CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA
UNIDADE DE PÓS-GRADUAÇÃO, EXTENSÃO E PESQUISA

MESTRADO PROFISSIONAL EM GESTÃO E DESENVOLVIMENTO DA EDUCAÇÃO PROFISSIONAL

RAPHAEL ANTONIO DE SOUZA

PREDIÇÃO DE EVASÃO ESCOLAR EM CURSOS DE EDUCAÇÃO PROFISSIONAL TÉCNICOS E TECNOLÓGICOS: ABORDAGEM COM AUTOML

São Paulo
Março/2024

RAPHAEL ANTONIO DE SOUZA

PREDIÇÃO DE EVASÃO ESCOLAR EM CURSOS DE EDUCAÇÃO PROFISSIONAL TÉCNICOS E TECNOLÓGICOS: ABORDAGEM COM AUTOML

Dissertação apresentada como exigência parcial para a obtenção do título de Mestre em Gestão e Desenvolvimento da Educação Profissional do Centro Estadual de Educação Tecnológica Paula Souza, no Programa de Mestrado Profissional em Gestão e Desenvolvimento da Educação Profissional, sob a orientação do Prof. Dr. Carlos Vital Giordano, na linha de pesquisa Políticas, Gestão e Avaliação

São Paulo
Março / 2024

FICHA ELABORADA PELA BIBLIOTECA NELSON ALVES VIANA
FATEC-SP / CPS CRB8-10894

Souza, Raphael Antonio de

S729p Predição de evasão escolar em cursos de educação profissional técnicos e tecnológicos : abordagem com AutoML / Raphael Antonio de Souza. – São Paulo: CPS, 2024.

108 f. : il.

Orientador: Prof. Dr. Carlos Vital Giordano

Dissertação (Mestrado Profissional em Gestão e Desenvolvimento da Educação Profissional) – Centro Estadual de Educação Tecnológica Paula Souza, 2024.

1. Abandono escolar. 2. Cursos técnicos. 3. Aprendizado de máquina. 4. Inteligência artificial. 5. AutoML. I. Giordano, Carlos Vital. II. Centro Estadual de Educação Tecnológica Paula Souza. III. Título.

RAPHAEL ANTONIO DE SOUZA

PREDIÇÃO DE EVASÃO ESCOLAR EM CURSOS DE EDUCAÇÃO PROFISSIONAL TÉCNICOS E TECNOLÓGICOS: ABORDAGEM COM AUTOML

Prof. Dr. Carlos Vital Giordano
Orientador - CEETEPS

Prof. Dr. Adilson Caldeira
Examinador Externo - UNIVERSIDADE PRESBITERIANA MACKENZIE

Prof. Dr. Paulo Roberto Prado Constantino
Examinador Interno - CEETEPS

São Paulo, 20 de março de 2024

Dedico este trabalho aos meus pais e à
minha amada esposa

AGRADECIMENTOS

Gostaria de expressar minha gratidão a todas as pessoas que contribuíram para a realização deste trabalho e tornaram possível a conclusão da qualificação.

Primeiramente, gostaria de agradecer ao meu orientador, Carlos Vital Giordano, pela orientação valiosa, apoio incansável e dedicação ao longo de todo o processo de pesquisa. Sua *expertise* e conhecimento foram fundamentais para o desenvolvimento do trabalho.

Gostaria de agradecer aos professores e pesquisadores do Programa de Pós-Graduação do Mestrado Profissional em Gestão e Desenvolvimento da Educação Profissional do Centro Paula Souza, cujas contribuições e ensinamentos enriqueceram minha compreensão e me inspiraram ao longo do curso.

Não posso deixar de agradecer à minha família, pelo amor incondicional, encorajamento constante e apoio emocional. Em especial minha esposa Marcela Leite. Sem vocês, nada disso seria possível.

Agradeço também os professores Adilson Caldeira e Paulo Roberto Prado Constantino pela participação em minha banca e por suas valorosas contribuições ao projeto.

Aos colegas do mestrado, que durante toda a jornada, permaneceram firmes, com incentivo e apoios, meus sinceros agradecimentos.

Agradeço ao Instituto Federal de São Paulo, pela oportunidade e incentivo à qualificação, valorizando sempre a capacitação docente, buscando melhorar cada vez mais o ensino público de qualidade.

Aos amigos do IFSP que me apoiaram e me incentivaram durante o curso de mestrado, deixo o meu muito obrigado pelo carinho e apoio de sempre.

A todos os mencionados e a todos aqueles que, de alguma forma, contribuíram para esta dissertação, meu sincero agradecimento. Seus esforços e apoio foram fundamentais para o sucesso deste trabalho.

A nova onda de inteligência artificial não nos traz propriamente a inteligência, mas um componente crítico dela: a previsão.

Ajay Agrawl

RESUMO

SOUZA, R. A. PREDIÇÃO DE EVASÃO ESCOLAR EM CURSOS DE EDUCAÇÃO PROFISSIONAL TÉCNICOS E TECNOLÓGICOS: ABORDAGEM COM AUTOML. 116 f. Projeto Mestrado Profissional em Gestão e Desenvolvimento da Educação Profissional. Centro Estadual de Educação Tecnológica Paula Souza, São Paulo, 2023.

O trabalho provém de estudos realizados na Linha de Pesquisa Políticas, Gestão e Avaliação, aderente ao projeto de pesquisa Gestão, Avaliação e Organização da Educação Profissional, da Unidade de Pós-Graduação, Extensão e Pesquisa do Centro Paula Souza – CEETEPS - e ao subprojeto desenvolvido junto ao grupo cadastrado no diretório CNPq que aborda práticas de ensino e aprendizagem condizentes com as realidades locais e regionais, inseridas no mundo informatizado, ligadas ao desenvolvimento de competências voltadas à formação profissional e que favoreçam a inserção social. O trabalho aborda a predição de evasão escolar em cursos técnicos e tecnológicos utilizando a abordagem com AutoML. A pesquisa objetiva analisar e desenvolver modelo preditivo que identifica antecipadamente a tendência de não conclusão de novos alunos matriculados nos cursos citados, por meio de IA, com o intuito de auxiliar as instituições de ensino na implementação de medidas preventivas e de intervenção. A fundamentação teórica da pesquisa se baseia em teorias de evasão escolar, aprendizado de máquina e autoaprendizado de máquina. A pesquisa foi realizada por meio da coleta de dados históricos de alunos, aplicação de técnicas de pré-processamento e engenharia de recursos nos dados, e treinamento e avaliação de modelos utilizando ferramentas de AutoML. Como resultado observou-se que o AutoML foi capaz de gerar diferentes algoritmos de aprendizado de máquina, todos capazes de predizer a evasão com acurácia superior a 70%. Também se observou que predições com dados acadêmicos são mais eficazes. O produto gerado por este trabalho é um *software* para geração de modelos de AM e predição de evasão, desenvolvido em Python, denominado EvaDetect e registrado no Instituo Nacional da Propriedade Industrial sob o número BR512024000219-7.

Palavras-chave: Abandono escolar. Cursos técnicos. Aprendizado de máquina. Inteligência artificial.

ABSTRACT

SOUZA, R.A. PREDICTING SCHOOL LEADERSHIP IN TECHNICAL AND TECHNOLOGICAL PROFESSIONAL EDUCATION COURSES: AN APPROACH WITH AUTOML**.** 116 p. Dissertation (Professional Master's Degree in Management and Development of Vocational Education and Training). São Paulo: State Center for Technological Education Paula Souza, 2023.

The work comes from studies carried out in the Policy, Management and Evaluation Research Line, part of the research project Management, Evaluation and Organization of Professional Education, of the Postgraduate, Extension and Research Unit of the Paula Souza Center – CEETEPS - and the subproject developed together with the group registered in the CNPq directory that addresses teaching and learning practices consistent with local and regional reality, inserted in the computerized world, linked to the development of skills specific to professional training and that favor social insertion. The research aims to analyze and develop a predictive model that anticipates the trend of non-completion among new students enrolled in the mentioned courses using AI. The purpose is to assist educational institutions in implementing preventive and intervention measures. The theoretical foundation of the research is based on dropout theories, machine learning and machine self-learning. The research will be carried out by collecting historical data from students, applying pre-processing and feature engineering techniques to the data, and training and evaluating models using AutoML tools. As a result, it was observed that AutoML was able to generate different machine learning algorithms, all capable of predicting dropout with an accuracy exceeding 70%. It was also noted that predictions using academic data are more effective. The product generated by this work is a software for generating ML models and predicting dropout, developed in Python, called EvaDetect, and registered with the National Institute of Industrial Property under the number BR512024000219-7.

Keywords: Dropout. Technical courses. Machine learning. Artificial Intelligence.

LISTA DE QUADROS

LISTA DE TABELA

LISTA DE FIGURAS

LISTA DE SIGLAS

AGI	Artificial General Intelligence
AM	Aprendizado de Máquina
ANI	Artificial Narrow Intelligence
ASI	Artificial Super Intelligence
ASR	Automatic Speech Recognition
AutoML	Automatic Machine Learning
CASH	Combined Algorithm Selection and Hyperparameter Optimization
CEETEPS	Centro Estadual de Educação Tecnológica Paula Souza
EDM	Educational Data Mining
FAQ	Frequently Asked Question
FN	Falso Negativo
FP	Falso Positivo
IBGE	Instituto Brasileiro de Geografia e Estatística
IA	Inteligência Artificial
IF	Instituto Federal de Educação Ciência e Tecnologia
IFSP	Instituto Federal de São Paulo
KDD	Knowledge Discovery in Databases
kNN	k-Nearest Neighbor
MEC	Ministério da Educação
MSE	Mean Squared Error
NLP	Natural Language Processing
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
PNP	Plataforma Nilo Peçanha
RFEPT	Rede Federal de Educação Profissional e Tecnológica
RNA	Rede Neural Artificial
RMSE	Root Mean Square Error
SETEC	Secretaria de Educação Profissional e Tecnológica
SVM	Suport Vector Machine
TDAH	Transtorno de Déficit de Atenção
TCU	Tribunal de Contas da União
TEA	Transtorno do Espectro Autista
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

XRT	Extremely Randomized Trees

Sumário

MEMORIAL

Esta é uma retrospectiva da minha vida profissional e acadêmica, revelando como cada experiência me conduziu ao mestrado profissional em Gestão de Sistemas Educacionais, com ênfase na pesquisa sobre Predição de Evasão Escolar em Cursos de Educação Profissional Técnica e Tecnológica: Abordagem com AutoML.

Minha jornada na área de tecnologia começou em 2004, quando me formei em Ciência da Computação. A graduação foi um período enriquecedor, repleto de aprendizados sobre programação, sistemas computacionais e a importância da tecnologia na transformação de diversos setores, incluindo a educação.

Após a graduação, busquei expandir meus conhecimentos e, movido pelo interesse em automação e robótica, cursei uma pós-graduação em Engenharia Eletrotécnica e de Computadores com ênfase nessa área. Durante esse período, aprofundei meus estudos em sistemas automatizados, controle de processos e integração de tecnologias, visando aplicar esses conhecimentos em projetos inovadores e na melhoria da eficiência dos processos industriais.

Minha trajetória na área de tecnologia e educação começou em 2010, quando tive a oportunidade de atuar como professor temporário do Centro Paula Souza, ministrando aulas na Etec (Escola Técnica Estadual). Durante os anos de 2010 a 2012, vivenciei o dinamismo da sala de aula e o desafio constante de despertar o interesse e a curiosidade dos estudantes em relação às disciplinas técnicas. Nesse período, aprimorei minhas habilidades pedagógicas e a capacidade de adaptar o conteúdo programático às necessidades específicas de cada turma, buscando sempre estimular o aprendizado e o desenvolvimento de habilidades práticas.

Foi então que, em 2014, iniciei minha jornada no Instituto Federal de São Paulo (IFSP) como professor da Educação Básica, Técnica e Tecnológica (EBTT).

As experiências como professor do IFSP e como Coordenador de Tecnologia da Informação (TI) no período de 2014 a 2016 reforçaram meu compromisso com a educação e me mostraram a importância de uma gestão eficiente para a melhoria dos processos educacionais e administrativos da instituição.

Ao assumir a Coordenação do Curso de Automação Industrial de 2019 a 2022, meu entusiasmo pela educação técnica e tecnológica se renovou, e a busca por aprimorar a qualidade do ensino e a formação dos alunos se tornou meu objetivo primordial. Durante esse período, participei ativamente da elaboração de Projetos Pedagógicos do Curso (PPCs) e coordenei reformulações curriculares, sempre buscando a atualização e a relevância das disciplinas em relação às demandas do mercado de trabalho.

Todas essas experiências ao longo da minha trajetória foram fundamentais para consolidar meu desejo de contribuir ainda mais para a área educacional. Com o passar dos anos, compreendi que a pesquisa acadêmica é uma ferramenta poderosa para impulsionar a inovação e a melhoria contínua dos processos educacionais.

Assim, em busca de aprimorar meus conhecimentos e habilidades, optei por seguir o caminho do mestrado profissional em Gestão e Desenvolvimento da Educação Profissional, focando minha pesquisa na "Predição de Evasão Escolar em Cursos de Educação Profissional Técnica e Tecnológica: Abordagem com AutoML". Acredito que a aplicação de técnicas avançadas de aprendizado de máquina, como a AutoML, fornecerá *insights* valiosos para o desenvolvimento de estratégias preventivas e personalizadas, visando a redução da evasão escolar e o aprimoramento do processo educacional.

1 INTRODUÇÃO

A Educação Profissional é um dos principais pilares para o desenvolvimento socioeconômico de uma nação. Por meio dela, é possível formar profissionais capacitados e qualificados para atuarem em diversas áreas, colaborando para o aumento da produtividade e da competitividade do mercado (Andersen; Werfhorst, 2010).

No entanto, de acordo com Neri (2009), a evasão e o abandono escolar é um dos principais desafios enfrentados pelas instituições de ensino, do nível básico ao superior, afetando também o ensino profissionalizante, comprometendo o processo de formação e afetando diretamente a vida dos alunos. Dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2023 (IBGE, 2023), apontam que 18% dos jovens (14 a 29 anos) não completaram a educação básica.

A conclusão dos estudos por parte dos alunos da Educação Profissional na Rede Federal de Educação Profissional e Tecnológica (RFEPT), entre os anos de 2009 e 2015, variou entre 11,4% e 15,4% no período segundo dados apresentados na pesquisa de Silva, Castioni e Martínez (2021) acerca da evasão escolar na RFEPT, no mesmo período indicado.

Analisando os dados da Plataforma Nilo Peçanha, que desde 2017 consolida de forma oficial os dados da RFEPT, no período de 2017 a 2019, encontra-se uma taxa de evasão de 47,3% no Instituto Federal de São Paulo, *lócus* desta pesquisa.

Segundo Barros et al. (2021) os valores estimados por meio de perdas geradas pela impossibilidade de melhores salários, chegam a R$ 214 bilhões. Perdas relacionadas à baixa atividade econômica, pela qualidade de vida inferior e pela possibilidade de envolvimento em crimes. Portanto a não terminalidade dos estudos acarretam prejuízos sociais e econômicos a toda nação.

A literatura sobre evasão escolar é vasta e o assunto é amplamente discutido pela comunidade acadêmica. As causas da evasão perpassam fatores sociais, núcleo familiar, social e unidade escolar conforme apontam os estudos de Filho e Araújo (2017) e corroborados pela PNAD de 2023 (IBGE, 2023). Uma vez que essas causas são identificadas a partir de alunos que não concluíram seus estudos, gestores educacionais acabam por ter dificuldades em tratar a evasão, melhorando assim a retenção dos discentes. Ou seja, o aluno com potencial de evadir, só é identificado após a evasão (Giordano; Souza, 2023). Nesse sentido, a presente dissertação propõe a utilizar a Inteligência Artificial (IA), na prevenção da evasão escolar em cursos de Educação Profissional.

O uso de algoritmos de Aprendizado de Máquina (AM) ganha destaque em diversas

áreas, dentre elas a educação. De acordo com Faceli et al. (2011) o AM é uma vertente da inteligência artificial que visa criar algoritmos e técnicas, permitindo que computadores possam aprender e tomar decisões a partir de dados, sem a necessidade de uma programação explícita.

Ainda de acordo com Faceli et al. (2011), por meio de modelos matemáticos e estatísticos, o aprendizado de máquina busca identificar padrões e relações nos dados, permitindo que sistemas computacionais realizem tarefas complexas, como otimizações, previsões, classificações e reconhecimento de padrões. O aprendizado ocorre por meio da exposição do sistema a exemplos e informações relevantes, permitindo que ele generalize esses conhecimentos para novos dados e contextos. O processo do aprendizado de máquina possui diversas abordagens, o que inclui aprendizado supervisionado, não supervisionado e por reforço, cada um com suas características e aplicações específicas, e tem se mostrado uma área de importância e impacto em diversas áreas, como medicina (Ben-Israel et al., 2020), finanças (Athey, 2018), ciência de dados e automação industrial (Wang; Siau, 2019).

Além do aprendizado de máquina convencional, uma outra abordagem surgiu nos últimos anos, o Aprendizado de Máquina Automatizado (AutoML do inglês *Automatic Machine Learning*), que busca automatizar o processo de construção e otimização de modelos de AM.

O AutoML envolve a aplicação de algoritmos e técnicas para automatizar tarefas como seleção de características, ajuste de hiperparâmetros, validação cruzada e até mesmo a construção completa de *pipelines* de aprendizado de máquina (Zöller; Huber, 2019). Essa abordagem se centra em reduzir a necessidade de expertise humana para projetar e implantar modelos de aprendizado de máquina, tornando o processo mais acessível e eficiente para usuários não especializados. Conforme apontado por Zöller e Huber (2019) e posteriormente por Chauhan et al. (2020), o AutoML é adotado em diferentes setores, permitindo às organizações aproveitarem os benefícios do aprendizado de máquina sem exigir um profundo conhecimento técnico.

Como mencionado anteriormente, a evasão de alunos é um dos principais problemas enfrentados pelas instituições de ensino, especialmente nos cursos técnicos, onde a complexidade dos conteúdos e a falta de motivação levam os alunos a abandonar os estudos. Identificar antecipadamente os alunos com maior potencial de evasão, significa possibilitar a implementação de medidas preventivas e de intervenção, visando mitigar esse problema conforme apontam Dore e Lüscher (2011). Nesse contexto, o uso da inteligência artificial tem despertado interesse como uma abordagem promissora para a previsão da evasão escolar, com dezenas de estudos publicados acerca da identificação de não conclusão de alunos em diversos

níveis e modalidades de ensino (Giordano; Souza, 2023)

Diante desse contexto, a presente pesquisa se guia na questão central: Como prever o potencial de evasões no ato da matrícula por meio de algoritmos de aprendizado de máquina, e de que forma essas previsões podem contribuir para intervenções pedagógicas e administrativas eficazes na minimização do problema?

Com base nessa questão, busca-se investigar a viabilidade e a eficácia dos algoritmos de IA na identificação precoce de alunos com risco de evasão, possibilitando intervenções pedagógicas e administrativas adequadas a fim de mitigar o problema.

O objetivo geral do projeto é desenvolver um sistema que utilize *framework* de AutoML, que seja capaz de analisar informações disponíveis no ato da matrícula e prever com precisão a probabilidade de evasão de cada estudante. Um dos intuitos também almejados é fornecer às instituições de ensino uma ferramenta eficaz para a identificação antecipada de alunos com potencial de evasão, possibilitando a adoção de estratégias preventivas e de intervenção adequadas para aumentar a taxa de conclusão dos estudantes.

Para atingir o objetivo geral proposto, estabelecem-se os seguintes objetivos específicos, alguns relacionados com o método a utilizar:

a) Aplicar e avaliar modelos de AutoML adequados para a previsão da evasão escolar;
b) Validar e comparar o desempenho dos modelos desenvolvidos, considerando métricas de avaliação apropriadas;
c) Propor recomendações e diretrizes para a aplicação prática dos resultados da pesquisa, visando auxiliar as instituições de ensino na implementação de estratégias efetivas de combate à evasão escolar.

A presente pesquisa se justifica pela importância da compreensão sobre a utilização de técnicas de aprendizado de máquina na área educacional, bem como a possibilidade de prevenção da evasão escolar por meio dessa tecnologia.

Além disso, o estudo também contribuirá para o aprimoramento das estratégias de gestão e acompanhamento dos alunos, permitindo que as instituições de ensino atuem de forma mais eficiente na redução da evasão escolar.

A pesquisa será realizada no Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - Campus Suzano, com foco na análise dos dados dos alunos matriculados no curso Técnico em Automação Industrial, utilizando a inteligência artificial (IA) para identificar possíveis casos de evasão.

O Instituto Federal de São Paulo (IFSP) é uma instituição vinculada ao Ministério da Educação (MEC) que oferece diversos cursos e modalidades de ensino. O Campus Suzano do

IFSP possui uma variedade de cursos em diferentes níveis de ensino, e a pesquisa concentrou-se no curso Técnico em Automação Industrial. Foram considerados critérios relacionados aos dados dos alunos, como o curso e o período analisado.

O período de análise irá abranger os anos de 2010 a 2019, excluindo os anos de 2020, 2021 e 2022 devido à pandemia de COVID-19 e às aulas remotas. São apresentadas tabelas que mostram a eficiência acadêmica do Campus Suzano e do IFSP como um todo, destacando os índices de conclusão, evasão e retenção.

No segundo capítulo, será apresentado a fundamentação teórica que norteia este trabalho, sendo dividido em seis subitens:

1) Breve histórico da educação profissional no âmbito federal e da importância da formação tecnológica no Brasil;

2) Trata o abandono escolar, onde destaca-se as causas, pesquisas e a necessidade de aprofundamento dos estudos sobre evasão na educação profissional e tecnológica;

3) Fundamentação teórica da inteligência artificial (IA) e dos benefícios que a IA agrega ao setor educacional;

4) Definição dos conceitos de mineração de dados e trata também da mineração de dados educacionais, um ramo de pesquisa relativamente recente;

5) Apresenta uma explicação sobre o aprendizado e máquina, descrevendo sua aplicação, funcionamento até chegar em modelos automatizados, como o H2O, ferramenta utilizada nesta pesquisa.

6) Revisão se encerra com uma explicação sobre as métricas de avaliação e desempenho em modelos de aprendizado de máquina, apresentando as métricas utilizadas neste estudo.

O terceiro capítulo aborda os procedimentos metodológicos que serão empregados e a caracterização do local de investigação, divididos em quatro subitens:

1) Caracterização do cenário da pesquisa, onde apresenta-se o Instituto Federal de São Paulo e o Campus Suzano;
2) Apresenta a método do estudo;
3) Apresenta o produto da pesquisa;
4) Trata dos materiais utilizados no estudo, passando pelos *softwares* de desenvolvimento até os conjuntos de dados que foram utilizados.

O quarto capítulo, apresenta a análise dos dados coletados e suas discussões. Destaca-se os resultados obtidos nos experimentos realizados com o EvaDetect e discute-se a eficácia dos

modelos utilizados, assim como as nuances identificadas nos diferentes contextos educacionais explorados.

Por fim, nas considerações finais, é apresentado os principais achados da pesquisa, oferecendo reflexões sobre sua relevância e contribuição para o campo da educação profissional. Além disso, sugere-se direções para futuras investigações, visando aprimorar o entendimento e a prevenção da evasão escolar.

2 REFERENCIAL TEÓRICO

Este capítulo se refere aos fundamentos teóricos nos quais se baseia a pesquisa. Considera-se o contexto da Educação Profissional Técnica de nível médio no Brasil no âmbito da Rede Federal de Educação Tecnológica, constituída pelos Institutos Federais, Colégio Pedro II, Centros Federais e Educação Tecnológica e colégios técnicos vinculados às universidades federais.

Explora o contexto da evasão escolar no Brasil e no contexto da educação tecnológica. Aborda ainda os conceitos de inteligência artificial, aprendizado de máquina, e mineração e dados que são o norte da pesquisa para prevenção da evasão nos cursos técnicos e tecnológicos.

2.1 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, considerado o primeiro estabelecimento criado pelo poder público, até a consolidação do ensino técnico-industrial no em 1906 (Brasil, 2009).

Apesar de observar-se ações de ensino profissionalizante ao longo do século 19, conforme destacado no relatório do MEC (Brasil, 2009) sobre o centenário da RFEPT, a Educação Profissional de nível técnico apenas se consolida em território nacional a partir do Decreto nº 787, de 11 de setembro de 1906, assinado pelo então presidente do Estado do Rio de Janeiro, Nilo Peçanha, que cria quatros escolas profissionais no estado.

O Brasil no final do século 19 e início do século 20 passava por um processo de reestruturação, urbanização e industrialização. Conforme apontado por Kunze (2009), mensagens presidenciais enviadas ao congresso nacional entre o período de 1891 e 1907 apontam a preocupação dos presidentes da época em institucionalizar e popularizar o ensino público. A partir do século 20, essas mensagens apresentam clara preocupação com o ensino técnico, dada à clara necessidade de qualificar mão de obra para a crescente industrialização nacional. A mensagem presidencial apresentada ao Congresso Nacional em 1907, pelo então Presidente da República, Afonso Augusto Moreira Pena diz: “[...]. Devemos cuidar com especial atenção do ensino profissional e técnico, tão necessário ao progresso da lavoura, do comércio, indústrias e artes” (Brasil, 1987, p. 40) .Assim, em 1909, Nilo Peçanha assina o

Decreto nº 7.566 que cria a Escola de Aprendizes Artífices em cada uma das capitais dos estados da nação (Brasil, 1909), dando início então à RFEPT.

As ações, em âmbito federal, no sentido de ampliar a oferta de ensino técnico se intensificam ao longo do século 20 conforme Quadro 1.

Quadro 1 - Evolução das políticas públicas, no âmbito federal, para oferta de cursos técnicos e tecnológicos

Ano	Ação
1909	Criação das Escolas de Aprendizes Artífices
1927	Projeto de Fidélis Reis – oferecimento obrigatório do ensino profissional no país.
1930	Criação do Ministério da Educação e Saúde Pública; Estruturada a Inspetoria do Ensino Profissional Técnico.
1937	A Constituição passa a tratar do ensino técnico, profissional e industrial; As escolas de Aprendizes e Artífices são transformadas em Liceus Profissionais.
1941	Ensino profissional passa a ser considerado de nível médio; Ingresso nas escolas industriais passa a depender de exame de admissão.
1942	Escolas de Aprendizes e Artífices são transformadas em Escolas Industriais e Técnicas; A formação profissional passa a ter equivalência ao secundário; Alunos formados nos cursos técnicos são autorizados a ingressar no ensino superior.
1959	Escolas Industriais e Técnicas são transformadas em autarquias e passam a se chamar Escolas Técnicas Federais.
1978	Três Escolas Técnicas Federais são transformadas em Centros Federais de Educação Tecnológica (CEFET); Os CEFET passam a formar engenheiros e tecnólogos.
1994	Lei 8.948 que transforma as Escolas Técnicas Federais em CEFET (Brasil, 1994).
1996	Leis de Diretrizes e Bases passa a ter um capítulo específico para a Educação Profissional.
2005	Lei 11.195 – Permite a ampliação da RFEPT (Brasil, 2005).
2008	Lei 11.892 – Criação dos Institutos Federais de Educação, Ciência e Tecnologia (IF) (Brasil, 2008).

