PROGNÓSTICO DA EVASÃO ESCOLAR EM INSTITUIÇÃO DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA POR MEIO DA INTELIGÊNCIA ARTIFICIAL

Carlos Vital Giordano

Doutor e professor no Mestrado Profissional em Gestão e Desenvolvimento da Educação Profissional do Centro Paula Souza – CEETEPS
giordanopaulasouza@yahoo.com.br | ORCID 0000-0002-5557-9529

Raphael Antonio de Souza

Mestrando no Mestrado Profissional em Gestão e Desenvolvimento da Educação Profissional do Centro Paula Souza - CEETEPS
raphael.souza@cpspos.sp.gov.br | ORCID 0000-0002-0952-1887

Resumo

A desistência escolar é problema preocupante para as instituições de ensino, a sociedade, os formuladores de políticas educacionais e alunos. A identificação precoce dos alunos com alta probabilidade de evasão é essencial no estabelecimento de ações de prevenção do distúrbio. A investigação objetiva verificar a opção da utilização de algoritmo de Aprendizado de Máquina (AM) na identificação de alunos com maior risco de evasão. Por meio de ferramenta de Autoaprendizado de Máquina (AutoML), analisou-se os dados de 1.222 alunos de curso técnico de IEPT (Instituição de Educação Profissional e Tecnológica). Após o pré-processamento dos dados, submeteu-se os dados purgados à ferramenta de Atol, e esta gerou modelo de algoritmo com acurácia superior a 90,0% quando da identificação de alunos com possibilidade de abandono. Os resultados do estudo demonstram a perspectiva favorável no uso da AM em dados educacionais.

Palavras-chave: Desistência; H2O; Auto ML; Aprendizado de Máquina.

Abstract

School dropout is a worrisome problem for educational institutions, society, educational policy makers, and students. The early identification of students with high dropout probability is essential in the establishment of actions to prevent the disorder.

The research aims to verify the option of using Machine Learning (ML) algorithm to identify students with higher dropout risk. Using an Auto Machine Learning (AutoML) tool, the data of 1,222 students from an IEPT technical course was analyzed. After pre-processing the data, the purged data was submitted to the AutoML tool, which generated an algorithm model with accuracy higher than 90.0% when identifying students with the possibility of dropping out. The study results demonstrate the favorable perspective in the use of AM in educational data.

Keywords: Dropout; H2O; AutoML; Machine Learning.

Introdução

A Instituição de Educação Profissional e Tecnológica (IEPT) em exame é uma autarquia federal de ensino, presente em diversos municípios com campus próprio ou avançado e mais de 62 mil alunos matriculados nos cursos técnicos de nível médio, superiores e pós-graduação *stricto e lato sensu.* Desses, de acordo com a Lei 11.892 de 29 de dezembro de 2008 (Brasil, 2008), 50% oferecem cursos técnicos nas modalidades concomitante/subsequente e integrado ao ensino médio. A IEPT conta com cursos técnicos nas modalidades concomitante/subsequente e integrado ao ensino médico, cursos superiores de tecnologia e bacharelados, licenciatura e pós-graduação *lato-sensu,* tendo registrado até o primeiro semestre de 2022, segundo dados do sistema acadêmico, 6.412 matrículas. Dessas, 2.731 (42,6%) se realizaram em cursos de nível técnico na modalidade concomitante/subsequente. Contudo, apenas 1.251 (45,8%) alunos concluíram ou estão matriculados na unidade. Os demais 1.480 alunos, ou seja, 54,2 % desistiram dos estudos, tendo seus *status* como trancado, evadido ou cancelado.

No ensino profissionalizante, assim como em outros níveis de ensino, a evasão escolar é um problema em destaque no Brasil. Dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2019 (IBGE, 2019), apontam que 20,2% dos jovens não completaram a educação básica. A pesquisa mostra que os indicadores de causas possíveis da evasão se centram na renda, na região de residência, no difícil deslocamento até as escolas, na cor ou raça, em fatores sociais, dentre outros. Desde a educação básica até o ensino superior o assunto é tema de pesquisas que apontam

propostas de discussão e solução para cada um dos fatores que desencadeiam o processo de evasão.

No último Relatório Sistêmico de Fiscalização do Tribunal de Contas da União (TCU, 2015), o órgão recomendou a instituição de plano voltado ao tratamento da evasão nos institutos federais, que contemple, entre outros aspectos, a identificação de alunos com maior propensão de abandono dos cursos e a alocação de profissionais para o acompanhamento escolar e social dos estudantes.

Identificar alunos com maior tendência a evadir-se do sistema escolar é tarefa que exige entendimento das características individuais e, portanto, é um processo requintado e às vezes, moroso. Nesse sentido o intuito da investigação é responder à seguinte pergunta: É possível empregar ferramentas de inteligência artificial e estatísticas avançadas a fim de identificar os discentes com alto risco de evasão em curso técnico, e assim proporcionar aos gestores educacionais subsídios para que minimizem o abandono escolar?

