Revista Científica em Educação a Distância

Predição de Evasão em Cursos EAD: um Modelo Baseado em Árvore de Decisão que Integra Causas Exógenas e Endógenas

Dropout Prediction in Distance Learning Courses: a Decision Tree-Based Model Integrating Exogenous and Endogenous Causes

Jordan Paulesky JULIANI*
https://orcid.org/0000-0001-7823-6644
University of Wisconsin-Stevens Point - 427 College of Professional Studies, Stevens Point, Estados Unidos
*jjuliani@uwsp.edu

Resumo. Embora a maioria das IES esteja desenvolvendo programas para controlar a evasão, muitos dos modelos preditivos existentes se baseiam em dados ou causas que não impactam diretamente os altos índices de evasão. Neste contexto, e considerando estudos recentes que recomendam a utilização da ciência de dados para desenvolver modelos preditivos, esta pesquisa propôs um modelo de predição de evasão em cursos de graduação EAD. Utilizando aprendizagem de máquina e um algoritmo de aprendizagem supervisionada chamado árvore de decisão, o modelo integrou causas exógenas e endógenas de evasão. A pesquisa, de natureza aplicada e abordagem qualitativa, foi classificada como exploratória e incluiu um estudo de caso e revisão bibliográfica. Os resultados revelaram a importância e a interação entre as variáveis de engajamento, idade e a distância entre o município de residência do aluno e o polo de EAD. Estas descobertas oferecem insights valiosos para as estratégias de retenção em instituições que oferecem educação a distância.

Palavras-chave: Educação a distância. Evasão. Modelos Preditivos. Ciência de dados. Aprendizagem supervisionada.

Abstract. Although most HEIs are developing programs to control dropout rates, many existing predictive models are based on data or causes that do not directly impact the high dropout rates. In this context, and considering recent studies that recommend the use of data science to develop predictive models, this research aimed to propose a predictive model for dropout in

undergraduate EAD courses. Using machine learning and a supervised learning algorithm called decision tree, the model integrated exogenous and endogenous causes of dropout. The research, applied in nature and qualitative in approach, was classified as exploratory and included a case study and literature review. The results revealed the importance and interaction between the variables of engagement, age, and the distance between the student's residence municipality and the EAD center. These findings provide valuable insights for retention strategies in institutions offering distance education.

Keywords: Distance education. Dropout. Predictive models. Data science. Supervised learning.

Recebido: 18/03/2025 Aceito: 13/05/2025 Publicado: 15/05/2025

Editores Responsáveis: Daniel Salvador/ Carmelita Portela

1. Introdução

A modalidade EAD, portanto, continua sendo uma alavanca para o desenvolvimento do ensino superior brasileiro (ABED, 2024). Nesta esteira de crescimento, a evasão não é só o maior motivo de preocupação, mas o maior problema enfrentado por essa modalidade de ensino. Ao realizar uma pesquisa, não exaustiva, em qualquer base de dados por artigos científicos, dissertações, teses, pelos termos evasão e EAD, encontramos como resultado centenas de relatos de pesquisadores apontando motivos para a evasão, propondo, entre outras temáticas: estratégias de permanência, estratégias de acompanhamento do curso, estudos relacionados aos sentimentos dos alunos e estilos de aprendizagem e modelos preditivos para evasão. Percebeu-se, nesta busca breve por documentos científicos sobre evasão, que a partir de 2019 temos o emprego das tecnologias de bigdata, mineração de dados, aprendizagem de máquina, e inteligência artificial, com o propósito de utilizar dados para compreender melhor o fenômeno da evasão na EAD e propor modelos preditivos.

Com base no exposto, temos a seguinte pergunta de pesquisa: como a ciência de dados, por meio das suas técnicas e algoritmos, pode ser empregada para o desenvolvimento de um modelo preditivo de evasão para a graduação EAD? Para responder a este questionamento, esta pesquisa terá como propósito investigar o emprego de um algoritmo de aprendizagem supervisionada, as árvores de decisão ou classificação, para desenvolvimento de um modelo preditivo de evasão que contemple causas exógenas e endógenas, e para tanto, utilize diferentes fontes de dados, para além daquelas que contemplam dados educacionais, a exemplo do sistema de gestão acadêmica e da plataforma de EAD, comumente utilizados pelas instituições de ensino.

2. Diferenciação e Impacto das Causas de Evasão nos Cursos EaD

A evasão corresponde à interrupção de um curso por parte do estudante, seja no início, no decorrer ou no final do percurso. Se houver desistência no processo, deve-se considerar que houve evasão. Há de se considerar a mobilidade do estudante de um curso para o outro. Para Vellozo (2019) a evasão deve ser vista como um problema relacionado à gestão de cursos a distância e não como algo inerente à modalidade EAD. "Cabe aos educadores e pesquisadores que lidam nessa esfera buscar identificar as causas e prevenir sua incidência, para tentar reduzi-la" (De Fátima Bruno-Faria; Lopes Franco, 2011).