Fonte: Adaptado de (Brasil, 2008, 2009)

As ações por parte do governo federal, para ampliar e garantir a oferta de cursos técnicos e tecnológicos, buscava, e ainda buscam, contribuir para qualificar as pessoas, garantindo a elas as oportunidades de melhores postos de trabalho, promovendo assim uma igualdade social. Até o ano de 2002, desde 1909, haviam sido construídos em território nacional 140 unidades da RFEPT (Brasil, 2009). A partir da promulgação da Lei 11.195 de 2005 e posteriormente a Lei 11.892 de 2008 que cria os IF, há uma vertiginosa expansão de toda a rede ampliando para 661 unidades em 64 instituições, sendo 38 Institutos Federais, 2 CEFET, 22 escolas técnicas

vinculadas a universidades, 1 universidade tecnológica (UTFPR) e o Colégio Pedro II. A RFEPT conta ainda com 11.005 cursos e oferece 845.523 vagas de ingresso anualmente, contato com mais de um milhão e 500 mil alunos.

Toda essa ampliação tanto no que se refere a abrangência quanto no número de vagas, tem como objetivo proporcionar um caminho que prepare os estudantes para o mundo do trabalho, oferecendo formação técnica e profissionalizante em diversas áreas. Como consequência, é desejável que os egressos dos cursos técnicos e tecnológicos, tenham sua inserção profissional e ganhos financeiros que justifiquem esse tipo de oferta.

De fato, Oliva et al. (2015) realizaram um estudo no Centro Estadual de Educação Tecnológica Paula Souza (CEETEPS), e observou-se que ex-alunos do CEETEPS têm maior probabilidade de emprego (3,5%), especialmente mulheres (5,6%). O estudo também apontou um aumento da probabilidade de emprego formal (2,7%) junto com um impacto positivo nos rendimentos (7,8%), principalmente nos homens (10,2%).

Davidis, Nogueira e Leal (2020) num estudo realizado no IF Brasília, apontaram uma correlação positiva entre o aumento de salário e maior inserção no mercado de trabalho, indicando para uma redução de 62,5% do número de desempregados antes e após a formação.

Essas pesquisas são reforçadas por Bastos, Carvalho e Macedo (2022), onde, a partir de dados do IBGE de 2018 referentes aos rendimentos *per capita*, rendimento total e anos de estudo, por meio de regressão linear, foi possível observar aumento de ganhos salariais entre os indivíduos que possuem acima de 10 anos de escolaridade.

Os estudos corroboram, portanto, para a importância da Educação Profissional. Apesar de proporcionar a possibilidade de ingressar no mercado de trabalho com mais rapidez e preparo, além do desenvolvimento de habilidades técnicas específicas e a ampliação das oportunidades de emprego e renda, ainda existem desafios a serem enfrentados na área da Educação Profissional, como a falta de investimento em infraestrutura e recursos didáticos, a desvalorização dos profissionais da área e a falta de articulação entre as diferentes instâncias de governo, instituições de ensino e empresas. Para além disso, ainda há o abandono e a evasão que tanto afligem as instituições de ensino.

Conforme dados extraídos da Plataforma Nilo Peçanha (PNP), no período de 2017 a 2019 a taxa de evasão a nível nacional na RFEPT, considerando apenas os Institutos Federais e o Colégio Pedro II, foi de 19,0% perfazendo um total de 566.050 matrículas consideradas como evadidas. Ao olhar os dados de evasão somente do Instituto Federal de São Paulo (IFSP), no mesmo período, a evasão sobre para 24,0% representando um total de 46.255 matrículas evadidas para um total de 186.886 matrículas efetuadas.

2.2 Abandono e Evasão Escolar

Ao abordarmos o tema Abandono e Evasão Escolar, faz-se necessário primeiro realizar a diferenciação dos termos evasão escolar e abandono escolar. Diferentes autores fazem leituras diversas dos termos. Enquanto alguns autores apontam que evasão se refere ao ato de não permanecer no sistema escolar e abandono quando o aluno se desliga da escola, mas retorna no ano seguinte, outros preferem tratar o termo abandono como análogo a evasão. Conforme aponta Filho e Araújo (2017) essa falta de definição clara dos termos, atrapalha tanto na quantificação quanto no tratamento e estudos das causas de ambos os fenômenos. Diante disso, no que tange a Educação Profissional, é razoável que se aceite ambos os termos para se referenciar ao aluno que ao matricular-se em um curso técnico/tecnológico e por algum motivo não o conclua.

A evasão escolar é, portanto, um problema sério que afeta a educação em todo o mundo. Conforme apontam Branco et al. (2020), há várias razões pelas quais os estudantes abandonam a escola, desde questões financeiras e familiares até problemas de aprendizagem e desmotivação. Independente dos motivos que levam a evasão e ao abandono, ainda de acordo com os autores, há um claro prejuízo no que tange as questões sociais, tolhendo os indivíduos de sua autonomia e busca por igualdade social. Ademais, elencam ainda os desperdícios de recursos públicos, falta de mão de obra especializada para o mercado de trabalho, aumento da criminalidade, entre outros.

O tema evasão não é nenhuma novidade. Em 1941, Filho (1941) e Freitas (1941) publicaram trabalhos, talvez os primeiros, no qual apontavam a evasão escolar das escolas primárias, chegando ao alarmante número de apenas 22 alunos formados a cada 100 matriculados. Contudo, as pesquisas de Filho e Freitas trouxeram à luz, apenas dados estatísticos referentes aos educandos e uma crítica ao sistema educacional da época, incapaz de manter e formar os cidadãos, resultando em um alto índice de analfabetismo. Conforme apontado pelo estudo da Secretaria de Educação Profissional e Tecnológica (SETEC), vinculada ao Ministério da Educação (MEC), a partir da década de 1970, os fatores que levam ao abandono escolar tornaram-se foco de estudos (SETEC, 2014).

De fato, ao utilizarmos apenas a *string* de busca “evasão escolar” no *software Publish or Perish* (Harzing, 2007), utilizando como base de busca o *Google Scholar*, no período de 1930 a 1969, obtemos singelos 23 resultados com pesquisas que mencionam a evasão. Nesse período, as pesquisas apenas se referem a evasão como uma problema, sem necessariamente buscar entender quais causas levaram os alunos a evadir-se dos bancos escolares, como

mostram as pesquisas de Rehder (1950), Freitas (1957) e Marinho (1958). Marinho ainda aponta que na transição dos anos 1956 e 1957, 55,2% dos alunos abandonaram a escola. Os altos índices de evasão apontados até então demonstram a importância de discutir-se as causas desse movimento. Uma das primeiras pesquisas a tentar entender as causas da evasão é de Anísio Teixeira em 1956, no qual entende que a reprovação contínua dos estudantes, desencorajava-os a continuar os estudos (Freitas, 1956).

Para efeitos de comparação, utilizando a mesma *string* de busca, “evasão escolar”, na mesma base e em um período de 40 anos, entre 1970 e 2010, mais de mil resultados foram retornados. Em sua grande maioria, as publicações a partir da década de 1980, não tratam apenas dos números da evasão escolar, mas também das causas e consequências deste abandono. Com o intuito de entender melhor esse crescimento, realizou-se recortes temporais menores, uma vez que o *software* utilizado retorna no máximo mil publicações.

A Tabela 1 apresenta o número de publicações sobre evasão desde a década de 1940 até 1969, depois os registros de dez em dez anos. A partir do ano 2000, foram necessários recortes de tempo ainda menores, dada a limitação do software, a saber: 2000-2002 (744 registros), 2003-2004 (770 registros), 2005 (572 registros), 2006 (805 registros), 2007 (981 registros), 2008 (985 registros) e 2009 (+1000 registros[1]).

Tabela 1 - Número de publicações encontradas como o termo "evasão escolar", por período

Período	Número de Artigos
1940 - 1969	23
1970 - 1979	51
1980 - 1989	193
1990 - 1999	707
2000 - 2009	+1000

Fonte: Autor (2024)

Os estudos referentes as causas da evasão, apontam, a necessidade de trabalho devido a dificuldades financeiras, o desinteresse, a falta de apoio das famílias e pouco investimento das escolas como possíveis motivos para a evasão e abandono escolar. Esses motivos são corroborado por Filho e Araújo (2017), no qual apontam como causas que levam a 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.

[1] O *software* retorna no máximo 1000 resultados. Portanto, não é possível precisar quantos registros foram encontrados para o ano de 2009.

Nesse sentido, de acordo com a Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2019, o principal motivo para o abandono escolar são: a necessidade de trabalhar, falta de interesse, gravidez e afazeres domésticos, quando respondentes do sexo feminino. Ainda de acordo com a pesquisa, necessidade de trabalhar e a falta de interesse, atingem 70,0% dos jovens das grandes regiões (IBGE, 2019). As mesmas causas apontadas nos estudos nacionais para o problema da evasão também são apontadas em estudos internacionais, conforme observado no estudo de Margiotta, Vitale e Santos (2014).

No entanto, grande parte das pesquisas bem como as informações do Instituto Brasileiro de Geografia e Estatística (IBGE) apontam dados referentes à formação de nível fundamental e médio. Alguns outros estudos referem-se ao ensino superior, mas quando o assunto é o ensino técnico, com evidente problema de evasão, faltam estudos que levem a um entendimento dos fatores ligados à evasão, tornando o trabalho de prevenção tarefa bastante difícil conforme apontado por Dore e Lüscher (2011). De fato, encontra-se na literatura centenas de artigos relacionados a evasão em cursos técnicos, contudo, cada estudo ainda se refere a uma unidade de ensino ou região específica. O fenômeno da evasão a nível nacional ainda é carente de estudos.

Silva, Castione e Martínez (2021) realizaram um estudo sobre evasão escolar no âmbito da RFEPT, compreendendo os anos de 2003 a 2015. A pesquisa tomou como base os documentos oficiais da SETEC, MEC e Tribunal de Contas da União (TCU), analisando os indicadores dos relatórios dos órgãos citados, como Relação de Concluintes por Matrícula Atendida (RCM) que mede a capacidade de alcançar o sucesso acadêmico, o Índice de Eficiência Acadêmica de Concluintes (EAC) que mede a capacidade de alcançar êxito entre os alunos que finalizam o curso e o Índice de Retenção do Fluxo Escolar (RFE) que visa medir a relação de alunos que não concluem seus cursos no período previsto. O RCM médio das instituições da RFEPT, no período de 2009 a 2015 foi de 13,3%, o seja, 13 alunos formados a cada 100 matrículas ofertadas.

O RCM ideal é 100,0%, ou seja, para cada matrícula atendida, um aluno será formado. O Relatório Anual de Análise dos Indicadores de Gestão das Instituições Federais de Educação Profissional, Científica e Tecnológica de 2019 (MEC, 2019) aponta isso ao indicar que a relação esperada para cursos de até 1 ano, seja de 100,0%. Porém, esse índice, segundo o mesmo relatório, cai para 20,0% para cursos com duração igual ou superior a 5 anos. Um percentual baixo, mas que ainda assim, é superior aos números encontrados na pesquisa de Silva, Castione e Martínez (2021).

Dore e Lüscher (2011) realizaram um estudo sobre a permanência e evasão na educação

técnica em Minas Gerais. Os dados coletados junto ao Programa de Educação Profissional do estado de Minas Gerais, apontaram que em 2008 a evasão no ensino técnico foi de 27,4%. Ainda nessa pesquisa, as autoras afirmam que "quando se trata da educação técnica não há pesquisas e/ou informações sistematizadas sobre a evasão". Contudo, Dore, Araújo e Mendes (2014), Dore e Lüscher (2011), Silva, Castione e Martínez (2021) são unânimes ao relacionar fatores individuais, institucionais e sistêmicos como formas de abordar o tema da evasão na educação técnica.

Logo, são necessárias ações concretas para abordar e tratar a evasão escolar. Com base numa extensa literatura acerca das causas do abandono, é possível chegar a formas diversas de tratar esse problema.

Uma das maneiras de abordar a evasão escolar é por meio da implementação de políticas públicas e programas educacionais que apoiem os alunos em risco de abandonar a escola, incluindo a oferta de aulas de reforço, aconselhamento, programas de mentoria, apoio financeiro para famílias de baixa renda e identificação precoce do risco de evasão do estudante. 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.

Além disso, é importante trabalhar para aprimorar a qualidade da educação oferecida nas instituições de ensino, promovendo aulas mais atrativas e relevantes para os estudantes. Isso inclui a adoção de recursos tecnológicos na educação, atividades práticas e projetos de grupo, que contribuam para manter o interesse e o engajamento dos alunos no seu processo de aprendizado. Conforme apontado por Pereira, Hahn e Bovo (2020) o uso de tecnologias em sala de aula apresenta possibilidades de deixar o processo de ensino aprendizagem mais ativo auxiliando assim o combate a evasão.

Um estudo promovido na Universidade Federal de Alagoas, nos cursos de matemática e química, apontou o ganho de aprendizado e engajamento da turma, indicando 0,0% de desistência nas disciplinas que utilizaram de estratégia de gamificação em sala de aula, indicando assim o valor da utilização de novas tecnologias no contexto da sala de aula para o efetivo combate a evasão (Pimentel; Ferreira; Freitas, 2020).

Por fim, é importante envolver os alunos em todo o processo de tomada de decisão sobre sua educação. Isso inclui a realização de pesquisas e fóruns de discussão para que os alunos possam expressar suas preocupações e ideias sobre como melhorar a escola. Ações como essas,

ajudam a aumentar o senso de pertencimento dos alunos à escola e incentivá-los a permanecer envolvidos em seu aprendizado. Esses achados corroboram com os apontamentos de Lüscher e Dore (2011) que indicam que a prevenção por meio da identificação e acompanhamento individual dos alunos com potencial de evasão.

A partir desses estudos, é perceptível a necessidade de identificação precoce dos alunos com risco de evasão para que ações possam ser tomadas e a causas, se possível, tratadas. Contudo, esse tipo de tarefa é, muitas vezes, complexo. Nesse sentido, conforme apontado por (Giordano; Souza, 2023), o uso de inteligência artificial possui potencial para auxiliar gestores educacionais na identificação antecipada de alunos com risco de evasão.

2.3 Inteligência Artificial

A Inteligência Artificial (IA) figura atualmente como uma das tendências na área de tecnologia. Nos últimos anos, inovações e avanços tecnológicos, outrora confinados à ficção científica, gradativamente evoluíram para o mundo real.

Os especialistas consideram a IA como um elemento integral da produção, que tem o potencial de introduzir novas fontes de crescimento e reconfigurar os métodos de execução em variados setores industriais. Em um relatório produzido pela PWC, há uma previsão que a IA seria capaz de contribuir com US$ 15,7 trilhões para a economia global até 2035 (Rao; Verweij, 2020). Países desenvolvido como China e os Estados Unidos estão preparados para se beneficiar ao máximo com o próximo *boom* da IA respondendo por quase 70,0% do impacto global.

Considerando a IA como um fator importante para a indústria, a qual agrega valor ao negócio, é justo se perguntar como a IA é capaz de fazer isso. A resposta está em seu funcionamento: a IA é desenvolvida por meio de modelos matemáticos e estatísticos que estudam os padrões do cérebro humano e analisam o processo cognitivo. Isso permite que um computador, um robô controlado por computador ou um *software* "pense" e aja de forma inteligente, como a mente humana. Em outras palavras, a IA é capaz de aprender e tomar decisões com base em informações coletadas e processadas por esses modelos, gerando *softwares* e sistemas inteligentes.

Embora se possam encontrar diversas definições, um dos responsáveis pela criação do termo IA a definiu como:

> É a ciência e a engenharia de fazer máquinas inteligentes, especialmente programas de computador inteligentes. Está relacionado à tarefa semelhante

> de usar computadores para entender a inteligência humana, mas a IA não precisa se limitar a métodos que são biologicamente observáveis. (Mccarthy, 2007).

Russel e Norving (2016), concentraram os estudos em quatro abordagens da IA diferenciando sistemas computacionais ancorados no pensamento e racionalidade *versus* ação. Segundo os autores, as quatro abordagens podem ser divididas em dois grupos, a abordagem humana que envolve sistemas que pensam e agem como humanos e a abordagem racionalista que envolve sistemas que pensam e agem racionalmente.

Os memos autores ainda trazem em seu livro os conceitos de IA fraca e IA forte. A IA fraca é conhecida como ANI do inglês *Artificial Narrow Intelligence,* e é treinada e focada para executar tarefas específicas. O conceito de fraco, não necessariamente indica que a IA é fraca, mas sim limitada. Neste caso, sistemas ANI permitem desenvolvimento de aplicativos muito robustos, como veículos autônomos e o sistema *IBM Watson,* por exemplo.

Já IA forte é composta pela *Artificial General Intelligence* (AGI) e pela *Artificial Super Intelligence* (ASI). A AGI (Inteligência Geral Artificial), também conhecida como IA geral, representa uma concepção teórica de inteligência artificial na qual uma máquina possuiria uma capacidade intelectual equivalente à humana. Isso implica na habilidade de resolver desafios, assimilar conhecimento e elaborar planos futuros, adaptando-se aos contextos específicos nos quais é aplicada. Já a ASI - conhecida como superinteligência - transcenderia tanto a inteligência quanto a capacidade do cérebro humano. Embora a ASI permaneça predominantemente no campo teórico e careça de exemplos práticos em uso atualmente, isso não impede que os pesquisadores no campo da IA estejam investigando seu desenvolvimento. As investigações em torno da ASI têm como objetivo a criação de máquinas que ultrapassem a capacidade humana, em todos os aspectos possíveis.

De fato, a IA é realmente uma façanha revolucionária da ciência da computação, destinada a se tornar um componente central de todos os *softwares* modernos nos próximos anos e décadas. Atualmente, existem inúmeras aplicações reais de sistemas de IA, como reconhecimento da fala, atendimento ao cliente, visão computacional, mecanismos de recomendação e negociação automatizada de ações.

O reconhecimento automático da fala (ASR do inglês *Automatic Speech Recognition*) é um sistema que usa o processamento de linguagem natural (NLP do inglês *Natural Language Processing*) para processar a fala humana em um formato escrito. Diversos modelos de Aprendizado de Máquina veem sendo utilizados, principalmente com os avanços do

aprendizado profundo (DL do inglês *Deep* Learning) (Padmanabhan; Johnson; Premkumar, 2015). Muitos dispositivos móveis incorporam o reconhecimento de fala em seus sistemas para realizar pesquisa por voz – por exemplo, Siri – ou fornecer mais acessibilidade em torno de mensagens de texto.

Em uma outra frente, os agentes virtuais online estão assumindo o papel dos agentes humanos nos sistemas de atendimento ao cliente virtual. Esses agentes virtuais são capazes de responder a perguntas frequentes (FAQ) sobre diversos tópicos, como remessas, além de fornecer conselhos personalizados, sugestões de produtos de venda cruzada e auxiliar os usuários com orientações sobre tamanhos. Essa mudança na abordagem está transformando a maneira como encaramos o envolvimento do cliente em sites de compras e mídias sociais (Ranoliya; Raghuwanshi; Singh, 2017). Na mesma linha de compras virtuais, é possível utilizar algoritmos de IA e dados de comportamento de consumo anterior, para descobrir tendências de dados relevantes que são utilizadas para desenvolver estratégias mais eficazes de vendas cruzadas. Essas recomendações adicionais são oferecidas aos usuários no processo de finalização de compras online, tornando a experiência de compra mais personalizada e atraente (Zhao; Keikhosrokiani, 2022).

A tecnologia de inteligência artificial também é utilizada no campo da visão computacional, permitindo que computadores e sistemas obtenham informações relevantes a partir de imagens, vídeos e outras entradas de dados visuais, utilizando essas informações para tomar ações adequadas. Essa capacidade de entregar sugestões e interações é o que a diferencia das tradicionais tarefas de reconhecimento de imagem. Utilizando redes neurais convolucionais, a visão computacional encontra aplicações em diversas áreas, como marcação de fotos em mídias sociais, interpretação de imagens de radiologia no campo da saúde, veículos autônomos na indústria automotiva e reconhecimento de imagens e padrões faciais (Alirezazadeh; Fathi; Abdali-Mohammadi, 2015)

E por fim, há também os sistemas de negociação de ações que lançam mão das tecnologias de inteligência artificial para aprimorar os portfólios de ações, executando milhares de negociações diariamente sem intervenção humana (Sezer; Ozbayoglu; Dogdu, 2017). Essas plataformas utilizam algoritmos avançados para analisar dados e identificar oportunidades de negociação em alta velocidade, proporcionando maior eficiência e rapidez nas operações do mercado financeiro.

2.3.1 Inteligência Artificial na Educação

A IA está transformando praticamente todos os aspectos de nossas vidas, e a educação não é exceção. Com as ferramentas e plataformas baseadas em IA se tornando mais predominantes na sala de aula, educadores, administradores e formuladores de políticas públicas precisam entender o cenário emergente da inteligência artificial na educação.

Por meio da IA é possível revolucionar a forma como aprendemos, ensinamos e administramos, oferecendo oportunidades sem precedentes para melhorar os resultados educacionais e apoiar o sucesso dos alunos. De experiências de aprendizado personalizadas a ferramentas de avaliação de ponta, a IA já está mudando profundamente a face da 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 relatório reafirma ainda as diversas pesquisas realizadas nos últimos anos quanto ao uso de IA em ambientes educacionais nas mais diversas áreas.

A aprendizagem personalizada, por exemplo, é uma abordagem que adapta o conteúdo educacional às necessidades específicas de cada aluno.

> Modelos de aprendizagem personalizada buscam se adaptar ao ritmo de aprendizagem e às estratégias de ensino, conteúdo e atividades usadas para melhor atender aos pontos fortes, fracos e interesses de cada aluno. A aprendizagem personalizada é sobre dar aos alunos o controle sobre sua aprendizagem, instruções diferenciadas para cada criança e fornecendo feedback em tempo real (Shemshack; Spector, 2020).

Neste sentido a IA é capaz de analisar vastas quantidades de dados sobre as preferências e hábitos de aprendizagem de cada aluno, permitindo que os educadores criem experiências de aprendizado personalizadas que possam atender às necessidades e habilidades exclusivas de cada aluno. De fato Chassignol et al. (2018), apontaram a crescente utilização de ferramentas de IA em universidades americanas na adaptação de conteúdos, promovendo a melhora na qualidade do processo educacional, auxiliando no desenvolvimento de conteúdo, nos métodos de ensino, na avaliação do aluno e na comunicação entre professor e alunos. Samarakou et al. (2016) realizaram o monitoramento contínuo e a avaliação de estudantes de engenharia em que

a IA foi utilizada para fornecer *feedback* personalizado e avaliar o desempenho usando dados qualitativos e quantitativos. Neto (2019) verificou em seu trabalho a utilização de algoritmos preditivos para prever o desempenho dos alunos em avaliações somativas, em ambiente virtual de aprendizagem, a partir de dados como idade, gênero, resultados, índice de entregas e entregas em atraso, número de interações em fóruns e tempo de uso da plataforma.

Outro exemplo de utilização da IA na educação é nas tecnologias assistivas. Ferramentas de IA são capazes de ajudar alunos com necessidades especiais a acessar uma educação mais igualitária, por exemplo, lendo passagens de texto para um aluno com deficiência visual ou substituindo a atividade escrita pela fala. Barua et al. (2022) realizaram um estudo sobre ferramentas que utilizam IA para melhorar a educação de crianças com distúrbios do neurodesenvolvimento, a saber, dislexia, transtorno de déficit de atenção (TDAH) e transtorno do espectro autista (TEA). Neste estudo, foram identificadas 5 ferramentas para auxiliar alunos com TDAH, que vão desde relógios para ajudar a enviar lembretes para os alunos se concentrarem nos trabalhos até *softwares* que substituem a atividade escrita pela fala.

Para alunos com dislexia, oito ferramentas foram levantadas, com ferramentas para melhorarem as habilidades de leitura, escrita, cognitivas, além de ferramentas que convertem textos em imagens para facilitar o aprendizado. Outras 13 ferramentas foram encontradas para auxiliar alunos com TEA, que vão desde a ajuda para o reconhecimento de emoções e ajuda na expressão delas até o desenvolvimento de habilidades diversas, como matemática e comunicação. Os autores concluíram que os estudos mostraram que as ferramentas assistidas por IA mostraram impactos positivos no aprendizado dos alunos e foram consideradas aceitáveis por professores, pais, educadores especiais e terapeutas e viáveis para implementar em suas práticas de ensino ou terapêuticas.

A educação infantil também se beneficia dos avanços da IA. Conforme sugerem Devi et al. (2022), é possível utilizar a IA para automatizar tarefas administrativas, liberando os professores para melhor interação com os alunos. As ferramentas de IA também são capazes de gerar sistemas de tutoria e apoio fora da sala de aula por meio de robôs e *chatbots* e ainda a utilização de *softwares* de reconhecimento facial, fornecendo aos professores informações sobre o comportamento dos alunos durante a aula, permitindo que eles se envolvam ou tomem alguma ação, bem como criem práticas centradas no aluno e aumentem a participação do aluno.

À medida que a tecnologia progride em direção à sala de aula, conforme exposto, o mesmo acontece a nível da administração escolar. A IA também está sendo utilizado para criar modelos preditivos e de diagnóstico para apoiar decisões e gerar *feedback*, tanto ao nível da unidade de ensino (escolas, universidades etc.) quanto ao sistema de ensino (cidade, estado,

país etc.). Bakhshinategh et al. (2018), apontaram em seus estudos aplicações alvo que se beneficiam do uso de tecnologias de IA no âmbito da administração escolar por meio da previsão de desempenho acadêmico, detecção de comportamento indesejável dos estudantes, criação de perfil e agrupamento, análise de redes sociais, fornecimento de relatórios, planejamento e programação. Vincent-Lancrin e Vlies (2020) vão ao encontro do trabalho de Bakhshinategh ao apontar o uso da IA como recurso necessário para prevenir a evasão escolar, por meio de alertas antecipados, indicando a gestores as características dos alunos e fornecendo intervenções apropriadas.

Sobre a evasão escolar, diferentes estudos promoveram soluções diversas, utilizando algoritmos de IA para ajudar no suporte da prevenção da evasão escolar. Sales; Balby e Cajueiro (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, alcançando uma precisão entre 82,0% e 89,0%.

Chen Hsieh e Do (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*.

Utilizando métodos de aprendizado de máquina, Chung e Lee (2019) utilizaram os dados de 165.715 estudantes de nível médio, coletados junto ao Sistema Nacional de Informação de Educação da Coréia do Sul 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, Silva e Xavier (2021), 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.

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. Além disso, a utilização das TIC tem sido fundamental para o avanço da mineração de dados educacionais. Por meio da análise de dados, as instituições educacionais conseguem identificar padrões e tendências que ajudam a melhorar a eficácia do ensino e da aprendizagem, bem como prevenir a evasão escolar. Utilizando mineração de dados é possível extrair informações valiosas de grandes conjuntos de dados educacionais. Como destacado anteriormente, as TIC possibilitaram a coleta e gerenciamento eficiente de dados, o que é um recurso valioso para a mineração de dados.