Objetiva-se em termos gerais com auxílio de ferramentas de inteligência artificial e estatísticas avançadas, esmiuçar os dados dos alunos matriculados, não concluídos e concluídos, com a finalidade de prognosticar a tendência de não conclusão dos novos alunos matriculados na IEPT, inscritos em um curso técnico.

Em complemento, estipulam-se os objetivos específicos:

a) Levantar os dados de todos os alunos já matriculados no campus da IEPT;

b) Identificar quais campos do registro de dados podem ser analisados e associados para a identificação do aluno concluinte e evadido;

c) Verificar o percentual de alunos concluintes;

d) Realizar pré-processamento no conjunto de dados;

e) Identificar os campos mais relacionáveis;

f) Aplicar ferramentas de análise de dados estatísticos e de inteligência artificial sobre os dados dos alunos, a fim de prognosticar os alunos com potencial de evasão;

g) Gerar modelo que possa ser replicado aos demais cursos da unidade e instituição.

Adiciona-se ainda, como hipótese H1: Algoritmos de aprendizado de máquina demonstram aptidão suficiente para predizer as eventuais evasões da unidade de ensino (caso da pesquisa).

O método empregado se baseia em pesquisa descritiva exploratória a partir de análise documental (dados primitivos), analisados por ferramenta de Autoaprendizado Máquina (AutoML – do inglês *Automated Machine Learning*).

Fundamentação Teórica

Inicia-se a fundamentação teórica pela breve revisão dos conceitos e das definições intrínsecos à investigação, tendo como um dos intuitos facilitar o entendimento dos temas pelo leitor.

Educação Profissional e Tecnológica

O ensino profissionalizante está arraigado no país desde o período colonial, passando por momentos marcantes ao longo dos primeiros séculos, como a criação dos Centros de Aprendizagem de Ofícios na Marinha do Brasil, a proibição de fábricas em todo território nacional em 1785 e a criação do Colégio de Fábricas em 1808, até a consolidação do ensino técnico-industrial em 1906 (MEC, 2007).

Em 2005, após a publicação da Lei 11.195 houve vertiginosa expansão da Rede Federal de Educação Profissional Ciência e Tecnologia (RFEPCT). Contudo, muitos dos alunos matriculados na Rede Federal, não concluem a formação. Conforme dados da Secretaria de Educação Profissional e Tecnológica (Setec) (SETEC, 2014), vinculada ao Ministério da Educação (MEC), a taxa de conclusão no ensino técnico na modalidade concomitante/subsequente é de 31,4%. O mesmo documento traz, portanto, orientações para superação da retenção e diminuição da evasão no âmbito da RFEPCT.

Evasão

A evasão escolar é um dos problemas que mais inquietam instituições de ensino, públicas ou privadas. O abandono escolar acarreta desperdícios sociais, acadêmicos e econômicos (Silva Filho et al., 2007). Entende-se como evasão escolar quando há a interrupção do ciclo escolar, seja do nível que for. Entender os fatores

que levam ao abandono escolar, tornaram-se foco de estudos desde a década de 1970, conforme apontados pelo estudo da SETEC (Setec, 2014).

Segundo Neri (2009), em estudo com cerne nos jovens de 15 a 17 anos, a falta de demanda, ausência de interesse por parte do educando e necessidade de trabalho, configuram-se como as principais causas do abandono escolar. Motivos esses corroborados por Filho e Araújo (2017) que apontam como causas que levam à evasão, a individual, que se relaciona ao estudante e as circunstâncias de seu percurso escolar; e a institucional, que se relaciona com a família, a escola, a comunidade e os grupos de amigos (Filho & Araujo, 2017). Informações essas, reafirmadas pela PNAD 2019, em que indica como principal motivo para o abandono escolar a necessidade de trabalhar, seguido por falta de interesse, gravidez e afazeres domésticos, quando respondentes do sexo feminino. Ainda de acordo com a PNAD 2019, necessidade de trabalhar e a falta de interesse, atingem 70% dos jovens nas grandes regiões (IBGE, 2019).

No entanto, a pesquisa do IBGE aponta apenas para dados referentes à formação de nível fundamental e médio. Segundo Dore e Lücher (2001) quando o assunto é o ensino de nível técnico, o problema da evasão é evidente, mas ainda faltam estudos que levem a um entendimento dos fatores ligados à evasão, tornando o trabalho de prevenção tarefa bastante difícil. De fato, encontra-se na literatura centenas de artigos relacionados a evasão em cursos técnicos, no entanto, cada estudo ainda se refere a uma unidade de ensino ou região específica. O fenômeno da evasão em nível nacional ainda é carente de estudos.

Igualmente, de acordo com Castro (2009) e Guzzo & Euzébios Filho (2005), faz-se necessário entender os contextos em que a evasão escolar, em diferentes níveis ocorre, objetivando não somente a retenção do discente na unidade de educação, mas também melhorando a qualidade de vida dos cidadãos por meio do desenvolvimento de uma sociedade mais justa e igualitária. Outrossim, o Insper em parceria com a Fundação Roberto Marinho, realizou extenso estudo acerca das consequências da não terminalidade dos estudos. O trabalho publicado em 2021 aponta que o ano poderia terminar com 575 mil estudantes sem terminar a educação básica. De forma direta, informa Barros et al. (2021), ainda baseado no estudo, os números acarretaram R$ 214 bilhões de prejuízo. Valores estimados por meio de perdas geradas pela impossibilidade de melhores salários, perdas devido à baixa

atividade econômica, pela qualidade de vida inferior e pela possibilidade envolvimento em crimes.