Muitas causas podem ser apontadas para a evasão na EAD, seja por fatores relacionados aos estudantes, como habilidades, saberes, desinformação do curso e da carreira e desencanto, como também fatores externos, como condições financeiras, de saúde, de trabalho e família, ou, ainda, internos à instituição de ensino, tais como: currículo, orientação, docentes, estrutura física, suporte acadêmico e questões tecnológicas (Habowski; Branco; Conte, 2020). Umekawa (2014) reforça esse entendimento ao afirmar que "as causas da não permanência de estudantes são múltiplas e correspondem tanto a variáveis internas à própria ação instrucional, quanto a elementos externos à mesma situação".

As variáveis externas, também denominadas como causas exógenas, são apresentadas por Bittencourt et al. (2014) no quadro 1.

Quadro 1: Causas de evasão exógenas

Grupo	Descrição do Grupo	Subgrupo	Descrição do subgrupo
Problemas com os professores ou tutores	Decorrem da ausência de contato pessoal entre alunos e/ou professores	Baixa interação com os tutores	Contato com professores
		Baixa interação com outros docentes	Contato entre colegas de cursos
		Problemas com os professores	Insatisfação com o tutor
Condição pessoal	Fatores relacionados a problemas de ordem pessoal do aluno e sua adequação ao curso	Saúde	Problemas de saúde
		Problemas pessoais	Problemas de ordem pessoal
		Educação básica ineficiente	Deficiências acumuladas ao longo da Educação Básica
		Condição financeira	Dificuldades financeiras
		Mobilidade	Deslocamento até o polo presencial
		Desmotivação ou desinteresse	Insatisfação com o curso; falta de perspectiva de trabalho; falta de interesse; e desânimo para a conclusão; modificação do

			interesse pessoal ou profissional
Falta de Tempo	Decorrem da indisponibilidade de tempo para realizar as atividades pessoais, profissionais e as relacionadas ao curso	Trabalho	Elevada carga horária semanal de trabalho; falta apoio da organização onde trabalha
		Organização pessoal	Tempo para estudar; cursar outro curso superior; adaptação ao sistema universitário; dificuldade de adequação a EAD
Contexto familiar	Relacionado às questões familiares que podem interferir nos estudos	Problemas familiares	Influência familiar
		Mudança de estado civil	Mudança de estado civil
Acesso à internet	Relacionado com a condição socioeconômica do aluno ou local onde reside não dispor do acesso adequado a internet	Pouco ou nenhum acesso à internet	Dificuldades de acesso à internet

Fonte: adaptado de Oliveira et al. (2021) .

As variáveis internas, também denominadas como causas exógenas, são apresentadas no quadro 2 (Bittencourt; Mercado, 2014).

Quadro 2: Causas de evasão endógenas

Grupo	**Descrição do Grupo**	**Subgrupo**	**Descrição do subgrupo**
Dificuldades acadêmicas	Decorrem de aspectos acadêmicos e institucionais do curso ou da instituição	Problemas com atividades e avaliações	Reprovação em disciplinas; tempo de respostas ou correção das atividades; conteúdo das disciplinas; quantidade de disciplinas cursadas simultaneamente; critérios de avaliação; grau de complexidade das atividades; exigência de provas ou atividades presenciais; dificuldade de assimilação da formação prática com a teórica.
		Problemas com o material didático	O material didático fornecido
Uso da plataforma	Decorrem da plataforma de ensino aprendizagem utilizada	Dificuldades com a gestão ou utilização da plataforma	Estrutura e organização do ambiente virtual, instabilidade da plataforma, tecnologia inadequada utilizada.
Gestão do curso	Relativo à gestão acadêmica do curso	Problemas estruturais	Infraestrutura do polo; problemas administrativos; gestão da

			instituição; qualidade do curso ofertado; falha na elaboração do curso.
		Fornecimento de informações prévias	Desconhecimento prévio do funcionamento da modalidade e/ou do curso; informações imprecisas sobre o curso; falsa expectativa de facilidade do curso à distância.
		Apoio aos estudantes	Apoio dos professores e tutores; apoio da coordenação do curso; apoio dos professores do polo presencial; motivação e incentivo do tutor.

Fonte: adaptado de Oliveira et al. (2021) .

Oliveira et al. (2021) desenvolveram uma meta-análise a partir de uma revisão sistemática baseada em 40 artigos científicos publicados em periódicos listados no Qualis CAPES que trataram da temática da evasão na EAD. Estes artigos trataram das causas de evasão em instituições de ensino no Brasil. Esta meta-análise tomou como base as causas exógenas e endógenas apresentadas nos quadros 1 e 2, respectivamente, encontradas na revisão de literatura, para então identificar a frequência de citações de cada uma das causas nas publicações revisadas. No que se refere às causas exógenas, o grupo que apresentou maior número de artigos que o citou foi a "falta de tempo", com 80 %, sendo o subgrupo "organização pessoal" e "trabalho" os mais citados com 50% e 45%, respectivamente. Na segunda posição, temos o grupo "condição pessoal" com 75% das citações, sendo o subgrupo "problemas pessoais" recebendo 40% das citações. Já com relação às causas endógenas, o grupo "dificuldades do curso" teve 52,5% das citações, sendo o grupo "problemas com atividades e avaliações" o mais citado com 45%. O segundo grupo mais citado foi "gestão do curso" com 47,5%, sendo que os subgrupos mais citados foram "problemas estruturais" e "apoio aos estudantes", ambos com 25% de citações.

