UNIVERSIDADE FEDERAL DE OURO PRETO

INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS

DEPARTAMENTO DE COMPUTAÇÃO

MARCELLA SILVEIRA CAMPOS

PROPOSTA E DESENVOLVIMENTO DE UMA ESTRATÉGIA PARA PREDIÇÃO DA EVASÃO DE ALUNOS EM ÁREAS DE TECNOLOGIA

Ouro Preto
2026

MARCELLA SILVEIRA CAMPOS

PROPOSTA E DESENVOLVIMENTO DE UMA ESTRATÉGIA PARA PREDIÇÃO DA EVASÃO DE ALUNOS EM ÁREAS DE TECNOLOGIA

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Guilherme Tavares de Assis

Ouro Preto
2026

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Marcella Silveira Campos

Proposta e desenvolvimento de uma estratégia para predição da evasão de alunos em áreas de tecnologia

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 20 de Julho de 2026.

Membros da banca

Guilherme Tavares de Assis (Orientador) - Doutor - Universidade Federal de Ouro Preto
Jadson Castro Gertrudes (Examinador) - Doutor - Universidade Federal de Ouro Preto
Andrea Gomes Campos (Examinadora) - Doutora - Universidade Federal de Ouro Preto

Guilherme Tavares de Assis, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 20/07/2026.

Documento assinado eletronicamente por **Guilherme Tavares de Assis**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 20/07/2026, às 17:37, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **1142649** e o código CRC **973728A9**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Dedico este trabalho a cada pessoa que não desistiu de mim, que solidarizou-se com minha condição de saúde e que disponibilizou-se a me ajudar. Vocês me deram força para que eu conseguisse fazer meu tratamento mesmo quando tudo parecia perdido e impossível.

Agradecimentos

Agradeço a Deus pela segunda chance de encerrar, enfim, esta longa jornada, por toda força que concedeu ao meu ser e por nunca ter me deixado sozinha.

Agradeço à minha mãe, que cuidou de mim com tanta dedicação e carinho, sempre tentando me animar e mostrar o lado positivo das coisas. Sem ela eu não estaria aqui.

Agradeço ao meu namorado, que jamais soltou minha mão, teve todo o cuidado do mundo e não mediu esforços para me ver bem.

Agradeço à minha tia-avó, meu pai e madrasta e meus irmãos pelo amor, apoio, cuidado e incentivo constantes.

Agradeço às minhas gatas, pela companhia nos momentos de estudo e todas as vezes que grudaram em mim ao me ver chorar.

Agradeço ao meu professor orientador Guilherme, por todas as palavras gentis, pela paciência, compreensão, orientação e apoio inegável.

Agradeço à Yasmin, pela amizade, incentivo e por sempre tentar arrancar um sorriso ou uma risada de mim.

Agradeço à República Eclipse, pela convivência, carinho, compreensão, atenção e companheirismo até o fim.

E por último, mas não menos importante, agradeço ao Doutor Sérgio Schusterschitz, por toda sabedoria, dedicação, paciência, comprometimento com minha melhora e cada um dos sorrisos que nunca foram escassos durante nossas consultas.

A todos, minha mais sincera gratidão. Essa vitória é nossa.

Resumo

A evasão estudantil no ensino superior é um dos principais desafios enfrentados por instituições de ensino, especialmente em cursos da área de tecnologia. Este trabalho, de natureza aplicada e caráter metodológico, tem como objetivo propor, desenvolver e validar uma estratégia de predição da evasão de alunos nas áreas de tecnologia, utilizando registros acadêmicos e informações sociodemográficas reais de alunos do curso de Ciência da Computação da Universidade Federal de Ouro Preto (UFOP). Para isso, foi definida uma arquitetura metodológica composta pelas etapas de coleta, pré-processamento dos dados, treinamento e validação de três algoritmos de aprendizado supervisionado: Floresta Aleatória, K-Vizinhos Mais Próximos (KNN) e *XGBoost*. A eficácia de cada modelo foi comparada por meio das métricas de acurácia, precisão, revocação e *F1-score*, estabelecendo-se o algoritmo Floresta Aleatória como o mais eficaz para compor a estratégia final de predição. O diferencial do trabalho está na proposição de uma estratégia capaz de estimar, individualmente, a probabilidade de evasão de cada aluno, permitindo que gestores e professores intervenham preventivamente, de forma direcionada, na tentativa de reduzir os índices de evasão. Assim, este trabalho contribui com o meio acadêmico ao oferecer um modelo sólido, replicável e potencialmente aplicável não apenas ao curso de Ciência da Computação da UFOP, mas também a outros cursos da área de tecnologia e, futuramente, de diferentes áreas do conhecimento.

Palavras-chave: Evasão estudantil. Mineração de dados educacionais. Aprendizado de máquina. Predição de dados.

Abstract

Student dropout in higher education is one of the main challenges faced by educational institutions, especially in technology programs. This work, applied in nature and methodological in character, aims to propose, develop, and validate a strategy for predicting student dropout in technology fields, using real academic records and sociodemographic information from Computer Science students at the Federal University of Ouro Preto (UFOP). To achieve this, a methodological architecture was defined, consisting of data collection, pre-processing, training, and validation steps for three supervised learning algorithms: Random Forest, K-Nearest Neighbors (KNN), and XGBoost. The effectiveness of each model was compared using accuracy, precision, recall, and F1-score metrics, establishing the Random Forest algorithm as the most effective to compose the final prediction strategy. The main contribution of this work lies in proposing a strategy capable of individually estimating the dropout probability of each student, allowing managers and professors to intervene preventively, in a targeted manner, in an attempt to reduce dropout rates. Therefore, this work contributes to the academic community by providing a solid, replicable model that is potentially applicable not only to the Computer Science program at UFOP but also to other programs in the technology field and, in the future, to different areas of knowledge.

Keywords: Student dropout. Educational data mining. Machine Learning. Data prediction.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

DECOM	Departamento de Computação
UFOP	Universidade Federal de Ouro Preto
KNN	K-Nearest Neighbors

Sumário

1 Introdução

A evasão estudantil é um fenômeno recorrente no ensino superior brasileiro. O nível de dificuldade, a qualidade do ensino básico e a escassez de recursos investidos nas universidades são alguns dos fatores que estimulam esse acontecimento (SACCARO; FRANÇA; JACINTO, 2019). Nas áreas da tecnologia, esse problema é particularmente preocupante, com altas taxas de abandono observadas ao longo dos semestres. A temática afeta diretamente a vida da sociedade brasileira, que não poderá desfrutar dos impactos positivos que esses profissionais poderiam vir a oferecer.