Portanto, cabe aos educadores e gestores escolares observar com responsabilidade os dados de evasão, promovendo ações de prevenção. A tarefa é longe de ser simples, uma vez que, o aluno do curso técnico que se evade da unidade, por vezes, não mostra sinais claros de que isso acontecerá. É uma análise subjetiva e realizar essas observações com milhares de estudantes torna o trabalho dos gestores educacionais fatigante, senão dubitável.

Essa tarefa, porém, pode ser auxiliada por recursos tecnológicos modernos. O crescente desenvolvimento das Tecnologias da Informação e Comunicação (TIC) proveram capacidade computacional de forma a permitir a análise de grandes volumes de dados de forma veloz. Um dos recursos disponíveis é a mineração de dados.

Mineração de dados

Na revolução tecnológica promovida pelas TIC há coleta e armazenamento de largos volumes de dados, mas se não se analisam esses dados com maiores detalhes, tornam-se nada além de enorme quantidade de dados. Com novos métodos e técnicas computacionais pode-se analisar esses dados para obter informações úteis. O método conhecido para isso é a mineração de dados.

Entende-se mineração de dados como o processo de descobrir novos padrões a partir de grande volume de dados, utilizando ferramentas computacionais a fim de extrair conhecimento (Silwattananusarn & Tuamsuk, 2012). É uma etapa essencial no processo de descoberta do conhecimento em bancos de dados (KDD do inglês Knowledge Discovery in Databases), que, de acordo Tan; Steinbach; Kumar (2005) é o processo de encontrar informações úteis a partir de dados brutos, conforme ilustrado na Figura 1.

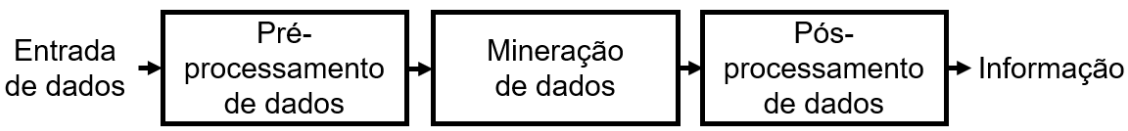

Figura 1 – Processo de descoberta do conhecimento de dados.

Fonte: Adaptado de Ta net al., 2005.

Para entender como a mineração de dados funciona, é importante compreender alguns conceitos fundamentais. Primeiro, as tarefas de mineração de dados dividem-se em duas categorias principais: tarefas preditivas e descritivas. As tarefas preditivas objetivam prever determinada saída com base nos atributos de entrada; as tarefas descritivas buscam correlações, tendências, agrupamentos e anomalias. Caracterizam-se como tarefas exploratórias que exigem explicação e pós-processamento (Tan et al., 2005).

Utilizam-se modelos preditivos para classificar, a partir de dados de treinamento, instâncias não rotuladas com base nas características dos dados de entrada. Esses métodos geram modelos de classificação e regressão (Faceli et al., 2011). Por exemplo: prever se um cliente fará uma compra é uma tarefa de classificação porque a variável de destino possui valor binário (0 ou 1). Por outro lado, prever o preço futuro de uma ação é uma tarefa de regressão uma vez que o preço é um atributo de valor contínuo. O objetivo de ambas as tarefas é aprender um modelo que minimiza o erro entre os valores previstos e os valores verdadeiros da variável de destino.

Segundo Ta net al. (2005) e Facelli et al. (2011), usam-se modelos descritivos a fim de identificar características em instâncias de diferentes classes, ou seja, utilizadas para determinar as semelhanças nos dados e encontrar padrões existentes. Esse método gera tipicamente modelos de Associação e Agrupamento, um deles representado pelos dados educacionais.

Mineração de dados educacionais

A mineração de dados é o processo de extrair informações e padrões ocultos e úteis de grandes conjuntos de dados. A sua aplicação em diversas áreas como: finanças, telecomunicações, saúde, marketing de vendas, etc., já é bastante conhecida. A possibilidade de adquirir novos conhecimentos a partir da análise de dados, abriu as portas para um novo ramo de estudo, a Mineração de Dados Educacionais (EDM do inglês Educational Data Mining), que se define como uma disciplina, preocupada com o desenvolvimento de métodos para explorar os tipos únicos de dados que vêm de ambientes educacionais e usar esses métodos para entender melhor os alunos e as configurações em que aprendem (Baker & Yacef, 2009). Ainda de acordo com os autores, os métodos de mineração de dados

educacionais envolvem mineração de dados, aprendizado de máquina, psicometria, estatística e modelagem computacional.

A EDM é o processo de transformação de dados brutos de amplos bancos de dados educacionais em informações úteis e significativas a serem usadas na melhor compreensão dos alunos e de suas condições de aprendizagem, melhorando o suporte ao ensino e a tomada de decisões nos sistemas educacionais. Aplicam-se, por fim, diversas técnicas de mineração de dados nos dados educacionais.