2.4 Mineração de Dados

Entende-se a mineração de dados como o processo de classificação de grandes conjuntos de dados com o intuito de identificar padrões e relacionamentos que permitem ajudar a resolver problemas de negócios por meio da análise de dados. As técnicas e ferramentas de mineração de dados permitem que as empresas prevejam tendências e tomem decisões de negócios com mais informações. A mineração de dados desempenha um papel essencial na análise de informações e é uma disciplina fundamental dentro da ciência de dados, empregando métodos avançados de análise para extrair *insights* valiosos de conjuntos de dados. A mineração de dados é parte integral do processo de descoberta de conhecimento em bancos de dados (KDD do inglês *Knowledge Discovery in Databases*), uma metodologia capaz de analisar dados brutos e transformá-los em informação útil (Tan; Steinbach; Kumar, 2005a). A Figura 1 ilustra o processo de descoberta do conhecimento de dados.

O pré-processamento de dados é a primeira etapa após a entrada de dados no sistema. Consiste no processo de análise dos dados, com o objetivo de deixá-los mais adequado à fase de mineração, pegando os dados brutos e os transformando em um formato válido para ser entendido e analisado por ferramentas de IA. Para tal, diversas técnicas e estratégias são aplicadas aos dados, como eliminação manual de atributos, integração de dados, limpeza,

transformação, redução de dimensionalidade e amostragem (Faceli et al., 2011; Tan; Steinbach; Kumar, 2005a). Essa etapa é de extrema importância, uma vez que, os dados brutos do mundo real na forma de texto, imagens, vídeo etc., são confusos, frequentemente contendo erros e inconsistências, sendo incompletos e não tendo um *design* regular e uniforme.

> Técnicas de pré-processamento de dados são frequentemente utilizadas para melhorar a qualidade de dados por meio da eliminação ou minimização dos problemas citados. Essa melhora pode facilitar o uso de técnica de AM, levar à construção de modelos mais fiéis à distribuição real dos dados, reduzindo sua complexidade computacional, tornar mais fáceis e rápidos o ajuste de parâmetros do modelo e seu posterior uso. (Faceli et al., 2011).

Portanto, dados bem estruturados e pré-processados se tornam ainda mais importantes do que os algoritmos mais poderosos, a ponto de os modelos de aprendizado de máquina treinados com dados ruins terem o potencial de prejudicar à análise, fornecendo resultados indesejáveis.

Figura 1 - Processo de descoberta do conhecimento de dados.

Fonte: Giordano e Souza (2023)

Depois que os dados são preparados, o cientista de dados ou analista de dados precisa escolher a técnica de mineração de dados apropriada e, em seguida, implementar um ou mais algoritmos para fazer a mineração. Em aplicativos de AM, os algoritmos geralmente devem ser treinados com conjuntos de dados de amostra, denominados conjunto de treinamento, para buscar as informações que estão sendo procuradas antes de serem executados no conjunto de teste, que é desconhecido pelo modelo. Conforme apontado por Tan, Steinbach e Kumar (2005), os principais elementos da mineração de dados incluem AM, IA, reconhecimento de padrões e análise estatística. A utilização de ferramentas de IA e AM automatizou ainda mais o processo e facilitou a extração de conjuntos de dados massivos, como bancos de dados de clientes, registros de transações e arquivos de *log* de servidores da *web*, aplicativos móveis e sensores.

Ainda de acordo com Tan et al. (2005) 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.

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 Tan, Steinbach e Kumar (2005) e Faceli 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.

2.4.1 Mineração de Dados Educacionais

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).

A EDM é uma área em constante crescimento, que vem sendo utilizada por instituições de ensino para analisar grandes conjuntos de dados educacionais a fim de descobrir padrões que possam melhorar a eficácia da aprendizagem e do ensino. Com o advento da tecnologia, o número de dados coletados pelas instituições educacionais tem aumentado exponencialmente, permitindo que a mineração de dados educacionais se torne uma poderosa ferramenta para identificar tendências, padrões e oportunidades de melhoria.

O principal objetivo da EDM é, portanto, melhorar a educação de maneira mais eficiente e efetiva, identificando áreas nas quais os alunos e professores necessitam de ajuda ou melhoria. Por meio da mineração de dados, é exequível que instituições de ensino identifiquem padrões e tendências nos dados dos alunos, como desempenho acadêmico, frequência, comportamento em sala de aula e informações demográficas, para ajudar a melhorar a qualidade do ensino e a efetividade do aprendizado (Dutt; Ismail; Herawan, 2017).

Uma das aplicações mais importantes da mineração de dados educacionais é a identificação de alunos em risco de desistência ou fracasso acadêmico. Através da análise de dados, as unidades de ensino conseguem identificar fatores que estão impedindo o sucesso acadêmico dos alunos, como a falta de frequência, desempenho abaixo do esperado ou dificuldades em determinadas disciplinas, conforme apontam Bakhshinategh et al. (2018). A partir dessas informações, as instituições se tornam aptas a desenvolver intervenções personalizadas para ajudar esses alunos a superarem esses obstáculos e melhorar seu desempenho.

Outra aplicação da mineração de dados educacionais é a personalização da aprendizagem. Os dados coletados sobre o desempenho dos alunos possuem potencial de utilização na identificação de necessidades individuais de aprendizagem, permitindo que os professores adaptem seus métodos de ensino para atender a essas necessidades. Isso inclui, por exemplo, ajustar o ritmo da aula, oferecer recursos adicionais ou fornecer feedback mais personalizado aos alunos.

De acordo com Baker e Yacef (2009), as técnicas de mineração de dados educacionais incluem análise de associação, classificação, agrupamento e análise de sequência, psicometria, estatística, modelagem computacional, entre outras. A análise de associação é utilizada para descobrir relações entre variáveis, como o desempenho acadêmico e a frequência escolar. A classificação é usada para identificar grupos de alunos com características semelhantes, enquanto o agrupamento é usado para agrupar alunos com base em suas características. A análise de sequência é usada para estudar a ordem em que as atividades são realizadas pelos alunos, como a ordem em que eles acessam recursos *on-line*.

Assim, é razoável que a mineração de dados educacionais seja aplicada a uma variedade de contextos educacionais, desde a educação básica até o ensino superior e a aprendizagem ao longo da vida. No entanto, a EDM também levanta preocupações sobre privacidade e segurança dos dados dos alunos. É importante que as instituições educacionais tenham políticas claras e proteções adequadas em vigor para proteger os dados dos estudantes.

2.5 Aprendizado de Máquina

O advento da quarta revolução industrial vem implicando em uma mudança na forma com que o sistema produtivo funciona. Uma automatização completa de fábricas por meio de tecnologias como a *internet* das coisas e a computação em nuvem permitirá a criação de fábricas inteligentes. Segundo Schwab (2016), a quarta revolução industrial baseia-se na revolução

digital, tendo como características uma maior distribuição e mobilidade da *internet*, diminuição de tamanho e aumento da capacidade de sensores, pela inteligência artificial e o aprendizado de máquina. Identificar falhas, detectar e classificar padrões em sistemas de monitoramento, previsão de manutenção, falhas e condições de trabalho são apenas algumas das aplicações possíveis em que há a possibilidade de aplicar métodos de aprendizado de máquina (Ahuett-Garza; Kurfess, 2018).

O dinamismo da Indústria 4.0 está levando organizações de todos os segmentos e tamanhos, a buscarem estratégias de melhoria e suporte à tomada de decisão. Em consequência disso, surge a utilização da ciência de dados, aperfeiçoando o entendimento de problemas e consequentemente melhorando a produtividade e o gerenciamento do negócio (Gokalp et al., 2016; Pinzone et al., 2017). O efeito direto dessas mudanças é a crescente necessidade de profissionais aptos a atuarem no gradativo ramo da ciência de dados. Conforme aponta a reportagem da Folha de São Paulo, as oportunidades de carreira para profissionais de ciências de dados têm aumentado significativamente. Somente nos EUA, entre 2012 e 2017, o número de vagas foi multiplicado por 6,5 (Folha, 2018).

Alguns estudos, como o de Fitsilis, Tsoutsa e Gerogiannis (2018) apontam as características dos profissionais aptos a trabalharem com ciência de dados, como possuir graduação, mestrado, doutorado ou experiência equivalente em ciência da computação, matemática ou estatística, experiência na linguagem de programação Python, e conhecimento de algoritmos e estruturas de dados. As características necessárias desses profissionais possuem relação com a complexidade envolvida nas atividades do cientista de dados. O trabalho de um cientista de dados envolve a coleta e análise de grandes volumes de dados, entendendo seu comportamento e utilizando para isso estatística e matemática aplicadas às técnicas de análise, como o aprendizado de máquina, por exemplo.

Realça-se que o trabalho de um cientista de dados envolve o conhecimento e habilidades em diferentes áreas do saber conforme aponta Aalst (2014), principalmente em computação, na qual o aprendizado de máquina se destaca como uma das principais ferramentas para mineração de dados.

O aprendizado de máquina é, portanto, uma linha de pesquisa da Ciência da Computação, mais precisamente um ramo de pesquisa dentro da inteligência artificial. Tal linha de pesquisa nasceu da crescente necessidade de processos computacionais mais autônomos e sofisticados, com menor dependência de especialistas (Faceli et al., 2011). Ainda de acordo com os autores, o aprendizado de máquina é o "processo de indução de uma hipótese (ou aproximação de função) a partir da experiência passada". Isso implica em disponibilizar ao

algoritmo de AM, dados para que possa, por meio de induções, realizar conclusões lógicas sobre o problema a ser tratado. Isso permite que o algoritmo aprenda o comportamento por meio de funções matemáticas, levando ao aprendizado.

Para Tan, Steinbach e Kumar (2005) a abordagem sistemática para aprender um modelo de classificação dado um conjunto de treinamento é conhecida como algoritmo de aprendizado. O processo de usar um algoritmo de aprendizado para construir um modelo de classificação a partir dos dados de treinamento é conhecido como indução.

Portanto, entende-se o aprendizado de máquina, como uma metodologia que busca o aprendizado a partir de análise de dados, automatizando a construção de modelos.

Diferentes modelos e algoritmos de aprendizado de máquina vêm sendo desenvolvidos nos últimos anos a fim de atender à crescente demanda do mundo moderno. Os algoritmos propostos para tarefas de aprendizado de máquina seguem dois paradigmas distintos, classificando-se em modelos preditivos e descritivos, segundo as definições de Faceli et al. (2011) e Tan, Steinbahc e Kumar (2005). Modelos preditivos são utilizados 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 por meio de paradigma de aprendizado supervisionado.

Já modelos descritivos são utilizados para identificar características em instâncias de diferentes classes. Esse método gera tipicamente modelos de associação e agrupamento por meio de paradigma de aprendizado não-supervisionado.

Independente do modelo a ser utilizado, estes precisam ser alimentados com informações ou conhecimento. É necessário então, que o conhecimento do mundo real, seja organizado e representado de forma adequada ao sistema computacional ou algoritmo a ser utilizado. Todo programa de computador possui, de certa forma, um conhecimento agregado que o faz ser útil à tarefa para a qual foi desenvolvido. Contudo, esse conhecimento não é explícito e tampouco suporta atualização ou manipulação com agilidade. Isto é, os requisitos levantados em fase de análise de sistemas e posteriormente codificados em alguma linguagem de programação, representam um conhecimento específico, mas que não são facilmente alterados, por exemplo.

Dentro da inteligência artificial, o conhecimento deve ser representado de forma que, independentemente do sistema desenvolvido, a máquina possa chegar a conclusões sobre o mundo a partir dessas representações. Dúvidas sobre como o conhecimento deve ser representado ou qual a maneira mais adequada para representá-lo são foco de estudos. Lógica de predicados, *frames*, redes semânticas e banco de dados são algumas das formas possíveis de

representarmos o conhecimento em sistemas de IA (Levesque, 1986).

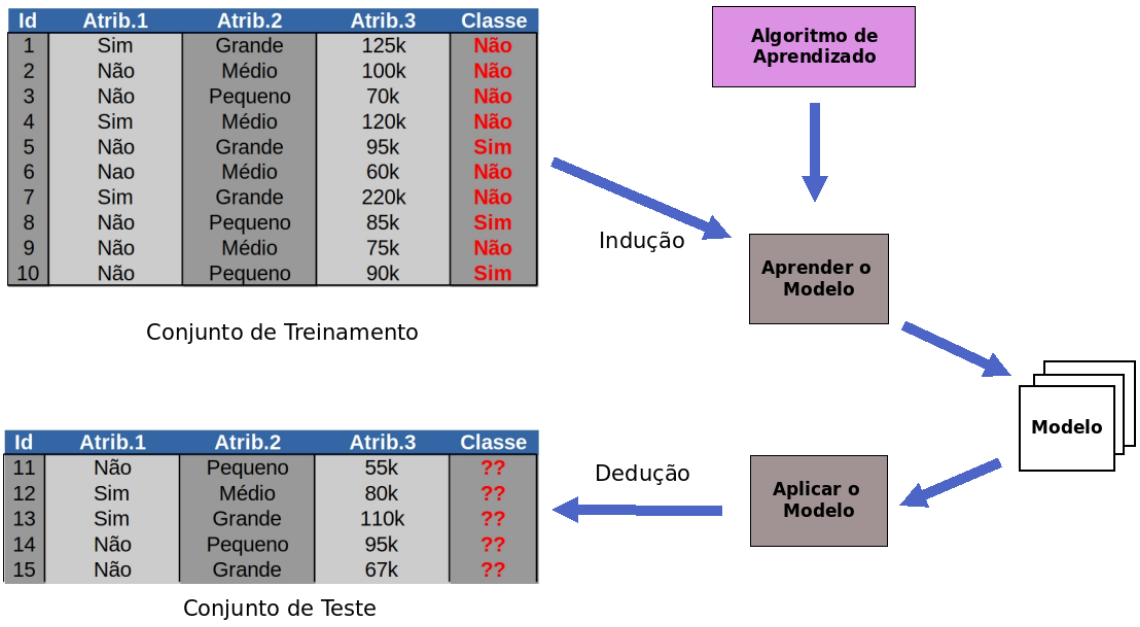
Para se representar o conhecimento do mundo real de forma computacional, faz-se necessário entender como é realizado o processo de classificação. Quando deseja-se classificar observações em categorias, como por exemplo, identificar se um veículo é do tipo leve (carros de passeio) ou pesado (caminhões e ônibus), ou ainda, se um alimento é vegetal ou animal, observamos suas características (atributos) e determinamos sua categoria. Esta é uma tarefa intuitiva aos humanos, que, com base no aprendizado ao longo dos anos e com centenas de milhares de observações diferentes, tornaram-se aptos a executar classificações quase que instantaneamente.

Trazer essa classificação e consequentemente, a representação do conhecimento, para o mundo computacional, implica em entender, a princípio, como se dá o processo de classificação. Segundo Tan et al. (2019) a classificação é um modelo que tem como objetivo determinar a classe de uma determinada observação tendo como base seus atributos, isto é, identificar a qual conjunto de categorias esta pertence. Ao aplicarmos essa definição em um modelo computacional, com o objetivo de darmos essa mesma capacidade de classificação a uma máquina, faz-se necessário que esta possua um classificador, que deve tentar aproximar uma função f' de uma função desconhecida f. Isso permite estimar o valor de da função f para cada observação nova de x em $D = \{(x_i, f(x_i)), i = 1, \dots, n\}$ (Faceli et al., 2011). A observação x é formada por uma tupla (y, z) onde y são atributos de qualquer tipo que descrevem o objeto x enquanto z assume apenas valores categóricos e representa a classe a qual x pertence. O modelo irá classificar uma observação de forma correta se $f(y) = z$ (Tan et al., 2019).

A partir destes conceitos, a classificação então, é dividida, segundo Faceli et al. (2011) em duas tarefas: preditivas e descritivas. A predição é o processo de utilizar um algoritmo de aprendizado para construir um modelo de aprendizado com base em um conjunto de dados de treinamento. Já a tarefa descritiva busca a exploração de um determinado conjunto de dados, tendo como objetivo, por exemplo, tarefas de agrupamento, como encontrar grupos de objetos semelhantes no espaço de busca. A Figura 2 ilustra o modelo de construção de um classificador. Conforme ilustrado pela Figura 2 um conjunto de treinamento deve ser repassado para a construção de um modelo, enquanto um conjunto de validação e outro de teste são separados para validar e testar o modelo construído. Normalmente, um conjunto de dados é dividido em um conjunto de treinamento, um conjunto de validação e um conjunto de teste onde pontos de dados no conjunto de treinamento são excluídos do conjunto de validação e teste.

O objetivo do aprendizado de máquina é criar um modelo para prever os dados do conjunto de teste. Logo, utiliza-se os dados de treinamento para ajustar o modelo e os dados de

Figura 2 - Modelo para construção de um classificador.



Fonte: Adaptada de Tan et al. (2019)

teste para testá-lo, ver Figura 2.

2.5.1 Autoaprendizado de Máquina

Desde a virada do século 21, o aprendizado de máquina vem ganhando mais atenção de pesquisadores e indústrias em todo o mundo. Uma vasta gama de problemas suportam soluções por meio de algoritmos de AM, como por exemplo, agricultura (Duro; Franklin; Dubé, 2012), educação (Lykourentzou et al., 2009), reconhecimento de imagens (Alirezazadeh; Fathi; Abdali-Mohammadi, 2015), matemática, diagnósticos médicos (Konenko, 2001), música (Dannenberg; Thom; Watson, 1997), processamento de linguagem natural (Collobert; Weston, 2008), robótica (Stone; Veloso, 1997), reconhecimento de fala (Padmanabhan; Johnson Premkumar, 2015), na administração pública em diversos setores (VEALE; BRASS, 2019) entre outras (Michalski; Carbonell; Mitchell, 1983). Grandes corporações, com alto poder de gerar dados, como *Amazon*, *Facebook (Meta)* e *Google*, entenderam o potencial do aprendizado de máquina, seja para entender o perfil de compra e apontando tendências de um determinado consumidor, até o reconhecimento de voz e montagem de legendas automáticas.

O avanço dos algoritmos de aprendizado de máquina entregou um novo ferramental aos cientistas de dados para o trabalho de analisar e entregar resultados sobre dados brutos. Contudo, para que um cientista de dados possa transformar dados brutos em informação útil,

segundo (Guyon et al., 2015), é necessário que:

a) Formalize uma pergunta para uma abordagem do problema;
b) Selecione os dados apropriados;
c) Projete um modelo;
d) Realize o treinamento;
e) Valide o treinamento com testes; e,
f) Interprete os resultados.

Em adição à dificuldade do trabalho de um cientista de dados, soma-se a crescente demanda desse tipo de profissional, além do rápido estabelecimento da indústria 4.0 e têm-se uma área de pesquisa que começa a ser explorada com maior interesse, o AutoML. O objetivo do AutoML é proporcionar às pessoas que não possuem um conhecimento avançado de inteligência artificial e programação, acesso a ferramentas de aprendizado de máquina para o auxílio na tomada de decisões de seus próprios negócios (Budjač et al., 2019; Jin; Song; Hu, 2019).

Segundo Feurer et al. (2015), quando da aplicação de um serviço de aprendizado de máquina, este 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.

Portanto, o processo de construção de uma solução, utilizando aprendizado de máquina passa por um processo interativo de refinamento e escolha do melhor método. Um processo interativo, abre a oportunidade de escolha para a automatização deste processo. Assim o conceito de AutoML deriva da ideia de que, se vários modelos de aprendizado de máquina devem ser construídos, usando uma variedade de algoritmos e várias configurações diferentes de hiperparâmetros, essa construção de modelo tem potencial para ser automatizada, bem como a comparação de desempenho e precisão do modelo.

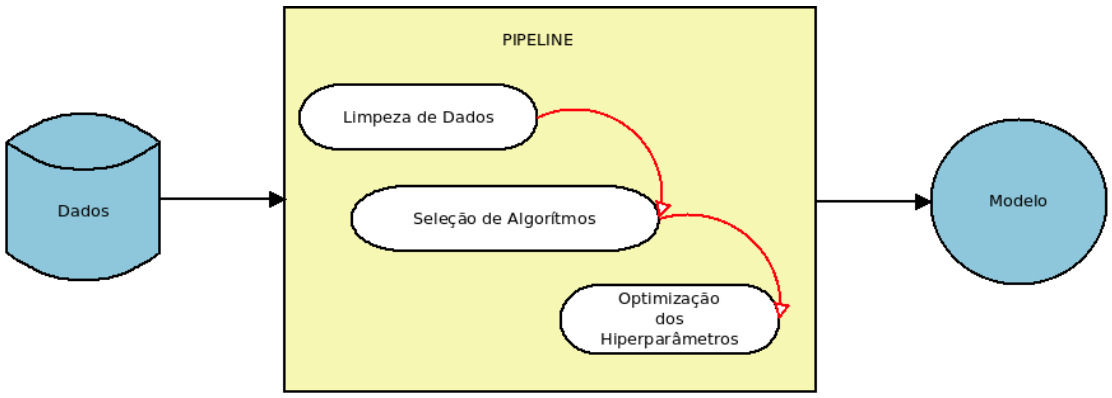
Zöller e Huber (2019) entendem o processo de automatização da construção de um modelo de aprendizado de máquina, definido então como AutoML, como o encapsulamento dos processos interativos (*pipeline*), possuindo assim uma etapa de limpeza e seleção de dados, uma etapa para a seleção de algoritmos e uma etapa para a otimização dos hiperparâmetros. A Figura 3 - Generalização de um *pipeline.* ilustra o modelo de generalização de um *pipeline.*

O objetivo é automatizar o fluxo de trabalho do aprendizado de máquina permitindo assim que dados sejam transformados e correlacionados em um modelo, sendo testados e avaliados para alcançar determinado resultado. Um *pipeline* bem estruturado torna a

implementação mais flexível pois permite ajustes em etapas independentes buscando alcançar melhores modelos.

Os *pipelines* são modelos interativos, onde cada etapa é repetida continuamente buscando melhorar o modelo. De acordo com Zöller e Huber (2019) o processo de construção de um *pipeline* começa na definição de sua estrutura. Alguns processos que podem fazer parte destes modelos são limpeza, extração de recursos, seleção de algoritmos, validação do modelo e visualização.

Figura 3 - Generalização de um *pipeline.*



Fonte: Adaptada de Zöller e Huber (2019)

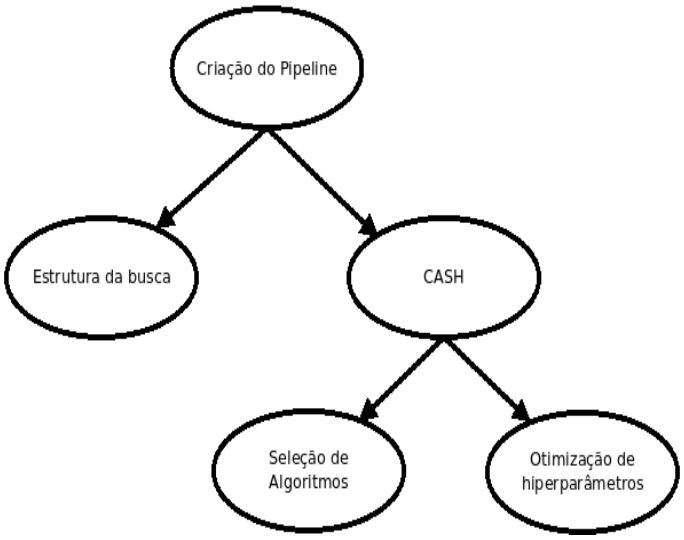
A busca pelo melhor modelo, que se adéque à necessidade de cada problema, é compreendida por Zöller e Huber (2019), como um problema de busca e otimização. Há nesse cenário, uma série de variáveis que influenciam o desempenho do modelo. O objetivo nesse tipo de problema é encontrar o conjunto de variáveis que possibilite o melhor desempenho possível para o problema em questão. Ou seja, dentro de um espaço de busca, encontrar a melhor solução dentre todas as possibilidades existentes (Tanomaru, 1995). Essa busca pelo modelo mais adequado e otimizado é conhecida como *Combined Algorithm Selection and Hyperparameter Optimization* (CASH) (Thornton et al., 2013).

O termo CASH foi utilizado por Thorton et al. (2013) pela primeira vez para definir o problema de seleção automática de um algoritmo de aprendizado de máquina ideal e suas respectivas configurações dos hiperparâmetros. Em seu trabalho a seleção dos algoritmos é abordada como um hiperparâmetro, tratando a seleção do algoritmo de aprendizado e a otimização dos hiperparâmetros simultaneamente. Para isso, o problema de seleção e otimização foi reformulado como uma otimização de parâmetros hierárquicos combinados com o espaço completo de hiperparâmetros para **n** algoritmos, que foi definido como $\Lambda = \Lambda^1 \cup \Lambda^2 \cup \ldots \cup \Lambda^k \cup \lambda_r$, onde o parâmetro λ_r possui relação com os demais parâmetros. Isso implica na

seleção do parâmetro Λ^k para algum algoritmo A^k pertencente ao conjunto $A = A^1, ...A^k$ (Thornton et al., 2013; Zöller; Huber, 2019).

Zöller e Huber (2019) interpretaram a proposta de Thorton como uma árvore de seleção, conforme exemplificado na **Erro! Fonte de referência não encontrada.**. A árvore de seleção é ser entendida como um problema de minimização da função proposta. No trabalho de Thornton, a minimização da função, ou seja, encontrar os hiperparâmetros hierárquicos, é realizada por meio de otimização bayesiana. Porém, outros métodos de otimização também são utilizados, como algoritmos genéticos, pesquisa em grade, pesquisa aleatória, gradiente descendente, entre outras (Bergstra et al., 2011; Zöller; Huber, 2019).

Figura 4 - Árvore representando os subproblemas de um pipeline



Fonte: Adaptado de Zöller e Huber (2019)

Além da abordagem proposta por Thorton et al. (2013), existem diversos métodos de AutoML que visam simplificar a seleção do algoritmo de aprendizado de máquina e a otimização dos hiperparâmetros simultaneamente. Esses métodos oferecem soluções práticas para lidar com o desafio de encontrar a combinação ideal de algoritmo e hiperparâmetros em problemas complexos de aprendizado de máquina. Entre os principais métodos de AutoML, destacam-se o Auto-WEKA, TPOT, Auto-Sklearn e H2O, cada um com suas próprias abordagens e estratégias de otimização (Zöller; Huber, 2019).

O Auto-WEKA (Kotthoff; Thornton; Hutter, 2017; Thornton et al., 2013) realiza a seleção de algoritmos e configuração de hiperparâmetros considerando todos os algoritmos de classificação candidatos, disponíveis na conhecida ferramenta de mineração de dados WEKA, que inclui algoritmos baseados em diferentes tipos de representações de conhecimento (ou

modelo) - por exemplo, árvores de decisão, regras de classificação se-então, classificadores de rede bayesiana, redes neurais, SVM, etc.

Já o TPOT é um método baseado em programação genética que procura encontrar o *pipeline* de classificação mais adequado para o conjunto de dados de entrada. Ele engloba (parte) os métodos disponíveis na biblioteca *scikit-learn* em seu espaço de pesquisa e permite diferentes maneiras de combinar os métodos de pré-processamento de dados (em sequência ou em paralelo) e os algoritmos de classificação (apoiando abordagens de grupo ou não). Embora o TPOT tenha sido projetado para classificação geral, ele alternativamente possui uma versão específica para estudos de bioinformática, denominada TPOT-MDR (Sohn; Olson; Moore, 2017).