3. Modelos Preditivos de Evasão no Ensino Superior a Distância

Kowalski et al. (2020), desenvolveram uma revisão sistemática de literatura sobre modelos preditivos de evasão para o ensino superior a distância. O quadro 3 sintetiza os oito modelos que foram encontrados pela revisão sistemática.

Quadro 3: Modelos preditivos de evasão para o ensino superior a distância

Autores do modelo	**Descrição do modelo**
Silva (2017)	Identificaram as variáveis que influenciam na evasão em cada curso pesquisado e, com base nessas variáveis, desenvolveram os modelos preditivos. A pesquisa descritiva com abordagem qualitativa usou o método de regressão logística binária.

Sales et al. (2012)	Apresentaram um controle acadêmico automatizado com auxílio de uma ferramenta de avaliação formativa e de medição contínua do desempenho dos alunos, denominada Learning Vectors (LV), ou Vetor de Aprendizagem, desenvolvido a partir da extensão e do reuso de códigos das ferramentas do próprio ambiente virtual de aprendizagem Moodle. O modelo é fundamentado na mediação iconográfica e em intervenções geradas pelo professor-tutor como uma maneira de se comunicar com seus alunos no ambiente virtual. Os LV também armazenam as notas das atividades presenciais e gerenciam a frequência dos alunos em ferramentas como fóruns, tarefas, wikis e chats.
Ramos et al. (2017)	Elaboraram um modelo de regressão logística baseado nos construtos da distância transacional desenvolvida por Michael Moore, que funcionam como preditores da evasão de alunos na EaD. O estudo usou dois tipos de análises quantitativas: análise multivariada de dados e mineração de dados educacionais
Sepúlvida (2016)	Geraram e validaram um modelo de predição de evasão baseado em KDD (Knowledge Discovery in Database). Os resultados de sua aplicação a um ambiente virtual de aprendizagem demonstraram que o modelo consegue predizer a evasão com precisão significativa.
Kampff et al. (2014)	Propuseram um sistema de alertas para ambientes virtuais de aprendizagem (AVA), configurável pelo próprio professor a partir de indicadores do AVA e de informações geradas a partir da mineração de dados educacionais. O sistema procura identificar perfis de evasão e mau desempenho de alunos em educação a distância.
Portal e Schlemmer (2015)	Utilizaram a mineração de dados e learning analytics para a concepção e criação de um sistema web, denominado GCWise, para prever e minimizar a evasão em educação a distância.
Wilges et al. (2010)	Propuseram um modelo conceitual preditivo da evasão em educação a distância modelado por uma arquitetura de sistema multiagentes (SMA) em um ambiente virtual de aprendizagem. O modelo consegue prever comportamentos pelo monitoramento constante e dinâmico da aprendizagem e pela coleta de informações, calculando, assim, o risco de evasão.
Lira et al. (2016)	Desenvolveram um módulo para um sistema multiagentes, cuja finalidade é acompanhar o comportamento dos alunos e identificar precocemente quando eles tendem à evasão. O SMA interage com os dados do ambiente virtual de aprendizagem Moodle, acessando seu banco de dados. Pela mineração de dados educacionais, foi possível identificar padrões de comportamento que refletem a tendência de evasão.

Fonte: Kowalski et al. (2020)

A ciência de dados emerge a partir de uma avalanche de dados que aumenta continuamente em complexidade, diversidade, velocidade e volume. Isso ocorre em virtude do desenvolvimento e evolução de diferentes tecnologias a exemplo do advento da internet, dos sistemas baseados na web, do streaming de dados, da computação nas nuvens, da internet das coisas, e das tecnologias vestíveis. Hoje, a quantidade de dados produzidas em um dia é superior a todos os dados criados até 2003 (Filatro, 2020).

De forma mais objetiva, Grus (2021) afirma que o propósito da ciência de dados é transformar dados em conhecimento, em outras palavras, extrair conhecimento de dados desorganizados.

O processo de ciência de dados é constituído de seis etapas a saber (Filatro, 2020):

1. Definição do problema: ter uma visão clara sobre o propósito e o contexto do problema, estabelecer os recursos a serem utilizados, a maneira como a análise será realizada e a lista de entregas dispostas em uma linha do tempo.
2. Coleta de dados: ter acesso aos dados, em vários formatos, e em diferentes fontes.
3. Preparação dos dados: transformar/tratar os dados de modo a torná-los utilizáveis, seja corrigindo erros (valores ausentes, duplicados, inválidos), seja mesclando dados de fontes diferentes.
4. Exploração dos dados: buscar correlações e padrões que permitam obter insights e compreensão dos dados; para tanto, poderão ser usadas técnicas visuais e descritivas para realizar a análise exploratória.
5. Modelagem dos dados: construir um modelo que, utilizando abordagens descritivas, preditivas ou prescritivas, permita responder ao problema.
6. Comunicação dos resultados: apresentar os resultados obtidos em diferentes formatos: relatórios detalhados, painéis de controle (dashboards), gráficos e mapas de calor (heatmaps), por exemplo.
7. Automatização da análise: quando necessário/possível, automatizar a análise.

Cabe ressaltar que as etapas descritas não são prescritivas. A ciência de dados raramente é linear; de fato trata-se de um ciclo iterativo, que pode e deve ser revisitado até que o objetivo seja alcançado.