No que ser refere aos estudantes, por meio de algoritmos de agrupamentos é possível analisar, por exemplo, a motivação, atitude e comportamento ou entender a forma como o aluno aprende (Dutt et al., 2017). Outra técnica possível se centra na utilização de modelos preditivos no intuito de prever o desempenho ou comportamentos indesejáveis (Bakhshinategh et al., 2018). Em relação à gestão escolar, os algoritmos permitem conceber informações que auxiliarão no processo de tomada de decisão por meio de alertas, feedbacks, recomendações e aprimoramento de materiais didáticos (Bakhshinategh et al., 2018). Processos e cômputos esses, auxiliados pela Inteligência Artificial (IA).

Uso da Inteligência Artificial na educação

A IA se apresenta como uma das mais promissoras ferramentas de tecnologia a empregar em diversas áreas do conhecimento, incluindo a educação. Conforme apontado no relatório da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) de 2020, a IA ajuda a alcançar algumas metas globais de educação identificadas nos Objetivos de Desenvolvimento Sustentável (ODS4), seja por meio das análises de aprendizagem, recomendações diversas ou fornecendo ferramentas para variados diagnósticos, envolvendo alunos, professores, administradores, pais e políticas públicas (Vincent-Lancrin & Vlies, 2020).

O recente relatório da OCDE reafirma as diversas pesquisas realizadas nos últimos anos quanto ao uso de IA em ambientes educacionais. Em estudo de 2011, Wei (2011) propôs modelo de Rede Neural Artificial (RNA) para predição do sucesso acadêmico de alunos. A partir de dados como horas de estudo após as aulas, tempo gasto na internet, anotações nas aulas, frequência nas aulas entre outros fatores, o modelo obteve 79,0% de eficácia na predição do sucesso acadêmico dos estudantes.

Chen et al. (2014) utilizaram algoritmos meta-heurísticos inspirados em pássaros cuco, como Cuckoo Search (CS) e Cuckoo Optimization Algorithm (COA) para otimizar os pesos sinápticos e os biases de uma RNA a fim de predizer o desempenho acadêmico de estudantes universitários. Utilizando como variáveis de entrada, dados do resultado do vestibular, pontuação média no exame de graduação do ensino médio, tempo decorrido entre a conclusão do ensino médio e a entrada na universidade, localização do ensino médio (região), tipo do ensino médio (público ou privado) e gênero, os autores encontraram desempenhos satisfatórios na predição do desempenho acadêmico dos estudantes quando utilizando como métrica de desempenho a raiz quadrática dos erros RMSE do inglês Root Means Squared Error.

Sales et al. (2016) promoveram estudo utilizando árvores de decisão para prever a evasão dos alunos de 76 cursos da Universidade Federal de Campina Grande. Os autores abordaram o problema criando dois classificadores diferentes, um para o semestre e um para cada curso/semestre. Analisaram pelos modelos os dados de 32.342 estudantes tendo como variáveis o identificador de curso, identificador de semestre, média do semestre, status do semestre, número de créditos concluídos entre outras, apresentando precisão entre 82,0% e 89,0%.

Ainda tratando da evasão escolar, as abordagens de aprendizado de máquina se apresentam com uma das soluções viáveis no auxílio ao combate do abandono escolar.

Chung e Lee (2019) utilizaram os dados de 165.715 estudantes de nível médio, coletados junto ao Sistema Nacional de Informação de Educação nos Estados Unidos para prever o abandono escolar. Utilizando características como: ausência não autorizada nas primeiras quatro semanas, atraso não autorizado nas primeiras quatro semanas, ausência não autorizada, licença antecipada não autorizada, ausência de aula não autorizada, atraso não autorizado, tempo de atividade autorregulada, tempo de atividade do clube, tempo de trabalho voluntário e tempo de desenvolvimento de carreira, aplicados em algoritmos de aprendizado de máquina, mais precisamente árvores de decisão, os autores conseguiram a impressionante marca de 95,0% de acurácia do modelo ao prever o risco de abandono. Também por meio de algoritmos de aprendizado de máquina, o trabalho de Bitencourt et al., obteve resultados superiores a 73,0% de acurácia ao identificar a permanência ou o abandono de alunos de quatro cursos superiores, de um campus do Instituto Federal de Minas Gerais (Bitencourt et al., 2021).

Portanto, o conhecimento incorporado na literatura apresenta o potencial de transformar a luta contra a evasão de reativa a proativa. Isso é mais viável agora do que nunca, pois as TIC transformaram a forma como se coletam e gerenciam-se os dados, o que é um recurso importante para o aproveitamento das informações.

Logo, a mineração de dados se mostra ferramenta analítica poderosa que permite que as instituições educacionais aloquem melhor recursos e equipes, gerenciem proativamente os resultados dos alunos e contribuam na melhora da eficácia do desenvolvimento dos alunos. Com a capacidade de descobrir padrões ocultos em grandes bancos de dados, instituições de ensino constroem modelos a fim de prever – com alto grau de precisão – o comportamento dos discentes. Ao utilizar modelos preditivos, as instituições educacionais abordam efetivamente questões que vão desde o desempenho acadêmico até o abandono escolar.