O Auto-Sklearn é uma biblioteca de código aberto de AutoML para *Python*. Ele faz uso das bibliotecas de aprendizado de máquina *Scikit-Learn* para transformações de dados e algoritmos de aprendizado, utilizando otimização bayesiana para encontrar com eficiência um *pipeline* com melhor desempenho para o conjunto de dados informado. Além disso o Auto-Sklearn aprende com os modelos que tiveram um bom desempenho em conjuntos de dados semelhantes e cria automaticamente um conjunto de modelos de alto desempenho como parte do processo de otimização.

E por fim o H2O é uma plataforma de aprendizado de máquina e IA de código aberto que oferece suporte a uma interface baseada na *web* chamada *Flow*. A ferramenta é utilizada para criar modelos de aprendizado de máquina sem escrever nenhum código, permitindo simplesmente selecionar e clicar para criar *pipelines* de aprendizado de máquina. Há bibliotecas disponíveis para R, *Python* e Scala (Ledell; Poirier, 2020).

Entre os métodos de AutoML apresentados, o H2O se destaca como uma escolha promissora para simplificar a seleção do algoritmo de aprendizado de máquina e a otimização dos hiperparâmetros simultaneamente. Além de ser uma plataforma de código aberto, apresenta também uma abordagem que permite criar modelos de aprendizado de máquina sem a necessidade de escrever códigos complexos. Por meio do H2O, é possível utilizar uma interface intuitiva, na qual é possível selecionar e clicar para criar *pipelines* de aprendizado de máquina(H2O.AI, 2020).

Além da sua facilidade de uso, a ferramenta também disponibiliza bibliotecas para R, Python e Scala, o que proporciona uma ampla gama de opções de integração com diferentes ambientes de desenvolvimento e *frameworks* de análise de dados. Isso facilita a adoção do H2O em diferentes cenários e ambientes de trabalho.

Outro aspecto relevante é que o H2O é uma plataforma de código aberto, o que significa

que é possível acessar o código-fonte, modificá-lo e adaptá-lo conforme necessário. Essa característica proporciona flexibilidade e possibilita a contribuição da comunidade para o aprimoramento contínuo.

Ao considerar esses pontos, a escolha do H2O como método de AutoML se justifica não apenas pela sua abordagem intuitiva e amigável, mas também pela sua capacidade de integração com diferentes linguagens de programação e pelo acesso ao código-fonte, permitindo uma maior customização e adaptação às necessidades específicas de cada projeto.

2.5.1.1 Hiperparâmetro

O hiperparâmetro é uma configuração externa ao modelo de aprendizado de máquina e diferentemente dos parâmetros, não é estimado a partir do conjunto de dados. Geralmente, são ajustados para auxiliar o modelo em um determinado problema.

Contudo, nem sempre se sabe o valor a ser atribuído aos hiperparâmetros. Copiar valores utilizados em outros problemas semelhantes ou encontrar o melhor valor por tentativa e erro são maneiras de abordar o problema. Ao calibramos um algoritmo de aprendizado de máquina para uma tarefa particular, estamos essencialmente otimizando os hiperparâmetros do modelo para identificar os conjuntos de parâmetros que levam a previsões mais precisas.

> Muitos modelos têm parâmetros importantes que não podem ser estimados diretamente a partir dos dados. Por exemplo, no modelo de classificação K-vizinho mais próximo. Esse tipo de parâmetro de modelo é referido como parâmetro de ajuste porque não existe uma fórmula analítica disponível para calcular um valor apropriado (Kuhn; Johnson, 2013, p.55).

Os hiperparâmetros do modelo são, portanto, propriedades que governam todo o processo de treinamento. Eles incluem variáveis que determinam a estrutura de uma rede neural, por exemplo, como número de camadas ocultas, as variáveis que determinam como a rede é treinada e a taxa de aprendizagem. Os hiperparâmetros do modelo são definidos antes do treinamento. Alguns exemplos de hiperparâmetros a serem configurados, em uma Rede Neural Artificial (RNA), por exemplo, incluem a taxa de aprendizagem, número de épocas, número de camadas ocultas e a função de ativação. Outros exemplos incluem os hiperparâmetros C e σ das máquinas de vetores de suporte ou *Suport Vector Machine* (SVM) e o parâmetro k do algoritmo k-Vizinhos mais próximos ou *k-Nearest Neighbor* (kNN).

Existem várias maneiras de ajustar os hiperparâmetros de um modelo, como pesquisa

em grade, busca aleatória ou por otimização *bayesiana*.

Na pesquisa em grade, os parâmetros são definidos e pesquisados exaustivamente. Uma vez que um modelo é construído para todas as combinações possíveis de parâmetros, esse processo torna-se caro e demorado do ponto de vista computacional.

A pesquisa aleatória realiza uma busca randômica sobre os parâmetros. Neste caso, cada configuração é amostrada a partir de uma distribuição sobre os valores de parâmetros possíveis. Ao contrário de pesquisa em grade, nem todos os parâmetros especificados serão tentados. Em comparação com a pesquisa exaustiva, este método mostra-se vantajoso pois há a opção de escolher o número máximo de tentativas desejadas para esta pesquisa. Os parâmetros a serem amostrados são especificados por meio de um dicionário.

Já na otimização bayesiana, os resultados obtidos de um experimento são usados para melhorar a amostragem para o próximo experimento. Este processo é repetido até que os parâmetros ótimos sejam obtidos. O método é baseado no Teorema de Bayes. É principalmente usado quando a função objetivo em questão é complexa ou computacionalmente cara para avaliar.

2.5.1.2 H2O

O H2O é um produto da H2O.ai, uma 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 aprendizado de máquina a mais pessoas, por meio de uma plataforma de fácil utilização e implementação. A proposta *open-source* apresenta produtos diversos, desde aplicações mais simples até soluções corporativas. É uma ferramenta que permite aos usuários automatizarem o processo de criação de modelos, desde a seleção do conjunto de dados até a seleção do modelo mais adequado, incluindo pré-processamento de dados, ajuste de hiperparâmetros e seleção de recursos. É uma ferramenta altamente automatizada que auxilia na redução do tempo e no esforço necessário para criar um modelo de aprendizado de máquina. Ele também tem recursos para lidar com dados ausentes, transformação de dados e seleção de recursos, e é usado para criar modelos para uma variedade de tarefas, como classificação, regressão e agrupamento.

O modelo desenvolvido pela empresa utiliza técnicas de compactação de memória, ou seja, por meio do gerenciamento de memória, o sistema reduz a quantidade de dados inativos na memória RAM para liberar o espaço não utilizado, permitindo que mais programas utilizem a mesma memória. Isso permite que o sistema lide com conjuntos de dados em grande escala

na memória mesmo com um *cluster* bastante pequeno (Nykodym et al., 2020). Esta é a razão pela qual o H2O é considerado uma plataforma “rápida” pois os dados estão sendo 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 um armazenamento distribuído do tipo chave/valor é usado para acessar e fazer referência a dados, modelos, objetos etc., em todos os nós e máquinas. Os algoritmos são implementados no topo de uma estrutura de mapa distribuída do H2O e utilizam o *framework Fork/Join* para paralelismo com o objetivo de permitir um dimensionamento horizontal simples para um determinado problema a fim de produzir uma solução com maior velocidade. Assim, o H2O constitui um *software* potencialmente utilizável para modelagem de dados e computação em geral, com um motor de processamento distribuído, paralelo e em memória (H2O.AI, 2020; Ledell; Poirier, 2020).

O paralelismo do H2O é dividido em duas categorias. O primeira lança um único nó (máquina local). O segundo lança vários nós em um *cluster*. A partir disso, o analisador de dados do H2O tem a capacidade de adivinhar o esquema do conjunto de dados que é importado e lido de várias fontes em vários formatos (H2O.AI, 2020; Ledell; Poirier, 2020). O H2O oferece uma boa qualidade junto com velocidade, facilidade de uso e implantação de modelo para os vários algoritmos supervisionados e não supervisionados. Assim, os modelos de aprendizado de máquina que são construídos no R ou Python são facilmente convertidos para o formato POJO e implementados em qualquer ambiente Java. Por fim, é possível realizar os experimentos simplesmente com o H2O, apenas iniciando o *framework* ou realizando os mesmos com Python ou R em um navegador.

O Quadro 2 apresenta os algoritmos de classificação e pré-processamento utilizado pelo H2O. O *framework* permite ao usuário escolher quais modelos deseja treinar senão todos.

Por fim, o H2O possui um conjunto de ferramentas de visualização e diagnóstico que auxiliam os usuários a entenderem seus dados e modelos. Oferece ainda recursos para ajustar os hiperparâmetros dos modelos, como taxa de aprendizado ou a profundidade de uma árvore, com o objetivo de melhorar a qualidade dos modelos.

2.6 Métricas de Avaliação e Desempenho em Modelos de Aprendizado de Máquina

Para entender o desempenho de modelos de AM, é necessário que estes passem por métodos de avaliação. Neste tópico, serão apresentadas diversas métricas, divididas em métricas de avaliação (AUC e AUCPR) e métrica de desempenho (Matriz de Confusão,

Acurácia, Precisão, Revocação e F1-Score).

Quadro 2 - Algoritmos de métodos de pré-processamento no H2O

Algoritmos de Classificação	Métodos de Pré-Processamento
Deep Neural Net	*Automatic Imputation*
eX. Gradient Boosting Classifier	*One-hot Enconding*
eX. Random Forest (XRT)	*Normalization*
Generalized Linear Model (GLM)	
H2O GBM	
Random Forest	

Fonte: H2O.AI (2020); Ledell, Poirier (2020)

2.6.1 AUC (Area Under the ROC Curve)

De acordo com (Tan; Steinbach; Kumar, 2005) AUC é uma métrica que fornece uma abordagem para avaliar a performance média de um modelo de classificação binária. Neste contexto, caso um modelo consiga diferenciar entre as classes positiva e negativa, de forma perfeita (100,0% de distinção), a sua AUC seria igual a 1. Isso significa que não há sobreposição nas distribuições das duas classes, resultando em uma separação perfeita. Contudo, caso o modelo esteja fazendo suposições aleatórias, a sua AUC seria igual a 0.5. Isso ocorre devido ao fato de que, metade das instâncias positivas seriam classificadas corretamente, enquanto a outra metade seriam classificadas incorretamente.

Assim, na comparação de dois modelos, o melhor será aquele com a maior AUC encontrada. Um maior índice AUC aponta para uma maior capacidade do modelo em discriminar entre as classes. Tan, Steinbach e Kumar (2015) enfatizam ainda que a AUC é uma medida que varia de 0 a 1, sendo 0.5 associado adivinhação aleatória e 1 associado à perfeição do modelo.

2.6.2 AUCPR (Área Sob a Curva da Precisão-Revocação)

O AUCPR (*Precision-Recall Area Under the Curve*) é uma métrica de AM utilizada para avaliar o desempenho de modelos de classificação binária, principalmente quando as classes estão desequilibradas. De acordo com Sofaer, Hoeting e Jarnevich (2019), ao contrário da AUC, que traça a taxa de verdadeiros positivos (TVP) em relação à taxa de falsos positivos (TFP), a AUCPR traça a precisão em relação à sensibilidade em diferentes configurações de limite.

A precisão é a proporção de previsões positivas verdadeiras de todas as previsões positivas feitas pelo modelo (vide equação 2), enquanto a sensibilidade (*recall*) é a proporção de previsões positivas verdadeiras de todas as amostras positivas reais no conjunto de dados (vide equação 3). A curva PR é criada variando o limite para prever um resultado positivo ou negativo e traçando a precisão em relação à sensibilidade para cada limite (Sofaer; Hoeting; Jarnevich, 2019).

Portanto, a AUCPR fornece uma avaliação mais precisa do desempenho do modelo do que métricas como precisão ou F1-*Score*, que costumam ser tendenciosas para a classe majoritária. Além disso, são capazes de fornecer informações sobre o compromisso entre precisão e sensibilidade e ajudar a identificar o limite ideal para fazer previsões (Davis; Goadrich, 2006).

2.6.3 Matriz de Confusão

A matriz de confusão é uma tabela que descreve o desempenho de um modelo de classificação em termos de resultados verdadeiros e falsos. É frequentemente utilizada em problemas de classificação binária, onde o modelo tem duas classes possíveis, geralmente denominadas positiva e negativa (Faceli et al., 2011). As classes, portanto, são representadas na forma da Tabela 2, onde:

VP: Verdadeiros Positivos (VP), são os casos corretamente identificados da classe positiva.

VN: Verdadeiros Negativos (VN), são os casos corretamente identificados da classe negativa.

FP: Falsos Positivos (FP), são os casos em que a classe real é negativa, mas foram erroneamente classificados como positivos

FN: Falsos Negativos (FN), são os casos em que a classe real é positiva, mas foram erroneamente classificados como negativos.

Tabela 2 - Matriz de confusão para classificação binária

		Classe Predita	
		+	-
Classe Verdadeira	+	VP	FN
	-	FP	VN

Fonte 1: (Faceli et al., 2011)

Ainda de acordo com Faceli et al. (2011), a partir da matriz de confusão, encontra-se a acurácia, precisão, sensibilidade e outras medidas, conforme descrito abaixo.

A acurácia é calculada a partir da soma dos acertos dividido pela soma de todos os valores da matriz. Indica a proporção total de acertos realizados pelo modelo conforme Equação 1.

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

A precisão indica quão preciso são as previsões positivas realizadas pelo modelo. É calculada apenas para a classe positiva (+) conforme Equação 2.

$$\mathrm{Prec} = \frac{\mathrm{VP}}{\mathrm{VP+FP}} \qquad (2)$$

A sensibilidade ou revocação (recall) indica a proporção de casos positivos reais que foram identificadas pelo modelo de forma correta. Como na sensibilidade, também é calculada apenas para a classe positiva (+) conforme Equação 3.

$$\mathrm{Sens} = \frac{\mathrm{VP}}{\mathrm{VP+FN}} \qquad (3)$$

A especificidade indica a proporção de casos negativos reais que foram identificados pelo modelo de forma correta. Neste caso, a especificidade é calculada para a classe negativa, conforme Equação 4.

$$\mathrm{Esp} = \frac{\mathrm{VN}}{\mathrm{VN+FP}} \qquad (4)$$

A taxa de erro da classe positiva indica a proporção de erros que o modelo classificou de forma incorreta da classe positiva, conforme Equação 5:

$$err_{+} = \frac{\mathrm{FN}}{\mathrm{VP+FN}} \qquad (5)$$

A taxa de erro da classe negativa indica a proporção de erros que o modelo classificou de forma incorreta da classe negativa, de acordo com a Equação 6:

$$err_{-} = \frac{\mathrm{FP}}{\mathrm{FP+VN}} \qquad (6)$$

A Equação 7, indica a taxa de erro geral, ou seja, o quanto o modelo errou ao classificar os elementos das classes positivas e negativas:

$$\text{err} = \frac{\text{FP+FN}}{\text{VP+VN+FP+FN}} \qquad (7)$$

Já o F1-*Score,* demonstrado na Equação 8, é a média harmônica entre precisão e a sensibilidade (*recall)*. Útil quando há desequilíbrio nas classes.

$$\text{F1} = \frac{2*(\text{Prec}*\text{Sens})}{\text{Prec+Sens}} \qquad (8)$$

3 MÉTODO

Inicia-se o capítulo pela caracterização do local de investigação: o Instituto Federal de Educação Ciência e Tecnologia de São Paulo – campus Suzano, além do instrumental metodológico empregado.

O perfil da pesquisa é derivado de estudos realizados na linha de Políticas, Gestão e Avaliação, no âmbito do projeto 'Concepções e Políticas da Educação Profissional' da Unidade de Pós-Graduação, Extensão e Pesquisa do CEETEPS.

3.1 Caracterização do cenário de pesquisa

A pesquisa teve como foco analisar por meio de IA os dados dos alunos matriculados, evadidos, cancelados, trancados entre outros, do curso Técnico em Automação Industrial do campus Suzano do IFSP, com o objetivo de identificar a possibilidade de evasão de novas matrículas.

O IFSP é uma autarquia do governo federal vinculada ao MEC, compondo a Rede Federal de Educação Profissional, Científica e Tecnológica. Criada em 2008 pela Lei nº 11.892 é composta por 38 Institutos Federais, 02 Centros Federais de Educação Tecnológica, a Universidade Tecnológica Federal do Paraná, 22 escolas técnicas vinculadas às universidades federais e o Colégio Pedro II. Ao todo, são 661 unidades distribuídas em todo território nacional (MEC, 2023).

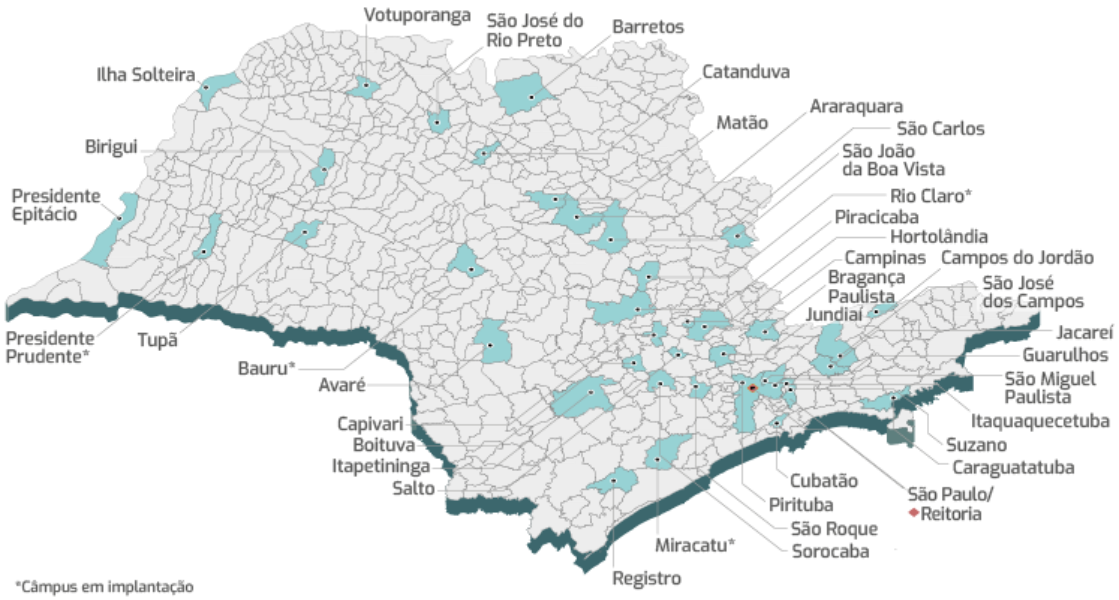
O IFSP oferece cursos nas modalidades técnico integrado, técnico concomitante/subsequente, cursos superiores de graduação em bacharelados, licenciaturas e de tecnologias, pós-graduação *lato* e *stricto sensu*, além de cursos de formação inicial e continuada. Representa hoje o maior órgão da Rede Federal. Segundo dados consolidados da Plataforma Nilo Peçanha, ano base 2022, o IFSP constava com 38 unidades, 763 cursos, 81.157 matrículas efetuadas, 64.038 novas vagas ofertadas e 15.020 concluintes(Brasil, 2023). Das

vagas ofertadas no IFSP, de acordo com a Lei 11.892, 50,0% são destinadas aos cursos técnicos e 20,0% para os cursos de licenciatura e para programas especiais de formação pedagógica (*stricto sensu* e *lato sensu*). A Figura 5 ilustra a distribuição dos *campi* do IFSP no estado.

O campus Suzano do IFSP oferece 10 cursos em 3 níveis de ensino, oferecendo 465 novas vagas todos os anos. Há 5 cursos de nível médio técnico, 2 na modalidade concomitante/subsequente (Automação Industrial e Administração), 1 Ensino de Jovens e Adultos (EJA) (Administração) e 2 cursos técnicos integrados ao ensino médio (Automação Industrial e Química). O nível superior oferta 1 Tecnólogo em Logística e Operações (em fase de fechamento), 1 Engenharia de Controle em Automação, 1 Bacharelado em Química Industrial, 1 Licenciatura em Química e 1 Bacharelado em Administração. Conta por fim, com 1 pós-graduação (*lato sensu*) em Logística.

Porém, a vasta gama de cursos e modalidades não reflete na formação de profissionais para o mercado de trabalho. Entre os anos de 2010 e 2019, segundo dados do sistema acadêmico do campus Suzano, foram efetuadas 4.393 novas matrículas em todos os cursos. Destas, 1.632 (37,2%) concluíram o curso, 421 (9,6%) permaneciam matriculados (em 2019) e 2.340 (53,3%) matrículas constavam como evadidos (em 2019). Os números encontrados no período são próximos aos indicadores da PNP no que se refere à eficiência acadêmica, nos ciclos 2017 – 2019, e que apontam um alto índice de evasão não só no campus Suzano, mas também em todo

Figura 5 - Mapa dos *campi* do IFSP distribuídas pelo Estado de São Paulo em 2023



Fonte: IFSP (2023)

o IFSP, conforme apontado nas Tabela 3 e Tabela 4.

A Tabela 3 apresenta os dados de eficiência acadêmica do campus Suzano do Instituto Federal de São Paulo (IFSP) nos anos de 2017, 2018 e 2019. A eficiência acadêmica é medida através de três indicadores: conclusão no ciclo, evasão no ciclo e retenção no ciclo. A "Conclusão no Ciclo" é um indicador que mostra a quantidade de alunos que concluíram o ciclo de estudos dentro do período estipulado. Por exemplo, em 2017, 786 alunos (53,6% do total) concluíram seus estudos dentro do ciclo previsto. Já o indicador "Evasão no Ciclo" se refere a quantidade de alunos que abandonaram ou interromperam seus estudos durante o ciclo. Em 2017, houve 675 casos de evasão (46,0% do total). E por fim a "Retenção no Ciclo" aponta a quantidade de alunos que permaneceram matriculados e não concluíram o ciclo no período estipulado. Em 2017, apenas 6 alunos (0,4% do total) se encaixavam nessa categoria.

Tabela 3 - Eficiência acadêmica do campus Suzano do IFSP

Ano	Conclusão no Ciclo	Evasão no Ciclo	Retenção no Ciclo
2017	786 (53,6%)	675 (46,0%)	6 (0,4%)
2018	476 (53,1%)	381 (42,6%)	38 (4,3%)
2019	346 (47,7%)	307 (42,3%)	73 (10,0%)

Fonte: PNP (2023)

A Tabela 4 apresenta os dados de eficiência acadêmica no Instituto Federal de São Paulo (IFSP) como um todo, nos mesmos anos de 2017, 2018 e 2019. Os indicadores são os mesmos da tabela anterior: conclusão no ciclo, evasão no ciclo e retenção no ciclo.

a) Conclusão no Ciclo: Em 2017, o IFSP teve 16.070 alunos (50,3% do total) que concluíram seus estudos dentro do ciclo previsto.

b) Evasão no Ciclo: Em 2017, houve 14.914 casos de evasão (46,7% do total) no IFSP.

c) Retenção no Ciclo: Em 2017, 958 alunos (3,0% do total) permaneceram matriculados, mas não concluíram o ciclo no período estipulado.

Tabela 4 - Eficiência acadêmica no Instituto Federal de São Paulo

Ano	Conclusão no Ciclo	Evasão no Ciclo	Retenção no Ciclo
2017	16.070 (50,3%)	14.914 (46,7%)	958 (3,0%)
2018	17.654 (49,4%)	16.972 (47,5%)	1.117 (3,1%)
2019	16.058 (49,2%)	15.565 (47,6%)	1.049 (3,2%)

Fonte: PNP (2023)

Essas tabelas mostram a eficiência acadêmica tanto do campus Suzano quanto do IFSP como um todo, nos anos mencionados. As tabelas são importantes para avaliar a taxa de conclusão dos cursos, a evasão de alunos e a retenção dos estudantes ao longo do tempo. Os

números apresentados nestas tabelas, indicam que o campus Suzano, não sofre de um fenômeno isolado, mas sim, que há um problema sistêmico em relação à evasão em todos o IFSP.

Consideraram-se alguns critérios para a pesquisa, relacionados aos dados dos alunos a serem analisados por meio de algoritmos de IA sendo:

a) Curso - O curso de Automação Industrial está presente no campus Suzano desde sua inauguração em 2010. Os demais cursos, sofreram alteração de nome, se encerraram ou iniciaram após 2010. O Quadro 3 mostra as alterações e início das atividades dos cursos. Portanto, o curso de Técnico Automação Industrial é o que possui mais matrículas, garantindo assim um maior conjunto de dados a serem analisados pelos algoritmos.

b) Período - A seleção do período buscou coletar a maior quantidade possível, de registros de alunos em conjunto com a seleção do curso a ser analisado. Os anos de 2020, 2021 e 2022 foram excluídos da análise e do cenário da pesquisa por se tratar do período mais letal da pandemia de COVID19 (2020 e 2021) onde as aulas foram ministradas de forma remota e o ano subsequente (2022) ser o primeiro ano de atividades presenciais após 2 anos de ensino remoto. Portanto, o período analisado na pesquisa corresponde aos anos de 2010 a 2019.

3.2 Método aplicado

Segundo Gil (2002, p. 41) é possível classificar as pesquisas em três grandes grupos: exploratórias, descritivas e explicativas. Nesse sentido, entende-se que a pesquisa é de cunho descritivo e exploratório, uma vez que a investigação busca analisar os dados de matrículas dos alunos do curso Técnico em Automação Industrial e a partir desses dados prever a possibilidade de evasão de novos alunos matriculados. O entendimento das relações entre as variáveis (características dos alunos) bem como o grau de influência que cada variável possui sobre a possibilidade de evasão do discente, está em acordo com a descrição de Gil (2002) sobre a pesquisa descritiva.

> As pesquisas descritivas têm como objetivo primordial a descrição das características de determinada população ou fenômeno ou, então, o estabelecimento de relações entre variáveis... Entre as pesquisas descritivas, salientam-se aquelas que têm por objetivo estudar as características de um grupo: sua distribuição por idade, sexo, procedência, nível de escolaridade, estado de saúde física e mental etc. (Gil, 2002, p.42).

Quadro 3: Ano de início e encerramento dos cursos no campus Suzano do IFSP

Curso	Modalidade	Início	Encerramento[2]
Automação Industrial	Técnico Concomitante/Subsequente	2010	-
Comércio	Técnico Concomitante/Subsequente	2010	2015
Eletroeletrônica	Técnico Concomitante/Subsequente	2011	2015
Administração	Técnico Concomitante/Subsequente	2016	-
Eletroeletrônica	Técnico Integrado ao Ensino Médio	2013	2013[3]
Automação Industrial	Técnico Integrado ao Ensino Médio	2015	-
Química	Técnico Integrado ao Ensino Médio	2015	-
Administração	Técnico Integrado ao Ensino Médio	2015	2018
Mecatrônica Industrial	Superior de Tecnologia	2015	2020
Logística	Superior de Tecnologia	2015	2022
Processos Químicos	Superior de Tecnologia	2015	2018
Eng. De Controle e Automação	Bacharelado	2021	-
Química Industrial	Bacharelado	2019	-
Química	Licenciatura	2015	-
Administração	PROEJA Integrado	2022	-
Logística e Operações	Pós-Graduação *Lato Sensu*	2014	-
Administração	Bacharelado	2023	-

Fonte: Autor (2024)

Quanto à metodologia, utilizando as classificações propostas por Gil (2002), classificou-se a pesquisa como *ex-post facto.* Conforme a descrição do autor, a pesquisa ex-*post facto* se assemelha à pesquisa experimental por permitir entender e verificar a existência de relação entre as variáveis do estudo. A diferença apontada é que neste caso, o pesquisador não possui controle sobre a variável independente.

