao ingresso (denominados *ex-post*). As variáveis *ex-post* são interativas ao longo do tempo, referindo-se ao período da vida acadêmica dos estudantes e que sofrem interferências de elementos do ambiente externo. Segundo esse mesmo autor, estudantes não integrados aos sistemas acadêmico e social da instituição podem ter seus níveis de comprometimento com a finalização do curso e com a instituição afetados, o que poderá levá-los à evasão.

No contexto brasileiro, conforme o Censo da Educação Superior, observa-se que a taxa de evasão nacional, em 2018, se comparado com o ano anterior, para cursos superiores presenciais, é de aproximadamente 22,5%. Taxa que não diverge muito da taxa média encontrada por Silva Filho (2017), entre 2011 a 2015, de aproximadamente de 22%. Essa estabilidade sugere que as ações adotadas para reduzir a evasão não se têm mostrado eficientes, corroborando os apontamentos de Figueiredo e Salles (2017), os quais sugerem que os esforços dos diferentes agentes ligados a Educação têm se mostrado insuficientes para garantir a permanência dos estudantes em seus cursos, o que amplia a necessidade da realização de trabalhos sobre a evasão que se traduzam em ações práticas.

Esse cenário, em que quase um quarto dos estudantes não conclui seus estudos no Ensino Superior, é preocupante, pois o insucesso na permanência dos estudantes vai na contramão da necessidade que o Brasil tem de possibilitar à população o acesso a maiores níveis de Educação e qualificação, elementos necessários ao desenvolvimento humano, social e econômico do país (SILVA; SANTOS, 2017), bem como reduzem os efeitos positivos do esforço para a ampliação do número de vagas observado nas últimas décadas (FIGUEIREDO; SALLES, 2017). Conforme Lima Junior *et al.* (2019), a ampliação de vagas na Educação Superior só converte-se em indicadores positivos para a economia e para a sociedade se essa ampliação vier acompanhada de ações que garantam a permanência dos estudantes, do ingresso à formatura. Ademais, a ampliação de vagas e a democratização do

acesso ao Ensino Superior permitiram o incremento da diversidade do perfil dos estudantes, ampliando, assim, a necessidade de aprofundamento e compreensão dos diferentes perfis e seu risco de evasão.

Buscando compreender melhor esse fenômeno, muitos autores propõem o uso de Aprendizagem de Máquina e Mineração de Dados, por meio de algoritmos de classificação, para prever o abandono escolar e para identificar fatores de risco associados com a evasão escolar (GOLDSCHMIDT; BEZERRA; PASSOS, 2015).

A natureza desses estudos divide-se entre a avaliação de desempenho de diferentes modelos de predição ou a classificação dos fatores de evasão. Alguns estudos buscam, ainda, propor sistemas de alerta sobre o risco de evasão (CHUNG; LEE, 2019; DIGIAMPIETRI; NAKANO; LAURETTO, 2016; KNOWLES, 2015; ROVIRA; PUERTAS; IGUAL, 2017).

Diante do exposto, o presente trabalho propõe-se a fazer uma análise da evasão escolar no âmbito da Educação Profissional de Nível Superior, mais especificamente no Instituto Federal de Minas Gerais – em um de seus 18 *campi*, na modalidade presencial. Para atingir ao objetivo proposto, serão adotadas técnicas de Mineração de Dados Educacionais com Aprendizado de Máquina. Acredita-se que o modelo preditivo construído permitirá a identificação de fatores que são significativos para a predição de estudantes com risco de evasão, levando à proposição de produto educacional que se constituirá em um processo de detecção e acompanhamento do risco de evasão, construído e baseado em informações já existentes nos sistemas de informação da instituição.

2 O estudo da evasão

A evasão representa uma fraqueza e uma preocupação para qualquer sistema educacional, pois resulta em desperdícios individuais e coletivos para a sociedade (TONTINI; WALTER, 2014).

Prestes e Fialho (2018) ressaltam que as políticas educacionais voltadas à evasão no Ensino Superior são recentes e incipientes no Brasil, tendo em vista seus efeitos devastadores, reforçando a necessidade de esforços para compreensão e redução desse fenômeno.

Para Silva Filho e Araújo (2017), o estudo da evasão pode ser realizado de forma abrangente ou baseada em dimensões específicas, tais como a temporalidade da

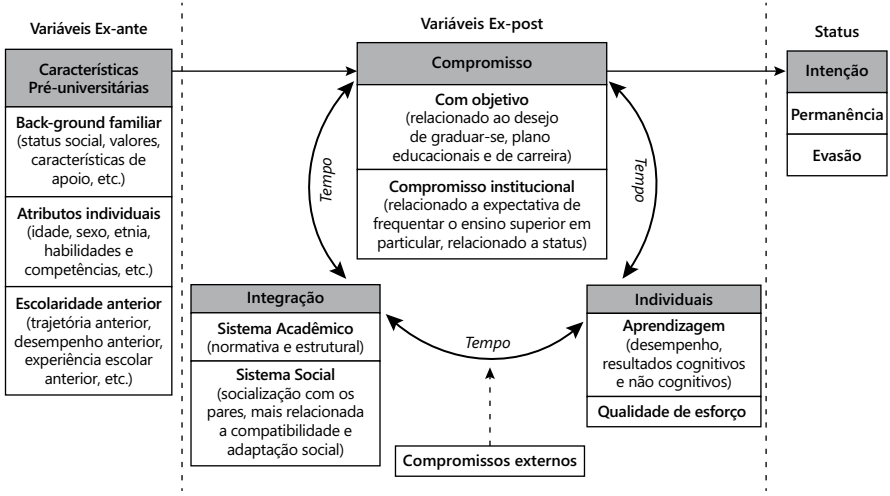
ocorrência, os tipos (sistema, instituição ou curso) e as razões (socioeconômicas, pessoais ou outras). Dessa forma, existem diferentes alternativas para compreensão da evasão, mas em todos é importante compreender quais variáveis podem ser determinantes desse abandono, viabilizando a identificação dos casos com alto risco antes que se efetivem.

Uma proposição teórica amplamente adotada em estudos sobre a evasão universitária é o modelo de impacto proposto por Tinto em 1975 e melhorado, pelo autor, em sucessivos estudos sobre o tema (MASSI; VILLANI, 2015; SANTOS JUNIOR; REAL, 2017). Conhecido como Teoria de Integração Estudantil, o modelo foi construído a partir de estudos realizados nos Estados Unidos, em analogia com os estudos de Durkheim sobre suicídio.

Conforme Tinto (1997), os estudantes ingressam no curso com características prévias (por exemplo, histórico familiar, atributos individuais, aptidão e outros), as quais foram chamadas neste estudo de *ex-ante*, pois são variáveis existentes antes mesmo do ingresso do estudante no curso, sendo muitas delas registradas pela instituição no ato da matrícula. Após seu ingresso, *ex-post* à matrícula, os estudantes interagem com o ambiente institucional e essas experiências com o ambiente escolar e com o ambiente externo influenciam seus compromissos e intenções. As variáveis *ex-post* existentes na maioria dos sistemas acadêmicos são, por exemplo, coeficiente de rendimento (CR), taxa de absenteísmo e retenção.