Portanto, técnicas de IA, em particular o Aprendizado de Máquina (AM), vêm sendo utilizadas com sucesso na área de mineração de dados.

Para que um cientista de dados possa transformar dados brutos em informação útil, é necessário: (1) formalização de uma pergunta para a abordagem do problema; (2) seleção dos dados apropriados; (3) projeção de modelo; (4) realização de treinamento; (5) validação do treinamento com testes; e, por fim, (6) interpretação dos resultados (Guyon et al., 2019). Com a crescente demanda de profissionais no trabalho de análise de dados, e sendo a IA área complexa, que exige longa formação específica, uma nova vertente da IA começou a ser explorada o Autoaprendizado de Máquina (AutoML do inglês Automated Machine Learning).

Autoaprendizado de Máquina (AutoML)

O objetivo do AutoML é proporcionar às pessoas que não possuem conhecimento avançado de IA e programação, acesso a ferramentas de AM para o auxílio na tomada de decisões de seus próprios negócios (Budjac et al., 2019; Jin et al., 2019).

Segundo Feurer et al. (2015), quando da utilização de um serviço de AM, o usuário precisa selecionar o algoritmo de aprendizado que melhor se adéque ao conjunto de dados, pré-processar ou não o conjunto inicial e por fim, definir os hiperparâmetros do algoritmo. Portando, o processo de construção de uma solução, utilizando AM passa por processo interativo de refinamento e escolha do melhor

método. Um processo interativo, abre a oportunidade de escolha na automatização do processo. Assim o conceito de AutoML deriva da ideia de que, se vários modelos de AM devem ser construídos, usando uma variedade de algoritmos e várias configurações diferentes de hiperparâmetros, essa construção de modelo pode ser automatizada, bem como a comparação de desempenho e precisão do modelo.

Entende-se o processo de automatização da construção de um modelo de aprendizado de máquina, definido então como AutoML, como um encapsulamento dos processos interativos (pipeline), possuindo assim uma etapa de limpeza e seleção de dados, uma etapa para a seleção de algoritmos e uma etapa para a otimização dos hiperparâmetros (Zöller & Huber, 2019). Uma das soluções que utilizam o AutoML é o H2O.

H2O

H2O é um produto da H2O.ai, companhia de software instalada em Mountain View, Califórnia. A empresa possui parceiros de mercado como IBM, Intel, Anaconda, AWS, Google, entre outras, com o intuito de promover a cultura maker possibilitando o acesso ao AM a cada vez mais pessoas por meio de plataforma de fácil utilização e implementação. A proposta open-source apresenta produtos diversos, desde aplicações mais simples até soluções corporativas.

O H2O utiliza técnicas de compactação de memória podendo lidar com conjuntos de dados em grande escala na memória mesmo com um cluster pequeno (Nykodym et al., [s.d.]). Essa é a razão pela qual o H2O é considerado uma plataforma “rápida” pois os dados são distribuídos pelo cluster e armazenados na memória na forma de colunas compactadas permitindo a paralelização dos dados.

O código principal do H2O é escrito em Java e o armazenamento distribuído do tipo chave/valor é usado para acessar e fazer referência a dados, modelos, objetos etc., em todos os nós e máquinas. O H2O constitui um software que pode ser utilizado para modelagem de dados e computação em geral, com a finalidade primária de um motor de processamento distribuído, paralelo e em memória. (H2O.AI, [s.d.]; Ledell & Poirier, [s.d.]).

O H2O oferece boa qualidade junto com velocidade, facilidade de uso e implantação de modelo para os vários algoritmos supervisionados e não supervisionados. Assim, constroem-se os modelos de AM no R ou Python e podem ser facilmente convertidos para o formato POJO, e implementados em qualquer

ambiente Java. Por fim, os experimentos podem ser feitos simplesmente com o H2O, apenas iniciando o framework e fazendo experimentos com Python ou R, isso em um navegador.

Método

Empregou-se o método de pesquisa descritiva, exploratória e com marca experimental, com abordagem quantitativa. A pesquisa exploratória se justifica, segundo Gil (2002), porque permite novas descobertas, maior flexibilidade na investigação bem como, a análise de exemplos do problema estudado. Já em relação aos objetivos específicos, a pesquisa descritiva é adequada, uma vez que, segundo o mesmo autor, objetiva principalmente entender as propriedades do grupo de estudo, além de possibilitar o entendimento das relações entre variáveis deste grupo, estando a pesquisa de acordo com a abordagem. Ainda segundo Gil (2002), a pesquisa experimental estabelece a identificação de um objeto de estudo, a seleção de varáveis que podem alterá-lo e a definição de formas de controle e observação. Tais características se enquadram nas propostas da atual investigação. Em relação a abordagem quantitativa se justifica, pois permite a avaliação dos dados estatísticos em relação às relações sociais.

Amostra

Para a investigação, utilizaram-se os dados dos alunos da IEPT, de um curso, no período entre o primeiro semestre de 2011 e o segundo semestre de 2019. (recorte escolhido por conveniência pelos autores por causa da sequente epidemia de COVID-19).