De fato, o objeto de estudo, registro de alunos, é fato passado. O treinamento da IA se dá a partir do registro acadêmico dos alunos, que concluíram ou não os estudos. Isso permite ao algoritmo ajustar o modelo matemático para que haja previsão de intencionalidade futura de evasão.

O registro dos alunos foi expedido pela Diretoria Adjunta Educacional mediante solicitação e com autorização da Direção Geral do campus. Para a investigação, utilizaram-se os dados dos alunos do campus Suzano, do curso Técnico em Automação Industrial, no período entre o primeiro semestre de 2010 e o segundo semestre de 2019.

[2] Refere-se ao ano de entrada da última turma.

[3] Houve a entrada de apenas uma turma.

O conjunto de dados foi submetido a um 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.

Empregou-se no intuito de identificar a tendência de não conclusão dos alunos os algoritmos de AutoML, da ferramenta H2O e o uso de ferramentas estatísticas avançadas. Para tal, o conjunto de dados foi dividido em subconjuntos de treinamento, teste e validação.

3.3 Produtos da pesquisa

Para além da dissertação, é esperada uma contribuição institucional à Administração campus e posteriormente à reitoria do IFSP por meio de relatório técnico, indicando os resultados e possibilidades de utilização e implementação das ferramentas de aprendizado de máquina, junto ao sistema acadêmico.

Criou-se o *software* denominado **EvaDetect,** devidamente registrado e elaborado manual de procedimentos na utilização da ferramenta na previsão de evasão.

O sistema foi desenvolvido com dois módulos principais:

1) O primeiro módulo foi criado com o propósito de gerar o modelo de aprendizado de máquina, o qual permite parametrização de acordo com as especificidades do usuário, além de receber o conjunto de dados de treinamento.
2) O segundo módulo foi desenvolvido para predizer a possibilidade de evasão dos alunos, a partir do modelo gerado no primeiro módulo e os dados dos novos alunos. O Apêndice B, apresenta o fluxograma de funcionamento do sistema bem como os algoritmos dos módulos criados.

Para ambos os módulos, optou-se pela utilização do framework H2O-AutoML. Conforme apresentado na seção 2.5.1, o AutoML apresenta-se como uma opção viável para a criação de modelo de AM, economizando tempo e esforço de desenvolvimento, além de ser uma opção viável de utilização por parte de não especialistas. Ademais, o H2O-AutoML, proporciona a possibilidade de criação de dezenas de modelos e configurações diferentes, levando a criação de modelos apropriados a cada tipo de problema.

Contudo, faz-se necessários algumas considerações em relação ao uso da IA no ambiente escolar. Vincent-Lancrin e Vlies (2020) apontam que os avanços na área de IA, implicam em benefícios para a sociedade, especialmente na educação, mas destacam a importância das questões relacionadas à segurança e proteção de dados. A coleta massiva de

dados, especialmente de estudantes menores de idade, pode acarretar riscos para a privacidade. Além disso, destacam que, embora o uso de dados pessoais aprimore a eficácia dos sistemas de IA na educação, também cria riscos de privacidade, especialmente relacionados ao uso comercial desses dados.

Este cenário é discutido por Boyd e Crawford (2012) em um trabalho no qual levantam preocupações sobre viés, discriminação, privacidade e consentimento no contexto da coleta e utilização de grandes conjuntos de dados. As autoras alertam para os desafios éticos e sociais associados à big data, destacando a necessidade de uma abordagem que vá além das considerações técnicas.

Dessa forma, a implementação de sistemas de IA na educação, como o EvaDetect, requer não apenas avanços tecnológicos, mas também uma atenção cuidadosa às implicações éticas, sociais e de privacidade. É importante que políticas públicas e práticas adequadas sejam formuladas de forma a garantir segurança e proteção dos dados dos alunos, além de promover uma maior compreensão do impacto da IA no ambiente educacional.

3.4 Material

Para o desenvolvimento deste trabalho, fez-se necessário, para além do desenvolvimento do *software* acima citado, o levantamento de dados dos alunos. Como o intuito de validar o sistema desenvolvido, três outros conjuntos de dados, além do conjunto principal, foram utilizados nesta pesquisa. A seguir, serão descritos os conjuntos e os procedimento utilizados em cada um deles.

3.4.1 Conjunto de Dados 1 *(*CD1)

Utilizou-se o conjunto de dados 1 (CD1) disponível na UCI *Machine Learning Repository*. A UCI é um repositório de conjuntos de dados para AM. Foi criado pela Universidade da Califórnia, Irvine (UCI), e é uma das fontes mais conhecidas e utilizadas para conjuntos de dados em diversos campos da ciência de dados e aprendizado de máquina. Além disso, o repositório possui uma ampla variedade de conjuntos de dados, o que inclui diferentes domínios, como biologia, medicina, finanças, ciências sociais, engenharia e outros. Esses conjuntos de dados são frequentemente utilizados para pesquisa acadêmica, desenvolvimento e teste de algoritmos de aprendizado de máquina, bem como para competições e projetos práticos.

O conjunto de dados utilizado neste trabalho foi criado pelo Instituto Politécnico de

Portalegre e abrange alunos matriculados em diversos cursos de graduação, como agronomia, design, educação, enfermagem, jornalismo, gestão, serviço social e tecnologias. Contém 4424 registros com 35 atributos que incluem informações conhecidas no momento da matrícula do aluno como dados demográficos, socioeconômicos, macroeconômicos e dados de desempenho acadêmico dos alunos no final do primeiro e segundo semestres. Martins et al. (2021) utilizaram o mesmo conjunto de dados para prever o insucesso acadêmico dos estudantes portugueses. No trabalho, cada registro foi classificado como Sucesso, Sucesso Relativo e Fracasso.

a) Sucesso: Indica que o estudante obteve o diploma no tempo previsto, sem atrasos significativos;
b) Sucesso Relativo: Indica que o estudante obteve o diploma, mas levou até três anos a mais para concluí-lo. Apesar do atraso, é considerado sucesso;
c) Fracasso: Indica que o estudante não obteve o diploma no tempo previsto ou levou mais de três anos para concluir, ou ainda, não obteve o diploma.

O conjunto de dados foi disponibilizado no UCI após pré-processamento de dados, nos quais foram eliminadas anomalias, *outliers* e valores ausentes. De acordo com as informações disponibilizadas no *site* da UCI, a sugestão de utilização é de 80,0% do conjunto para treinamento e 20,0% para testes. Informações estas em acordo com os dados do artigo publicado por Martins et al. (2021).

O objetivo do EvaDetect é identificar os alunos com potencial de não conclusão ou fracasso, a partir da matrícula do aluno. Portanto, para o conjunto de dados da UCI, o atributo alvo “Target” que possui três possíveis *status (Dropout, Graduate e Enrolled)*, foi reduzido a apenas dois, 0 e 1, atribuindo zero para os alunos com *status Dropout* e 1 para os alunos com *status Graduate* e *Enrolled.* Além disso, esse conjunto foi preparado de duas formas, sendo denominados CD1-1 e CD1-2 para os experimentos com o EvaDetec.

a) Para o CD1-1 utilizou-se o conjunto, com todos os atributos, apenas com a redução do atributo alvo. O objetivo desse conjunto é validar o funcionamento geral do EvaDetect;
b) Para o CD1-2, foram removidos os atributos referentes ao desempenho acadêmico dos alunos. O objetivo desse conjunto é identificar os alunos com potencial de não conclusão, no ato da matrícula, onde não existem informações de desempenho acadêmico. De acordo com Realinho et al. (2022) o conjunto de dados está subdivido em 6 classes, sendo elas: Demográficos, sociodemográficos, macroeconômicos, acadêmicos no ato da matricula, acadêmicos ao final do 1º semestre e acadêmicos ao final do segundo semestre. As duas classes com dados acadêmicos, que se referem ao desempenho dos alunos no primeiro e no segundo semestre do curso, possuem 6

atributos cada, que foram então devidamente removidas, reduzindo o conjunto a 23 atributos.

A etapa seguinte foi normalizar os rótulos das colunas, removendo caracteres como apóstrofos, parênteses, traços, pontos, aspas, acentos, entre outros. Os conjuntos CD1-1 e CD1-2 foram então divididos conforme sugerido anteriormente, 80,0% para treinamento e 20,0% para testes. Para isso, gerou-se um número aleatório, entre 0 e 1, para cada registro. Posteriormente, organizou-se os registros por ordem crescente, tendo como base o número aleatório gerado e em seguida, realizou-se a divisão dos dados, separando os 3.540 primeiros registros (80,0%) para treinamento e os 884 (20,0%) registros restantes para teste. Realizada a divisão, foram excluídos dos conjuntos os números aleatórios utilizados para organizar e dividir os dados. Para o conjunto de testes, também foi excluída a coluna com o atributo alvo (*Target*), uma vez que o objetivo do conjunto de testes é apresentar dados desconhecidos ao sistema para que este possa realizar a previsão.

3.4.2 Conjunto de Dados 2 *(*CD2)

O segundo conjunto de dados utilizado, denominado CD2, refere-se aos dados dos alunos dos cursos superiores do campus Suzano, inclusos, licenciatura, bacharelado e de tecnologia. O objetivo com o CD2 foi a verificação do EvaDetect em realizar previsões de não conclusão, com alunos de cursos superiores do IFSP campus Suzano, na forma como realizada no CD1.

O CD2 original, possui 1744 registros acadêmicos dos alunos dos cursos superiores do campus Suzano apontados no Quadro 3 da seção 3.1. Após análise do conjunto, todos os atributos referentes a informações pessoais, que não teriam relação com a possibilidade de conclusão ou não do aluno, foram removidos, mantendo o CD2 com 17 atributos, conforme Quadro 4.

Dos 17 atributos selecionados, 3 deles foram escolhidos com o objetivo de transformá-los em dados não existentes. Os atributos Data de Matrícula e Data de Nascimento foram utilizados para calcular a idade do aluno no ato da matrícula. Após a realização do cálculo da idade, os atributos de data foram removidos, mantendo o novo atributo criado, denominado idade.

Quadro 4 - Atributos previamente selecionados dos alunos dos cursos superiores do campus Suzano do IFSP

Atributo	**Tipo**
Matrícula	Categórico Nominal
Ano de Conclusão do Ensino Anterior	Numérico
Cidade	Categórico Nominal
Data da Matrícula	Data
Data de Nascimento	Data
Descrição do Curso	Categórico Nominal
Endereço	Categórico Nominal
Estado Civil	Categórico Nominal
Etnia/Raça	Categórico Nominal
Forma de Ingresso	Categórico
Meio de Transporte	Categórico Nominal
Renda Bruta	Quantitativo
Renda *Per Capta*	Quantitativo
Sexo	Categórico Nominal
Tipo de Escola de Origem	Categórico Nominal
Zona Residencial	Categórico Nominal
Situação no Curso	Categórico Nominal

Fonte: Autor (2024)

O atributo endereço foi utilizado para calcular a distância da residência do aluno, até o IFSP. Com o auxílio das bibliotecas Geopy e Geopanda, para Python, foi possível obter as coordenadas (latitude e longitude) de cada um dos endereços dos alunos e posteriormente calcular a distância até o campus Suzano do IFSP. 298 registros não puderam ter a distância calculada. Outros 13 registros tiveram distâncias calculadas acima de 150km, por se tratar de cidades de outros estados ou do interior do estado de São Paulo. Esses 13 registros tiveram suas distâncias apagadas, por se tratar de informações que não são reais em caso de frequência dos estudantes. Um total de 311 de 1744 registros ficaram sem distância calculada.

Portanto, dos 17 atributos, 3 foram removidos e outros dois adicionados, resultando ao final 16 atributos. O atributo “Situação no Curso”, possuía 9 descrições. Quatro delas referentes a possibilidade de conclusão e 5 que indicavam não conclusão, conforme Quadro 5. Como o objetivo do trabalho é identificar os alunos que, por algum motivo, venham a não concluir o curso, as 9 descrições foram reduzidas a apenas 2: Não Conclusão (0) e Conclusão (1). Isso reduziu o problema a tarefas de classificação binária, permitindo que o algoritmo construído identifique os alunos em apenas duas categorias, contendo 848 alunos não concluintes e 896 concluintes ou matriculados.

Quadro 5 – Quantidade de alunos por situação no curso

Situação no Curso	Número de Aluno	Indica Conclusão ou Possibilidade de Conclusão?
Matriculado	647	Sim
Matrícula Vínculo Institucional	37	Sim
Transferido Interno	75	Sim
Formado	137	Sim
Cancelado	257	Não
Cancelamento Compulsório	21	Não
Evasão	496	Não
Trancado	26	Não
Trancado Voluntariamente	48	Não

Fonte: Autor (2024)

Uma vez que os dados referentes à situação dos alunos, foram transformados em um atributo numérico (0 e 1), optou-se, por conveniência e melhor treinamento dos modelos de AM, transformar todos os atributos em numéricos. O Apêndice A traz as tabelas com os atributos, dados e código de conversão. Como nenhum dos atributos apresenta ordem de acontecimento, optou-se por numerá-los de 0 até o limite de informações necessárias para cada um deles.

Por fim, o modelo foi dividido entre conjunto de treinamento e conjunto de testes. Não há na literatura uma definição padrão sobre como deve-se dividir os conjuntos. A divisão dos conjuntos deve ser realizada de acordo com a necessidade do modelo, de acordo com o tamanho do conjunto de dados e ajustada caso haja necessidade. Portanto, a divisão foi realizada separando-se 70,0% (1220 registros) do conjunto para treinamento do modelo e 30,0% (524 registros) para testes. No conjunto de testes, para efeitos de verificação, o atributo alvo foi retirado, para posterior comparação de desempenho.

3.4.3 Conjunto de Dados 3 *(*CD3)

O terceiro conjunto de dados utilizado, denominado CD3, refere-se aos dados dos alunos dos cursos técnicos, modalidade concomitante e/ou subsequente do campus Suzano, incluso também os alunos do curso Técnico em Automação Industrial. O objetivo com o CD3 foi a

verificação do EvaDetect em realizar previsões de não conclusão, com alunos de cursos técnicos, uma vez que o sistema já foi submetido a outros dois testes (CD1 e CD2). Sendo portando, os resultados obtidos no CD3, satisfatórios, avança-se então para o objeto final do estudo.

De forma similar ao realizado no CD2, após uma primeira análise, 20 atributos foram selecionados, sendo 3 deles descartados posteriormente: data de nascimento e data de matrícula, que foram utilizados para calcular a idade; e o endereço, que foi utilizado para calcular a distância da casa do aluno até a unidade do campus Suzano. Os mesmos procedimentos de conversão de dados aplicados no CD2 foram aplicados no CD3, uma vez que os dados pertencem ao mesmo sistema e os atributos dos registros são os mesmos.

O CD3 possui 2.731 registros de todos os alunos dos cursos técnicos, no período de 2010 a 2019. Destes 1.466 registros contam em situação de não conclusão e os demais 1.265 em situação de conclusão ou matriculados. Da mesma maneira como realizado no CD2, o CD3 foi dividido em conjunto de treinamento (70,0% - 1911 registros) e teste (30,0% - 820 registros), por meio de sorteio aleatório.

3.4.4 Conjunto de Dados 4 *(*CD4)

O quarto e último conjunto é o objetivo final do trabalho. Identificar a possibilidade de não conclusão no curso, por parte dos alunos do curso Técnico em Automação Industrial do campus Suzano. O CD4 é, portanto, um subconjunto do CD3 com todas as suas modificações. O CD3 possui 1222 registros, sendo que 712 alunos constam com situação de não conclusão e os demais 510 com situação de conclusão ou matriculados. O CD4 também foi dividido utilizando os mesmos critérios do CD2 e CD3: conjunto de treinamento com 855 (70%) e 367 registros no conjunto de teste (30%).

4 ANÁLISE E DISCUSSÕES

Foram realizados 5 experimentos com o sistema EvaDetect. Para cada experimento foram utilizados conjuntos de dados distintos (CD1-1, CD1-2, CD2, CD3 e CD4), os quais foram detalhados na seção anterior.

4.1 Experimento 1

Para o experimento 1, foi utilizado o CD1-1. O objetivo do experimento foi validar o funcionamento do EvaDetect, uma vez que o conjunto de dados já havia sido utilizado no trabalho de Martins et al. (2021).

O EvaDetect foi configurado da seguinte forma:

Tempo Máximo De Execução: 6000 (segundos)

Máximo De Modelos A Serem Gerados: 10

Tipo De Divisão Do Conjunto De Treinamento: 1

Pontos De Validação: -1

Tempo Máximo Por Modelo: 600

Métrica De Validação: Auc

Semente De Aleatoriedade: 1234

Número De Threads: 20

Máximo De Memória: 8

Balanceamento De Classes: False

Após o treinamento do modelo, com o conjunto de treinamento, previamente separado, gerou-se o modelo de AM. O melhor modelo encontrado pelo sistema foi o StackedEnsemble_BestOfFamily, utilizando o AUC como métrica de comparação de desempenho dos modelos. A Tabela 5 apresenta o desempenho do conjunto de treinamento nas etapas de construção do modelo, o que inclui validação cruzada e teste. O Quadro 19 do Apêndice C apresenta o relatório parcial, com as informações mais relevantes, da saída do AutoML após a geração do modelo.

Tabela 5 - Desempenho do CD1 nas etapas de Treinamento, Validação Cruzada e Teste

	CONJUNTOS		
Métricas	**Treinamento**	**Validação Cruzada**	**Teste**
AUC	0.9685	0.9236	0.9203
AUCPR	0.9823	0.9434	0.9567

Fonte: Autor (2024)

Destaca-se aqui que os conjuntos utilizados nas etapas apresentados na Tabela 5, são derivados apenas do conjunto de treinamento e foram subdivididos pelo próprio H2O. O conjunto de testes, contendo 20,0% dos dados originais, foi previamente separado para um teste final e comparação dos resultados. Assim, utilizando-se da funcionalidade principal do EvaDetec, que é a previsão de evasão, o conjunto de teste separado anteriormente, foi

submetido à análise de previsão de evasão, por meio do modelo gerado (StackedEnsemble_BestOfFamily).

Ao colocarmos os dados previstos lado a lado com o conjunto original separado como conjunto de testes (20,0%), sem a remoção do atributo alvo, identificou-se uma queda de desempenho do modelo em relação à classe positiva. Ainda assim, o modelo de forma geral, obteve uma acurácia de 88,7%, precisão de 87,2% e um erro geral de 12,5%. A Tabela 6 apresenta a matriz de confusão gerada a partir dos dados do conjunto de teste, quando submetidos à previsão.

Tabela 6 - Matriz de confusão gerada a partir das previsões do conjunto de teste

	0	**1**	**Erro**	**Taxa**
0	191	83	0.3029	83/274
1	28	582	0.0459	68/610
		TOTAL	0.1255	111/884

Fonte: Autor (2024)

A partir da matriz de confusão apresentada na Tabela 6, calculamos as métricas de desempenho: acurácia, precisão, *recall*, sensibilidade e medida F1, para avaliar o desempenho do modelo no conjunto de teste, conforme apresentados na Tabela 7. Essas métricas oferecem informações sobre a eficácia do modelo em diferentes aspectos, como a capacidade de prever corretamente, evitar falsos positivos ou capturar todos os casos positivos. A comparação dessas métricas com os dados gerados durante o treinamento do modelo nos fornece uma visão da capacidade geral do modelo de generalização com novos dados.

Tabela 7 - Comparação de desempenho entre os conjuntos de treinamento e teste

Métricas	**Conjunto de Teste**	**Conjunto de Treinamento**
Acurácia	0.8874	0.925
Precisão	0.8721	0.936
Recall	0.6971	0.953
Especificidade	0.9541	0.870
F1-Score	0.7748	0.886

Fonte: Autor (2024)

A partir destas medidas, é possível avaliarmos os resultados obtidos com os resultados de Martins et al. (2021). Utilizando o mesmo conjunto de dados e com mesma proporção de divisão, a melhor performance obtida no trabalho de Martins et al. (2021), foi com o modelo *Extreme Gradient Boosting*, com acurácia de 73,0% e medida F1-Média de 65,0%. Ao comparamos essas medidas com as medidas apresentadas pelo modelo gerado pelo H2O

percebe-se que o modelo do EvaDetec apresenta uma acurácia mais alta (88,7%) em comparação com o modelo do artigo (73,0%). Isso indica que o modelo gerado teve um desempenho superior em termos de previsões globais corretas. Além disso, ao avaliarmos os resultados do EvaDetec com o estudo de Chung e Lee (2019), que empregou métodos de aprendizado de máquina para prever o abandono escolar, notamos uma convergência nos resultados. Chung e Lee (2019) obtiveram uma notável acurácia de 95,0%, destacando a eficácia dessas abordagens na antecipação de situações de evasão. Sales, Balby e Cajueiro (2016) obtiveram resultados muito próximos ao encontrado neste estudo, com precisão entre 82,0% e 89,0%, enquanto o EvaDetec obteve precisão de 87,2% com conjunto de dados com características semelhantes.

Quando se compara a métrica F1, o modelo do EvaDetec obteve 77,4%, enquanto o modelo do artigo atingiu uma média de F1-score de 65,0%. Essa medida demonstra um equilíbrio entre precisão e recall, resultando em uma métrica F1 mais alta.

Além disso a especificidade (95,4%) é alta, indicando uma boa capacidade de identificar corretamente casos negativos.

Importante destacar que o objetivo do Experimento 1 foi verificar a capacidade do EvaDetect, por meio de H2O, de gerar um modelo de AM capaz de prever a não conclusão de curso. Ao comparar o modelo gerado com um modelo de trabalho publicado, confirmamos essa possibilidade.

4.2 Experimento 2

Para o experimento 2, foi utilizado o CD1-2. O objetivo do experimento foi verificar a capacidade do EvaDetec em prever a não conclusão de alunos, mesmo sem a utilização de dados acadêmicos. O Experimento 1 apontou a capacidade do sistema em gerar modelos de AM para classificação binária. Contudo, o conjunto de dados possuía dados acadêmicos dos alunos, referentes aos primeiro e segundo semestres do curso, o que é um indicativo da viabilidade da capacidade de predição. Portando, o experimento 2 visa investigar, com o mesmo conjunto de dados, dessa vez sem os dados acadêmicos, a capacidade do modelo de predizer o abandono.

O sistema foi configurado com os mesmos parâmetros do Experimento 1. O melhor modelo gerado pelo EvaDetec foi o StackedEnsemble_AllModels. A diferença do StackedEnsemble_AllModels e o StackedEnsemble_BestOfFamily (gerado no Experimento 1), diz respeito à forma com que os modelos Ensemble selecionam os algoritmos. O AllModels utiliza todas as previsões dos modelos individuais gerados pelo AutoML, onde cada modelo

básico contribui com suas previsões para o *ensemble*, independentemente de seu desempenho individual. Já o BestOfFamily, seleciona apenas o melhor modelo de cada tipo de algoritmo. O Quadro 20 do Apêndice C apresenta o relatório parcial, com as informações mais relevantes, da saída do AutoML após a geração do modelo.

Após a criação do modelo, o conjunto de teste foi submetido à análise. A matriz de confusão apresentada na Tabela 8 indica o desempenho do modelo aplicado ao conjunto de testes de 20% separado anteriormente.

Observando a Tabela 8 é possível encontrar as informações que nos permitirão avaliar o modelo, portanto, a Tabela 9 apresenta as métricas de desempenho do modelo, calculadas a partir da matriz de confusão. A principal observação refere-se à acurácia, atingindo 80,3% para a classe positiva. Esse valor é inferior ao modelo do Experimento 1, mas ainda assim, superior aos números encontrados no trabalho de Martins et al. (2021). Contudo, destaca-se a baixa sensibilidade do modelo, o que aponta que o modelo não está identificando bem as instâncias positivas reais, dados corroborados quando olhamos o erro da classe positiva.

Tabela 8 - Matriz de confusão gerada a partir das previsões do conjunto de teste sem Dados Acadêmicos

	0	1
0	110	147
1	27	600

Fonte: Autor (2024)

Contudo, apesar dos pontos negativos, o erro da classe negativa e a especificidade do modelo estão muito bem ajustados. De fato, é possível melhorar o desempenho do modelo por meio do ajuste de t*hreshold*. O *threshold* é utilizado em problemas de classificação, onde o intuito é classificar os objetos analisado entre as classes positiva e negativa. Um modelo de classificação, após realizar a previsão de uma determinada instância ou objeto, retorna uma probabilidade de pertencimento a umas das classes. O *threshold* portanto, é uma probabilidade que divide essas previsões. Se a probabilidade estimada for maior que o *threshold*, a instância é classificada como positiva, caso contrário, é classificada como negativa.

Uma vez que o Experimento 2, teve como objetivo verificar a capacidade do sistema em gerar modelos de AM capazes de prever a evasão, mesmo sem dados acadêmicos, entende-se que o modelo teve sucesso.

Tabela 9 - Métricas de Desempenho - Conjunto de Teste

Métricas	**Desempenho**

Acurácia	0.8032
Precisão	0.8029
Sensibilidade	0.4280
Especificidade	0.9569
Err+	0.5720
Err-	0.0431
err	0.1968

Fonte: Autor (2024)

4.3 Experimento 3

O terceiro experimento deste projeto, centrou-se nos dados dos alunos dos cursos superiores IFSP Suzano, CD2. Como demonstrado nos experimentos 1 e 2, o EvaDetect foi capaz de predizer com boa acurácia geral, a conclusão ou não dos alunos de cursos superiores, com e sem dados acadêmicos. Portanto, o objetivo deste experimento é provar que o sistema desenvolvido é capaz de predizer o insucesso acadêmico, desta vez focado em alunos dos cursos superiores da instituição foco desta pesquisa.

Para este experimento, o EvaDetec foi configurado da seguinte forma:

Tempo máximo de execução: 120000 (segundos)

Máximo de modelos a serem gerados: 20

Tipo de divisão do conjunto de treinamento: 2

Pontos de validação: 10

Tempo máximo por modelo: 600

Métrica de validação: Auc

Semente de aleatoriedade: 1234

Número de threads: 20

Máximo de memória: 8

Balanceamento de classes: false

Não se aplica. Balanceamento de classes desabilitado.

Stopping rounds: desabilitado

A mudança de configuração em relação aos experimentos anteriores, teve como objetivo buscar outros modelos, para além da família *Ensemble.* Modelos *Ensemble* não entregam resultados considerando a importância das variáveis e estes dados possivelmente são de interesse de análise. Como resultado, o EvaDetec entregou um GBM, ou seja, um modelo de *Gradient Boosting Machine* (GBM). O GBM é uma técnica de aprendizado de máquina que

constrói uma sequência de árvores de decisão, onde cada árvore corrige os erros das anteriores. Cada árvore é ajustada para prever os resíduos dos modelos anteriores, resultando em um modelo mais poderoso (Ayyadevara, 2018).

O modelo encontrado pelo sistema, apresentou boas métricas de desempenho conforme Tabela 10. O índice AUC de 79,9%, associadas ao RMSE (do inglês *Root Mean Squared Error*) de 21,6% e ao MSE (do inglês *Mean Squared Error*) de 19,5%, sugere que modelo tem um desempenho geral sólido, fornecendo previsões precisas para a previsão de não conclusão do curso. O relatório de saída do modelo com informações complementares consta no Quadro 21 Apêndice C.