A Figura 1 ilustra uma representação construída a partir das interações propostas no modelo de Tinto (1997). Observa-se por meio dela que o estudante, ao ingressar, tem seu compromisso afetado apenas pelas informações *ex-ante*. A seguir, no decorrer da sua vida acadêmica, surgem novas variáveis (*ex-post*) e interações institucionais, sociais e externas. Essas interações geram um efeito plástico e adaptativo aos patamares de compromisso do estudante, o que determinará o seu interesse em permanecer ou se evadir-se, ao longo do tempo, permitindo a compreensão destes fatores sobre a predisposição do estudante à permanência ou ao abandono.

Figura 1 - Esquematização da integração das variáveis



Fonte: Adaptado de Tinto (1997)

Com a evolução da Tecnologia de Informação, tornou-se possível analisar os registros existentes na escola, que antes eram apenas um repositório de dados, e que passaram a ser valiosos, quando transformados em informações úteis, criando um campo vasto para a aplicação de Mineração de Dados Educacionais (COLPANI, 2018), contribuindo não somente para o entendimento teórico, mas apoiando a adoção de ações práticas de combate à evasão por parte das instituições de ensino.

Muitos dos estudos aplicam técnicas de mineração em base de dados governamentais, enquanto outros focam em minerar dados de instituições específicas. Nesse quesito, os estudos internos podem ser mais vantajosos por permitir um maior detalhamento dos dados e adoção de estratégias específicas à instituição (DIGIAMPIETRI; NAKANO; LAURETTO, 2016), embora a análise de bases oficiais do governo forneçam uma visão do sistema educacional, como um todo, e mostrem-se úteis para tomada de decisão (HOFFMANN; NUNES; MULLER, 2019),

Na literatura internacional, vários autores estudam formas para construir sistemas de acompanhamento do risco de evasão, a partir do ingresso do estudante na instituição, principalmente, para o Ensino Médio. Knowles (2015) propõe o uso de um comitê de máquina entre diferentes métodos de Aprendizado de Máquina para melhorar a previsão do modelo adotado pelo Estado de Wisconsin, enquanto Robison *et al.* (2017) propõem o uso de Regressão Logística para estudo do tema.

Rovira, Puertas e Igual (2017) usaram Regressão Logística, *Gaussian Naive Bayes*, SVM, Floresta Aleatória, e *Boosting* Adaptativo propondo um sistema para prever notas e evasão para a Universidade de Barcelona. Chung e Lee (2019) propõem um modelo eficiente de predição para o Ensino Médio na Coreia do Sul, usando Floresta Aleatória. São exemplos que demonstram a viabilidade do uso de técnicas de Aprendizagem de Máquina na análise de evasão.

3 Método

Nesta seção as operações de pré-processamento que envolvem seleção de dados, limpeza, enriquecimento, normalização, correção de prevalência e partição dos dados (GOLDSCHMIDT; BEZERRA; PASSOS, 2015) foram aplicadas na base de dados. Tanto o pré-processamento quanto a construção de modelos de predição via técnicas de Aprendizado de Máquina foram implementadas utilizando pacotes da linguagem R, no ambiente R Studio®. Nesse texto, os termos atributo e variável devem ser considerados com o mesmo sentido, sendo adotados de acordo com a conveniência.

3.1 Pré-processamento dos dados

Foram selecionados dados socioeconômicos, de assiduidade e rendimento acadêmico, do período de 2013 a 2019, de estudantes dos cursos de graduação de um *campus* do Instituto Federal de Minas Gerais (IFMG), extraídos do Sistema de Registro e Controle Acadêmico da instituição. As informações sobre o itinerário formativo de cada curso foram extraídas dos respectivos Projetos Pedagógicos de Curso (PPC).

A opção de iniciar as análises em 2013 decorre do fato de que nesse ano ocorreu o primeiro processo seletivo após a Lei nº 12.711/2012, que dispõe sobre a reserva de vagas em universidades e institutos federais, como meio de inclusão.

Para trabalhar apenas as informações relevantes, variáveis consideradas inelegíveis para a modelagem foram excluídas. Aos valores faltantes, optou-se por imputá-los de maneira automática na fase de limpeza, via algoritmo *k-Nearest Neighbour*, que selecionou um valor representativo substituto dentre os dez vizinhos mais próximos. Inconsistências foram corrigidas e atributos foram uniformizados e, em alguns casos, reduzidos a uma quantidade menor de níveis categóricos.

Novos atributos foram criados, derivados do cruzamento entre os dados originais e informações extraídas do PPC, resultando nos indicadores Coeficiente de Rendimento do Período (CR), Nota Média do Período (NMP), Frequência Relativa

no Período (FRP), Dependências Acumuladas ao Longo dos Períodos (DAP) e Percentual de Aproveitamento de Créditos (PAC), que sumarizam informações sobre assiduidade, esforço individual e desempenho do estudante, período a período. Esses indicadores serão detalhados na Tabela 1. Ao final dessa etapa, os dados foram reduzidos para 14 atributos originais e sete atributos derivados, que se repetem do primeiro ao quinto período. A identificação do período, ao qual o atributo se refere, é realizada pelo acréscimo do número correspondente ao período e à letra P (PAC3P, por exemplo, que representa a variável PAC no terceiro período).

As variáveis numéricas contínuas e discretas foram normalizadas usando-se *z-score*. A descrição da base de dados final consta na Tabela 1.

Tabela 1 - Base de dados utilizada

Nº	Atributos	Descrição	Domínio
1	Curso	Nome do curso do estudante	• Administração • Ciência da Computação • Engenharia Elétrica • Gestão Financeira • Matemática
2	Idade	Idade no instante da matrícula, em anos completos	Variável discreta
3	Cotas	Caso o estudante ingresse por algum sistema de cotas	• Cotista (independente da modalidade) • Não cotista
4	Tipo de ingresso	Indica a forma de acesso	• Instrumento interno de seleção (vestibular, transferência ou obtenção de novo título) • Sistema de Seleção Unificada (SISU)
5	Sexo	Gênero do estudante	• Feminino • Masculino
6	Raça	Raça declarada pelo estudante	• Branca • Negra/Parda/Amarela/Outros
7	Trabalha	Indica se o estudante trabalha	• Sim • Não
8	Onde estudou	Natureza da escola onde cursou até o Ensino Médio	• Somente na rede pública • Total ou parcialmente na rede privada