Empregou-se no intuito de identificar a tendência de não conclusão dos alunos os algoritmos de AutoML e o uso de ferramentas estatísticas avançadas.

Submeteu-se o conjunto de dados a pré-processamento para que, de acordo com Faceli et al. (2011, p.29), “melhorar a qualidade dos dados por meio da eliminação ou minimização dos problemas”, como ruídos, valores incorretos, inconsistentes, ausentes, entre outros.

Para o uso de ferramentas de AutoML, dividiu-se o conjunto de dados em subconjuntos de treinamento, teste e validação.

Ao fim se gerou modelo de algoritmo de aprendizado de máquina com acurácia para previsão de evasão de novas matrículas.

Desenho do estudo

Por meio de pré-processamento de dados, realizou-se a eliminação manual de atributos considerados irrelevantes. Segundo Faceli et al. (2011, p. 30) “quando um atributo claramente não contribui para a estimativa do valor do atributo alvo, é considerado irrelevante”. Portanto, removeu-se 68 atributos do conjunto de dados inicial mantendo no conjunto 19 atributos: Ano de Conclusão do Ensino Anterior, Cidade, Estado Civil, Etnia/Raça, Forma de Ingresso, Meio de Transporte, Município de Residência, Nacionalidade, Naturalidade, Nível de Ensino Anterior, Percentual de Progresso, Período de Ingresso, Renda Bruta Familiar (R$), Renda Per Capita, Sexo, Tipo de Escola de Origem, Turno, Zona Residencial e Situação no Curso (atributo alvo).

O atributo Situação no Curso apresenta no conjunto de dados original dados como: concluído, trancado voluntariamente, evasão, cancelado, matriculado, jubilado, cancelamento compulsório. Para este estudo, os dados que possuem relação a não conclusão do curso, como trancado voluntariamente, evasão, cancelado, jubilado e cancelamento compulsório foram substituídos pela condição Não Concluído. Assim para o conjunto constam três condições para o aluno: Concluído, Não Concluído e Matriculado.

A demais técnicas de pré-processamento, como tratamento de dados desbalanceados, ruídos, incompletude dos dados, redundância e conversão de dados categóricos em numéricos não se realizaram por se tratar de problema que será trabalhado pelo algoritmo de autoaprendizado de máquina.

Instrumento de avaliação

Usou-se como método de AutoML o H20, completando com a validação cruzada.

O conjunto de dados possui 1.222 registros de alunos, sendo 304 (24,9%) concluídos, 718 (58,8%) não concluídos e 200 (16,3%) matriculados.

Como parâmetros do método de AutoML, configurou-se o tempo máximo de 10.000 segundos de execução do algoritmo (parâmetro obrigatório), o número máximo

de 30 modelos a gerar (parâmetro obrigatório), número de 4 threads a utilizar e tamanho máximo 4 gigas bytes de memória RAM respectivamente.

Análises e Discussão

Utilizando validação cruzada, com parâmetro k = 10 (número de subdivisões do conjunto de dados), o modelo entregou como melhor algoritmo um Floresta Aleatória Distribuída (DRF do inglês Distributed Random Forest). Uma floresta aleatória é uma poderosa ferramenta de classificação e regressão baseada em árvores de decisão. Essa técnica agrega um conjunto de árvores de decisão permitindo reduzir o overfiting (sobreajuste) e o erro devido ao viés, produzindo melhores resultados. As árvores de decisão dividem um problema complexo em problemas mais simples de forma recursiva. Tipicamente utilizada em problemas de classificação, como é o caso do problema tratado por esse artigo. Classificar o aluno entre as categorias concluído, matriculado e não concluído. Já o overfiting diz respeito ao desempenho excelente do modelo em relação ao conjunto de treinamento, porém ao utilizar os dados do conjunto de teste o resultado não é o mesmo. Isso ocorre porque o modelo se ajusta aos dados de treinamento e acaba “decorando” o que deveria ser feito.

A matriz de confusão (Tabela 1) apresenta os resultados encontrados aplicando o DRF ao conjunto de testes. Este conjunto representa 29,8% do conjunto inicial. É possível observar que o modelo obteve acurácia de 90,1% quando da previsão da entrada de dados desconhecidos (conjunto de teste). A acurácia do modelo é dada pela razão das predições corretas e todas as predições do modelo. Ou seja, (81+50+197)/364 = 0,901).

Tabela 1 - Matriz de Confusão - Conjunto de Testes.