As razões que levam à evasão no ensino superior são tão diversas e complexas que se torna difícil enumerá-las com precisão. Almeida, Soares e Ferreira (2002) afirmam que os alunos, ao ingressarem na universidade, deparam-se com uma série de desafios que extrapolam o conteúdo acadêmico, incluindo questões pessoais, conflitos interpessoais, pressões familiares e obstáculos institucionais. Esses fatores, muitas vezes refletindo nas notas e presenças do aluno, impactam diretamente o percurso formativo e contribuem para o abandono do curso. Diante desse cenário, é desejável que os professores tenham instrumentos mais eficazes para identificar sinais de risco e atuar preventivamente.

Entretanto, para os professores, não é possível alterar a bagagem com que o aluno chega no ensino superior. O que pode ser analisado é seu desempenho, uma vez já ingressado no curso. Atualmente, a tecnologia pode ajudar a investigar essa problemática. Vários estudos têm usado inteligência artificial no âmbito educacional, e essa área vem crescendo exponencialmente nos últimos 25 anos, como visto em Roll e Wylie (2016). A inteligência artificial aplicada à educação tem se mostrado uma aliada promissora na criação de ambientes de aprendizagem mais personalizados, eficientes e responsivos às demandas contemporâneas do ensino superior. Ela pode ser usada para encontrar soluções, oferecer opções antes não avaliadas e encontrar padrões não vistos. Em um meio tão amplo quanto a educação universitária, a tecnologia pode sim ajudar consideravelmente os pesquisadores com seus respectivos trabalhos.

Com essa perspectiva, tecnologias têm sido propostas para analisar grandes quantidades de dados e tentar encontrar um padrão entre eles. Assim, uma rede neural pode ser treinada para encontrar algo que se repita com frequência em dado resultado. Diante disso, torna-se relevante investigar estratégias que permitam antecipar quais alunos apresentam maior risco de evadir, possibilitando ações preventivas por parte das coordenações e docentes.

Logo, a proposta deste trabalho consiste em utilizar, inicialmente, algoritmos como Floresta Aleatória, *XGBoost* e K-Vizinhos Mais Próximos (KNN) para identificar padrões em dados acadêmicos e sociodemográficos que possam estar relacionados à evasão de alunos em cursos de graduação na área de tecnologia. A aplicação prática dos modelos busca auxiliar os

docentes na identificação precoce de alunos em risco, permitindo intervenções como conversas individualizadas, suporte acadêmico e encaminhamento para atividades extracurriculares.

1.1 Justificativa

A evasão de alunos em cursos de graduação na área de tecnologia tem se mostrado um desafio persistente e preocupante para instituições de ensino superior no Brasil. Embora essas carreiras estejam entre as mais promissoras do mercado atual, com alta demanda por profissionais qualificados, os cursos costumam apresentar elevadas taxas de abandono. As turmas frequentemente iniciam com um número expressivo de estudantes, mas apenas uma pequena parcela consegue concluir a graduação, evidenciando uma grande perda de capital humano e intelectual ao longo do percurso formativo.

Esse paradoxo, entre o alto potencial da área tecnológica e o número significativo de desistências, justifica a urgência em investigar os fatores que levam os alunos a evadir, bem como em propor soluções para mitigar esse problema. A formação de profissionais em ciência da computação, engenharia, sistemas de informação, entre outros, é estratégica para o desenvolvimento econômico e tecnológico do país. Portanto, compreender os padrões que antecedem a evasão pode permitir às instituições a adoção de políticas mais eficazes de permanência, resultando em maior eficiência acadêmica e social.

Nesse cenário, o uso de Inteligência Artificial, especialmente por meio de modelos de aprendizado de máquina, tem se mostrado uma ferramenta poderosa para a análise de dados educacionais (SACCARO; FRANÇA; JACINTO, 2019). Dentre os diversos algoritmos disponíveis, os modelos de predição destacam-se pela capacidade de capturar relações complexas entre variáveis, mesmo quando essas relações não são evidentes ou lineares. Ao serem treinados com dados históricos de estudantes, esses modelos podem identificar padrões de comportamento e desempenho que precedem o abandono, oferecendo uma previsão acurada do risco de evasão.

A proposta de aplicar técnicas preditivas baseadas em algoritmos de classificação, como Floresta Aleatória, *XGBoost* e KNN, busca justamente unir o potencial dessas ferramentas com a realidade dos cursos de tecnologia. Esses algoritmos podem fornecer uma avaliação individualizada do risco de evasão. Isso representa um avanço importante em relação aos métodos tradicionais, que geralmente atuam apenas após o abandono já ter ocorrido.

Portanto, este trabalho justifica-se não apenas pela relevância acadêmica do tema, mas também por sua aplicabilidade prática. Ao desenvolver uma estratégia de predição baseada em Inteligência Artificial, voltada especificamente para cursos da área de tecnologia, esta pesquisa pretende contribuir com o debate sobre a permanência estudantil, além de propor uma ferramenta que possa ser utilizada por instituições de ensino para enfrentar um dos maiores obstáculos da formação superior: a evasão.

1.2 Objetivos

O objetivo principal deste trabalho é propor e validar uma estratégia de predição capaz de identificar, com antecedência, a probabilidade de evasão de um aluno em cursos da área de tecnologia, utilizando modelos de aprendizado de máquina, baseado em dados anônimos de alunos do curso de Ciência da Computação da Universidade Federal de Ouro Preto.

Como objetivos específicos, este trabalho propõe:

- a definição de uma base de dados tratados composta por informações acadêmicas e sociodemográficas reais de alunos para fornecer aos modelos de predição escolhidos;
- o oferecimento de indicadores úteis como, por exemplo, "probabilidades >75% são uma alta probabilidade de evasão", que possibilitem que os professores possam intervir antes do aluno evadir, sendo com conversas pessoais, atividades extracurriculares ou pontos extras;

1.3 Método do Trabalho

Para alcançar o objetivo principal descrito, foram elaboradas duas arquiteturas complementares. A primeira abrange o processo de treinamento, validação e comparação de três modelos supervisionados (Floresta Aleatória, *XGBoost* e K-Vizinhos Mais Próximos (KNN)) utilizando dados acadêmicos e sociodemográficos reais de alunos do curso de Ciência da Computação da Universidade Federal de Ouro Preto. Essa arquitetura visa selecionar o modelo com melhor desempenho na predição da evasão. A segunda arquitetura, por sua vez, utiliza o modelo selecionado para estabelecer uma estratégia operacional, na qual os dados de cada novo aluno são processados individualmente, permitindo estimar de forma precisa a probabilidade de evasão para cada caso específico, auxiliando assim na identificação de estudantes em potencial risco e possibilitando ações preventivas.

Baseada em tais arquiteturas, a estratégia foi desenvolvida e experimentos iniciais de validação foram conduzidos, considerando os padrões gerados pelos modelos e a comparação de suas eficácias. Como estudo de caso, visando validação da estratégia proposta, a base de dados incluiu informações como idade, sexo, cidade de origem, notas, reprovações e demais registros acadêmicos de alunos do curso de Ciência da Computação da UFOP. Ademais, foram utilizadas métricas como acurácia, precisão, revocação e *F1-Score*, permitindo avaliar o potencial preditivo de cada abordagem.