Continua

Continuação

Nº	Atributos	Descrição	Domínio
9 e 10	Grau de instrução dos pais	Último grau de escolaridade obtido pelo pai e pela mãe	• Até o Ensino Fundamental (sem estudo, com Ensino Fundamental incompleto ou completo e Ensino Médio incompleto) • Ensino Médio (Ensino Médio completo ou Ensino Superior incompleto) • Ensino Superior completo (Ensino Superior completo, pós-graduação incompleta ou completa)
11	Tipo de residência	Natureza contratual do tipo de residência	• Própria • Outros
12	Área de procedência	Localidade da origem	• Urbana • Rural
13	Renda Familiar	Renda familiar, em salários mínimos	Variável contínua
14	Pessoas na família	Número de pessoas que integram o núcleo familiar	Variável discreta
15	Períodos cursados	Total de períodos cursados, sendo usados para derivar outros atributos	Variável Inteira Positiva Não Nula
16	PAC: Percentual de Aproveitamento de Créditos	Aproveitamento da carga horária, de acordo com PPC vigente no ato da matrícula (razão entre a carga horária acumulada cumprida com êxito e a acumulada prevista)	Variável contínua
17	CR: Coeficiente de Rendimento por Período	Coeficiente de rendimento a partir das disciplinas cursadas, com ou sem aprovação, no período. Média ponderada da nota pelo total de créditos	Variável contínua
18	DAP: Dependências Acumuladas	Número de Dependências Acumuladas até o período	Variável discreta
19	FRP: Frequência Relativa no Período	Taxa de frequência global para as disciplinas matriculadas no período	Variável contínua
20	NMP: Nota Média no Período	Média simples das notas obtidas para cada disciplina no período	Variável contínua
21	FLAG: Indicador de Evasão	Flag que indica o *status* do estudante no período. Usado para filtrar registros inválidos	• Evade-se • Permanece • Não Computado

Fonte: Elaborada pelos autores (2020)

Os atributos de nº 1 a 14 mapeiam variáveis *ex-ante* ao ingresso e são coletados no ato da matrícula. A partir do término do primeiro período cursado, as variáveis de desempenho, assiduidade e esforço individual, representadas pelas variáveis *ex-post* de nº 15 a 20, são incluídas na base de dados. A variável nº 21 é utilizada exclusivamente para selecionar registros válidos, por período.

O desbalanceamento de casos positivos e negativos foi amenizado pelo uso da técnica ROSE (*random oversampling example)* (FERNANDEZ HILARIO *et al*., 2018), que utiliza *bootstrap* suavizado na vizinhança de casos positivos para gerar novos exemplos da classe minoritária.

A seguir, o conjunto final de dados foi particionado em dois subconjuntos, um para treinamento (70% dos exemplos), e outro para teste/validação (30% dos exemplos), respeitando-se a proporção de exemplos positivos (evasão) e negativos (permanência) existente na base após balanceamento.

3.2 Mineração de Dados Educacionais (MDE) por Aprendizado de Máquina

Segundo Goldschmidt*,* Bezerra, Passos*,* (2015), a MDE é um conjunto de esforços para a descoberta de conhecimentos, a partir de bases de dados educacionais, transformando dados brutos em informações úteis, que contribuem tanto para a pesquisa em Educação quanto para a prática do processo educacional, podendo ser utilizado para descrição e identificação de características estudantis, via classificação ou clusterização.

O processo de classificação pode ser realizado via Aprendizado de Máquina, que é uma área de pesquisa da Inteligência Artificial, cujo objetivo é desenvolver programas de computador capazes de aprender regras de decisão a partir de sua experiência.

Em Aprendizado de Máquina, os algoritmos aprendem a induzir uma função ou hipótese capaz de resolver um problema a partir de um conjunto de dados (FACELI *et al*., 2011). Empregam principalmente dois tipos de técnicas: o aprendizado supervisionado e não supervisionado. No presente trabalho, o foco é o aprendizado supervisionado, que utilizará o conjunto de dados preparado para aprender e identificar os fatores significantes para a evasão, construindo um modelo de predição que permita a identificação dos estudantes com potencial risco de evasão antes que ela se concretize.

Assim, para predizer casos de estudantes em situação de evasão (positivo) ou de permanência (negativo), em seis janelas temporais distintas (da matrícula até o quinto período), foram utilizados quatro algoritmos de classificação: Máquina de Vetores de Suporte (SVM, do inglês*: Support Vector Machine*), Floresta Aleatória (RF, do inglês: *Random Forest*), *Gradient Boosting Machines* (GBM) e um comitê de máquina (*Ensemble*) com os três algoritmos anteriores. A partição de dados de treinamento foi utilizada para ajustar os modelos via validação cruzada 10-*fold* com dez repetições.

A avaliação de modelos é feita através da matriz de confusão, que registra a quantidade de Verdadeiros Positivos (VP, quando o estudante em situação de evasão é classificado corretamente pelo modelo), Verdadeiros Negativos (VN, quando o estudante em situação de permanência é classificado corretamente pelo modelo), Falsos Positivos (FP, casos em que o modelo prevê que um estudante evade quando ele, na verdade, permanece no curso) e Falsos Negativos (FN, é quando o modelo classifica a situação do estudante como permanência, quando na realidade ele evade). A partir dessas quatro medidas, é possível calcular as métricas Acurácia, Especificidade, Sensibilidade, Precisão, Taxa de Falsos Positivos e Coeficiente Kappa, cuja interpretação é apresentada no Quadro 1.

Quadro 1 - Métricas e sua interpretação no contexto do trabalho

Métrica	Cálculo	Interpretação
Acurácia (ACC)	$\frac{VP + VN}{VP + VN + FP + FN}$	Percentual de classificações corretas de evasão ou de permanência, em relação ao total de classificações.
Sensibilidade (SEN)	$\frac{VP}{VP + FN}$	Representa a capacidade do modelo em identificar corretamente os indivíduos que evadem.
Especificidade (SPE)	$\frac{VN}{FP + VN}$	Capacidade do modelo em identificar corretamente os indivíduos que permanecem no curso.
Precisão (PRE)	$\frac{VP}{VP + FP}$	Probabilidade de ocorrência de evasão quando o modelo sugere caso positivo.
Taxa de Falso Positivo (TFP)	1 – Especificidade	Proporção de falsos positivos dentre todos os exemplos cuja classe esperada é negativo.
Coeficiente Kappa	$\frac{ACC_{Classificador} - ACC_{Aleatória}}{1 - ACC_{Aleatória}}$	Compara a acurácia do classificador proposto com a acurácia de um classificador aleatório (*baseline*).

Fonte: Elaborado pelos autores (2020)

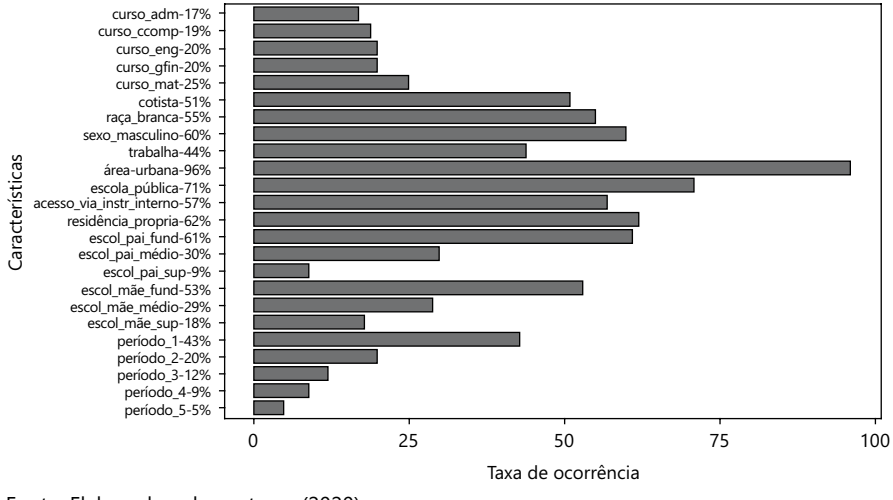
Além das métricas apresentadas, pode ser utilizado os Gráficos ROC (*Receiver Operating Characteristics*), que segundo Faceli *et al*. (2011), evidenciam o contraste entre a inabilidade do modelo de evitar falsos alarmes e sua a habilidade em detectar a classe positiva corretamente. É formado pela representação gráfica entre a Taxa de Falso Positivo pela Sensibilidade, com uma reta diagonal denominada "linha sem discriminação", que separa o ROC em espaço de boa classificação (acima da linha) ou de classificação ruim (abaixo da linha).