	Concluído	Matriculado	Não Concluído	Erro	Taxa
Concluído	81a	4c	0c	0,05	4/85
Matriculado	3b	50a	10c	0,21	13/63
Não Concluído	0b	19b	197a	0,09	19/216
	84	73	207	0,10	36/364

a: Verdadeiro Positivo; b: Falso Positivo c: Falso Negativo

Fonte: Base de dados e autores, 2023

Analisando a precisão somente dos alunos condição “não concluído”, foco do estudo, o modelo obteve precisão de 95,0% e revocação de 91,0%. Precisão e revocação indicam o quanto o modelo é eficiente e eficaz podendo ser calculas de acordo com as Equações 1 e 2, em que: VP: Verdadeiro Positivo; FP: Falso Positivo; e, FN: Falso Negativo.

$$Precisão = \frac{VP}{VP+FP} \quad (1)$$

$$Revocação = \frac{VP}{VP+FN} \quad (2)$$

A Tabela 2 apresenta os resultados da precisão, revocação, medida f para cada estado e F1 total. A medida f é calculada de acordo com a Equação 3.

$$F = 2 * \frac{Precisão * Revocação}{Precisão + Revocação} \quad (3)$$

Tabela 2 - Precisão, Revocação e Medida F.

	Precisão	Revocação	Medida f
Concluído	0,96	0,95	0,95
Matriculado	0,68	0,79	0,73
Não Concluído	0,95	0,91	0,93
		F1*	0,87

* Média das medidas f

Fonte: Base de dados e autores, 2023.

A Tabela 3 apresenta a relação de importância dos atributos de entrada onde pode-se observar que o percentual de progresso representa grande impacto na predição do modelo. Dentro do aprendizado de máquina, a seleção de atributos é

fundamental para direcionar equipes de desenvolvimento na seleção de variáveis que são mais eficientes e eficazes para um determinado sistema de aprendizado de máquina. Além disso, a seleção adequada de atributos auxilia na redução da dimensionalidade e do overfiting. Portanto, a importância da variável é determinada pelo cálculo da influência relativa de cada variável, ou seja, se essa variável foi selecionada para ser dividida durante o processo de construção da árvore e quanto o erro quadrado melhorou como resultado.

Tabela 3 - Importância das variáveis para a predição do modelo.

Variável	Importância (%)
Percentual de Progresso	48,42 %
Renda Bruta Familiar	8,77 %
Renda Per Capita	7,96 %
Estado Civil	6,24 %
Forma de Ingresso	6,17 %
Zona Residencial	5,38 %
Naturalidade	4,82 %
Ano de Conclusão do Ensino Anterior	2,75 %
Nível de Ensino Anterior	2,16 %
Meio de Transporte	2,11 %
Etnia / Raça	1,32 %
Cidade	0,96 %
Município de Residência	0,93 %
Tipo de Escola de Origem	0,58 %
Sexo	0,48 %
Turno	0,47 %
Período de Ingresso	0,44 %
Nacionalidade	0,04 %

Fonte: Base de dados e autores, 2023.

Resultados

Para os dados utilizados no experimento, o atributo Percentual de Progresso, que indica o quanto o aluno já progrediu no curso, apresenta elevada relação de dependência com o atributo alvo, Situação no Curso.

Apresentando menores graus de dependência, mas ainda assim consideráveis para a capacidade do modelo em predizer os resultados: Renda Bruta Familiar, Renda Per Capita, Estado Civil, Forma de Ingresso e Zona Residencial (> 5%).

Em menor importância com relação às variáveis mais consideráveis, os atributos de entrada com menor impacto (< 1%): Cidade (0,96%), Município de Residência (0,93%), Tipo de Escola de Origem (0,58%), Sexo (0,48%), Turno (0,47%), Período de Ingresso (0,44%) e Nacionalidade (0,04%).

Considerações Finais

Analisar grandes volumes de dados é um desafio que aumenta a cada dia em diversos tipos de negócios. Na educação não é diferente. Entender a demografia escolar, por meio de sua população, através de dados como idade, nível escolar, frequência, região, sexo e evasão, por exemplo, é desafio imposto aos gestores educacionais.

Relativo à hipótese H1: Algoritmos de aprendizado de máquina demonstram aptidão suficiente para predizer as eventuais evasões de unidade de ensino, evidencia-se, pelas Tabelas 1, 2 e 3, o seu aceite.

Possibilitou-se, por meio de ferramenta de AutoML, obter modelo de aprendizado de máquina, o DRF, com acurácia geral de 90,1%. Ademais, a taxa de acertos do modelo ao predizer apenas os alunos não concluídos foi de 91,3%. Assim, os resultados apontam que o uso de algoritmos de aprendizado de máquina, em especial, o AutoML, se configuram promissores e apontam decerto a uma área propícia de ser explorada.

No modelo gerado, a variável de maior importância foi o percentual de progresso. Isso se deve ao fato de que, para o conjunto de dados analisados, foram considerados os *status* concluído, não concluído e matriculado. Uma vez que todos os alunos analisados possuem um percentual de conclusão, variando de 0% a 100%, entende-se a relação dessa variável.

Para futuras pesquisas, outros dados devem ser relacionados a fim de melhor precisão e efetividade do modelo, como, por exemplo, a distância entre a residência do aluno e a unidade de ensino.

Também se sugere a remoção dos atributos de pouca relevância para as predições do modelo, verificando-se posteriormente a acurácia geral.

Além disso, outros modelos de AutoML devem ser testados com o objetivo de analisar modelos mais eficientes e, ao fim, mais eficazes.

Referências Bibliográficas

Baker, R. et al. (2009). O estado da mineração de dados educacionais em 2009: uma revisão e visões futuras. *Revista de mineração de dados educacionais, 1*(1), 3-17.