Portanto, de uma forma geral, a proposta metodológica compreende as seguintes etapas: (i) compreensão e organização de uma base de dados; (ii) implementação e aplicação dos modelos selecionados (Floresta Aleatória, *XGBoost* e KNN); (iii) análise dos resultados obtidos nos treinos e testes para encontrar o modelo que teve o melhor desempenho, com as métricas de avaliação mais altas; (iv) desenvolvimento de uma estratégia de predição definitiva para identificar

alunos com maior probabilidade de evasão; (v) validação dessa estratégia proposta com dados que ela ainda não teve acesso; (vi) análise dos resultados obtidos para avaliar a eficácia do modelo. O trabalho contou com implementação prática, visando testar os modelos e discutir sua aplicabilidade em cenários reais, além de fornecer meios para o uso da estratégia por professores no acompanhamento preventivo de seus alunos.

1.4 Estrutura da Monografia

O restante desta monografia está organizado da seguinte maneira: o Capítulo 2 apresenta a revisão de literatura e o embasamento teórico necessários para a compreensão do trabalho, abordando conceitos fundamentais, tecnologias associadas e estudos relacionados. No Capítulo 3, descreve-se a metodologia adotada, detalhando as características da proposta e as arquiteturas confeccionadas. O Capítulo 4 apresenta toda a etapa de experimentação do trabalho. O Capítulo 5 expõe as conclusões extraídas a partir deste estudo e indica perspectivas para trabalhos futuros.

2 Revisão Bibliográfica

Este capítulo apresenta a revisão bibliográfica relativa ao trabalho proposto, sendo uma etapa fundamental de qualquer pesquisa científica, pois oferece o embasamento necessário para compreender o tema investigado, identificar lacunas no conhecimento existente e justificar a relevância da proposta de estudo. Logo, este capítulo encontra-se delineado como se segue. A Seção 2.1 aborda a fundamentação teórica necessária para o desenvolvimento deste trabalho e a Seção 2.2 apresenta trabalhos diretamente relacionados.

2.1 Fundamentação Teórica

Nesta seção, são expostos alguns assuntos fundamentais para a maior compreensão deste trabalho, encontrando-se delineada como se segue. A Subseção 2.1.1 descreve a evasão, termo de grande importância para este trabalho. A Subseção 2.1.2 apresenta modelos de predição, que são usados para propor a estratégia de predição da evasão de alunos em áreas da tecnologia. A Subseção 2.1.3 busca formalizar conceitos essenciais para a compreensão deste trabalho.

2.1.1 Evasão

A evasão no ensino superior é um fenômeno com várias formas de manifestação, cuja definição varia conforme a abordagem adotada. A evasão é compreendida de forma ampla, sem delimitação de tempo, sendo considerada como a saída definitiva do estudante de seu curso, independentemente da etapa em que essa saída ocorra, desde que implique a não conclusão da formação (UTIYAMA; BORBA, 2003). Por outro lado, Abbad, Carvalho e Zerbini (2007) definem evasão como a desistência definitiva do aluno em qualquer fase do curso, embora não esclareçam se essa definição abarca os estudantes que apenas se matricularam e não chegaram a iniciar as atividades do curso. Esta abordagem já considera o abandono como um evento que pode ocorrer tanto no início quanto em etapas posteriores da formação. Já Maia e Meirelles (2005) ampliam a compreensão do conceito, incluindo não só os estudantes que não completaram o curso, mas também aqueles que se matriculam e desistem antes mesmo de começar as atividades.

Para os fins deste trabalho, é adotada a definição de evasão como a desistência definitiva de um estudante que chegou a iniciar efetivamente o curso, independentemente da etapa em que essa desistência ocorra.

Além das definições conceituais, é importante destacar que a evasão no ensino superior, especialmente em cursos da área de tecnologia, pode ser provocada por uma ampla variedade de fatores, que abrangem desde dificuldades acadêmicas até questões emocionais e socioculturais. De acordo com o estudo de Santos e Marczak (2023), que realizou um mapeamento sistemático

da literatura sobre a participação feminina na Computação, foram identificados 88 fatores que contribuem para a evasão de mulheres nesses cursos. Entre os fatores mais recorrentes estão a baixa representatividade feminina e a ausência de modelos de sucesso na área, a existência de estereótipos de gênero reforçados por comentários depreciativos de colegas e professores, a falta de apoio institucional em momentos como gravidez e maternidade, a dificuldade com disciplinas de base como matemática, lógica e programação, e ainda aspectos emocionais, como ansiedade, estresse e sentimento de incapacidade. Esses elementos demonstram que, no caso das mulheres, a decisão de abandonar o curso muitas vezes extrapola o desempenho acadêmico, estando fortemente relacionada ao contexto social e emocional no qual estão inseridas.

Complementarmente, o estudo de Jesus, Rodriguez e Junior (2021), ao analisar a evasão em um curso de Licenciatura em Computação, apontou que muitos alunos enfrentam sérias dificuldades nas disciplinas introdutórias de programação, especialmente no primeiro ano do curso. A ausência de conhecimento prévio, as dificuldades com raciocínio lógico e a falta de estratégias pedagógicas eficazes contribuem para elevadas taxas de reprovação, trancamento e abandono nessas matérias. Em diversas disciplinas práticas de programação, o número de reprovações e desistências superou o número de aprovações. Assim, evidencia-se que tanto fatores acadêmicos, como o desempenho em disciplinas específicas, quanto fatores emocionais e socioculturais estão entre as principais causas de evasão em cursos da área de tecnologia, sendo imprescindível considerá-los na construção de estratégias de prevenção e intervenção.

2.1.2 Modelos de Predição

Modelos de predição baseados em Inteligência Artificial têm se mostrado eficazes no contexto educacional ao utilizarem algoritmos para analisar dados como notas, frequência e perfil socioeconômico. Segundo Giraffa e Kohls-Santos (2023), essas técnicas permitem prever resultados como desempenho e evasão escolar, além de personalizar estratégias pedagógicas e auxiliar na tomada de decisões mais assertivas pelas instituições de ensino. Esta seção encontra-se delineada como se segue. A Subseção 2.1.2.1 descreve o algoritmo de aprendizado de máquina Floresta Aleatória, deixando claro seu funcionamento. A Subseção 2.1.2.2 apresenta o algoritmo *XGBoost*, explicando seu comportamento. A Subseção 2.1.2.3 mostra o algoritmo K-Vizinhos Mais Próximos (KNN), apresentando seu procedimento.