4 Análise dos resultados

O perfil sociodemográfico dos estudantes é caracterizado por conjunto de informações *ex-ante* ao ingresso do estudante no curso e que sofrem pouco ou nenhuma mudança durante a vida escolar do estudante. Características que influenciam o compromisso inicial do estudante, tanto com a instituição quanto com a conclusão do curso, são elementos que podem interferir no sentimento de adesão e pertencimento, e, portanto, importantes para analisar como a singularidade de cada estudante pode levá-lo a permanecer ou evadir do curso (MASSI; VILLANI, 2015).

A Figura 2 apresenta o percentual de estudantes, dentre os evadidos, segundo curso, cota, raça, sexo, área de procedência, natureza da escola nos níveis anteriores, número de membros da família, se trabalhava no instante do ingresso, escolaridade dos pais e períodos cursados antes de evadir. Para os atributos categóricos com níveis binários apenas um dos níveis foi apresentado, de forma que o restante do percentual é atribuído ao nível complementar.

Figura 2 - Perfil dos estudantes evadidos segundo variáveis *ex-ante*



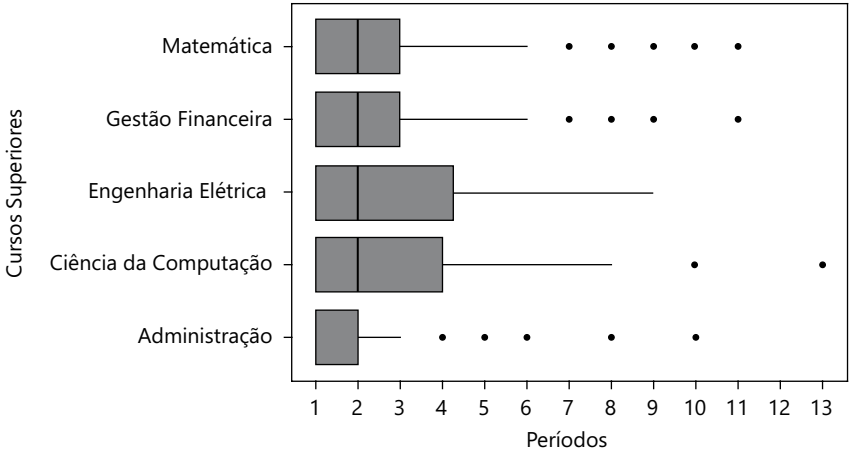
Fonte: Elaborada pelos autores (2020)

A evasão total observada foi de 614 registros dos 1.429 em análise, representando, aproximadamente, 43%, valor superior ao encontrado para a média nacional que, de acordo Silva Filho (2017), é de 22%. Dentro desses 614 registros, é possível verificar similaridade na ocorrência de evasão nos cursos em análise, sendo que a Licenciatura em Matemática apresentou a maior taxa de evasão dentre os cursos (25%). Os cursos de Administração, Ciência da Computação, Engenharia Elétrica e Gestão Financeira correspondem a aproximadamente 17%, 19%, 20% e 20% dos estudantes em situação de evasão, respectivamente.

Taxas maiores de evasão no curso de Licenciatura em Matemática também foram observadas em outros estudos. Conforme Souto (2016), existem tanto a evasão do educando no curso de Licenciatura em Matemática quanto do educador na profissão. Segundo o autor, embora a profissão tenha notória importância, ela não é valorizada no Brasil, o que pode ser expresso pelos baixos salários e condições de trabalho, elementos que afetam negativamente o compromisso do estudante com o curso.

Nas análises, constatou-se que 63% das evasões ocorrem entre o primeiro e segundo períodos. A análise do Diagrama de Caixa (Figura 3), construído a partir da estratificação dos dados por período e curso, demonstra que, para o curso de Administração, metade das evasões ocorrem já no primeiro período, enquanto que para os demais cursos esse fato é observado no segundo período.

Figura 3 - Diagrama de Caixa do número de períodos cursados antes da evasão, por curso e período



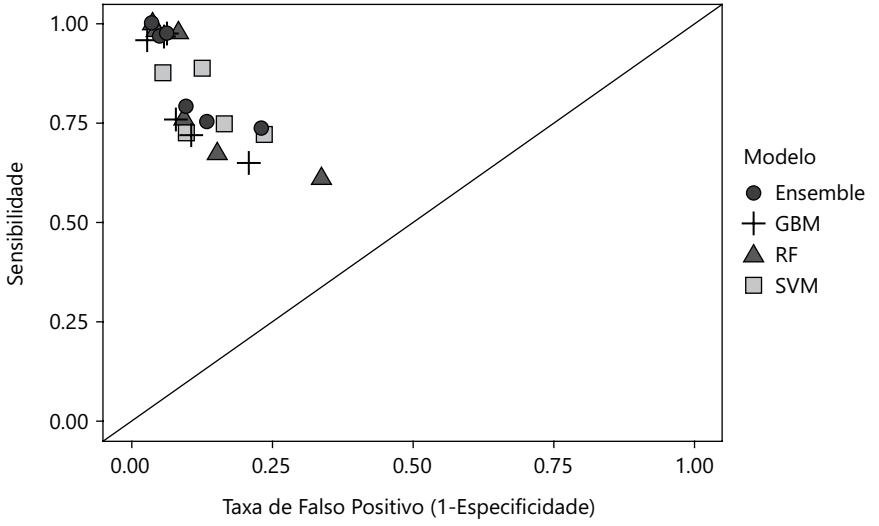
Fonte: Elaborada pelos autores (2020)

Ainda pelo Diagrama de Caixa, nota-se que 75% das evasões ocorrem até o quinto período, de forma que esses serão o número máximo de períodos a serem considerados na análise. Definir esse intervalo limite permite que os esforços da instituição se concentrem em janelas temporais nas quais a evasão é mais frequente. Essas medidas demonstram alinhamento como os apontamentos feitos por Tinto (2006) e Matta, Lebrão e Heleno (2017), que indicam o primeiro ano do estudante na Educação Superior como o mais crítico.

Segundo Tinto (2006), buscar o envolvimento do estudante no primeiro ano da Educação Superior é importante, devendo a instituição implementar ações que enriqueçam a experiência do estudante na escola. Conforme Matta, Lebrão e Heleno (2017), ao ingressar no Ensino Superior, o estudante sofre um impacto desse novo nível/modelo de ensino, ocasionando mudanças, tanto positivas quanto negativas, nesses estudantes. Ainda segundo esses autores, a instituição deve elaborar ações educacionais preventivas e de orientação aos estudantes, buscando facilitar integração, a permanência e, sobretudo, o sucesso acadêmico.