A ciência de dados se presta a resolver diferentes tipos de problemas, entre eles o de classificação, que possibilita a predição, objeto desta pesquisa. Para tanto, uma grande variedade de técnicas pode ser aplicada, entre elas, o aprendizado de máquina (*machine learning*). Aprendizado de máquina pode ser definido como um campo de inteligência artificial que se preocupa com o desenvolvimento de algoritmos e técnicas que permitem que um computador aprenda e ganhe inteligência com a experiência. No aprendizado de máquina, modelos aprendem com dados históricos que podem ser primários dados ou dados secundários (Kaur & Kumari, 2022). Um modelo, por sua vez, resulta da aplicação de um algoritmo a uma base de dados. Assim sendo, um algoritmo, quando aplicado a diferentes conjuntos de dados, gera modelos diferentes. Parâmetros são características (variáveis) de um modelo. Tomando como exemplo a função y = a + bx, a e b são parâmetros a serem estimados a partir dos dados. Denominam-se de hiperparâmetros as características (variáveis) que são definidas pelo analista, antes da execução do algoritmo; são variáveis estabelecidas para algoritmo, que guiam a sua execução (Sicsú, 2023).

Os algoritmos de machine learning podem ser classificados em: algoritmos de aprendizagem supervisionada e algoritmos de aprendizagem não supervisionada. Os algoritmos de

aprendizagem supervisionada são aplicados quando uma base de dados é formada por variáveis denominadas previsoras (X1, X2, ..., Xp) e uma variável Y, denominada variável alvo. Estes algoritmos tem como propósito detectar a relação entre as variáveis previsoras em à variável alvo. Uma vez identificada a relação, é possível prever ou classificar a variável Y, a partir das variáveis predecessoras. Consideremos, como exemplo, um conjunto de dados relativo a apartamentos, com valor (preço do imóvel) conhecido e algumas variáveis, tais como: número de cômodos, área útil, bairro, andar, entre outros. A partir dessa base de dados, pode ser aplicado um algoritmo de aprendizagem supervisionada que possa prever o preço de novos apartamentos, baseando-se nessas mesmas características.

De acordo com Sicsú (2023)

> os algoritmos supervisionados de classificação são utilizados para classificar uma nova observação em uma das categorias da variável qualitativa Y, ou seja, para preverem em qual categoria deve ser classificada uma nova observação.

Entre os algoritmos de aprendizagem supervisionada, dedicados à classificação, destacam-se: a regressão logística, as árvores de decisão ou classificação, as randomforests, XGBoost e o SVM-Supervised Vector Machine.

As árvores de decisão produzem regras de decisão de forma implícita. A construção da árvore ocorre por meio de um processo denominado particionamento recursivo (Bressan & Endo, 2021). Este processo consiste no particionamento gradativo do conjunto de dados em subconjuntos cada vez mais "homogêneos" que o conjunto original. A partição continua até satisfazer um critério de parada previamente definido (Sicsú, 2023).

4. Metodologia

A caracterização da pesquisa está relacionada com os procedimentos metodológicos e operacionais que definem como os dados serão coletados, analisados e tratados para a solução do problema de pesquisa.

Quanto à natureza, a pesquisa é classificada como pesquisa aplicada, uma vez que tem como propósito a resolução de um problema concreto. A abordagem empregada será qualitativa em função de que os dados coletados serão analisados por meio da interpretação subjetiva. Quanto aos objetivos, caracteriza-se como exploratória. Uma pesquisa exploratória tem o propósito de examinar um tema ou problema de investigação pouco estudado, sobre o qual se têm muitas dúvidas. A pesquisa caracteriza-se como exploratória pois tem como missão identificar causas exógenas e endógenas de evasão, coletar os dados e criar um modelo preditivo baseado no algoritmo de árvores de decisão. Finalmente, com relação aos procedimentos técnicos, se enquadra como pesquisa bibliográfica e estudo de caso. A classificação como estudo de caso deve-se ao fato da pesquisa ter sido desenvolvida no CEAD/UDESC, tendo como base dados provenientes da disciplina intitulada "Empreendedorismo e Inovação", ofertada nos anos de

2020, 2021 e 2022, no curso de bacharelado interdisciplinar em ciência e tecnologia, na modalidade EAD pelo departamento de educação científica e tecnológica (DECT).

Os procedimentos metodológicos basicamente contemplam três etapas. Serão realizados dois levantamentos bibliográficos, o primeiro sobre causas exógenas e endógenas de evasão, e o segundo, envolvendo ciência de dados e o algoritmo de árvore de decisão. Este levantamento bibliográfico ofereceu subsídio para a criação do modelo preditivo, considerando as peculiaridades dos cursos de graduação EAD da UDESC. Cabe destacar a intenção de executar o modelo durante o semestre em que a disciplina Empreendedorismo e Inovação está sendo ofertada, de modo a tornar possível tomar ações proativas contra a evasão.

A escolha das árvores de decisão como algoritmo para a predição se deu em função desses algoritmos se destacarem pela facilidade de visualização e interpretação dos resultados obtidos. Ao analisar uma árvore de classificação, é simples identificar a relação entre as variáveis previsoras e a variável alvo.