2.1.2.1 Floresta Aleatória

A Floresta Aleatória (*Random Forest*) é um algoritmo de aprendizado supervisionado pertencente à classe dos métodos de conjunto, amplamente utilizado em tarefas de classificação e regressão. De acordo com Breiman (2001), o método foi desenvolvido como uma extensão do *bagging* (*bootstrap aggregating*), com o objetivo de reduzir a alta variância associada às árvores de decisão individuais, mantendo um baixo viés preditivo.

Uma árvore de decisão funciona por meio da divisão recursiva do espaço de atributos, selecionando, em cada nó, a variável e o ponto de corte que melhor separam os dados em relação à variável resposta. Apesar de serem modelos interpretáveis e intuitivos, árvores isoladas apresentam elevada sensibilidade a pequenas variações nos dados de treinamento, o que pode resultar em modelos instáveis e fortemente ajustados aos dados observados.

A Floresta Aleatória busca contornar essas limitações combinando múltiplas árvores de decisão independentes, cada uma treinada a partir de uma amostra diferente dos dados. O algoritmo introduz dois mecanismos fundamentais de aleatoriedade. O primeiro é a amostragem com reposição, conhecida como *bootstrap*, na qual cada árvore é treinada a partir de uma amostra aleatória do conjunto original, permitindo que algumas observações sejam repetidas e outras não sejam selecionadas. Esse procedimento garante diversidade entre as árvores do conjunto.

O segundo mecanismo consiste na seleção aleatória de apenas um subconjunto dos atributos disponíveis a cada divisão da árvore. Em vez de avaliar todos os preditores possíveis em cada nó, o algoritmo escolhe aleatoriamente um número reduzido de variáveis candidatas, avaliando apenas essas para definir a melhor divisão. Essa restrição impede que variáveis muito dominantes sejam escolhidas repetidamente nas primeiras divisões de todas as árvores, o que aumentaria a correlação entre elas e reduziria a eficácia do método de conjunto.

Durante o treinamento, cada árvore cresce de forma independente, explorando diferentes combinações de observações e atributos. Na fase de predição, cada árvore fornece uma decisão individual para uma nova instância. Em problemas de classificação, a decisão final do modelo é obtida por meio do voto majoritário entre as previsões das árvores; em problemas de regressão, utiliza-se a média aritmética das previsões individuais.

Segundo Breiman (2001), o desempenho da Floresta Aleatória tende a melhorar à medida que o número de árvores aumenta, sem risco significativo de sobreajuste, uma vez que o acréscimo de árvores reduz a variância do modelo agregado. Além disso, o método fornece medidas de importância das variáveis, que indicam o quanto cada atributo contribuiu para a redução da impureza ao longo das divisões, sendo úteis para análise e interpretação dos resultados.

No contexto da predição de evasão estudantil, a Floresta Aleatória é particularmente adequada, pois consegue lidar com dados heterogêneos, não exige relações lineares entre as variáveis e é robusta a ruídos e valores ausentes. Variáveis como coeficiente acadêmico, número de reprovações, idade e características sociodemográficas podem ser combinadas de forma eficiente, permitindo identificar padrões complexos associados ao abandono do curso.

De forma geral, conforme descrito por Breiman (2001), o funcionamento da Floresta Aleatória pode ser resumido nos seguintes passos:

1. Definir o número total de árvores que comporão a floresta.
2. Para cada árvore do conjunto:

a) Gerar uma amostra aleatória com reposição a partir da base de dados original.

b) Construir uma árvore de decisão utilizando essa amostra.

c) Em cada nó da árvore:

   i. Selecionar aleatoriamente um subconjunto dos atributos disponíveis.

   ii. Avaliar as possíveis divisões apenas nesse subconjunto.

   iii. Escolher a divisão que maximiza o ganho de pureza.

3. Agregar as previsões das árvores:

   - Classificação: seleção da classe mais votada.
   - Regressão: média das previsões individuais.

2.1.2.2 XGBoost

O XGBoost é um modelo de predição baseado no conceito de impulsionamento (*boosting*), uma técnica de aprendizado de máquina na qual modelos simples são combinados de forma sequencial para formar um preditor mais preciso. Embora o termo *XGBoost* represente uma implementação moderna e otimizada, seu fundamento teórico está diretamente associado aos métodos de impulsionamento com árvores de decisão apresentados por Chen e Guestrin (2016).

Diferentemente dos métodos de conjunto baseados em agregação paralela, como a Floresta Aleatória, o impulsionamento constrói os modelos de maneira sequencial. O processo inicia-se com um modelo simples, geralmente uma constante, que fornece uma predição inicial para todas as observações. Em seguida, novas árvores são adicionadas uma a uma, cada qual ajustada para corrigir os erros cometidos pelo modelo atual.

Em cada iteração, calcula-se a diferença entre os valores reais observados e as previsões produzidas até aquele momento. Esses erros residuais são utilizados como variável resposta para o treinamento de uma nova árvore de decisão. Dessa forma, cada árvore subsequente concentra-se em regiões do espaço de atributos onde o modelo anterior apresentou maior dificuldade, permitindo um refinamento progressivo das previsões.

Um aspecto central do impulsionamento é o uso de um fator de redução, conhecido como taxa de aprendizado (*learning rate*). Esse parâmetro controla o quanto a contribuição de cada nova árvore influencia o modelo final. Valores menores tornam o aprendizado mais lento e gradual, porém aumentam a estabilidade do modelo e reduzem o risco de sobreajuste, especialmente quando combinados com um número maior de árvores.

Segundo Chen e Guestrin (2016), o desempenho do método depende fortemente de três parâmetros principais: o número total de árvores, a complexidade de cada árvore (controlada, por exemplo, pela profundidade máxima) e a taxa de aprendizado. O equilíbrio entre esses parâmetros permite construir modelos altamente expressivos sem comprometer a capacidade de generalização.

O modelo final obtido pelo impulsionamento com árvores consiste na soma ponderada das contribuições de todas as árvores geradas ao longo das iterações. Essa abordagem permite capturar relações não lineares complexas entre as variáveis explicativas e a variável resposta.

No contexto da predição de evasão estudantil, o *XGBoost* mostra-se adequado por sua capacidade de modelar interações complexas entre variáveis acadêmicas e socioeconômicas. Fatores como histórico de reprovações, desempenho ao longo do curso e características individuais do aluno podem interagir de maneira não trivial, sendo capturados eficientemente pelo modelo.

De acordo com Chen e Guestrin (2016), o funcionamento geral do método de impulsionamento com árvores pode ser descrito pelos seguintes passos:

1. Inicializar o modelo com uma predição simples, geralmente constante.
2. Para cada iteração do algoritmo:
    a) Calcular os resíduos entre os valores reais e as previsões atuais.
    b) Ajustar uma nova árvore de decisão utilizando os resíduos como variável resposta.
    c) Atualizar o modelo somando a contribuição da nova árvore, ponderada pela taxa de aprendizado.
3. Repetir o processo até atingir o número máximo de árvores ou um critério de parada.
4. Utilizar a soma das árvores ajustadas para gerar a predição final.