Isso posto, é imprescindível que a instituição acompanhe, período a período, os estudantes com potencial risco de evasão, atuando com antecedência, na tentativa de evitar o abandono. No presente trabalho, propõe-se que a sinalização dos estudantes que precisam ser acompanhados ocorra via predição, utilizando os algoritmos propostos, construídos a partir de base de dados existentes na instituição. Desta forma, investigou-se a capacidade preditiva dos algoritmos SVM, GBM, RF e comitê de máquina para identificar estudantes em situação de abandono. Os modelos ajustados foram aplicados ao conjunto de teste para os diferentes cursos e períodos, sendo que a comparação de desempenho será feita pelo gráfico ROC apresentado na Figura 4.

Figura 4 - Gráfico ROC para os classificadores



Fonte: Elaborada pelos autores (2020)

Pela Figura 4, observa-se que o comitê de máquina concentrou melhor os valores de sensibilidade e taxa de falsos positivos no canto superior esquerdo, com menor dispersão comparativamente aos demais métodos. Em função da melhor adequação do comitê de máquina para a classificação dos dados de evasão e permanência, todas as demais construções e análise serão realizadas com este preditor. O Quadro 2 apresenta as matrizes de confusão com os resultados da validação do comitê de máquina para cada período. O cruzamento das linhas S com colunas S representam os VPs; o cruzamento das linhas N com colunas N representam os VNs; o cruzamento das linhas S com colunas N são os FPs, e o cruzamento das linhas N com colunas S são os FNs. As medidas de desempenho do comitê de máquina são reportadas na Quadro 3, período a período.

Quadro 2 - Matriz de confusão resultante do uso do comitê nos dados teste

Matriz de Confusão	*Ex-ante*		*Ex-post*									
	Matrícula		Período 1		Período 2		Período 3		Período 4		Período 5	
	S	N	S	N	S	N	S	N	S	N	S	N
S	150	50	162	19	105	20	122	6	98	3	85	7
N	53	175	41	206	42	137	5	129	0	102	2	88

Fonte: Dados da pesquisa (2020)

Quadro 3 - Desempenho do comitê de máquina para os cinco primeiros períodos

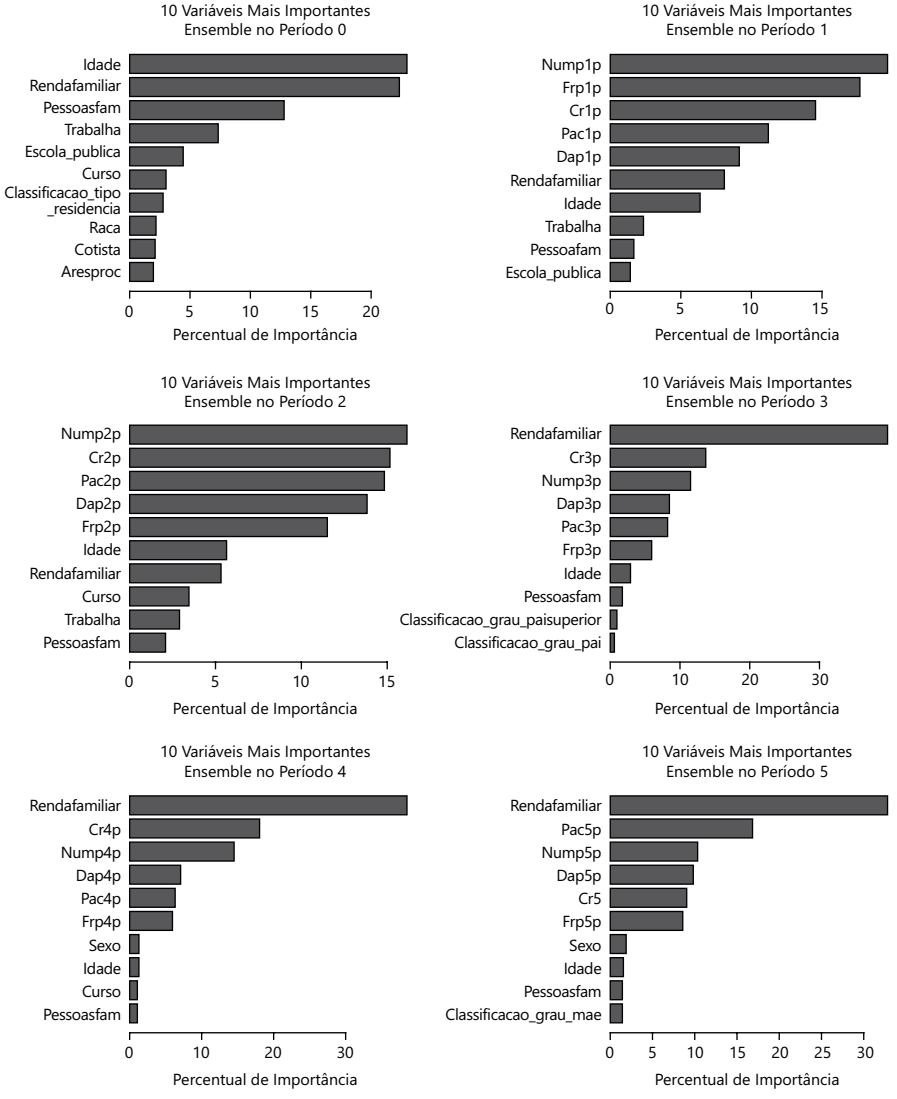
Medida	*Ex-ante*	*Ex-post*				
	Matrícula	**Período 1**	**Período 2**	**Período 3**	**Período 4**	**Período 5**
ACC (%)	75,93	85,98	79,61	95,80	98,52	95,05
SEN (%)	73,89	79,80	71,43	96,06	100,00	97,70
SPE (%)	77,78	91,56	87,26	95,56	97,14	92,63
PRE (%)	75,00	89,50	84,00	95,31	97,03	92,39
KAPPA (%)	51,71	71,74	58,97	91,60	97,04	90,11

Fonte: Dados da pesquisa (2020)

A capacidade do preditor comitê de máquina em identificar corretamente um caso positivo (SEN%) foi de 73,89% a 100% e para identificar os estudantes que permanecem no curso (SPE%) de 77,78% a 97,14%. A taxa de sucesso global (ACC%) e a precisão (PRE%) em indicar um estudante em situação de abandono no período com apenas informações *ex-ante,* foram superiores a 75%. A precisão (PRE%) foi de 84% a partir do momento que as variáveis de desempenho, assiduidade e esforço individual passaram a ser incorporadas ao modelo.

Como resultado da predição, a Figura 5 reporta as dez variáveis mais importantes para determinar o perfil do estudante com maior risco de evasão, calculadas pelo comitê.