O MSE é uma medida que avalia a qualidade das previsões feitas por um modelo em problemas de regressão. Ele fornece uma medida do quão próximas as previsões do modelo estão dos valores reais. O MSE é calculado pela média dos quadrados das diferenças entre as previsões e os valores reais(Faceli et al., 2011). Um valor de MSE igual a 0, indica um modelo perfeito. Valores baixos, como o apresentado pelo modelo em questão, indicam um bom ajuste do modelo aos dados.

Tabela 10 - Desempenho do melhor modelo após treinamento para o CD2

Métricas de Desempenho	GBM_grid
AUC	0.799306
LogLoss	0.589974
AUCPR	0.722834
MEAN_PER_CLASS_ERROR	0.216021
RMSE	0.442009
MSE	0.195372

Fonte: Autor (2024)

Já o RMSE é uma medida popular para avaliar a precisão de um modelo de regressão. É uma versão modificada do MSE, onde a raiz quadrada é aplicada à média dos quadrados das diferenças entre as previsões do modelo e os valores reais (Chai; Draxler, 2014). Da mesma forma como no MSE, quanto menor o valor do RMSE, melhor. Isso indica que as previsões do modelo estão, em média, mais próximas dos valores reais.

Como nos demais experimentos, o modelo gerado foi aplicado ao conjunto de testes, previamente separado, conforme descrito na seção 3.4.2. Para a predição dos valores, foi ajustado o *threshold* em 0.53564, valor estimado para melhor acurácia durante a fase de validação do modelo (vide Quadro 21 Apêndice C). O desempenho do modelo com o conjunto de testes é apresentado na Tabela 11 e Tabela 12.

Os resultados deste experimento são semelhantes aos obtidos por Fernández-García et al. (2021), que também obtiveram uma acurácia, precisão e sensibilidade de 68,2%, 62,1% e 72,3% respectivamente, utilizando o GBM em dados de alunos de cursos superiores, no ato da matrícula, ou seja, sem dados de desempenho acadêmico. Já Bitencourt, Silva e Xavier (2021) obtiveram acurácia superior (75,9%) na identificação de alunos com potencial de evasão no ato da matrícula.

Tabela 11 - Matriz de Confusão. Conjunto de teste do CD2

	0	**1**	**Erro**	**Taxa**
0	169	68	0.2869	68/237
1	92	195	0.3206	92/287
Total	261	263	0.3053	160/524

Fonte: Autor (2024)

Tabela 12 - Desempenho do GBM com o conjunto de teste para o CD2

Métricas	**GBM**
Acurácia	0.6947
Precisão	0.6475
Sensibilidade	0.7131
Especificidade	0.6794
Err+	0.2869
Err-	0.3206
err	0.3053

Fonte: Autor (2024)

Este experimento focou nos dados dos alunos dos cursos superiores IFSP Suzano, CD2, utilizando o EvaDetect para prever o sucesso ou insucesso acadêmico. Os experimentos anteriores demonstraram a capacidade do EvaDetect em predizer com boa acurácia a conclusão ou não dos alunos, considerando dados acadêmicos.

A comparação com resultados semelhantes da literatura, como os obtidos por Fernández-García et al. (2021) e Bitencourt, Silva e Xavier (2021), valida a eficácia do modelo, indicando que o EvaDetect, ao utilizar o GBM, apresenta resultados consistentes e comparáveis.

4.4 Experimento 4

Os experimentos 1, 2 e 3 demonstraram que o sistema desenvolvido consegue executar sua função principal, ou seja, a previsão de evasão, de forma satisfatória, com alunos de cursos superiores. Assim, o quarto experimento deste projeto, buscou verificar os modelos de AM

gerados no EvaDetect, para alunos de cursos técnicos. Como no experimento 3, configurou-se o sistema para que pudesse encontrar algoritmos diferentes dos experimentos 1 e 2, e por isso, manteve-se a configuração do experimento 3, passando como conjunto de treinamento o CD3 descrito na seção 3.4.3. O Quadro 22 do Apêndice C apresenta o relatório parcial de saída do AutoML com as informações mais relevantes.

O modelo encontrado pelo sistema foi um XRT (do inglês *Extremely Randomized Trees*), uma técnica proposta por Geurts, Ernst e Wehenkel (2006) para construção de modelos de árvores de decisão. Essa abordagem introduz uma camada de aleatoriedade no processo de tomada de decisão das árvores durante a construção de florestas aleatórias. Ao construir cada árvore, são escolhidos subconjuntos aleatórios de atributos para determinar os pontos de divisão, e a seleção dos pontos também é realizada de maneira aleatória, utilizando o conjunto completo de atributos. A dupla abordagem de aleatoriedade tem o benefício de reduzir o *overfitting*, tornando os modelos mais robustos e menos propensos a memorizar os dados de treinamento.

O modelo gerado apresentou um índice AUC médio, durante as etapas de treinamento, validação e teste de 77,5% com um erro médio (MSE) nas três etapas de 19,5%, valores semelhantes aos resultados do experimento 3 com o CD2 (média de 76,92 % e 20,6% respectivamente).

O XRT foi aplicado ao conjunto de teste, que foi previamente separado, seguindo a mesma abordagem adotada nos experimentos anteriores. Nesta etapa, de maneira similar ao Experimento 3, a escolha do *threshold* desempenhou um papel crucial na avaliação do modelo. O *threshold* selecionado foi 0,478838, valor este identificado na saída do AutoML como aquele que resultou na máxima acurácia durante a validação do modelo (vide Quadro 22 Apêndice C).

Vale ressaltar que durante a análise do modelo, outros *thresholds* foram considerados, incluindo o valor indicado na performance do modelo (0,305605) e o valor associado à máxima acurácia durante o treinamento (0,444444). Entende-se que o *threshold* atua como um ponto de equilíbrio sensível entre as classes, influenciando diretamente na precisão do modelo. Ao ajustar esse limiar, é natural observar alterações nos erros de uma classe em detrimento da outra: aumentar ou diminuir o valor do *threshold* impacta diretamente na diminuição dos erros de uma classe e no aumento dos erros da outra, e vice-versa. O desempenho do modelo no conjunto de testes está apresentado nas Tabela 13 e Tabela 14.

Tabela 13 - Matriz de Confusão. Conjunto de teste do CD3

	0	**1**	**Erro**	**Taxa**

0	324	110	0.2534	110/434
1	111	275	0.2875	111/386
Total	435	385	0.2695	221/820

Fonte: Autor (2024)

Pode-se observar por meio dos dados apresentados que o sistema se saiu bem em prever a conclusão ou não do curso, atingindo acurácia de 73,1%. Apesar do desempenho na etapa de treinamento ser muito parecido com o do experimento 3, na fase de testes, o experimento 4 teve desempenho ligeiramente superior, com todas as métricas de desempenho acima de 70,0%. O erro de 23,35% da classe positiva também é destaque, uma vez que, essa classe contém os alunos evadidos, indicando sucesso ao predizer o insucesso acadêmico. O desempenho ligeiramente superior do experimento 4 em relação ao experimento 3 pode ter relação com o *threshold* mais bem ajustado no experimento 4. É importante ressaltar que o EvaDetect, ao fazer uso do AutoML H2O para otimizar os processos de criação de modelos de AM para predição da evasão escolar, não elimina a necessidade de ajustes finos no sistema, monitoramento dos dados e conhecimento real do problema. O trabalho do analista de dados se torna facilitado por uma ferramenta que entrega um modelo de AM com as capacidades necessárias para prever a evasão. Contudo, faz-se necessário um olhar cuidadoso do especialista, que poderá por meio de análise detalhada da saída dos resultados do AutoML, ajustar o sistema de acordo com suas necessidades ou eventualmente, treinar novos modelos para uma maior assertividade.

Tabela 14 - Desempenho do XRT com o conjunto de teste para o CD3

Métricas	**XRT**
Acurácia	0.7305
Precisão	0.7448
Sensibilidade	0.7465
Especificidade	0.7124
Err+	0.2335
Err-	0.2876
err	0.2695

Fonte: Autor (2024)

Portanto, os resultados até agora apresentados, indicam a capacidade do sistema em adaptar-se a diferentes contextos educacionais, apontando sua promissora aplicabilidade na identificação e prevenção de evasões, tanto em cursos superiores quanto técnicos.

4.5 Experimento 5

Seguindo o sucesso dos experimentos anteriores, nos quais o sistema demonstrou sua eficácia na previsão de evasão em cursos superiores e técnicos, o quinto e último experimento focou em um cenário mais específico. Este último estágio da pesquisa concentra-se em avaliar a performance do EvaDetect em um contexto único, a predição de evasão no curso técnico de Automação Industrial.

Neste experimento, utilizou-se o CD4 descrito na seção 3.4.4 e manteve-se as configurações dos experimentos 3 e 4. Após o processo de treinamento do modelo, o sistema entregou como melhor algoritmo de AM um GBM, mesmo modelo do experimento 3. Reforça-se que o GBM do experimento 3 difere do modelo gerado neste experimento e, portanto, apesar de serem o mesmo algoritmo, suas configurações diferem devido ao processo de treinamento e o conjunto de dados para o qual foi modelado. Enquanto o GBM do experimento 3 possui uma profundidade média das árvores de 12,69 e uma média do número de folhas de 71,19, o GBM do experimento 5 possui médias de 6 e 23,23 respectivamente. Os Quadros 23 e 24 do Apêndice C apresentam o relatório parcial de saída do AutoML com as informações mais relevantes.

Como nos experimentos anteriores, o processo incluiu a aplicação do *threshold* ao conjunto de teste. O melhor desempenho do modelo na fase final, com o conjunto de teste, foi encontrado com a configuração do *threshold* em 0,131166. Esse valor é o indicado no relatório de saída do AutoML, durante a fase de validação cruzada, como o melhor valor para o máximo de medida F1 (vide Quadro 23 Apêndice C). Lembrando que a medida F1 é uma medida que visa equilibrar precisão e sensibilidade. As Tabela 15 e 16 apresentam o desempenho geral do modelo por meio da matriz de confusão e métricas de desempenho.

Tabela 15 - Matriz de Confusão. Conjunto de teste do CD4

	0	**1**	**Erro**	**Taxa**
0	141	71	0.3349	71/212
1	35	120	0.2258	35/155
Total	176	191	0.2888	106/367

Fonte: Autor (2024)

O GBM obteve uma acurácia de 71,1% que demonstra a proporção geral de previsões corretas em relação ao total de instâncias observadas no conjunto de teste. A precisão para a classe positiva (indicando alunos que evadiram) alcançou 80,1%, destacando a capacidade do modelo em identificar corretamente aqueles que não concluíram o curso.

Tabela 16 - Desempenho do GBM com o conjunto de teste para o CD4

Métricas	GBM
Acurácia	0.7112
Precisão	0.8011
Sensibilidade	0.6651
Especificidade	0.7742
Err+	0.3349
Err-	0.2258
err	0.2888

Fonte: Autor (2024)

A especificidade, indicando a capacidade do modelo em identificar verdadeiros negativos, apresentou um valor de 77,4%, indicando que o modelo foi eficiente em reconhecer também, alunos que não evadiram. Em relação aos erros, a taxa de erro positivo (err+), representando a proporção de casos preditos erroneamente como conclusão, foi de 33,4%. A taxa de erro negativo (err-), que representa a proporção de casos preditos erroneamente como evasão, foi de 22,5%. O erro global do modelo (err) foi de 28,8%, demonstrando a eficácia geral do EvaDetect na predição de evasão dos alunos do curso Técnico em Automação Industrial.

Os resultados obtidos no experimento anterior, destinado aos cursos técnicos como um todo, fornecem uma importante base de comparação. Ao comparar os resultados dos dois modelos, XRT do Experimento 4 e GBM do Experimento 5, é possível observar algumas diferenças significativas em suas métricas de desempenho na predição de evasão para o curso Técnico em Automação Industrial, conforme Tabela 17.

Em termos de acurácia, o modelo GBM alcançou uma taxa de 71,1%, enquanto o XRT obteve uma acurácia um pouco superior, atingindo 73,1%. Isso indica que, globalmente, o modelo XRT apresentou uma ligeira vantagem em termos de previsões corretas em comparação ao modelo GBM.

Quanto à precisão, o modelo GBM apresentou um desempenho superior, atingindo 80,1%, enquanto o XRT registrou uma precisão de 74,5%. Isso mostra que o modelo GBM foi mais eficaz em identificar corretamente os casos de alunos que evadiram.

Em relação à sensibilidade, o XRT alcançou 74,6%, superando a sensibilidade do GBM, que foi de 66,5%. Isso indica que o modelo XRT foi mais eficiente em capturar casos reais de evasão.

Tabela 17 - Comparação de desempenho dos modelos destinados aos cursos técnicos.

Métricas	XRT (Exp. 4)	GBM (Exp. 5)
Acurácia	0.7305	0.7112
Precisão	0.7448	0.8011
Sensibilidade	0.7465	0.6651
Especificidade	0.7124	0.7742
Err+	0.2335	0.3349
Err-	0.2876	0.2258
err	0.2695	0.2888

Fonte: Autor (2024)

Na especificidade, ambos os modelos apresentaram desempenho acimas dos 70,0%, mas o GBM alcançou um valor um pouco superior (77,4%) em comparação ao XRT (71,2%). Isso sugere que o GBM foi mais eficiente em reconhecer alunos que não evadiram.

Analisando as taxas de erro, o modelo XRT teve uma taxa de erro global (err) de 28,8%, enquanto o GBM registrou uma taxa de erro de 26,9%. Portanto, o modelo GBM apresentou um desempenho geral ligeiramente superior, indicando menor proporção de previsões incorretas em relação ao total de instâncias observadas.

A análise da importância das variáveis nos modelos XRT (Tabela 18 - Experimento 4) e GBM (Tabela 19 - Experimento 5) revela sutilezas distintas na predição de evasão para alunos de cursos técnicos. No contexto geral, abrangendo todos os cursos técnicos, incluindo automação, o XRT destaca a distância da residência do aluno para o Instituto Federal como o fator mais significativo, seguido por características como ano de conclusão do ensino anterior, forma de ingresso e outros.

Já no cenário específico do curso Técnico em Automação Industrial, o GBM mantém a distância como variável mais influente, mas destaca variáveis específicas do curso, como renda *per capta* e etnia/raça. Essa diferenciação sugere que, ao focalizar em um curso específico, certas características ganham destaque na predição de evasão. Essa abordagem, validada pelos resultados dos modelos, está alinhada com a perspectiva de Bakhshinategh et al. (2018), que destacam a necessidade de análise de dados para identificar fatores que impactam o sucesso acadêmico, permitindo intervenções personalizadas.

Em ambos os modelos, a variável renda bruta aparece, sendo que no segundo modelo aparece a variável renda *per capta*. Essas variáveis podem estar relacionadas às questões socioeconômicas dos estudantes e vão ao encontro do trabalho de Dore, Araújo e Mendes (2014), onde os autores apontam que o abandono por questões de trabalho e emprego, sendo estas ligadas diretamente às questões socioeconômicas, são a principal causa de evasão nos

cursos técnicos no estado de Minas Gerais.

No mesmo estudo, o transporte e mudança de endereço aparecem entre os doze motivos mais relevantes para a evasão. A variável transporte aparece no primeiro modelo, contudo, se considerarmos que a distância apontada nos modelos aqui apresentados, possui ligação direta com o tipo de deslocamento que o aluno deverá ter até o campus do IF ou a município de residência, pode-se inferir que os achados nos modelos estão em concordância com a literatura.

Tabela 18 - Importância das Variáveis no modelo XRT - Experimento 4

Variável	Importância Relativa	Escala de Importância	Porcentagem
Distância do IF	5779.83	1	0.613505
Ano de Conclusão do Ens. Anterior	805.616	0.139384	0.0855128
Forma de Ingresso	518.344	0.0896816	0.0855128
Nível de Ensino Anterior	370.719	0.0641401	0.0855128
Zona Residencial	369.727	0.0639685	0.0855128
Meio de Transporte	333.236	0.057655	0.0353716
Renda Bruta Familiar	322.858	0.0558595	0.0342701
Idade	228.706	0.0395697	0.0242762

Fonte: Autor (2024)

Tabela 19 - Importância das Variáveis no modelo GBM - Experimento 5

Variável	Importância Relativa	Escala de Importância	Porcentagem
Distância do IF	437.623	1	0.72996
Zona Residencial	43.3002	0.098944	0.0722252
Ano de Conclusão do Ens. Anterior	30.8565	0.0705092	0.0514689
Renda Per Capita	25.0189	0.05717	0.0417318
Nível de Ensino Anterior	20.4026	0.0466213	0.0340317
Idade	14.7128	0.0336198	0.0245411
Renda Bruta Familiar	12.4662	0.0284862	0.0207938
Etnia/Raça	4.10898	0.00938931	0.00685382

Fonte: Autor (2024)

E por fim, entende-se que embora ambos os modelos tenham apresentado resultados sólidos, o GBM se destacou em precisão, enquanto o XRT teve uma sensibilidade ligeiramente superior. A escolha entre os modelos dependerá das prioridades específicas do problema em questão, considerando a importância relativa de identificar corretamente casos de evasão versus evitar falsos positivos ou falsos negativos. No entanto, ao focar em um curso técnico específico, busca-se avaliar se as adaptações e ajustes feitos no sistema resultam em melhorias

significativas na precisão das previsões.

4.6 Conclusão e Síntese dos Experimentos

Ao avaliar o desempenho do EvaDetec ao longo dos experimentos, é possível observar uma consistência nos resultados apresentados, conforme Tabela 20. O Experimento 1 destaca-se em seu desempenho superior, contudo, o conjunto de dados utilizado possui para além de informações sociodemográficas dos alunos, dados de desempenho acadêmico ao longo de três semestres de curso. Ao desconsiderarmos os dados de desempenho acadêmico, no experimento 2, observa-se uma queda de rendimento em relação ao experimento 1, mas ainda assim, superior aos demais experimentos. As diferenças justificam-se por se tratar de conjuntos de dados diferentes e, portanto, os resultados serão diferentes. Outro fato a se considerar, são o número de atributos do CD1-2 em relação aos CD2, CD3 e CD4. Enquanto o primeiro possui 23 atributos os demais possuem 16, 18 e 18 atributos respectivamente. De fato, o pior desempenho ficou por conta do Experimento 3 realizado com o CD2, conjunto este com a menor quantidade de atributos. A diferença no número de atributos, portanto, indica uma relação de precisão dos modelos, uma vez que, quanto mais características (atributos) o sistema possuir, melhor serão os modelos treinados.

O Quadro 6 apresenta um resumo dos experimentos realizados neste estudo, mais precisamente no âmbito do EvaDetect, fornecendo uma visão do principal foco, do melhor modelo identificado e uma breve descrição de cada experimento. No Experimento 1, voltado para cursos superiores, o foco foi explorar a previsão de evasões, utilizando um conjunto de dados validado na literatura, e o modelo StackedEnsemble_BestOfFamily destacou-se ao combinar várias técnicas. No Experimento 2, também direcionado para cursos superiores, como o mesmo conjunto de dados, mas excluindo dados de desempenho acadêmico, o modelo StackedEnsemble_AllModels também demonstrou eficácia ao empregar a mesma abordagem de *ensemble* do Experimento 1.

O Experimento 3 concentrou-se nos cursos superiores do IFSP, campus Suzano. O modelo gerado pelo EvaDetect foi o GBM, que constrói sequências de árvores de decisão e proporcionou previsões sólidas para evasões nessa modalidade de ensino. Já o Experimento 4 investigou a evasão em cursos técnicos do IFSP, campus Suzano, e o modelo XRT destacou-se ao trazer uma camada de aleatoriedade para construir modelos robustos, com ênfase na identificação de alunos propensos a evadir. Por fim, no Experimento 5, explorou-se a predição de evasões no curso Técnico em Automação Industrial no IFSP, campus Suzano, e mais uma

vez, o GBM demonstrou sua eficácia na identificação precisa de evasões.

Como produto deste trabalho, projetou-se o EvaDetect, um *software* de IA, parametrizável, capaz de entregar modelos de AM com capacidade de prever evasões acadêmicas. Detalhes adicionais sobre o *software*, incluindo sua arquitetura, funcionalidades e processos, podem ser encontrados no Apêndice B.

Tabela 20 - Comparação de desempenho dos modelos com diferentes conjuntos de dados

	Experimento 1	**Experimento 2**	**Experimento 3**	**Experimento 4**	**Experimento 5**
Acurácia	0.8874	0.8032	0.6947	0.7305	0.7112
Precisão+	0.8721	0.8029	0.6475	0.7448	0.8011
Sensibilidade+	0.6971	0.4280	0.7131	0.7465	0.6651
Especificidade	0.9541	0.9569	0.6794	0.7124	0.7742
err+	0.3029	0.5720	0.2869	0.2335	0.3349
err-	0.0459	0.0431	0.3206	0.2876	0.2258
err	0.1255	0.1968	0.3053	0.2695	0.2888

Fonte: Autor (2024)

Quadro 6 - Resumo dos experimentos realizados no EvaDetect.

Experimento	**Principal Foco**	**Melhor Modelo**	**Descrição**
1	Cursos Superiores (CD1-1) - Explorar a previsão de evasões em cursos superiores com conjunto de dados já validado na literatura.	StackedEnsemble_BestOfFamily	Esse modelo combina várias técnicas para prever evasões em cursos superiores
2	Cursos Superiores (CD1-2) - Analisar a previsão de evasões em cursos superiores, sem dados de desempenho acadêmico, comparado com resultados do experimento 1.	StackedEnsemble_AllModels	Este modelo utiliza uma abordagem de *ensemble*, combinando vários algoritmos para melhorar a precisão na previsão de evasões em cursos superiores, considerando diferentes perspectivas de aprendizado de máquina.
3	Cursos Superiores, IFSP Suzano (CD2) - Enfocar na previsão de evasões em cursos superiores	GBM	Este modelo destaca-se por construir sequências de árvores de decisão, corrigindo erros anteriores,

Experimento	Principal Foco	Melhor Modelo	Descrição
	específicos do IFSP Suzano, após a validação dos modelos nos experimentos 1 e 2. O objetivo foi verificar a aplicabilidade do sistema em conjunto de dados diferente.		proporcionando previsões sólidas para evasões em cursos superiores no IFSP Suzano.
4	Cursos Técnicos, IFSP Suzano (CD3) – Tendo o *software* obtido bons desempenhos com conjuntos de dados diferentes, investigou-se a evasão em cursos técnicos do IFSP Suzano. Contexto e dados acadêmicos diferente dos experimentos anteriores	XRT	Este modelo é eficaz na previsão de evasões em cursos técnicos, trazendo uma camada de aleatoriedade para construir modelos robustos, com ênfase na identificação de alunos que podem evadir.
5	Curso Técnico de Automação Industrial, IFSP Suzano (CD4) – Uma vez que o EvaDetect se provou eficaz em predizer evasão em contextos diferentes, explorou-se a predição de evasões em um contexto mais específico: o curso técnico de Automação Industrial no IFSP Suzano	GBM	Novamente o GBM demonstra sua eficácia ao focar no curso técnico de Automação Industrial, presentando um modelo robusto que se destaca na identificação precisa de evasões nesse contexto específico.

Fonte: Autor (2024)

CONSIDERAÇÕES FINAIS

A evasão dos cursos técnicos é um problema que exige atenção e soluções, pois não afeta somente as instituições de ensino, mas principalmente os estudantes, causando impactos significativos em suas vidas profissional e acadêmica. Durante o ano de 2022, havia no Brasil 3.607.900 alunos matriculados em cursos técnicos de nível médio, um público considerável que precisa de atenção para que o abandono escolar seja, dentro do possível, minimizado. Portanto, o estudo buscou contribuir para que as instituições de ensino possam compreender como as tecnologias de inteligência artificial contribuem no desafio de reduzir a evasão escolar, entregando como resultado um sistema de IA eficaz de auxílio no combate ao abandono escolar.

Diante desse contexto, a presente pesquisa se orientou pela seguinte questão: Como prever o potencial de evasões no ato da matrícula por meio de algoritmos de aprendizado de máquina, e de que forma essas previsões contribuem para intervenções pedagógicas e administrativas eficazes na minimização do problema? Com base nessa questão, investigou-se a viabilidade e eficácia de algoritmos de AM na prevenção da evasão, até chegar à consideração do uso de ferramentas de AutoML.

O objetivo do estudo portanto, foi encontrar uma solução para previsão da evasão, no ato da matrícula, por meio do desenvolvimento e validação de um *software*, que poderá ser replicado não somente a alunos de cursos técnicos, mas também a alunos de cursos tecnológicos, bacharelados, licenciaturas e nível médio. Assim, alinhado à questão de pesquisa e ao objetivo proposto, o EvaDetect, desenvolvido e validado ao longo dos Experimentos 1 a 5, surge como uma ferramenta promissora para lidar com a evasão escolar. O produto gerado por este trabalho é um *software* de predição de evasão, desenvolvido em Python, com o *framework* H2O-AutoML, denominado EvaDetect, devidamente registrado no Instituto Nacional da Propriedade Industrial (INPI) sob o número BR512024000219-7 conforme apresentado no Anexo B.

As previsões geradas pelos algoritmos entregues pelo EvaDetect desenvolvido no estudo, oferecem ferramenta importante para orientar intervenções pedagógicas e administrativas direcionadas à minimização do problema da evasão escolar. Ao antecipar potenciais casos de evasão no ato da matrícula, as instituições de ensino ganham a capacidade de identificar alunos em risco antes mesmo de seu envolvimento completo no curso. A identificação precoce objetiva permitir que a instituição de ensino implemente estratégias pedagógicas adaptadas às necessidades específicas de cada estudante, como programas de tutoria, acompanhamento personalizado, concessão de bolsas de estudo e auxílio acadêmico,

orientação profissional e serviços de apoio psicológico. Destaca-se ainda que o EvaDetect em seu relatório final, entrega apenas as matrículas dos alunos com as probabilidades de conclusão ou não do curso, preservando assim a identidade e privacidade dos estudantes. A partir desse ponto, é de responsabilidade da equipe sócio pedagógica, realizar as intervenções de maneira adequada, preservando os alunos e seus dados.

Ao observar os resultados dos experimentos, evidenciou-se que a evasão é um fenômeno complexo, influenciado por uma gama de fatores interrelacionados. Contudo, o EvaDetect, demonstrou capacidade de adaptar-se a diferentes cenários, identificando nuances específicas de cursos técnicos e superiores.

Considera-se após os Experimento 4 e 5 que a distância percorrida pelo aluno, de sua residência até o Campus Suzano do IFSP, ano de conclusão do ensino anterior e renda *per capta*, constituem-se variáveis cruciais na predição da evasão. Obviamente que somente essas variáveis não são fatores determinantes para a identificação precoce, mas ao associarmos as demais variáveis, em um conjunto de dados robusto e com uma menor quantidade da dados ausentes, é possível predizer com maior precisão a possibilidade de evasão.

Outra consideração importante, é a relação das predições com os dados acadêmicos. Ao retirar os dados acadêmicos no Experimento 2, o modelo perde acurácia, sensibilidade e precisão. Junta-se a isso o fato de haver raras pesquisas que utilizam apenas dados de entrada dos alunos, conclui-se que modelos treinados com dados acadêmicos, melhoram significativamente o seu desempenho. Sendo a ciência de dados um aprendizado constante, modelos diferentes devem ser treinados para cada situação de análise. Realizar o acompanhamento dos alunos ao longo dos semestres, com dados atualizados, permitirá aos modelos de AM maior assertividade, possibilitando políticas de permanência e êxito mais acertadas.