2.1.2.3 K-Vizinhos Mais Próximos

O algoritmo dos K-Vizinhos Mais Próximos (KNN) é um método de aprendizado supervisionado utilizado em tarefas de classificação e regressão, cujo princípio fundamental está baseado na similaridade entre observações. Conforme descrito por Fix e Jr. (1989), o KNN pertence à classe dos métodos baseados em instâncias, pois não constrói explicitamente um modelo durante uma fase de treinamento. Em vez disso, o processo de aprendizagem ocorre implicitamente por meio do armazenamento dos dados de treinamento.

A ideia central do KNN é que observações que apresentam características semelhantes tendem a possuir respostas semelhantes. Assim, para realizar uma predição, o algoritmo compara uma nova observação com todas as instâncias do conjunto de treinamento, avaliando o grau de proximidade entre elas. A decisão final é tomada com base nas observações mais próximas da instância de interesse.

A proximidade entre as observações é determinada por uma métrica de distância. A distância euclidiana é a mais comumente utilizada, especialmente quando os dados são quantitativos. Para duas observações $\vec{x} = (x_1, x_2, \ldots, x_p)$ e $\vec{y} = (y_1, y_2, \ldots, y_p)$, a distância euclidiana é definida pela Equação 2.1:

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2} \tag{2.1}$$

Fix e Jr. (1989) destacam que atributos em escalas distintas podem distorcer significativamente o cálculo da distância, fazendo com que variáveis com maior amplitude numérica dominem o processo de comparação. Por essa razão, a padronização ou normalização dos dados é uma etapa fundamental na aplicação do KNN.

Após o cálculo das distâncias, o algoritmo identifica os k exemplos do conjunto de treinamento mais próximos da nova observação. No caso de problemas de classificação, cada um desses vizinhos contribui com um voto para sua respectiva classe, e a classe mais frequente entre os k vizinhos é atribuída à nova instância. Em problemas de regressão, o valor predito corresponde à média dos valores observados entre os vizinhos selecionados.

A escolha do parâmetro k exerce influência direta no comportamento do modelo. Valores pequenos de k resultam em modelos mais flexíveis, altamente sensíveis a ruídos e variações locais nos dados. Por outro lado, valores maiores de k produzem decisões mais suavizadas, reduzindo a variância do modelo, porém aumentando o viés. Segundo Fix e Jr. (1989), a seleção de um valor adequado para k envolve um compromisso entre viés e variância, sendo frequentemente realizada por meio de validação cruzada.

Uma característica importante do KNN é a ausência de um processo de treinamento explícito. Embora isso torne o algoritmo conceitualmente simples, implica em um custo computacional elevado no momento da predição, pois é necessário calcular a distância entre a nova observação e todas as instâncias do conjunto de treinamento. Esse fator limita a aplicação do KNN em bases de dados muito grandes, a menos que sejam utilizadas técnicas de otimização, como estruturas de busca espacial ou redução de dimensionalidade.

No contexto da predição de evasão estudantil, o KNN permite comparar um aluno com outros estudantes que apresentam perfis acadêmicos e socioeconômicos semelhantes. A partir de informações como coeficiente de rendimento, histórico de reprovações, idade e características demográficas, o algoritmo pode inferir o risco de evasão com base nos padrões observados em alunos similares. Por exemplo, um estudante cujo histórico assemelha-se ao de outros alunos que abandonaram o curso tende a ser classificado como de maior risco, auxiliando na implementação de ações preventivas.

De acordo com Fix e Jr. (1989), o funcionamento geral do algoritmo dos k-Vizinhos Mais Próximos pode ser descrito pelos seguintes passos:

1. Pré-processar os dados, aplicando padronização ou normalização aos atributos.
2. Definir o número de vizinhos k e a métrica de distância a ser utilizada.

3. Para cada nova observação:

    a) Calcular a distância entre a observação e todas as instâncias do conjunto de treinamento.

    b) Selecionar os k exemplos mais próximos com base nas menores distâncias.

    c) Agregar as respostas dos vizinhos selecionados:

    - Classificação: atribuir a classe mais frequente entre os vizinhos.
    - Regressão: calcular a média dos valores dos vizinhos.

2.1.3 Medidas de performance para modelos de classificação

- **Acurácia:** métrica que representa a proporção de previsões corretas realizadas pelo modelo em relação ao total de previsões (vide Equação 2.2), sendo útil quando as classes estão balanceadas, mas pode ser enganosa em conjuntos de dados desbalanceados.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.2}$$

  Onde VP representa verdadeiro positivo, VN verdadeiro negativo, FP falso positivo e FN falso negativo.

- **Precisão:** razão entre o número de verdadeiros positivos e o total de elementos classificados como positivos por um modelo (vide Equação 2.3), medindo a confiabilidade das previsões positivas e sendo especialmente importante em contextos onde falsos positivos têm alto custo.

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{2.3}$$

- **Revocação:** razão entre o número de verdadeiros positivos e o total de elementos que pertencem à classe positiva (vide Equação 2.4), avaliando a capacidade do modelo de identificar corretamente os casos positivos e sendo crucial quando o custo de falsos negativos é elevado.

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{2.4}$$

- **F1-score:** métrica que corresponde à média harmônica entre precisão e revocação, proporcionando um equilíbrio entre ambas, sendo especialmente útil em cenários de classes desbalanceadas, pois avalia simultaneamente a capacidade do modelo de identificar corretamente os casos positivos e a confiabilidade dessas previsões.

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{2.5}$$

2.2 Trabalhos Relacionados

A evasão no ensino superior, especialmente nos cursos da área de tecnologia, tem sido um tema recorrente de pesquisas das últimas décadas, dada sua relevância social e impacto direto na eficiência das instituições de ensino. Diversos autores propuseram modelos preditivos com diferentes bases de dados, técnicas de mineração e objetivos específicos. Esta seção apresenta uma síntese de estudos relacionados, destacando seus métodos, objetivos e como seus resultados se articulam com o presente trabalho.

Uma das abordagens mais assertivas no uso de redes neurais artificiais foi aplicada por pesquisadores da Universidade do Estado do Amazonas (JESUS; RODRIGUEZ; JUNIOR, 2021), que estruturaram um modelo preditivo focado na evasão de alunos da Licenciatura em Computação. O estudo partiu da hipótese de que o histórico acadêmico, especialmente o desempenho em disciplinas práticas de programação, poderia ser decisivo para prever o risco de evasão. Utilizando uma rede neural de múltiplas camadas e dados extraídos de registros institucionais, o modelo alcançou 98% de precisão na classificação de alunos com risco de abandono, tornando-se uma referência de eficácia nesse tipo de modelagem.