Figura 5 - Importância das variáveis para o perfil de evasão por período



Fonte: Elaborada pelos autores com base na descrição das variáveis da Tabela 1 (2020)

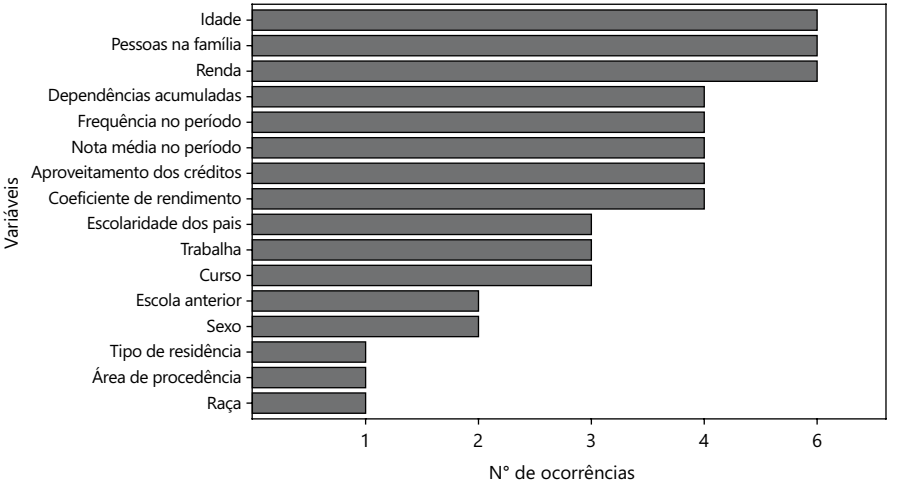
Compreender as variáveis que são significativas ao comprometimento do estudante em concluir um curso ou dele evadir permite a elaboração de um processo de triagem para identificação de estudantes com potencial risco de evasão (TINTO, 1997). Uma vez conhecidas as variáveis mais importantes, retomam-se as

informações constantes na Figura 2, para construção do perfil do estudante que deverá ser monitorado e receber algum tipo de intervenção pela instituição.

Os gráficos apresentados na Figura 5 indicam que as variáveis *ex-ante* são essenciais para explicar a evasão de estudantes que acabam de ingressar na instituição (Período 0). Contudo, as variáveis *ex-post* passam a figurar entre as variáveis de maior importância, assim que o primeiro período é efetivamente encerrado, fato observado até o quinto período e que corrobora com a melhoria da capacidade preditiva do comitê, como observado nas métricas apresentadas no Quadro 3.

Para sumarizar os resultados reportados pela Figura 5, um novo gráfico foi construído destacando dezessete variáveis que figuraram ao menos uma vez dentre as dez mais importantes nas seis janelas temporais analisadas. Ele mostra o número de ocorrência dessas variáveis, independentemente de sua posição na classificação e período. O gráfico resultante é apresentado na Figura 6 e dará suporte para discutir a caracterização de um perfil do estudante em situação de abandono.

Figura 6 - Variáveis importantes, segundo sua ocorrência para todos os períodos



Fonte: Elaborada pelos autores (2020)

Avaliando as variáveis categóricas eleitas como mais importantes pelo comitê de máquinas, e associando-as à frequência com que seus valores ocorrem dentro da classe de estudantes evadidos (Figura 2), é possível traçar o perfil daqueles

com maior risco de evasão, que, em sua maioria, é caracterizado por estudantes brancos (55%), que não trabalham (66%), do sexo masculino (60%), que residem em casa própria (62%), na área urbana (96%), com pais com escolaridade até o Ensino Fundamental (pai 61% e mãe 53%) e que são oriundos de escolas públicas em toda sua formação acadêmica anterior (71%). A variável cotista foi elencada como importante pelo modelo preditivo, mas foi desconsiderada por apresentar pouca diferenciação para ingresso por cotas (51% cotista e 49% não cotista).

De forma complementar à análise das variáveis categóricas, faz-se necessária a análise das variáveis numéricas. Diferentemente das variáveis categóricas, as quais eram compostas, em sua totalidade, por variáveis *ex-ante*, dentre as variáveis numéricas, apenas três são atributos oriundos da matrícula do estudante: renda, pessoas por família e idade. Como as variáveis *ex-ante* sofrem pouca ou nenhuma alteração ao longo da vivência acadêmica do estudante na instituição, elas são traçadas uma única vez, mantendo-se constantes ao longo das janelas temporais.

Em relação às variáveis *ex-ante* numéricas, observam-se que estudantes com renda familiar de até 3 salários mínimos, com família de 3 a 5 pessoas e idade de 18 a 23 anos representam, aproximadamente 70% dos estudantes em situação de evasão. Embora a renda *per capita* não figure entre as variáveis modeladas, ela foi calculada a partir dos dados da renda e do número de pessoas da família, permitindo identificar que, aproximadamente, 80% dos estudantes que evadiram são de famílias com renda *per capita* de até um salário mínimo. Essa análise mostrou-se relevante, pois a variável renda e número de pessoas na família foram elencadas como importantes em todas as janelas temporais em estudo, sugerindo que a renda *per capita* também impacta na chance de evasão.

Dada a dinâmica das variáveis *ex-post,* que resulta do processo interativo do estudante e sua vida acadêmica, existirá para essas variáveis um perfil específico de risco de evasão a cada janela temporal em avaliação, são dados observados na Tabela 2 e que evidenciam o evasor como tendo perfil de baixo desempenho acadêmico e elevadas taxas de absenteísmo.

Tabela 2 - Perfil de risco de Evasão, segundo variáveis *ex-post*

Janela Temporal	Perfil com maior risco de evasão segundo informações *ex-post*
Período 1 (266 evadidos)	Em 80% dos casos, o CR é inferior a 20%, a nota média até 15 pontos, PAC 0, e com 5 a 6 dependências. Apenas 55% dos evadidos frequentaram até 75% da carga horária.

Continua

Continuação

Janela Temporal	Perfil com maior risco de evasão segundo informações *ex-post*
Período 2 (120 evadidos)	Em 80% dos casos, o CR é inferior a 30%, a nota média até 25 pontos, PAC máximo de 37% e 9 dependências. Aproximadamente 65% frequentaram até 75% da carga horária.
Período 3 (71 evadidos)	Em 80% dos casos, o CR é inferior a 18%, a nota média até 21 pontos, PAC de até 41% e 11 dependências. Aproximadamente 55% frequentaram até 75% da carga horária.
Período 4 (57 evadidos)	Em 80% dos casos, o CR é inferior a 28%, a nota média até 28 pontos, PAC até 45% e 15 dependências. Aproximadamente 67% frequentaram até 75% da carga horária.
Período 5 (32 evadidos)	Em 80% dos casos, o CR é inferior a 40%, a nota média até 21 pontos, PAC máximo de 55% e com 18 a 19 dependências. Aproximadamente 65% frequentaram até 75% da carga horária.

Fonte: Dados da pesquisa (2020)

A renda familiar, número de pessoas na família e idade são atributos com alto grau de relevância, assim como os atributos relacionados ao desempenho acadêmico, esforço individual e assiduidade, pois foram indicadas como variáveis importantes em todos os períodos nos quais existem seus registros. Para Fritsch, Rocha e Vitelli (2015), o bom desempenho acadêmico tem impacto positivo na integração acadêmica, assim como a boa frequência e o cumprimento de créditos não descartam a necessidade de avaliação das demais variáveis, e que podem ajudar a instituição a priorizar ações.