O desenvolvimento do modelo foi realizado por meio do Google Colab[1].

5. Resultados e Discussão

Os resultados serão apresentados tomando como base o processo de ciência de dados apresentado por Filatro (2020), com exceção da primeira etapa, de definição do problema, contextualizada na introdução deste artigo, e da última, a automatização da análise.

Coleta de dados: A seleção dos dados (parâmetros) para a criação do modelo partiu da análise das causas exógenas e endógenas de evasão definidas por Bittencourt et al. (2014). Buscou-se nas fontes de dados disponíveis nos sistemas institucionais da UDESC, mais especificamente o sistema acadêmico e a plataforma Moodle[2], dados que fossem compatíveis com os subgrupos das causas de evasão apontados pelo referido autor. Fontes externas também foram usadas para localizar dados para compor o modelo. Para o subgrupo "desmotivação e desinteresse", buscou-se na plataforma Moodle o nível de acesso às atividades previstas na disciplina de Empreendedorismo e Inovação, relativas aos dois primeiros tópicos (dos quatro tópicos existentes). As atividades correspondem ao acesso aos conteúdos (livros e lições), participação em fóruns e demais atividades avaliativas, e acesso aos encontros síncronos por videoconferência. O motivo para esse recorte temporal é permitir predizer a evasão do aluno durante a oferta da disciplina, de modo a possibilitar a realização de ações de mitigação em tempo (durante a realização da disciplina). Para o subgrupo mobilidade, recorreu-se ao sistema

[1]Google Colab é um produto do Google da área de pesquisas científicas. Trata-se de uma ambiente de desenvolvimento que permite escrever e executar código Python pelo navegador. É um serviço gratuito que tem por objetivo incentivar a pesquisa sobre aprendizado de máquina e inteligência artificial.
[2]Moodle é uma plataforma de aprendizagem online que permite criar ambientes de aprendizagem personalizados.

acadêmico que armazena o município de residência do aluno e o município onde se localiza o polo de EAD a que o aluno está vinculado. Com esses dados, calculou-se a distância entre os municípios, utilizando o serviço disponível no site rotamapas.com.br. Para os subgrupos "pouco ou nenhum acesso à internet" e "educação básica ineficiente" utilizou-se o Índice de Desenvolvimento Humano Municipal (IDHM), disponível na base de dados ofertada pelo Programa das Nações Unidas para o Desenvolvimento (PNUD). Para o subgrupo "problemas com atividades e avaliações", três dados foram coletados dos sistema acadêmico da UDESC: a nota final obtida pelo aluno na disciplina, representado no modelo pelo parâmetro "NOTA_FINAL", se o aluno foi aprovado ou reprovado, definido pelo parâmetro "SituacaoResultadoDisciplina" do modelo, e se ele evadiu ou não no curso, identificado pelo parâmetro"SituacaoAtualNoCurso". Mais dois dados foram coletados, sem uma correspondência direta com as causas de evasão listadas por Bittencourt et al. (2014): a idade do aluno e o tipo de ingresso do aluno na Universidade. O quadro 4 resume os parâmetros coletados, suas fontes e a correspondência com as causas da evasão.

Quadro 4: Parâmetros coletados para o modelo preditivo de evasão.

Causa de evasão	**Grupo**	**Subgrupo**	**Fonte de dados**	**Nome do parâmetro**	**Tipo do parâmetro**
Exógenas	Condição pessoal	Desmotivação ou desinteresse	Moodle da UDESC	engajamento	numérico
Exógenas	Condição pessoal	Mobilidade	Sistema acadêmico da UDESC e rotamapas.com.br	Distancia_polo_residencia	numérico
Exógenas	Acesso à internet	Pouco ou nenhum acesso à internet	IDHM - *United Nations Development Programme*	idh	numérico
Exógenas	Condição pessoal	Educação básica ineficiente	IDHM - *United Nations Development Programme*	idh	numérico
Endógenas	Dificuldades acadêmicas	Problemas com atividades e avaliações	Moodle da UDESC	SituacaoResultadoDisciplina	categórico
Endógenas	Dificuldades acadêmicas	Problemas com atividades e avaliações	Sistema acadêmico da UDESC	NOTA_FINAL	numérico

Endógenas	Dificuldades acadêmicas	Problemas com atividades e avaliações	Sistema acadêmico da UDESC	SituacaoAtualNoCurso	categórico
Exógenas	Sem equivalência	Sem equivalência	Sistema acadêmico da UDESC	idade	numérico
Exógenas	Sem equivalência	Sem equivalência	Sistema acadêmico da UDESC	Tipo_Cota_Ingresso	categórico

Preparação dos dados: Os dados provenientes do sistema acadêmico foram fornecidos pela Pró-reitoria de Ensino (PROEN) da UDESC; os dados advindos do Moodle foram obtidos por meio da funcionalidade de extração de dados do próprio software. Os dados de IDHM foram obtidos a partir do sítio do PNUD (https://www.undp.org/pt/brazil/idhm-munic%C3%ADpios-2010), e para obter os dados de distância entre o município em que o aluno reside e o polo onde estuda, foram capturados por meio do sítio https://www.rotamapas.com.br; um software foi especificamente desenvolvido para esse fim. Os dados extraídos das fontes supracitadas foram armazenados em arquivos de formato CSV. A junção dos dados foi realizada usando o software *Microsoft Excel*, por meio da função PROCV. Para esse processo, utilizou-se como chaves os nomes dos alunos e o nome dos municípios. Finalizada a junção dos dados, ocorreu a limpeza de registros (linhas), com campos faltantes ou duplicados, a anonimização (remoção dos nomes dos estudantes), além da normalização dos dados numéricos usando a escala de Min-Max. Este procedimento transformou os dados em uma escala de 0 a 1 pela equação: X' = (X - Xmin) / (Xmax - Xmin), onde X' é o valor resultante da normalização, X é o valor da célula na planilha a ser normalizada, e Xmin e Xmax, são, respectivamente, os valores mínimo e máximo do conjunto de dados.