Explorando o contexto do curso de Sistemas de Informação, a análise de Noetzold e Pertile (2021) conduzida na Universidade Federal de Santa Maria abordou a evasão a partir de fatores individuais e acadêmicos dos alunos. O estudo adotou técnicas de mineração de dados para tratar os registros institucionais e aplicou árvores de decisão como principal mecanismo preditivo. O trabalho revelou que reprovações e baixo desempenho foram elementos-chave na composição do perfil dos alunos evadidos, contribuindo com uma visão mais granular sobre os indicadores de risco e reforçando a importância de intervenções precoces.

Outra iniciativa que se destaca pelo rigor metodológico foi desenvolvida com base em dados de cursos presenciais de instituições públicas. A proposta de Kantorski et al. (2016) consistiu em aplicar diferentes algoritmos de aprendizado de máquina e combinar informações pessoais, socioeconômicas e de desempenho acadêmico. A metodologia foi inspirada no modelo CRISP-DM e validada com diferentes simulações. Como resultado, obteve-se uma acurácia de 98% na previsão geral e mais de 70% na identificação efetiva de evasões, evidenciando a eficácia de abordagens multivariadas para este tipo de problema.

Com um viés inovador, o estudo de Oliveira (2023), realizado no cenário da Universidade Federal da Paraíba, propôs o uso de dados de autoavaliação institucional como parâmetros para modelos de predição. Ao considerar variáveis subjetivas fornecidas pelos próprios alunos, como percepção sobre disciplinas, intenção de permanência e desempenho individual, o modelo conseguiu atingir métricas de desempenho bastante consistentes, com precisão e revocação superiores a 90%. Essa abordagem amplia o escopo tradicional da predição, incluindo aspectos comportamentais que muitas vezes escapam dos registros acadêmicos convencionais.

Um dos trabalhos que se destaca pela proposta de implementação prática é a criação de

uma ferramenta computacional de predição de evasão escolar ((REIS; SOUZA; SANTANA, 2018)), desenvolvida com foco na gestão institucional. A partir da extração de dados acadêmicos estruturados, o estudo aplicou algoritmos de aprendizado de máquina, como *Naive Bayes* e Floresta Aleatória, sobre atributos como reprovações, rendimento semestral e número de trancamentos. A aplicação do modelo foi integrada a um protótipo de *software web*, capaz de gerar relatórios de risco por discente, com o objetivo de auxiliar ações pedagógicas por parte das coordenações. Além da alta taxa de acurácia obtida (cerca de 91,8% com Floresta Aleatória), o diferencial está na viabilidade de uso da ferramenta em tempo real, aspecto que dialoga diretamente com a proposta deste trabalho ao sugerir soluções que extrapolam a análise quantitativa e projetam-se para a aplicação institucional concreta.

A análise conjunta desses trabalhos evidencia a diversidade de abordagens na predição da evasão universitária, reforçando a relevância da escolha criteriosa dos dados de entrada, da metodologia de modelagem e das métricas de avaliação. O presente trabalho busca contribuir com essa discussão por meio da construção de um modelo preditivo voltado especificamente à realidade dos cursos de tecnologia, considerando variáveis específicas desses cursos e buscando um equilíbrio entre acurácia e aplicabilidade institucional.

Portanto, considerando os trabalhos relacionados apresentados nesta seção e o trabalho proposto nesta pesquisa, a Tabela 2.1 apresenta um comparativo entre os estudos, destacando as características utilizadas em cada um. Esse levantamento foi feito a partir de critérios como coeficiente de rendimento, reprovação por falta, reprovação por nota, sexo, idade, cota e cidade de origem, com o objetivo de evidenciar semelhanças e diferenças nas variáveis analisadas.

Tabela 2.1 – Trabalhos relacionados e características consideradas

Autores	**Características**						
	Coeficiente de rendimento	**Reprovação por falta**	**Reprovação por nota**	**Sexo**	**Idade**	**Cota**	**Natural da cidade da faculdade**
JESUS; RODRIGUEZ; JUNIOR, 2021	X	X	X	X			
NOETZOLD; PERTILE, 2021	X		X	X			
KANTORSKI et al., 2016	X	X	X	X	X		
OLIVEIRA, 2023	X		X	X			
REIS; SOUZA; SANTANA, 2018	X		X	X			
Presente Trabalho	X	X	X	X	X	X	X

Fonte: Elaborado pela autora.

De acordo com a Tabela 2.1, observa-se que os trabalhos analisados utilizam diferentes combinações de variáveis para tratar a evasão estudantil. Nota-se que alguns estudos exploram apenas um subconjunto limitado dessas características, enquanto outros ampliam a análise para múltiplos aspectos acadêmicos e sociodemográficos. O presente trabalho, entretanto, diferencia-se ao integrar todas as variáveis identificadas, oferecendo uma abordagem mais abrangente para a construção e validação de modelos preditivos, o que potencialmente aumenta a precisão e a utilidade prática da estratégia proposta.

3 Desenvolvimento

Como visto, este trabalho possui, como objetivo principal, a proposta, o desenvolvimento e a validação de uma estratégia para predição do risco de evasão de alunos em áreas de tecnologia, baseada em informações acadêmicas e sociodemográficas reais do curso de Ciência da Computação da Universidade Federal de Ouro Preto. Este capítulo descreve a metodologia adotada em duas etapas complementares: (i) treinamento, validação e comparação de três modelos supervisionados (Floresta Aleatória, *XGBoost* e K-Vizinhos Mais Próximos (KNN)) a partir de dados rotulados (vide Seção 3.1); e (ii) estabelecimento da estratégia de predição operacional, que aplica o modelo de melhor eficácia para estimar a probabilidade de evasão de um aluno em particular (vide Seção 3.2).

3.1 Treinamento, validação e comparação dos modelos

A Figura 3.1 ilustra a arquitetura de funcionamento para treinamento, validação e comparação dos modelos Floresta Aleatória, *XGBoost* e K-Vizinhos Mais Próximos desde a coleta e o pré-processamento dos dados até a aplicação de modelos de aprendizado de máquina, sendo capaz de estimar o percentual da evasão de alunos com dados rotulados. Os modelos foram avaliados com base nas métricas de acurácia, precisão, revocação e *F1-score*.