Prestes e Filho, (2018) sugerem que as experiências de sucesso escolar se iniciam antes mesmo do ingresso do estudante no Ensino Superior. Lima Junior *et al.* (2019) sugerem que ações internas e definidas pelas próprias instituições podem impactar positivamente na taxa de permanência dos estudantes e, consequentemente, reduzirem o risco de evasão.

Conforme Pereira (2019), a redução da taxa de evasão deve ser um objetivo, mas a solução não é simples e nem sempre óbvia, principalmente porque a evasão é um processo gradual.

As contribuições obtidas por meio das técnicas quantitativas são valiosas, mesmo considerando que a natureza da evasão pode ser associada a diferentes fatores que, em alguns casos, extrapolam os parâmetros disponíveis nos sistemas das escolas/governo ou que não são passíveis de modelagem. Assim, mesmo sabendo da limitação de modelos em sua capacidade de representar a realidade em sua

totalidade, o monitoramento e o controle, a fim de estabelecer as razões pelas quais um estudante não tem êxito na conclusão do seu curso, só é viável se a instituição usar os dados que já fazem parte da rotina escolar e que já constam em sistemas internos.

Diante do desempenho do comitê de máquina, da identificação de variáveis importantes que permitem construir o perfil do estudante em risco de evasão, e da disponibilidade de informações prévias que constam de sistemas acadêmicos institucionais, propõem-se a implantação de um processo de detecção e acompanhamento do risco de evasão, representado na Figura 7 e que envolve quatro etapas: (1) aquisição: coleta periódica de dados do sistema de registro e controle acadêmico; (2) identificação dos estudantes com maior risco de evasão: treinamento do modelo preditivo e classificação dos indicados de acordo com valores críticos, segundo análise de variáveis importantes; (3) monitoramento dos estudantes indicados e ações correspondentes ao perfil: para cada perfil, ações mitigadoras devem ser tomadas e acompanhadas pela instituição; (4) revisão do processo: ações e boas práticas devem ser revisitadas periodicamente, em função da experiência vivenciada.

Figura 7 - Processos de detecção e acompanhamento do risco de evasão proposto

Fonte: Elaborada pelos autores (2020)

A efetividade do processo proposto depende de três pressupostos: (i) o processo deve ser realizado a cada período, pois o suporte contínuo dado ao estudante poderá afetar sua decisão em permanecer ou evadir; (ii) as medidas a serem tomadas devem ter articulação da estratégia ao planejamento institucional da escola; e (iii) o modelo de predição precisa ser revisado periodicamente, pois mudanças no padrão dos dados podem torná-lo ineficaz para predição.

5 Conclusão

O presente trabalho estudou o fenômeno de evasão de estudantes de cinco cursos de graduação do IFMG, através de Mineração de Dados Educacionais por meio de quatro técnicas de Aprendizado de Máquina, a saber: SVM, GBM, RF e comitê de máquina. Os modelos preditivos foram aplicados a 1.429 registros de estudantes matriculados entre os anos de 2013 a 2019, cujo período analisado é da matrícula até o quinto período cursado. Os resultados apontam que o comitê de máquina apresentou desempenho superior aos demais métodos.

O modelo preditor proposto permitiu a identificação da importância das variáveis no risco de evasão, viabilizando a obtenção de perfil geral do estudante evasor da instituição. Para o caso em estudo, o perfil é caracterizado, em sua maioria, por estudantes brancos, que não trabalham, do sexo masculino, que residem em casa própria, na área urbana, com pais com escolaridade até o Ensino Fundamental, oriundos de escolas públicas em toda sua formação acadêmica anterior, renda familiar de até três salários mínimos, com família de três a cinco pessoas, idade de 18 a 23 anos, baixo desempenho acadêmico e elevadas taxas de absenteísmo.

Conhecer esse perfil do estudante evasor é importante para criar estratégias de mitigação do risco de evasão, sendo que o acompanhamento, período a período, pode trazer maior efetividade a essas estratégias, visto que o comprometimento do estudante em permanecer ou evadir de um curso é dinâmico ao longo do tempo.

Dada a adequação do modelo preditivo ao estudo realizado, o produto educacional resultante desta investigação é uma proposta de um processo de detecção e acompanhamento do risco de evasão. A instituição que desejar reproduzir as análises do presente estudo e adaptá-la à sua realidade poderá acessar os materiais disponibilizados em https://bit.ly/2VVVmWH. Ressalta-se que a implantação do processo sugerido, incorre com baixo ou nenhum custo adicional, uma vez que foram utilizados softwares livres, com foco em dados já usualmente registrados em sistemas de informação institucionais das escolas.

May Artificial Intelligence support actions against school dropout?

Abstract

School dropout is a world-level concern due to the negative consequences that it brings to society, so it is important to investigate it to understand and act to mitigate dropout risk. This work proposes the use of Educational Data Mining with Machine Learning to identify variables that are important to characterize the student profile in risk. Support Vector Machine, Gradient Boosting Machine, Random Forest and Ensemble were applied to 1,429 records of undergraduate students in a campus of the IFMG, between 2013 and 2019. The results suggest that Ensemble had the best performance, so it was used to compute the variable importance related to dropout prediction. We used the importance of tracing the student profile of dropout, and proposing a detection and monitoring process to avoid school dropout.

Keywords: *Dropout. Machine Learning. Undergraduate Students.*

¿Puede la inteligencia artificial apoyar acciones contra la deserción escolar universitaria?

Resumen

La deserción escolar es una preocupación global por sus consecuencias negativas para la sociedad en su conjunto, y es necesario investigarla para comprenderla y actuar con anticipación, mitigando su riesgo de ocurrencia. Este trabajo propone el uso de técnicas de Minería de Datos Educativos con técnicas de Aprendizaje de Máquina para identificar las variables que son importantes para caracterizar el perfil del estudiante en riesgo de deserción. . Las técnicas de Support Vector Machine, Gradient Boosting Machine, Random Forest y Machine Committee se aplicaron a 1.429 registros de estudiantes de cursos de educación superior en uno de los campus de IFMG, entre 2013 y 2019. Los resultados obtenidos sugieren un desempeño superior del comité de máquinas, a través del cual se obtuvo la importancia de las variables sobre el fenómeno en estudio, lo que permitió trazar el perfil del alumno desertor, por período. Estos resultados permitieron proponer un proceso de detección y seguimiento de estos estudiantes.

Palabras clave: *Deserción. Aprendizaje de Máquina. Estudiantes Universitarios.*

Referências

BAGGI, C. A. S.; LOPES, D. A. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação*, Campinas, v. 16, n. 2, p. 355-374, jul. 2011. https://doi.org/10.1590/S1414-40772011000200007

CHUNG, J. Y.; LEE, S. Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, Elmsford, v. 96, p. 346-353, Jan. 2019. https://doi.org/10.1016/j.childyouth.2018.11.030

COLPANI, R. Mineração de Dados Educacionais: um estudo da evasão no ensino médio com base nos indicadores do Censo Escolar. *Informática na Educação: Teoria & Prática*, Porto Alegre, v. 21, n. 3, p. 143-157, set.dez. 2018.