Ao se concentrar no Experimento 5, destinado ao curso técnico de Automação Industrial, percebe-se a adaptabilidade do EvaDetect. O modelo GBM alinhado com o ajuste do *threshold* destacado nesse experimento, ilustra como a personalização do sistema para cenários específicos otimiza significativamente sua capacidade de identificar casos de evasão. Essa abordagem específica evidencia que a prevenção da evasão não é uma solução única, mas sim um processo adaptativo que requer constante refinamento e treinamento de novos modelos.

A avaliação comparativa entre os Experimentos 4 e 5 fornece informações importantes para a tomada de decisões práticas por parte da equipe de gestão. Embora o modelo XRT tenha apresentado uma acurácia ligeiramente superior, o GBM se destacou em precisão, indicando que a escolha entre esses modelos depende das prioridades específicas da instituição de ensino.

A flexibilidade do EvaDetect em oferecer opções, abre portas para estratégias personalizadas de acordo com os objetivos de cada instituição.

Além disso, ao se analisar o desempenho geral do EvaDetect ao longo de todos os experimentos, nota-se consistência, com acurácia média de 76,1%. O desempenho mais baixo ficou com o experimento 3. O Experimento 1 se destacou, em termos de desempenho geral, mas esse desempenho possui relação direta com a variáveis de desempenho acadêmico, essas retiradas para o experimento 2. Assim sendo, considera-se que as variações nos resultados entre os experimentos podem ser atribuídas às nuances dos conjuntos de dados utilizados, reforçando a importância de adaptar o sistema a diferentes realidades.

A ligação entre os resultados obtidos com o EvaDetect e a literatura existente sobre evasão revela uma convergência significativa. Fatores socioeconômicos, como transporte, renda e local de residência, emergem como postos-chaves, corroborando estudos anteriores. Essa consistência fortalece a validade e relevância das considerações finais da pesquisa, indicando que o EvaDetect não apenas complementa, mas também enriquece a compreensão já existente sobre a evasão em contextos educacionais específicos.

O EvaDetec demonstrou ao longo dos experimentos uma boa precisão na identificação precoce de alunos com risco de evasão, sustentando assim a eficácia do modelo. Além disso, o sistema também se mostrou capaz de adaptar-se a diferentes cenários, entregando modelos adequados a cada necessidade. Por meio do ajuste de *threshold* foi possível otimizar a eficácia dos modelos, ajustando-os às especificidades de diferentes contextos.

Contudo, os experimentos mostraram uma alta dependência dos dados acadêmicos, que ao serem retirados, a partir do experimento 2, impactaram significativamente o desempenho do modelo, principalmente no que se refere ao erro da classe negativa. Além disso, o sistema desenvolvido não descarta a necessidade de um especialista em análise de dados ou em IA. Apesar do EvaDetect otimizar as tarefas de criação de modelos de AM, o especialista é necessário para realizar a filtragem dos dados, avaliação dos modelos, ajustes e interpretação dos resultados.

Por fim, considera-se que o EvaDetec entrega não apenas um modelo preditivo, mas a visão de um futuro educacional mais inclusivo e adaptável. O EvaDetect, ao ser integrado à gestão acadêmica, não oferece somente uma solução tangível para a evasão, mas também promove abordagem proativa na identificação e enfrentamento desse desafio. Ao capacitar as instituições de ensino com uma ferramenta personalizável, a pesquisa busca direcionar mudanças tangíveis na abordagem da evasão escolar nos cursos técnicos, apontando para a educação mais eficaz e equitativa.

Como pesquisas futuras, entende-se linhas de investigação que aprofundem a análise da influência das variáveis acadêmica e sociodemográficas, em diferentes regiões e como elas impactam a permanência dos estudantes. Sugere-se a linha de investigação que compreenda como as instituições de ensino podem incorporar de forma efetiva o EvaDetect e quais estratégias de intervenção a serem aplicadas a partir da identificação precoce dos estudantes.

REFERÊNCIAS

AALST, W. M. P. VAN DER. Data Scientist: The Engineer of the Future. Enterprise Interoperability VI, 41–51. https://doi.org/10.1007/978-3-319-04948-9t: The Engineer of the Future. **Enterprise Interoperability VI**, p. 41–51, 2014.

AHUETT-GARZA, H.; KURFESS, T. A brief discussion on the trends of habilitating technologies for Industry 4.0 and Smart manufacturing. **Manufacturing Letters**, v. 15, p. 60–63, 2018.

ALIREZAZADEH, P.; FATHI, A.; ABDALI-MOHAMMADI, F. A Genetic Algorithm-Based Feature Selection for Kinship Verification. **IEEE Signal Processing Letters**, v. 22, n. 12, p. 2459–2463, 2015.

ANDERSEN, R.; WERFHORST, H. G. V. D. Education and occupational status in 14 countries: the role of educational institutions and labour market coordination. **The British Journal of Sociology**, v. 61, n. 2, p. 336–355, 1 jun. 2010.

ATHEY, S. The impact of machine learning on economics. Em: **The economics of artificial intelligence: An agenda**. University of Chicago Press, 2018. p. 507–547.

AYYADEVARA, V. K. Gradient Boosting Machine. Em: **Pro Machine Learning Algorithms : A Hands-On Approach to Implementing Algorithms in Python and R**. Berkeley, CA: Apress, 2018. p. 117–134.

BAKER, R. S. J. D; YACEF, K. The State of Educational Data Mining in 2009: A Review and Future Visions. **Journal of Educational Data Mining**, v. 1, n. 1, p. 3–17, 1 out. 2009.

BAKHSHINATEGH, B. et al. Educational data mining applications and tasks: A survey of the last 10 years. **Education and Information Technologies**, v. 23, n. 1, p. 537–553, 1 jan. 2018.

BARROS, R. P. DE et al. **Consequências da Violação do Direito à Educação**. 1º ed. Rio de Janeiro: Autografia, 2021.

BARUA, P. D. et al. Artificial Intelligence Enabled Personalised Assistive Tools to Enhance Education of Children with Neurodevelopmental Disorders—A Review. **International Journal of Environmental Research and Public Health**, v. 19, n. 3, p. 1192, jan. 2022.

BASTOS, E. V. P.; CARVALHO, M. DA S.; MACEDO, M. A. ANÁLISE DA RELAÇÃO ENTRE A REMUNERAÇÃO DO TRABALHO E O NÍVEL DE INSTRUÇÃO: UMA ABORDAGEM ESTATÍSTICA.: ANALYSIS OF THE RELATIONSHIP BETWEEN WORK REMUNERATION AND LEVEL OF EDUCATION: A STATISTICAL APPROACH. **Revista Contexto & Educação**, v. 37, n. 116, p. 226–238, 3 jan. 2022.

BEN-ISRAEL, D. et al. The impact of machine learning on patient care: A systematic review. **Artificial Intelligence in Medicine**, v. 103, p. 101785, 2020.

BERGSTRA, J. et al. Algorithms for hyper-parameter optimization. **Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011**, p. 1–9, 2011.

BITENCOURT, W. A.; SILVA, D. M.; XAVIER, G. DO C. Pode a inteligência artificial apoiar ações contra evasão escolar universitária? **Ensaio: Avaliação e Políticas Públicas em Educação**, 29 nov. 2021.

BOYD, D.; CRAWFORD, K. CRITICAL QUESTIONS FOR BIG DATA. **Information, Communication & Society**, v. 15, n. 5, p. 662–679, 1 jun. 2012.

BRANCO, E. P. et al. EVASÃO ESCOLAR: DESAFIOS PARA PERMANÊNCIA DOS ESTUDANTES NA EDUCAÇÃO BÁSICA. **Revista Contemporânea de Educação**, v. 15, n. 34, p. 133–155, 29 dez. 2020.

BRASIL. **Decreto nº 7.566**. Rio de Janeiro: Presidência da República, 23 set. 1909. Disponível em: <https://www2.camara.leg.br/legin/fed/decret/1900-1909/decreto-7566-23-setembro-1909-525411-publicacaooriginal-1-pe.html>. Acesso em: 28 fev. 2023.

BRASIL. A educação nas mensagens presidenciais (1890-1986). 1987.

BRASIL. **Lei nº 8.948**. Brasília: Presidência da República, 8 dez. 1994. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l8948.htm>. Acesso em: 18 jun. 2022.

BRASIL. **Lei nº 11.195**. Brasília: Presidência da República, 18 nov. 2005. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/lei/l11195.htm>. Acesso em: 18 jun. 2022.

BRASIL. **Lei nº 11.892**. Brasília: Presidência da República, 29 dez. 2008. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11892.htm>. Acesso em: 15 abr. 2022.

BRASIL. **Plataforma Nilo Peçanha**. Disponível em: <http://plataformanilopecanha.mec.gov.br/>. Acesso em: 29 mar. 2023.

BRASIL, M. **CENTENÁRIO DA REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA**. Brasília: Ministério da Educação, 2009. Disponível em: <http://portal.mec.gov.br/setec/arquivos/centenario/historico_educacao_profissional.pdf>. Acesso em: 28 abr. 2022.

BUDJAČ, R. et al. Automated Machine Learning Overview. **Research Papers Faculty of Materials Science and Technology Slovak University of Technology**, v. 27, n. 45, p. 107–112, 31 ago. 2019.

CHAI, T.; DRAXLER, R. Root mean square error (RMSE) or mean absolute error (MAE)? **Geosci. Model Dev.**, v. 7, jan. 2014.

CHASSIGNOL, M. et al. Artificial Intelligence trends in education: a narrative overview. **Procedia Computer Science**, v. 136, p. 16–24, 2018.

CHAUHAN, K. et al. **Automated Machine Learning: The New Wave of Machine Learning**. 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). **Anais**...2020.

CHEN, J.-F.; HSIEH, H.-N.; DO, Q. H. Predicting Student Academic Performance: A Comparison of Two Meta-Heuristic Algorithms Inspired by Cuckoo Birds for Training Neural

Networks. **Algorithms**, v. 7, n. 4, p. 538–553, dez. 2014.

CHUNG, J. Y.; LEE, S. Dropout early warning systems for high school students using machine learning. **Children and Youth Services Review**, v. 96, p. 346–353, 1 jan. 2019.

COLLOBERT, R.; WESTON, J. A unified architecture for natural language processing. p. 160–167, 2008.

DANNENBERG, R. B.; THOM, B.; WATSON, D. A machine learning approach to musical style recognition. **Proceedings of the 1997 International Computer Music Conference**, p. 344–347, 1997.

DAVIDIS, N. DA L.; NOGUEIRA, J. M.; LEAL, C. P. Ensino Técnico, Mercado de Trabalho e Incremento de Renda: evidências dos egressos do Campus Ceilândia do Instituto Federal de Brasília. **Educação em Foco**, v. 23, n. 41, p. 357–375, 18 dez. 2020.

DAVIS, J.; GOADRICH, M. **The Relationship between Precision-Recall and ROC Curves**. Proceedings of the 23rd International Conference on Machine Learning. **Anais**...: ICML '06.New York, NY, USA: Association for Computing Machinery, 2006. Disponível em: <https://doi.org/10.1145/1143844.1143874>

DEVI, J. S. et al. A path towards child-centric Artificial Intelligence based Education. **International Journal of Early Childhood**, v. 14, n. 03, p. 2022, 2022.

DORE, R.; ARAÚJO, A. C. DE; MENDES, J. DE S. . **Evasão na educação: estudos, políticas e propostas de enfrentamento**. Brasília: Instituto Federal de Brasília, 2014.

DORE, R.; LÜSCHER, A. Z. Permanência e evasão na educação técnica de nível médio em Minas Gerais. **Cadernos de Pesquisa**, v. 41, p. 770–789, dez. 2011a.

DORE, R.; LÜSCHER, A. Z. Permanência e evasão na educação técnica de nível médio em Minas Gerais. **Cadernos de Pesquisa**, v. 41, p. 770–789, dez. 2011b.

DURO, D. C.; FRANKLIN, S. E.; DUBÉ, M. G. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. **Remote Sensing of Environment**, v. 118, p. 259–272, 2012.

DUTT, A.; ISMAIL, M. A.; HERAWAN, T. A Systematic Review on Educational Data Mining. **IEEE Access**, v. 5, p. 15991–16005, 2017.

FACELI, K. et al. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. 1º ed. Rio de Janeiro: LTC, 2011.

FERNÁNDEZ-GARCÍA, A. J. et al. A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data. **IEEE Access**, v. 9, p. 133076–133090, 2021.

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

FILHO, L. A EVASÃO ESCOLAR NO ENSINO PRIMÁRIO BRASILEIRO. **Revista**

Brasileira de Estatística, v. 2, n. 7, p. 539–552, set. 1941.

FILHO, R. B. S.; ARAÚJO, R. M. DE L. Evasão e abandono escolar na educação básica no Brasil: fatores, causas e possíveis consequências. **Educação Por Escrito**, v. 8, n. 1, p. 35–48, 29 jun. 2017.

FITSILIS, P.; TSOUTSA, P.; GEROGIANNIS, V. Industry 4 . 0 : Required personnel competences. **International Scientific Journal "Industry 4.0"**, v. 133, n. 3, p. 4, 2018.

FOLHA. **Oportunidades aumentam para quem estuda ciência da dados**. Folha de São Paulo, , 2018. Disponível em: <https://www1.folha.uol.com.br/sobretudo/carreiras/2018/06/1971998-oportunidades-aumentam-para-quem-estuda-ciencia-de-dados.shtml?origin=folha>. Acesso em: 2 abr. 2022

FREITAS, M. A. T. AINDA A EVASÃO ESCOLAR NO ENSINO PRIMÁRIO BRASILEIRO. **Revista Brasileira de Estatística**, v. 2, n. 7, p. 553–642, 1941.

FREITAS, M. A. T. Bases para uma programação da educação primária no Brasil. **Revista do Serviço Público**, p. 37–52, 1956.

FREITAS, M. A. T. A municipalização do ensino primário. **Revista do Serviço Público**, p. 347–364, 1957.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine Learning**, v. 63, n. 1, p. 3–42, 1 abr. 2006.

GIL, A. C. **Como Elaborar Projetos de Pesquisa**. 4º ed. São Paulo: Atlas, 2002.

GIORDANO, C. V.; SOUZA, R. A. DE. Prognóstico da Evasão Escolar em Instituição de Educação Profissional e Tecnológica por meio da Inteligência Artificial. **Revista Interacções**, v. 19, n. 66, p. 1–20, 2023.

GOKALP, M. O. et al. **Big Data for Industry 4.0: A Conceptual Framework**. 2016 International Conference on Computational Science and Computational Intelligence (CSCI). **Anais**...dez. 2016.

GUYON, I. et al. Design of the 2015 ChaLearn AutoML challenge. **Proceedings of the International Joint Conference on Neural Networks**, v. 2015- Septe, 2015.

H2O.AI. **AUTOML: Automatic Machine Learning**. Disponível em: <https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html>. Acesso em: 15 nov. 2022.

HARZING, A. W. **Publish or Perish**. https://harzing.com/resources/publish-or-perish, 2007. Disponível em: <https://harzing.com/resources/publish-or-perish>

IBGE. **PNAD CONTÍNUA: EDUCAÇÃO 2019**. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv101736_informativo.pdf>. Acesso em: 2 abr. 2022.

IBGE. **PNAD CONTÍNUA: EDUCAÇÃO 2023**. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv102002_informativo.pdf>. Acesso em: 15 jan. 2024.

JIN, H.; SONG, Q.; HU, X. Auto-keras: An efficient neural architecture search system. **Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 1946–1956, 2019.

KONENKO, I. Machine learning for medical diagnosis: History, state of the art and perspective. **Artificial Intelligence in Medicine**, v. 23, n. 1, p. 89–109, 2001.

KOTTHOFF, L.; THORNTON, C.; HUTTER, F. **User Guide for Auto-WEKA version 2.6**. [s.l: s.n.].

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. 1. ed. Springer, 2013.

KUNZE, N. C. O surgimento da rede federal de educação profissional nos primórdios do regime republicano brasileiro. **Revista Brasileira da Educação Profissional e Tecnológica / Ministério da Educação, Secretaria de Educação Profissional e Tecnológica**, v. 2, n. 2, p. 8–24, nov. 2009.

LEDELL, E.; POIRIER, S. H2O AutoML: Scalable Automatic Machine Learning. 2020.

LEVESQUE, H. J. Knowledge Representation and Reasoning. **Annual Review of Computer Science**, v. 1, n. 1, p. 255–287, 1986.

LÜSCHER, A.; DORE, R. Política educacional no Brasil: educação técnica e abandono escolar. **Revista Brasileira de Pós-Graduação**, v. 8, 31 dez. 2011.

LYKOURENTZOU, I. et al. Dropout prediction in e-learning courses through the combination of machine learning techniques. **Computers and Education**, v. 53, n. 3, p. 950–965, 2009.

MARGIOTTA, U.; VITALE, G.; SANTOS, J. S. DOS. O FENÔMENO DO ABANDONO ESCOLAR NA EUROPA DO NOVO MILÊNIO: Dados, políticas, intervenções e perspectivas. **Cadernos CEDES**, v. 34, p. 349–366, dez. 2014.

MARINHO, I. P. **Boletim Informativo do Departamento Nacional de Educação. Campanha de Educação de Adolescentes e Adultos**. Rio de Janeiro: Departamento Nacional de Educação, 1958.

MARTINS, M. V. et al. Early Prediction of student's Performance in Higher Education: A Case Study. **Trends and Applications in Information Systems and Technologies**, Advances in Intelligent Systems and Computing series. v. 1, p. 602, 2021.

MCCARTHY, J. What is Artificial Intelligence? **Stanford University**, 12 nov. 2007.

MEC. **RELATÓRIO ANUAL DE ANÁLISE DOS INDICADORES DE GESTÃO DAS INSTITUIÇÕES FEDERAIS DE EDUCAÇÃO PROFISSIONAL, CIENTÍFICA E TECNOLÓGICA**. Brasilia: [s.n.]. Disponível em: <http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=142721-relatorio-anual-de-gestao-2018-r&category_slug=2020&Itemid=30192>. Acesso em: 13 abr. 2023.

MEC. **Rede Federal de Educação Profissional, Científica e Tecnológica**. Disponível em: <http://portal.mec.gov.br/rede-federal-inicial/apresentacao-rede-federal>. Acesso em: 11 abr. 2023.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine Learning as a Science**. v. 2

NERI, M. **Motivos da Evasão Escolar**. Disponível em: <https://bibliotecadigital.fgv.br/dspace/handle/10438/21964>. Acesso em: 30 maio. 2022.

NETO, H. R. P. **Avaliação de discentes na modalidade de ensino a distância em cursos de educação profissional em nível superior, por meio de algoritmos preditivos**. São Paulo: Centro Estadual de Educação Tecnológica Paula Souza - Unidade de Pós-Graduação, Extensão e Pesquisa, abr. 2019.

NYKODYM, T. et al. Generalized Linear Modeling with H2O. 2020.

OLIVA, B. T.; RIBEIRO, F. G.; SOUZA, ANDRÉ P. F. DE. O retorno da educação profissional no mercado de trabalho: evidências a partir de dados longitudinais. **Working Paper Series**, 29 jun. 2015.

PADMANABHAN, J.; JOHNSON PREMKUMAR, M. J. Machine Learning in Automatic Speech Recognition: A Survey. **IETE Technical Review**, v. 32, n. 4, p. 240–251, 4 jul. 2015.

PEREIRA, D. DA C.; HAHN, F. A.; BOVO, M. C. A Sala de Aula Invertida como possibilidade no combate à evasão escolar. **Multitemas**, p. 51–72, 12 mar. 2020.

PIMENTEL, F. S. C.; FERREIRA, A. R.; FREITAS, R. DE O. GAMIFICAÇÃO COMO ESTRATÉGIA PEDAGÓGICA NO COMBATE À EVASÃO: POTENCIALIDADES DA IMPLEMENTAÇÃO NO ENSINO SUPERIOR. **Anais do CIET:EnPED:2020 - (Congresso Internacional de Educação e Tecnologias | Encontro de Pesquisadores em Educação a Distância)**, 28 ago. 2020.

PINZONE, M. et al. **Jobs and Skills in Industry 4.0: An Exploratory Research**. (H. Lödding et al., Eds.)Advances in Production Management Systems. The Path to Intelligent, Collaborative and Sustainable Manufacturing. **Anais**...Cham: Springer International Publishing, 2017.

RANOLIYA, B. R.; RAGHUWANSHI, N.; SINGH, S. **Chatbot for university related FAQs**. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). **Anais**... Em: 2017 INTERNATIONAL CONFERENCE ON ADVANCES IN COMPUTING, COMMUNICATIONS AND INFORMATICS (ICACCI). 13 set. 2017.

RAO, A. S.; VERWEIJ, G. **Sizing the prize What's the real value of AI for your business and how can you capitalise?** Disponível em: <https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf>. Acesso em: 1 mar. 2023.

REALINHO, V. et al. Predicting Student Dropout and Academic Success. **Data**, v. 7, n. 11, 2022.

REHDER, W. Pesquisa de adaptabilidade profissional de alunos do SENAI. **Arquivos Brasileiros de Psicotécnica**, 1950.

RUSSEL, S.; NORVING, P. **Artificial Intelligence: A modern Approach**. 3. ed. Pearson, 2016. v. 1

SALES, A.; BALBY, L.; CAJUEIRO, A. Exploiting Academic Records for Predicting Student Drop Out: a case study in Brazilian higher education. **Journal of Information and Data Management**, v. 7, n. 2, p. 166–166, 2016.

SAMARAKOU, M. et al. Evaluation of an intelligent open learning system for engineering education. **Knowledge Management & E-Learning: An International Journal**, p. 496–513, 15 set. 2016.

SCHWAB, K. **A Quarta Revolução Industrial**. Edipro, 2016.

SETEC. **DOCUMENTO ORIENTADOR PARA A SUPERAÇÃO DA EVASÃO E RETENÇÃO NA REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL, CIENTÍFICA E TECNOLÓGICA**. Brasília: Ministério da Educação, 2014. Disponível em: <https://avr.ifsp.edu.br/images/pdf/Comissoes_Outros/PermanenciaExito/Documento-Orientador-SETEC.pdf>. Acesso em: 8 jul. 2022.

SEZER, O. B.; OZBAYOGLU, A. M.; DOGDU, E. **An Artificial Neural Network-Based Stock Trading System Using Technical Analysis and Big Data Framework**. Proceedings of the SouthEast Conference. **Anais**...: ACM SE '17.New York, NY, USA: Association for Computing Machinery, 2017. Disponível em: <https://doi.org/10.1145/3077286.3077294>

SHEMSHACK, A.; SPECTOR, J. M. A systematic literature review of personalized learning terms. **Smart Learning Environments**, v. 7, n. 1, p. 1–20, dez. 2020.

SILVA, D. B. M.; CASTIONI, R.; MARTÍNEZ, R. T. Evasão Escolar e os Indicadores da Rede Federal de Educação Profissional no Brasil entre 2003 e 2015. **Vértices (Campos dos Goitacazes)**, v. 23, n. 2, 2021.

SOFAER, H. R.; HOETING, J. A.; JARNEVICH, C. S. The area under the precision-recall curve as a performance metric for rare binary events. **Methods in Ecology and Evolution**, v. 10, n. 4, p. 565–577, 2019.

SOHN, A.; OLSON, R. S.; MOORE, J. H. Toward the automated analysis of complex diseases in genome-wide association studies using genetic programming. **CoRR**, v. abs/1702.01780, 2017.

STONE, P.; VELOSO, M. Multiagent Systems : A Survey from a Machine Learning Perspective 1 Introduction 2 Multiagent Systems. **Autonomous Robots**, v. 8, n. 3, p. 345–383, 1997.

TAN, P.-N. et al. Classification: Basic Concepts, and Techniques. **Introduction to Data Mining**, p. 839, 2019.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction do DATA MINING**. 1º ed. Addison-Wesley Professional, 2005.

TANOMARU, J. **Motivação, fundamentos e aplicações de algoritmos genéticos**. . Em: II CONGRESSO BRASILEIRO DE REDES NEURAIS. 1995. Disponível em: <https://www.dca.ufrn.br/~meneghet/FTP/Motiva%E7%E3o%20Fundamentos%20e%20Aplica%E7%F5es%20de%20Algoritmos%20Gen%E9ticos.pdf>. Acesso em: 10 mar. 2023

TCU. **Relatório Sistêmico de Fiscalização da Educação - Exercício de 2014**. Brasilia: [s.n.].

Disponível em: <https://portal.tcu.gov.br/biblioteca-digital/fisc-educacao-relatorio-sistemico-de-fiscalizacao-exercicio-2014.htm>. Acesso em: 2 abr. 2022.

THORNTON, C. et al. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. **Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, v. Part F1288, p. 847–855, 2013.

VEALE, M.; BRASS, I. Administration by algorithm? Public management meets public sector machine learning. **Public management meets public sector machine learning**, 2019.

VINCENT-LANCRIN, S.; VLIES, R. VAN DER. **Trustworthy artificial intelligence (AI) in education: Promises and challenges**. OECD Publishing, abr. 2020. Disponível em: <https://ideas.repec.org/p/oec/eduaab/218-en.html>.

WANG, W.; SIAU, K. Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work and Future of Humanity: A Review and Research Agenda. **Journal of Database Management (JDM)**, v. 30, n. 1, p. 61–79, 2019.

ZHAO, X.; KEIKHOSROKIANI, P. Sales Prediction and Product Recommendation Model Through User Behavior Analytics. **Computers, Materials \& Continua**, v. 70, n. 2, 2022.

ZÖLLER, M.-A.; HUBER, M. F. Survey on Automated Machine Learning. 2019.

APÊNDICE A – QUADROS DE TRANSFORMAÇÃO DE DADOS

Neste Apêndice A são apresentadas as tabelas de conversão dos dados categóricos em numéricos, dos atributos dos alunos dos cursos superiores e técnico do campus Suzano do IFSP.