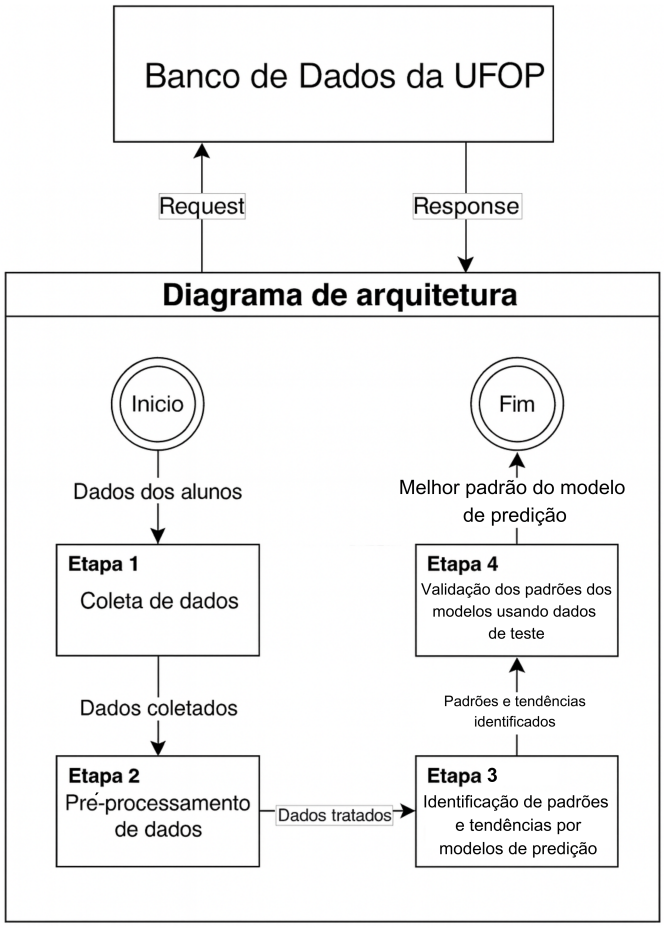

Figura 3.1 – Arquitetura de funcionamento para seleção do modelo.

De acordo com a Figura 3.1, o funcionamento da estratégia de predição proposta é composto por quatro etapas: (a) a primeira etapa, descrita na Subseção 3.1.1, consiste em coletar dados acadêmicos e sociodemográficos reais dos alunos por meio do acesso ao Banco de Dados da UFOP, gerando, para a próxima etapa, um conjunto de dados coletados; (b) a segunda etapa, apresentada na Subseção 3.1.2, consiste em realizar o pré-processamento desses dados, produzindo um conjunto de dados tratados; (c) a terceira etapa, descrita na Subseção 3.1.3, consiste em aplicar os três modelos estabelecidos, considerando os dados tratados advindos da etapa anterior, no intuito de identificar padrões e tendências; (d) a quarta etapa, apresentada na Subseção 3.1.4, consiste em validar os padrões e tendências identificados por meio de um conjunto de dados de teste para estabelecer o modelo que apresentou uma eficácia mais satisfatória como parte da estratégia para predição da evasão de um aluno na área de tecnologia.

3.1.1 Coleta de dados

Os dados utilizados neste trabalho foram obtidos a partir de registros acadêmicos reais do curso de Ciência da Computação da Universidade Federal de Ouro Preto, com todos os estudantes computados pelo censo. Foram coletados a partir de um Pedido de Acesso à Informação (NUP nº 23546.100776/2025-23) feito em 12/09/2025 pelo FALABR, ferramenta gratuita disponível no site do Governo Federal (gov.br), e a solicitação foi atendida em 15/12/2025. A base contém informações relacionadas ao desempenho acadêmico e dados pessoais dos alunos, sendo os atributos de interesse: (a) Coeficiente de Rendimento (CR); (b) reprovação por falta; (c) reprovação por nota; (d) sexo; (e) idade; (f) se ingressou por meio de cotas; e (g) se é natural da cidade onde está localizada a universidade.

Com o Coeficiente de Rendimento (CR), busca-se avaliar o desempenho global do aluno ao longo de todas as disciplinas cursadas, considerando a média de suas notas. Essa abordagem é importante para que a análise não fique restrita a matérias específicas nas quais o estudante possa apresentar mais facilidade ou dificuldade, já que a aptidão varia de indivíduo para indivíduo. Dessa forma, o CR funciona como um indicador amplo de engajamento e desempenho acadêmico.

Com a reprovação por falta, observa-se quantas disciplinas o estudante deixou de frequentar ao ponto de ultrapassar o limite máximo de ausências permitido pela instituição. Esse indicador revela situações em que o aluno, por questões pessoais, desmotivação, problemas de transporte ou outras circunstâncias, deixou de comparecer às aulas, resultando em reprovação automática, independentemente do seu rendimento acadêmico.

Já a reprovação por nota permite identificar disciplinas nas quais o aluno, mesmo tendo frequência suficiente, não alcançou a nota mínima exigida para aprovação, que no caso é 6,0. Esse dado indica deficiências no aprendizado ou dificuldades com o conteúdo, metodologia de ensino ou avaliação, o que pode contribuir para um processo de desmotivação e, em casos prolongados, para a evasão.

O atributo sexo é utilizado para investigar se existem diferenças significativas nos índices de evasão entre homens e mulheres, possibilitando identificar padrões ou desigualdades que possam estar relacionados a questões socioculturais, estereótipos de gênero ou condições específicas enfrentadas por cada grupo dentro do ambiente acadêmico.

A idade é analisada para verificar se a faixa etária exerce influência direta na probabilidade de evasão, considerando que alunos mais jovens podem enfrentar desafios diferentes dos mais velhos, como adaptação à vida universitária, responsabilidades familiares ou equilíbrio entre estudos e trabalho.

A variável ingresso por meio de cotas serve para verificar se estudantes que ingressaram por meio de políticas afirmativas enfrentam obstáculos adicionais que possam impactar na permanência, como lacunas de formação prévia, barreiras socioeconômicas ou falta de suporte institucional, aumentando ou não o risco de evasão.

Por fim, a naturalidade do aluno, que se refere à sua cidade ou estado de origem, permite avaliar se o fato de estudar fora do seu local de residência habitual, exigindo deslocamento, adaptação cultural ou mudança de ambiente, tem relação com a probabilidade de desistência do curso. Essa análise também poderá indicar se estudantes locais apresentam maior permanência devido à rede de apoio e familiaridade com a região.

3.1.2 Pré-processamento dos dados

Antes da aplicação de um algoritmo de predição, é necessário realizar o pré-processamento dos dados, etapa fundamental para garantir que os resultados obtidos sejam confiáveis e representem adequadamente a realidade do problema estudado. De acordo com Zelaya (2019), muitas das decisões que influenciam diretamente o comportamento preditivo de um modelo de aprendizado de máquina são tomadas nessa fase inicial, pois é no pré-processamento que se definem as transformações e ajustes nos dados que serão usados na modelagem. Assim, um conjunto de dados não tratados adequadamente pode levar o modelo a apresentar vieses, perda de desempenho ou interpretações equivocadas.

O pré-processamento contribui para reduzir ruídos, lidar com lacunas, tornar as variáveis comparáveis e evitar que certos atributos tenham peso desproporcional nos cálculos. Além disso, segundo o estudo de Zelaya (2019), mudanças no pré-processamento podem alterar significativamente as previsões para determinados indivíduos, demonstrando que essa etapa não é apenas preparatória, mas parte integrante da lógica de decisão do modelo.