Bakhshinategh, B. et al. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies, 23*(1), 537–553.

Barros, R. P. et al. (2021). *Consequências da Violação do Direito à Educação* (1.ª ed.). Autografia.

Bitencourt, W. A., Silva, D. M., & Xavier, G. C. (2021). Pode a inteligência artificial apoiar ações contra evasão escolar universitária? *Ensaio: Avaliação e Políticas Públicas em Educação, 30*(116), 669-694.

Brasil. (2022) Lei no 11.892. Brasília: Presidência da República, 29 dez. 2008. http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11892. htm.

Budjač, R. et al. (2019). Automated Machine Learning Overview. *Research Papers Faculty of Materials Science and Technology Slovak University of Technology, 27*(45), 107–112.

Castro, J. A. (2009). Evolução e desigualdade na educação brasileira. *Educação & Sociedade, 30(*108), 673–697.

Chen, J.-F., Hsieh, H.-N., & Do, Q. H. (2014). Predicting Student Academic Performance: A Comparison of Two Meta-Heuristic Algorithms Inspired by Cuckoo Birds for Training Neural Networks. *Algorithms, 7*(4), 538–553.

Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, 96, 346–353.

Dore, R., & Lüscher, A. Z. (2011). Permanência e evasão na educação técnica de nível médio em Minas Gerais. *Cadernos de Pesquisa*, 41, 770–789.

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*, 5, 15991–16005.

Faceli, K. et al. (2011). *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina* (1.ª ed.). LTC.

Feurer, M. et al. (2015). Efficient and robust automated machine learning. *Advances in Neural Information Processing Systems*, 2015- Janua, 2962–2970.

Filho, R. B. S., & Araújo, R. M. DE L. (2017). Evasão e abandono escolar na educação básica no Brasil: fatores, causas e possíveis consequências. *Educação Por Escrito, 8*(1), 35–48.

Gil, A. C. (2002). *Como Elaborar Projetos de Pesquisa* (4.ª ed.). Atlas.

Guyon, I. et al. (2019). Analysis of the AutoML Challenge Series 2015–2018. p. 177–219.

Guzzo, R. S. L., & Euzébios Filho, A. (2005). Desigualdade social e sistema educacional brasileiro: a urgência da educação emancipadora. *Escritos sobre Educação, 4*(2), 39–48.

H2O.AI. (2022). *AutoML*: Automatic Machine Learning — H2O 3.36.1.3 documentation. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html.

IBGE. (2022). *PNAD contínua*: educação 2019. [s.l: s.n.]. https://biblioteca.ibge.gov.br/visualizacao/livros/liv101736_informativo.pdf.

Jin, H., Song, Q., & Hu, X. (2019). Auto-keras: An efficient neural architecture search system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1946–1956.

Ledell, E., & Poirier, S. [s.d.]. *H2O AutoML: Scalable Automatic Machine Learning*.

MEC. (2022). *Centenário da rede federal de educação profissional e tecnológica*. Ministério da Educação, 2007. http://portal.mec.gov.br/setec/arquivos/centenario/historico_educacao_profissional.pdf.

Neri, M. (2022) Motivos da Evasão Escolar. https://bibliotecadigital.fgv.br/dspace/handle/10438/21964.

Nykodym, T., et al. [s.d.]. *Generalized Linear Modeling with H2O*.

Sales, A., Balby, L., & Cajueiro, A. (2016). Exploiting Academic Records for Predicting Student Drop Out: a case study in Brazilian higher education. *Journal of Information and Data Management, 7*(2), 166–166.

SETEC. (2022). *Documento orientador para a superação da evasão e retenção na rede federal de educação professional, científica e tecnológica*. Ministério da Educação, 2014. https://avr.ifsp.edu.br/images/pdf/Comissoes_Outros/PermanenciaExito/ Documento-Orientador-SETEC.pdf.

Silva Filho, R. L. L. E. et al. (2007). A evasão no ensino superior brasileiro. *Cadernos de Pesquisa, 37*(132), 641–659.

Silwattananusarn, T., & Tuamsuk, K. (2012). *Data Mining and Its Applications for Knowledge Management:* A Literature Review from 2007 to 2012.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction do DATA MINING* (1.ª ed.). [s.l.] Addison-Wesley Professional.

TCU. (2022). *Relatório Sistêmico de Fiscalização da Educação* - Exercício de 2014. Brasilia: [s.n.]. https://portal.tcu.gov.br/biblioteca-digital/fisc-educacao-relatorio-sistemico-de-fiscalizacao-exercicio-2014.htm.

Vincent-Lancrin, S., & Vlies, R. V. D. (2020). *Trustworthy artificial intelligence (AI) in education: Promises and challenges*. OECD Education Working Papers, n. 218.

Wei, X. (2022). Student Achievement Prediction Based on Artificial Neural Network. In International Conference on Internet Computing and Information Services. Hong Kong, 2011. https://ieeexplore.ieee.org/abstract/document/6063304/.

Zöller, M.-A., & Huber, M. F. (2019). *Survey on Automated Machine Learning*.