DIGIAMPIETRI, L. A.; NAKANO, F.; LAURETTO, M. S. Mineração de dados para identificação de alunos com alto risco de evasão: um estudo de caso. *Revista de Graduação USP*, São Paulo, v. 1, n. 1, p. 17-23, jul. 2016. https://doi.org/10.11606/issn.2525-376X.v1i1p17-23

FACELI, K. *et al*. *Inteligência artificial*: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

FERNÁNDEZ HILARIO, A. *et al*. *Learning from imbalanced data sets*. Berlin: Springer, 2018.

FIGUEIREDO, N. G. S.; SALLES, D. M. R. Educação profissional e evasão escolar em contexto: motivos e reflexões. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 25, n. 95, p. 356-392, abr. 2017. https://doi.org/10.1590/S0104-40362017002500397

FRITSCH, R.; ROCHA, C. S.; VITELLI, R. F. A evasão nos cursos de graduação em uma instituição de ensino superior privada. *Revista Educação em Questão*, Natal, v. 52, n. 38, p. 81-108, ago. 2015. https://doi.org/10.21680/1981-1802.2015v52n38ID7963

GOLDSCHMIDT, R.; BEZERRA, E.; PASSOS, E. *Data mining*: conceitos, técnicas, algoritmos, orientações e aplicações. Rio de Janeiro: Elsevier, 2015.

HOFFMANN, I. L.; NUNES, R. C.; MULLER, F. M. As informações do Censo da Educação Superior na implementação da gestão do conhecimento organizacional sobre evasão. *Gestão & Produção*, São Carlos, v. 26, n. 2, p.e2852, 2019. https://doi.org/10.1590/0104-530X-2852-19

KNOWLES, J. E. Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining*, Sydney, v. 7, n. 3, p. 18-67, July 2015. https://doi.org/10.5281/zenodo.3554725

LIMA JUNIOR, P. *et al*. Taxas longitudinais de retenção e evasão: uma metodologia para estudo da trajetória dos estudantes na educação superior. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 27, n. 102, p. 157-178, jan./mar. 2019. https://doi.org/10.1590/S0104-40362018002701431

MASSI, L.; VILLANI, A. Um caso de contratendência: baixa evasão na licenciatura em química explicada pelas disposições e integrações. *Educação e Pesquisa*, São Paulo, v. 41, n. 4, out./dez. 2015. https://doi.org/10.1590/s1517-9702201512135667

MATTA, C. M. B.; LEBRÃO, S. M. G.; HELENO, M. G. V. Adaptação, rendimento, evasão e vivências acadêmicas no ensino superior: revisão da literatura. *Psicologia Escolar e Educacional*, São Paulo, v. 21, n. 3, p. 583-591, set./dez. 2017. https://doi.org/10.1590/2175-353920170213111118

PEREIRA, M. C. Evasão escolar: causas e desafios. *Revista Científica Multidisciplinar Núcleo do Conhecimento*, São Paulo, v. 4, n.2, p. 36-51, fev. 2019.

PRESTES, E. M. T.; FIALHO, M. G. D. Evasão na educação superior e gestão institucional: o caso da Universidade Federal da Paraíba. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 26, n. 100, p. 869-889, jul./set, 2018. https://doi.org/10.1590/S0104-40362018002601104

ROBISON, S. *et al*. Correlates of educational success: predictors of school dropout and graduation for urban students in the Deep South. *Children and Youth Services Review*, Elmsford, v. 73, p. 37-46, Feb. 2017. https://doi.org/10.1016/j.childyouth.2016.11.031

ROVIRA, S.; PUERTAS, E.; IGUAL, L. Data-driven system to predict academic grades and dropout. *PLoS ONE*, San Francisco, v. 12, n.2, e0171207, 2017. https://doi.org/10.1371/journal.pone.0171207

SANTOS JUNIOR, J. S.; REAL, G. C. M. O acesso à educação superior na Universidade Federal da Grande Dourados: trajetória de estudantes ingressantes entre 2006-2009. *Revista Brasileira de Política e Administração da Educação*, v. 33, n. 2, p. 467-492, jan./abr. 2017. https://doi.org/10.21573/vol33n22017.71081

SILVA, A. M.; SANTOS, B. C. S. Eficácia de políticas de acesso ao ensino superior privado na contenção da evasão. *Avaliação*, Campinas, v. 22, n. 3, p. 741-757, nov. 2017. https://doi.org/10.1590/S1414-40772017000300009

SILVA FILHO, R. B.; ARAÚJO, R. M. L. Evasão e abandono escolar na educação básica no Brasil: fatores, causas e possíveis consequências. *Educação por Escrito*, Porto Alegre, v. 8, n. 1, p. 35-48, 2017. https://doi.org/10.15448/2179-8435.2017.1.24527

SILVA FILHO, R. L. L. A evasão no ensino superior brasileiro: novos dados. *Estadão*, São Paulo, 7 out. 2017. Disponível em: https://educacao.estadao.com.br/blogs/roberto-lobo/497-2/. Acesso em: nov. 2019.

SOUTO, R. M. A. Egressos da licenciatura em matemática abandonam o magistério: reflexões sobre profissão e condição docente. *Educação e Pesquisa*, São Paulo, v. 42, n. 4, p. 1077-1092, out./dez. 2016. https://doi.org/10.1590/S1517-9702201608144401

TINTO, V. Classrooms as communities: exploring the educational character of student persistence. *Journal of Higher Education*, Philadelphia, v. 68, n. 6, p. 599-623, Nov./Dec. 1997. https://doi.org/10.2307/2959965

TINTO, V. Research and practice of student retention: What next? *Journal of College Student Retention: Research, Theory & Practice*, Amityville, v. 8, n. 1, p. 1-19, 2006. https://doi.org/10.2190/4YNU-4TMB-22DJ-AN4W

TONTINI, G.; WALTER, S. A. Pode-se identificar a propensão e reduzir a evasão de alunos?: ações estratégicas e resultados táticos para instituições de ensino superior. Avaliação, Campinas, v. 19, n. 1, p. 89-110, mar. 2014. https://doi.org/10.1590/S1414-40772014000100005

Informações sobre os autores

Wanderci Alves Bitencourt: Mestra em Administração pela Universidade Federal de Lavras. Professora do Instituto Federal de Minas Gerais. Contato: wanda.bitencourt@ifmg.edu.br

http://orcid.org/0000-0002-9509-7786

Diego Mello Silva: Mestre em Ciência da Computação pela Universidade Federal de Minas Gerais. Professor do Instituto Federal de Minas Gerais. Contato: diego.silva@ifmg.edu.br

http://orcid.org/0000-0001-8600-5598

Gláucia do Carmo Xavier: Doutora em Linguística e Língua Portuguesa pela Pontifícia Universidade Católica de Minas Gerais. Professora do Instituto Federal de Minas Gerais. Contato: glaucia.xavier@ifmg.edu.br

http://orcid.org/0000-0003-3133-7354