Exploração dos dados: Ao iniciar a modelagem, observando a intenção de construir um modelo capaz de identificar o risco de evasão de um estudante durante o semestre em que a disciplina Empreendedorismo e Inovação está sendo ofertada, percebeu-se que os parâmetros “NOTA_FINAL”, que armazena a média final da disciplina obtida pelo aluno, e “SituacaoResultadoDisciplina”, parâmetro categórico que indica se o aluno foi aprovado ou reprovado por nota, não fariam sentido de serem considerados. Ambos os parâmetros são obtidos somente ao final do semestre letivo. Por esse motivo, foram excluídos.

Modelagem dos dados: Para a construção do modelo foram utilizadas as bibliotecas *Numpy* e *Pandas* para a manipulação de dados; para a visualização de dados, as bibliotecas *Matplotlib* e *Seaborn*, e o algoritmo *DecisionTreeClassifier* da biblioteca *Sklearn* de *Machine Learning*. O conjunto de dados utilizado para a execução do modelo e suas características são apresentadas na figura 1.

Figura 1 - Conjunto de dados utilizado para a execução do modelo.

```
RangeIndex: 146 entries, 0 to 145
Data columns (total 6 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   engajamento             146 non-null    float64
 1   idh                     146 non-null    float64
 2   Distancia_polo_residência  146 non-null    float64
 3   idade                   146 non-null    float64
 4   Tipo_Cota_Ingresso      146 non-null    object
 5   SituacaoAtualNoCurso    146 non-null    object
dtypes: float64(4), object(2)
```

Fonte: Imagem do autor.

Foram 146 linhas e seis parâmetros, sendo o parâmetro SituacaoAtualNoCurso (com valores evadiu e não evadiu) o parâmetro alvo. Durante a execução do modelo observou-se também a necessidade de tratar o parâmetro categórico Tipo_Cota_Ingresso. Esse parâmetro armazena três valores possíveis: Alunos não Cotistas, Alunos com Cotas por Ensino Público e Alunos com Cotas por Raça. A técnica utilizada para transformar esse parâmetro categórico em numérico foi a one hot encoding, que transforma cada possível valor em uma nova coluna, sendo que esta técnica, caso haja correspondência, atribui o valor 1 para a coluna específica e 0 para as outras colunas.

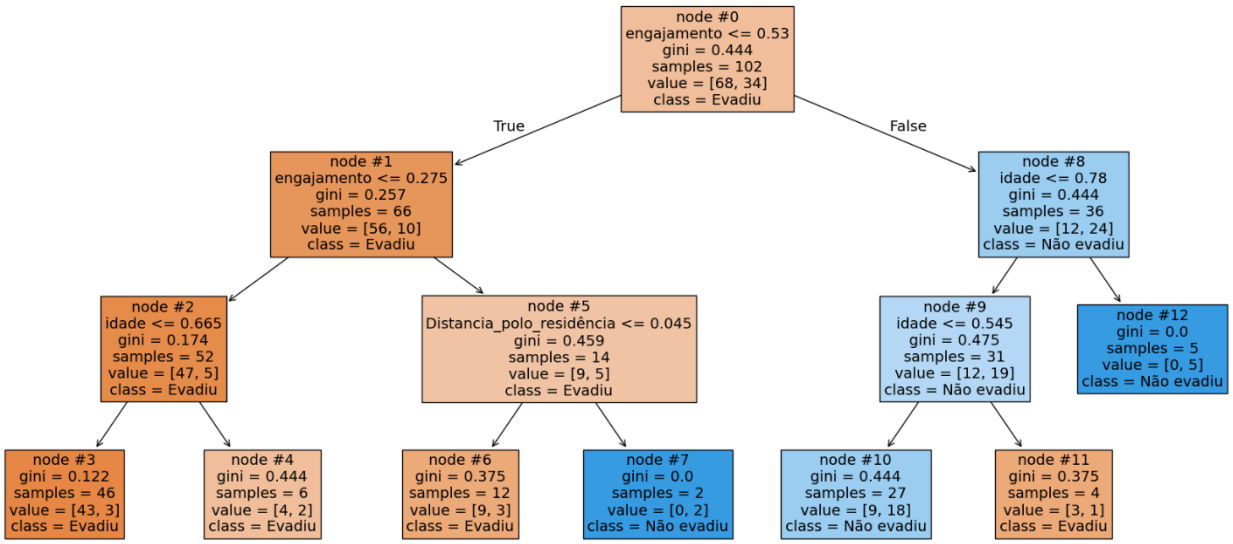
A divisão do conjunto de dados em conjunto de treinamento e conjunto de teste respeitou a proporção de 70% e 30%, respectivamente. Para o treinamento do modelo foi aplicada a técnica de poda (pruning). O modelo foi inicialmente treinado sem estabelecer nenhum mecanismo de limitação da profundidade da árvore criada. Em um segundo teste foi adicionado o argumento "max_depth=3", estabelecendo o limite de 3 níveis para a árvore de decisão criada: DecisionTreeClassifier(random_state=1, max_depth=3). Essa estratégia permitiu aferir melhores resultados ao modelo. A implementação do modelo está disponível no sítio:

https://colab.research.google.com/drive/1BZCKp_La_kFoInahFDkCmBm1SC3wa2DO?usp=sharing.

Comunicação dos resultados: No tocante à comunicação dos resultados, basicamente adotou-se a apresentação da árvore gerada por meio da classe *Tree*.A execução do método *plot_tree*, gerou a árvore de decisão apresentada na figura 2.

Figura 2– Árvore de decisão do modelo preditivo implementado.

Figura 2 - Árvore de decisão do modelo preditivo implementado.



Fonte: Imagem do autor.

Os resultados alcançados com o modelo desenvolvido, são apresentados na tabela 1.

Tabela 1 - Resultados alcançados com o modelo desenvolvido.

Métrica	Base de Dados de Treinamento	Base de Dados de Teste
Acurácia	0.8235	0.6591
Recall	0.8676	0.7586
Precisão	0.8676	0.7333
F1-Score	0.8676	0.7458

Destaque para a baixa acurácia da base de testes, possivelmente causada pela quantidade baixa de exemplos (linhas) do conjunto de dados submetido ao modelo.

Quanto à significância de cada parâmetro obteve-se que as variáveis engajamento, idh, Distancia_polo_residencia, idade, Tipo_Cota_Ingresso alcançaram os seguintes valores: 0.72758534, 0, 0.101195, 0.17121965, 0, respectivamente, ou seja as variáveis idh, e Tipo_Cota_Ingresso não geraram regras na árvore de decisão; ao mesmo tempo, o parâmetro engajamento representa alta importância no modelo.

Observando a árvore de decisão criada, duas regras chamam a atenção: a) se engajamento é alto (> 0,53) e os alunos são mais experientes (idade entre 0,54 e 0,78 ou idade > 0,78), não há

evasão. Percebe-se, nessa regra, a relevância da idade para o modelo; b) se o engajamento é baixo (entre 0,25 e 0,53), mesmo o aluno residindo próximo (distancia_polo_residência<= 0,045) do polo em que estuda, há evasão. Era esperado que o efeito de residir próximo ao polo fosse contribuir para o aluno não evadir, mesmo na condição de baixo engajamento.

A baixa acurácia sugere que o modelo pode ser melhorado com a adição de mais linhas (novos registros); ao mesmo tempo, a simplicidade da árvore gerada, mesmo com a limitação de profundidade, suscita que o modelo pode ser melhorado com o incremento de novos parâmetros, cobrindo um número maior de causas de evasão exógenas e endógenas.

4. Conclusão

A temática da evasão escolar tem mobilizado cientistas de dados a desenvolverem modelos preditivos, em especial na modalidade EAD do ensino superior, que sofre com altos percentuais de evasão em seus cursos.

Percebeu-se, ao desenvolver essa pesquisa, a criticidade das três primeiras etapas do processo de data science: a definição do problema, a coleta de dados e a preparação dos dados. Faz-se necessário que o cientista de dados concentre boa parte dos seus esforços nessas três etapas.

Iniciando-se pela necessidade de ter uma visão clara do contexto do problema, e, principalmente, do estabelecimento dos recursos, em especial, de dados, necessários para a criação do modelo. Muitos dos dados necessários para a criação do modelo, em se tratando da modalidade EAD, certamente não estarão disponíveis nas bases de dados dos sistemas acadêmicos, dos sistemas de avaliação institucional, ou da plataforma de EAD (como o Moodle). Portanto, demandará de uma estratégia, um plano, para a coleta desses dados. No que se refere à coleta dos dados, observou-se também uma certa dificuldade em obter dados da plataforma de EAD, cujo modelo de dados é extremamente extenso e complexo. Outro aspecto a ser considerado é a eventual dificuldade de acesso aos dados acadêmicos por parte dos pesquisadores, visto se são dados sensíveis para a organização, protegidos pela lei geral de proteção de dados (LGPD). Finalmente, deve-se considerar a relevância da preparação dos dados. Quanto maior a qualidade dos dados, maior a assertividade do modelo.

Mesmo enfrentando dificuldades para o desenvolvimento dessa investigação, em função da impossibilidade de coletar os dados, tomando como base os elencados como causas endógenas e exógenas de evasão, entende-se que o objetivo da pesquisa foi alcançado. Desenvolveu-se um modelo preditivo utilizando aprendizado de máquina, por meio de árvores de decisão. O modelo desenvolvido apresenta fragilidades em função da quantidade de exemplos e da pouca quantidade de parâmetros, já que não haviam dados disponíveis relativos às causas de evasão apontadas nas referências bibliográficas utilizadas.