Quadro 7 - Lista de Cidades e Códigos de Conversão

Cidade	Código	Cidade	Código	Cidade	Código
Arujá	1	Araras	25	Embu	49
Ferraz de Vasconcelos	2	Aurora	26	Diadema	50
Franco de Rocha	3	Barueri	27	Porto Feliz	51
Guararema	4	Bertioga	28	Taboão da Serra	52
Guarulhos	5	Biritiba Mirim	29	Santos	53
Itapetininga	6	Brasília	30	Franco da Rocha	54
Itaquaquecetuba	7	Caieiras	31	Sorocaba	55
Ituporanga	8	Campinas	32	Itapevi	56
Marapoama	9	Carapicuíba	33	Sumaré	57
Mauá	10	Nova Lima	34	Salesópolis	58
Mogi das Cruzes	11	Jacareí	35	Marília	59
Poá	12	São José dos Campos	36	Cariacica	60
Ribeirão Pires	13	Mairiporã	37	Capivari	61
Rio Grande da Serra	14	Cotia	38	Valinhos	62
Santo André	15	Teolândia	39	Ouro Branco	63
São Paulo	16	Francisco Morato	40	Jundiaí	64
Suzano	17	Hortolândia	41	Praia Grande	65
Cajamar	18	Novo Oriente	42	Itapecerica da Serra	66
Caraguatatuba	19	São José do Rio Preto	43	Pardinho	67
Indaiatuba	20	Itanhaém	44	Santa Bárbara D'Oeste	68
Osasco	21	Santa Isabel	45	Jandira	69
Sobral	22	Três Pontas	46		
Teresina	23	São Bernardo do Campo	47		
Vila Velha	24	São Vicente	48		

Fonte: Autor (2024)

Quadro 8 - Lista de Estado Civil e Códigos de Conversão

Estado Civil	Código
Solteiro	0
Casado	1
Não Informado	3
Divorciado	2

Fonte: Autor (2024)

Quadro 9 - Lista de Meios de Transporte e Códigos de Conversão

Meio de Transporte	Código	Meio de Transporte	Código	Meio de Transporte	Código
A pé	0	Bicicleta, Automóvel e Transporte coletivo	10	Transporte cedido por prefeitura e A pé	20
Automóvel	1	Bicicleta, Automóvel, Transporte coletivo e A pé	11	Outro	21
Automóvel e A pé	2	Bicicleta, Moto, Automóvel, Transporte coletivo, A pé e outro	12	Transporte cedido por prefeitura	22
Automóvel e Transporte coletivo	3	Bicicleta, Transporte coletivo e A pé	13	Transporte coletivo	23
Automóvel, Transporte coletivo e A pé	4	Bicicleta, Transporte coletivo, A pé e outro	14	Transporte coletivo e A pé	24
Bicicleta	5	Bicicleta, Transporte coletivo, A pé e Transporte locado	15	Transporte coletivo e Transporte cedido por prefeitura	25
Bicicleta e A pé	6	Moto e Automóvel	16	Transporte coletivo, A pé e outro	26
Bicicleta e Automóvel	7	Moto e Transporte cedido por prefeitura	17	Transporte coletivo, A pé e Transporte locado	27
Bicicleta e Transporte coletivo	8	Moto e Transporte coletivo	18	Transporte coletivo, Transporte cedido por prefeitura e A pé	28
Bicicleta, Automóvel e A pé	9	Moto, Automóvel, Transporte coletivo e A pé	19	Transporte Locado	29

Fonte: Autor (2024)

Quadro 10 - Lista de Cursos e Códigos de Conversão

Curso	Código
Técnico Em Eletroeletrônica	1198
Técnico Em Automação Industrial	1199
Técnico Em Administração	1201
Tecnologia Em Processos Químicos	36000
Licenciatura Em Química	37101
Tecnologia Em Logística	37000
Bacharelado Em Química Industrial	110700
Tecnologia Em Mecatrônica Industrial	3700
Bacharelado Em Engenharia De Controle E Automação	37200

Fonte: Autor (2024)

Quadro 11 - Lista das Formas de Ingresso e Códigos de Conversão

Forma de Ingresso	Código	Forma de Ingresso	Código	Forma de Ingresso	Código
Ampla Concorrência (Geral)	0	Processo Seletivo Simplificado - Escola Pública/Renda (L1)	10	Edital de Seleção - Aluno especial / Não regular	20
Ampla Concorrência (Vestibular)	1	Seleção Geral Graduação (SiSU) (Inativa)	11	Seleção Geral Graduação (SiSU) (Inativa)	21
Escola Pública	2	SiSU L1 (SGC L1) - Candidatos com renda familiar bruta per capita igual ou inferior a 1,5 salário-mínimo que tenham cursado integralmente o ensino médio em escolas públicas (Lei nº 12.711/2012)	12	Ampla Concorrência SISU	22
Escola Pública + PPI	3	SiSU L2 (SGC L2) - Candidatos autodeclarados pretos, pardos ou indígenas, com renda familiar bruta per capita igual ou inferior a 1,5 salário-mínimo e que tenham cursado integralmente o ensino médio em escolas públicas (Lei nº 12.711/2012)	13	SiSU L10 (SGC L7) - Candidatos com deficiência autodeclarados pretos, pardos ou indígenas, COM renda familiar bruta per capita igual ou inferior a 1,5 salário-mínimo, que tenham cursado integralmente o ensino médio em escolas públicas (Lei nº 12.711/2012)	23
Escola Pública + Renda	4	SiSU L5 (SGC L3) - Candidatos que, independente de renda (art. 14, II, Portaria Normativa nº 18/2012), tenham cursado integralmente o ensino médio em escolas públicas (Lei nº 12.711/2012).	14	SiSU L9 (SGC L5) - Candidatos com deficiência que tenham renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o ensino médio em escolas públicas (Lei nº 12.711/2012)	24
Escola Pública + Renda + PCD	5	SiSU L6 (SGC L4) - Candidatos autodeclarados pretos, pardos ou indígenas que, independente da renda (art. 14, II, Portaria Normativa nº 18/2012), tenham cursado integralmente o ensino médio em escolas públicas (Lei nº12.711/2012)	15	SiSU L13 (SGC L6) - Candidatos com deficiência que, independente da renda(art. 14, II, Portaria Normativa nº 18/2012), tenham cursado integralmente o ensino médio em escolas públicas (Lei nº 12.711/2012)	25

Escola Pública + Renda + PPI	6	Transferência de Curso (Interno - mesmo campus)	16	SiSU L14 (SGC L8) - Candidatos com deficiência autodeclarados pretos, pardos ou indígenas que, independentemente da renda (art. 14, II, Portaria Normativa nº 18/2012), tenham cursado integralmente o ensino médio em escolas públicas (Lei nº 12.711/2012)	26
Escola Pública + Renda + PPI + PCD	7	Transferência Intercampi	17	Ingresso de portador de diploma de graduação	27
Processo Seletivo Simplificado - Ampla Concorrência	8	Reingresso	18	Processo Seletivo Simplificado - Escola Pública/Renda/Etnia (L2)	28
Processo Seletivo Simplificado - Escola Pública (L3)	9	Processo Seletivo Simplificado - Escola Pública/Etnia (L4)	19	Transferência Facultativa	29

Fonte: Autor (2024)

Quadro 12 - Lista de Sexo e Códigos de Conversão

Sexo	Código
M	0
F	1

Fonte: Autor (2024)

Quadro 13 - Situação no Curso e Códigos de Conversão

Situação no Curso	Código
Concluído ou Matriculado	1
Não concluído	0

Fonte: Autor (2024)

Quadro 14 - Lista de Etnias e Códigos de Conversão

Etnia	Código
Amarela	0
Branca	1
Indígena	2
Não Declarado	3
Parda	4
Preta	5

Fonte: Autor (2024)

Quadro 15 - Lista de Níveis de Ensino e Código de Conversão

Nível Ensino Anterior	Código
Fundamental	0
Médio	1
Superior	2

Fonte: Autor (2024)

Quadro 16 - Tipos de Escola de Origem e Códigos de Conversão

Tipo Escola de Origem	Código
Pública	0
Privada	1

Fonte: Autor (2024)

Quadro 17 - Lista de Turnos e Código de Conversão

Turno	Código
Noturno	0
Vespertino	1
Matutino	2

Fonte: Autor (2024)

Quadro 18 - Zona Residencial e Códigos de Conversão

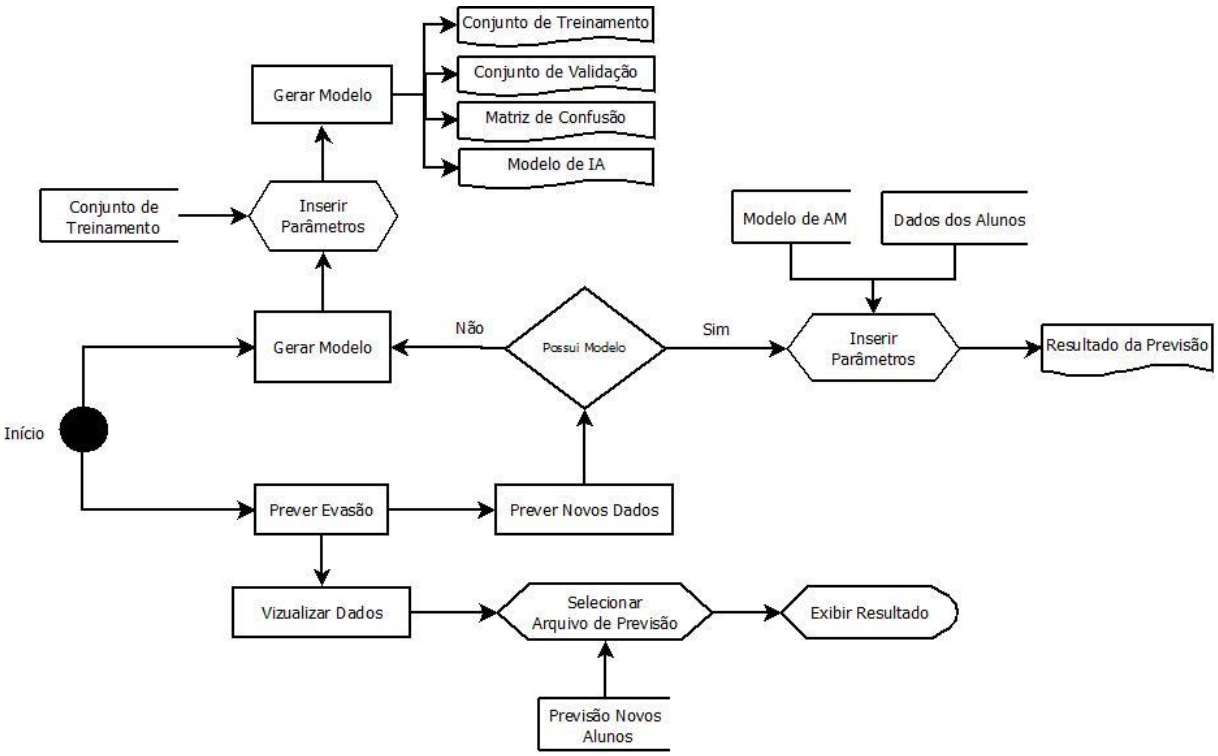
Zona Residencial	Código
Rural	0
Urbana	1

Fonte: Autor (2024)

APÊNDICE B – FLUXOGRAMA E ALGORITMOS DO EVADETECT

O Apêndice B apresenta o fluxograma de processo dos dois módulos principais do EvaDetect, além dos algoritmos dos respectivos módulos (Gerar Modelo e Prever Evasão)

Figura 6 - Fluxograma EvaDetect



Fonte: Autores (2023)

ALGORITMO MÓDULO 1 – GERAR MODELO

1. **Cópia do Arquivo:**
    - Copia o arquivo de dados de treinamento para um diretório temporário.
2. **Configurações Iniciais e Verificações:**
    - Exibe informações iniciais sobre o conjunto de dados, tempo máximo de execução, número máximo de modelos etc.
    - Cria diretórios necessários se não existirem.
    - Verifica e preenche valores padrão para semente, número de threads, tamanho máximo de memória etc., se não forem informados.
3. **Configuração do Ambiente H2O:**
    - Configura o ambiente H2O, com número de *threads*, tamanho máximo de memória e demais parâmetros.
    - Inicializa o H2O.

4. **Execução do H2O AutoML:**
    - Inicializa o processo de AutoML para cada conjunto de dados.
    - Configura e treina o modelo H2OAutoML com os parâmetros fornecidos.
    - Exibe informações sobre *o leaderboard* e o melhor modelo.
    - Exporta o modelo, conjuntos de treinamento e teste, e salva a matriz de confusão em um diretório específico.
5. **Finalização e Limpeza:**
    - Exibe mensagens de encerramento.
    - Exclui o arquivo temporário.
    - Salva a saída do AutoML em um arquivo.
    - Desliga o cluster H2O.
6. **Tratamento de Exceções:**
    - Lida com possíveis erros durante a execução.

ALGORITMO MÓDULO 2 – PREVER EVASÃO

1. **Configuração do Logger:**
    - Inicia um sistema de registro para exibir mensagens de informação, aviso e erro na interface.
2. **Configuração do Threshold:**
    - Obtém e verifica se o usuário forneceu um valor válido para o threshold.
3. **Carregamento do Modelo:**
    - Inicializa o sistema H2O.
    - Carrega o modelo de aprendizado de máquina.
4. **Realização de Previsões:**
    - Importa os dados de teste.
    - Realiza previsões com o modelo carregado.
    - Ajusta e salva os resultados.
5. **Atualização dos Dados de Teste Original:**
    - Adiciona as previsões ao conjunto de dados original.
6. **Log de Informações:**
    - Registra informações sobre o modelo, dados de teste e resultados.
    - Exibe mensagens informativas para o usuário.

APÊNDICE C – SAÍDA DO EVADETEC UTILIZANDO O H2O AUTOML

Quadro 19 - Relatório parcial de saída do AutoML para o CD1-1 após treinamento do modelo

LEADERBOARD: model_id	**auc**	**logloss**	**aucpr**	**mean_per_class_error**	**rmse**	**mse**
StackedEnsemble_BestOfFamily_1_AutoML_1_20231213_121212	0.923656	0.309774	0.943491	0.153361	0.302666	0.0916069
StackedEnsemble_AllModels_1_AutoML_1_20231213_121212	0.92206	0.310576	0.942374	0.148067	0.303089	0.0918629
GBM_2_AutoML_1_20231213_121212	0.919143	0.32912	0.94359	0.155466	0.311479	0.0970193
GLM_1_AutoML_1_20231213_121212	0.917953	0.319484	0.943945	0.16313	0.307914	0.094811

Model Summary for Stacked Ensemble:

key	value
Stacking strategy	cross_validation
Number of base models (used / total)	5/5
# GBM base models (used / total)	1/1
# GLM base models (used / total)	1/1
# DRF base models (used / total)	2/2
# DeepLearning base models (used / total)	1/1
Metalearner algorithm	GLM

ModelMetricsBinomialGLM: stackedensemble

**** Reported on train data. ****

MSE: 0.06157337956495357

RMSE: 0.24813983873000636

LogLoss: 0.21821374449678912

AUC: 0.9685036601437479

AUCPR: 0.9823133870303902

Gini: 0.9370073202874958

Null degrees of freedom: 2852

Residual degrees of freedom: 2847

Null deviance: 3627.7767581420558

Residual deviance: 1245.127626098679

AIC: 1257.127626098679

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.5580492608060791

	0	1	Error	Rate
0	803	145	0.153	(145.0/948.0)
1	73	1832	0.0383	(73.0/1905.0)
Total	876	1977	0.0764	(218.0/2853.0)

Fonte: Autor (2024)

Quadro 20 - Relatório parcial de saída do AutoML para o CD1-2 após treinamento do modelo

LEADERBOARD: model_id	**auc**	**logloss**	**aucpr**	**mean_per_class_error**	**rmse**	**mse**
StackedEnsemble_AllModels_1_AutoML_1_20231211_194045	0.81672	0.469355	0.882097	0.298322	0.390344	0.152368
GBM_2_AutoML_1_20231211_194045	0.814909	0.474259	0.881514	0.313867	0.392715	0.154225
StackedEnsemble_BestOfFamily_1_AutoML_1_20231211_194045	0.814627	0.473128	0.881104	0.318614	0.392588	0.154125
GBM_3_AutoML_1_20231211_194045	0.812778	0.472834	0.875179	0.277487	0.391037	0.15291

Model Summary for Stacked Ensemble:

key	value
Stacking strategy	cross_validation
Number of base models (used / total)	8/10
# GBM base models (used / total)	6/6
# DRF base models (used / total)	0/2
# GLM base models (used / total)	1/1
# DeepLearning base models (used / total)	1/1
Metalearner algorithm	GLM
Metalearner fold assignment scheme	Random
Metalearner nfolds	5

ModelMetricsBinomialGLM: stackedensemble

**** Reported on train data. ****

MSE: 0.09194855836423747

RMSE: 0.30323020687958757

LogLoss: 0.3159585173283729

AUC: 0.9562376885827919

AUCPR: 0.9764591654189801

Gini: 0.9124753771655838

Null degrees of freedom: 2494

Residual degrees of freedom: 2486

Null deviance: 3182.1791369167227

Residual deviance: 1576.633001468581

AIC: 1594.633001468581

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.575449320789583

	0	1	Error	Rate
0	1071	287	0.2113	(287.0/1358.0)
1	95	2087	0.0435	(95.0/2182.0)
Total	1166	2374	0.1079	(382.0/3540.0)

Fonte: Autor (2024)

Quadro 21 - Relatório parcial de saída do AutoML para o CD2 após treinamento do modelo

O MELHOR MODELO: Model Details

H2OGradientBoostingEstimator : Gradient Boosting Machine

Model Key: GBM_grid_1_AutoML_1_20240109_130234_model_4

Model Summary:

	number_of_trees	number_of_internal_trees	model_size_in_bytes	min_depth	max_depth	mean_depth	min_leaves	max_leaves	mean_leaves
	26	26	63229	10	15	12.6923	50	81	71.1923

ModelMetricsBinomial: gbm

**** Reported on cross-validation data. ****

MSE: 0.19819802823584537

RMSE: 0.44519437129847605

LogLoss: 0.586907730200412

Mean Per-Class Error: 0.2943699407998067

AUC: 0.7739972212154163

AUCPR: 0.7530111408082163

Gini: 0.5479944424308325

Maximum Metrics: **Maximum metrics at their respective thresholds**

metric	threshold	value	idx
max f1	0.478838	0.748815	250
max f2	0.0951677	0.860315	371
max f0point5	0.724873	0.715241	134
max accuracy	**0.53564**	0.714286	211
max precision	0.963149	1	0
max recall	0.0951677	1	371
max specificity	0.963149	1	0

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.4061660048321181

	0	1	Error	Rate
0	200	156	0.4382	(156.0/356.0)
1	56	316	0.1505	(56.0/372.0)
Total	256	472	0.2912	(212.0/728.0)

Fonte: Autor (2024)

Quadro 22 - Relatório parcial de saída do AutoML para o CD3 após treinamento do modelo

O MELHOR MODELO: Model Details

H2ORandomForestEstimator : Distributed Random Forest

Model Key: XRT_1_AutoML_1_20240110_144808

Model Summary:

number_of_trees	number_of_internal_trees	model_size_in_bytes	min_depth	max_depth	mean_depth	min_leaves	max_leaves	mean_leaves
21	21	103112	13	20	17.90475	162	108.286	

**** Reported on cross-validation data. ****

MSE: 0.18977273187645505

RMSE: 0.4356291219333885

LogLoss: 0.5735923504784589

Mean Per-Class Error: 0.28349606474606476

AUC: 0.7891731016731017

AUCPR: 0.7417971373833113

Gini: 0.5783462033462035

Maximum Metrics: Maximum metrics at their respective thresholds

metric	**threshold**	value	idx
max f1	0.279962	0.719533	262
max f0point5	0.610522	0.7091	145
max accuracy	**0.478838**	0.730298	194
max precision	1	1	0
max recall	0.00651269	1	398
max specificity	1	1	0

A PERFORMANCE: ModelMetricsBinomial: drf

** Reported on test data. **

MSE: 0.19938920400837015

RMSE: 0.44653018263984146

LogLoss: 0.5936309960389055

Mean Per-Class Error: 0.2875

AUC: 0.7644940476190477

AUCPR: 0.744092517523961

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.30560589404333205

	0	1	Error	Rate
0	120	72	0.375	(72.0/192.0)
1	35	140	0.2	(35.0/175.0)
Total	155	212	0.2916	(107.0/367.0)

Fonte: Autor (2024)

Quadro 23 - Relatório parcial de saída do AutoML para o CD4 após treinamento do modelo

O MELHOR MODELO: Model Details

==========

H2OGradientBoostingEstimator : Gradient Boosting Machine

Model Key: GBM_grid_1_AutoML_1_20240111_142847_model_8

Model Summary:

	number_of_trees	number_of_internal_trees	model_size_in_bytes	min_depth	max_depth	mean_depth	min_leaves	max_leaves	mean_leaves
	26	26	22307	6	6	6	14	31	23.2308

ModelMetricsBinomial: gbm

** Reported on cross-validation data. **

MSE: 0.2114330112719852

RMSE: 0.45981845468835325

LogLoss: 0.638371778173425

AUC: 0.75599128540305

AUCPR: 0.679012992383722

Confusion Matrix (Act/Pred) for **max f1 @ threshold = 0.1311661880162796**

	0	1	Error	Rate
0	196	119	0.3778	(119.0/315.0)
1	44	160	0.2157	(44.0/204.0)
Total	240	279	0.3141	(163.0/519.0)

A PERFORMANCE: ModelMetricsBinomial: gbm

** Reported on test data. **

MSE: 0.268477853188335

RMSE: 0.5181484856567035

Mean Per-Class Error: 0.32376126126126126

AUC: 0.6926379504504504

AUCPR: 0.6031547316842033

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.09712288654758286

	0	1	Error	Rate
0	52	44	0.4583	(44.0/96.0)
1	14	60	0.1892	(14.0/74.0)
Total	66	104	0.3412	(58.0/170.0)

Fonte: Autor (2024)

Quadro 24 - Relatório parcial de saída do AutoML para o CD4 após treinamento do modelo - Variáveis mais importantes

Variable Importances:

variable	relative_importance	scaled_importance	percentage
Distancia atÃ© o IF	437.623	1	0.72996
Zona Residencial	43.3002	0.098944	0.0722252
Ano de ConclusÃ£o do Ensino Anterior	30.8565	0.0705092	0.0514689
Renda Per Capita	25.0189	0.05717	0.0417318
NÃ-vel de Ensino Anterior	20.4026	0.0466213	0.0340317
IDADE	14.7128	0.0336198	0.0245411
Renda Bruta Familiar (R$)	12.4662	0.0284862	0.0207938
Etnia/RaÃ§a	4.10898	0.00938931	0.00685382
Estado Civil	3.19754	0.0073066	0.00533353
Cidade	2.72343	0.00622322	0.0045427
Turno	1.77241	0.00405009	0.0029564
Meio de Transporte	1.12259	0.0025652	0.0018725
Sexo	0.775578	0.00177225	0.00129367
PerÃ-odo de Ingresso	0.596099	0.00136213	0.0009943
Tipo de Escola de Origem	0.445507	0.00101801	0.00074311
Forma de Ingresso	0.393829	0.000899926	0.00065691

Fonte: Autor (2024)

ANEXO A – AUTORIZAÇÃO DE ACESSO AOS DADOS DOS ALUNOS

INSTITUTO FEDERAL
São Paulo

Processo Eletrônico
23437.000300.2022-88

Data 05/04/2022 15:51:03	**Tipo** Ensino: Outros
Setor de Origem SZN - AUT-SZN	**Assunto** Solicitação de acesso a dados dos alunos - Pesquisa de Mestrado
Situação Finalizado	**Interessados** Wagner Roberto Garo Junior

Últimos Trâmites

07/04/2022 14:13
Recebido por: AUT-SZN: Raphael Antonio de Souza

07/04/2022 11:52
Enviado por: DRG/SZN: Eugenio de Felice Zampini

07/04/2022 11:43
Recebido por: DRG/SZN: Eugenio de Felice Zampini

07/04/2022 11:18
Enviado por: DAE-SZN: Wagner Roberto Garo Junior

07/04/2022 11:17
Recebido por: DAE-SZN: Wagner Roberto Garo Junior

05/04/2022 15:53
Enviado por: AUT-SZN: Raphael Antonio de Souza

MINISTÉRIO DA EDUCAÇÃO
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CÂMPUS SUZANO
COORD DE CURSO TECNICO EM AUT INDUSTRIAL

OFÍCIO CIRCULAR N.º 1/2022 - AUT-SZN/DAE-SZN/DRG/SZN/IFSP

Suzano, 5 de abril de 2022.

Ao Diretor Adjunto Educacional,

Wagner Roberto Garo Junior

Assunto: Solicitação de acesso aos dados dos discentes

Eu, Raphael Antonio de Souza, SIAPE 2085134, professor de ensino básico técnico e tecnológico, lotado no campus Suzano, venho por meio deste, solicitar acesso ou os dados, de todos os discentes já matriculados no campus Suzano, em todos os cursos, na plataforma SUAP.

Motivo: Meu projeto de mestrado, em desenvolvimento no Centro Estadual de Educação Tecnológica Paula Souza, sob orientação do professor Carlos Vital Giordano, tem como proposta avaliar a possibilidade de uso de inteligência artificial para previsão de evasão. Para isso, se faz necessário acessar e conhecer os dados de alunos concluinte e não concluintes.

Adianto aqui que irei pautar as pesquisas no princípio da ética e da legalidade, não expondo dados sensíveis de alunos. Afirmo ainda que os dados serão utilizados somente após autorização do comitê de ética da instituição do mestrado.

Respeitosamente / Atenciosamente,

Documento assinado eletronicamente.

Raphael Antonio de Souza
Professor EBTT

Cópia de despacho #868806 digital impresso por Raphael Souza (2085134) em 16/01/2024 11:56.

7 de abril de 2022

Despacho:

Ciente e de acordo com a solicitação do servidor. Encaminho o pedido para Direção Geral para análise e disposição.

Assinatura:

Despacho assinado eletronicamente por:
- Wagner Roberto Garo Junior, DIRETOR(A) ADJUNTO(A) - CD4 - DAE-SZN, DAE-SZN, em 07/04/2022 11:18:06.

Cópia de despacho #868911 digital impresso por Raphael Souza (2085134) em 16/01/2024 11:56.

7 de abril de 2022

Despacho:

Prezado Professor. Considerando as declarações contidas no Ofício Circular Nº 1/2022 - AUT-SZN; Considerando que o professor somente poderá apresentar dados agrupados; Considerando que o professor deverá respeitar a Lei Geral de Proteção de Dados; Autorizamos o acesso aos dados solicitados. Atenciosamente

Assinatura:

Despacho assinado eletronicamente por:
- Eugenio de Felice Zampini, DIRETOR(A) GERAL - CD2 - DRG/SZN, DRG/SZN, em 07/04/2022 11:52:40.

SERVIÇO PÚBLICO FEDERAL

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO

DADOS DA FINALIZAÇÃO DO PROCESSO 23437.000300.2022-88

Interessados:	Wagner Junior
Assunto:	Solicitação de acesso a dados dos alunos - Pesquisa de Mestrado
Usuário responsável:	Raphael Souza
Matrícula SIAPE:	2085134
Data Finalização:	07/04/2022 14:13:38

Motivo da finalização

Autorização concedida

ANEXO B – REGISTRO DE PROGRAMA DE COMPUTADOR

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DO DESENVOLVIMENTO, INDÚSTRIA, COMÉRCIO E SERVIÇOS
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512024000219-7**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1° de janeiro subsequente à data de 23/01/2024, em conformidade com o §2°, art. 2° da Lei 9.609, de 19 de Fevereiro de 1998.

Título: EvaDetect

Data de publicação: 23/01/2024

Data de criação: 20/12/2023

Titular(es): RAPHAEL ANTONIO DE SOUZA

Autor(es): RAPHAEL ANTONIO DE SOUZA; CARLOS VITAL GIORDANO

Linguagem: PYTHON

Campo de aplicação: ED-01; ED-03

Tipo de programa: AP-02; FA-01; IA-01; TC-03

Algoritmo hash: SHA-512

Resumo digital hash:
90c0e04cc76a52c284d088d050d5802ae1400613ee082d26ef590f6ffc4581b4f15f7d4f2b3dd6a3f0080c346106035a1dcb66b457737d0cafe27ebf269a6ffc

Expedido em: 30/01/2024

Aprovado por:
Carlos Alexandre Fernandes Silva
Chefe da DIPTO