Assim sendo, recomenda-se como pesquisas futuras, desenvolver uma revisão sistemática de literatura focada em identificar o estado da arte das causas de evasão, e, a partir desse levantamento, criar os meios para que esses dados sejam coletados, sistematicamente, junto aos acadêmicos.

Ao mesmo tempo cabe aplicar outros algoritmos de aprendizagem de máquina dedicados a classificar dados permitindo a predição, de forma a permitir comparar os resultados. Da mesma forma, é valido adaptar o modelo proposto com o ajuste de hiperparâmetros, com intuito de melhorar a sua performance.

Referências Bibliográficas

ABED – Associação Brasileira de Educação a Distância. **Censo ead.br [livro eletrônico]: relatório analítico da aprendizagem a distância no Brasil 2023 = censo ead.br: analytic report of distance learning in Brazil 2023**. Edição bilíngue: português/inglês. Tradução por Camila Rosa. Curitiba, PR: InterSaberes, 2024. 1 recurso online (PDF; 2 MB). ISBN 978-85-227-1596-1.

AMARAL, F. **Introdução à ciência de dados: mineração de dados e big data**. 1. ed. Starlin Altas Book, 2016.

BITTENCOURT, I. M.; MERCADO, L. P. L. **Evasão nos cursos na modalidade de educação a distância: estudo de caso do Curso Piloto de Administração da UFAL/UAB**. p. 465-504, 2014.

BRESSAN, G. M.; ENDO, W. **Métodos de machine learning para classificação da temperatura no processo de mistura em uma planta industrial**. Disponível em: https://revistas.uepg.br/index.php/ret/article/view/18484/209209215370 Acesso em: 6 maio 2024.

CARVALHO, P. S. A.; ALMEIDA, E. A.; CAVALCANTI, M. da C. M. **Análise da evasão na Educação a Distância: um estudo no curso de bacharelado em Administração Pública do IFPB/UAB**, 2018.

CURTY, R. G.; SERAFIM, J. D. S. **A formação em ciência de dados: uma análise preliminar do panorama estadunidense. Informação & Informação**, v. 21, n. 2, p. 307, 2016. Disponível em: https://doi.org/10.5433/1981-8920.2016v21n2p307 Acesso em: 10 maio 2024.

DE FÁTIMA BRUNO-FARIA, M.; LOPES FRANCO, A. **Causas da evasão em curso de graduação a distância em Administração em uma universidade pública federal**, 2011.

DE OLIVEIRA, C. V. S. B.; BEZERRA, D. H. D. **Revisão sistemática da literatura sobre as causas de evasão da Educação a Distância no Brasil**, 2021.

FILATRO, A. C. **Data science da educação**. Editora Saraiva, 2020.

GRUS, J. **Data Science do Zero**. Editora Alta Books, 2021.

HABOWSKI, A. C.; BRANCO, L. S. A.; CONTE, E. **Evasão na EAD: perspectivas de prevenção. Perspectiva**, v. 38, n. 3, p. 1-20, 2020. Disponível em: https://doi.org/10.5007/2175-795x.2020.e62978 Acesso em: 7 maio 2024.

KAUR, H.; KUMARI, V. **Predictive modelling and analytics for diabetes using a machine learning approach**. Applied Computing and Informatics, v. 18, n. 1-2, p. 90-100, 2022. Disponívelem: https://doi.org/10.1016/j.aci.2018.12.004 Acesso em: 16 junho 2024.

KOWALSKI, A. *et al*. **Evasão no Ensino Superior a Distância: Revisão da Literatura em Língua Portuguesa**. EaD Em Foco, v. 10, n. 2, 2020. Disponível em: https://doi.org/10.18264/eadf.v10i2.983 Acesso em: 16 maio 2024.

RAUTENBERG, S.; CARMO, P. R. V. do. **Big data e ciência de dados. Brazilian Journal of Information Science: Research Trends**, v. 13, n. 1, p. 56-67, 2019. Disponível em: https://doi.org/10.36311/1981-1640.2019.v13n1.06.p56 Acesso em: 20 julho 2024.

SICSÚ, A. L. **Técnicas de machine learning**. Editora Blucher, 2023. Disponível em: https://app.minhabiblioteca.com.br/#/books/9786555063974/ Acesso em: 20 julho 2024.

UMEKAWA, E. E. R. **Preditores de fatores relacionados à evasão e à persistência discente em ações educacionais a distância**, 2014. Disponível em: https://doi.org/10.11606/D.59.2014.tde-23032014-115420 Acesso em: 02 abril 2024.

VELLOZO, S. R. G. *et al*. **Evasão na educação a distância: uma revisão sistemática**. Revista EDaPECI, v. 19, n. 3, p. 85-94, 2019. Disponível em: https://doi.org/10.29276/redapeci.2019.19.312213.85-94 Acesso em: 02 abril 2024.

COMO CITAR ESTE TRABALHO

ABNT: JULIANI, J. P. Predição de Evasão em Cursos EAD: um Modelo Baseado em Árvore de Decisão que Integra Causas Exógenas e Endógenas. **EaD em Foco**, v. 15, n. 1, e2504, 2025. doi: https://doi.org/10.18264/eadf.v15i1.2504

*