Neste trabalho, as etapas de pré-processamento foram implementadas através de um fluxo automatizado, compreendendo:

- **Padronização Textual e Limpeza**: foi realizada a remoção de acentos e caracteres especiais, além da conversão de strings para caixa alta em colunas como "sexo"e "naturalidade". Essa etapa garante a uniformidade dos dados e evita que variações ortográficas sejam interpretadas como categorias distintas.
- **Engenharia de Atributos (Feature Engineering)**: para os atributos de reprovação, desenvolveu-se uma lógica de contagem para transformar registros textuais de disciplinas em valores inteiros. Além disso, criou-se a variável binária "naturalidade_op"para identificar especificamente alunos naturais da sede da instituição.
- **Tratamento de valores faltantes**: utilizou-se a estratégia de imputação por média e mediana para preencher lacunas em variáveis numéricas, assegurando que o modelo pudesse processar o conjunto completo de instâncias sem descartar registros relevantes.
- **Codificação Categórica**: aplicou-se a técnica de *One-Hot Encoding* para converter variáveis qualitativas, como "cota"e "sexo", em representações numéricas binárias compatíveis com os algoritmos de aprendizado de máquina.

- **Normalização**: utilizou-se o *StandardScaler* para reescalonar os atributos numéricos (CR e idade), garantindo que todos possuíssem média zero e desvio padrão unitário. Essa prática é crucial para algoritmos sensíveis à escala, como o KNN, prevenindo distorções no cálculo de distâncias entre instâncias.

Ao adotar essas etapas, aumentou-se a precisão dos modelos e garantiu-se maior transparência no processo, tal como enfatizado no estudo de Zelaya (2019), no qual se verificou que ajustes feitos nesta fase podem modificar de maneira significativa as classificações finais.

3.1.3 Identificação de padrões e tendências por modelos de predição

Após o pré-processamento, a próxima etapa consistiu na identificação de padrões e tendências nos dados tratados, com o objetivo de desenvolver a estratégia de predição capaz de prever, com precisão, a probabilidade de evasão de alunos das áreas de tecnologia. Essa fase foi fundamental para transformar dados brutos e tratados em conhecimento estruturado, permitindo que os modelos aprendessem relações complexas entre variáveis independentes (atributos) e a variável dependente, o risco de evasão.

Para esta tarefa, foram empregados três algoritmos de aprendizado supervisionado: K-Vizinhos Mais Próximos (KNN), *XGBoost* e Floresta Aleatória. A escolha desses algoritmos justificou-se por sua complementaridade na análise de dados de classificação e pela robustez demonstrada em problemas educacionais semelhantes. Em substituição ao planejamento inicial que previa o uso do Naive Bayes, optou-se pelo XGBoost devido à sua superior eficiência em lidar com dados tabulares complexos e desbalanceados através do empilhamento sequencial de árvores de decisão.

O KNN foi utilizado por sua capacidade de capturar padrões baseados em proximidade entre instâncias, identificando perfis de alunos semelhantes quanto a desempenho e características pessoais. Para este modelo, a busca exaustiva por parâmetros via *GridSearchCV* identificou que a melhor configuração envolvia o uso de 7 vizinhos próximos e a métrica de distância euclidiana.

O *XGBoost* foi selecionado por sua alta performance em capturar relações não lineares complexas. Por meio da otimização realizada no código, estabeleceu-se o uso de 300 estimadores e uma taxa de aprendizado (*learning rate*) de 0,05, garantindo que o modelo realizasse um refinamento progressivo das predições para minimizar os erros residuais de iterações anteriores.

Por fim, o Floresta Aleatória foi aplicado por sua robustez e capacidade de reduzir o risco de sobreajuste ao combinar múltiplas árvores de decisão independentes. Os hiperparâmetros otimizados para este modelo incluíram o uso de 100 árvores (estimadores) e uma profundidade máxima de 10 níveis para cada árvore, assegurando estabilidade nas previsões e permitindo a análise da importância de cada variável no resultado final.

O treinamento de cada modelo foi realizado utilizando o conjunto de dados tratados e a

técnica de validação cruzada estratificada (*Stratified K-Fold*), buscando identificar padrões que correlacionassem os atributos analisados com o risco de evasão. Os modelos resultantes foram, então, preparados para a etapa de validação final e aplicação operacional.

3.1.4 Validação dos padrões dos modelos usando dados de teste

A última etapa do processo envolveu a validação dos padrões dos modelos de predição e sua aplicação no conjunto de teste, com o intuito de avaliar sua capacidade de generalização e estimar o percentual de risco de evasão do aluno. Essa etapa foi fundamental para garantir que a eficácia obtida durante o treinamento não fosse fruto de sobreajuste e que o modelo fosse capaz de oferecer previsões consistentes para novos dados. Para isso, a base de dados foi dividida de forma que 20% dos registros fossem destinados exclusivamente ao teste, garantindo uma avaliação imparcial.

As métricas de avaliação utilizadas foram: acurácia, precisão, revocação e F1-score. A acurácia mediu a proporção de previsões corretas em relação ao total de casos. A precisão indicou a confiabilidade do modelo ao classificar um aluno como evadido, enquanto a revocação mediu a capacidade de identificar todos os casos reais de evasão, sendo essencial para evitar que alunos em risco passassem despercebidos. Por fim, a *F1-score*, calculada como a média harmônica entre precisão e revocação, foi a métrica central para o ajuste dos modelos, dada sua relevância em cenários de desbalanceamento de classes, como o da evasão estudantil na UFOP.

Após a validação, o modelo com a melhor eficácia média foi selecionado para integrar a estratégia de predição operacional. O padrão estatístico gerado por este modelo permite estimar o percentual de risco de cada discente, servindo de base para intervenções preventivas, como suporte acadêmico ou encaminhamentos pedagógicos, contribuindo para o aumento dos índices de formação nas áreas de tecnologia.

3.2 Estabelecimento da estratégia de predição operacional

Nesta seção, é apresentada a estratégia de predição proposta. Uma vez validado e estabelecido o padrão do modelo de melhor eficácia na etapa anterior (vide Seção 3.1), pode-se aplicar o fluxo operacional da estratégia de predição. Nesse fluxo (vide Figura 3.2), cada novo aluno é tratado individualmente: seus dados acadêmicos e sociodemográficos são submetidos ao mesmo processo de pré-processamento aplicado durante o treinamento e, em seguida, os dados tratados do aluno são usados como entrada para o modelo estabelecido. O resultado obtido é a probabilidade de evasão associada àquele aluno, expressa em percentual. Essa abordagem permite a utilização prática da estratégia em cenários reais, fornecendo oportunidade de intervenções preventivas direcionadas caso a caso.

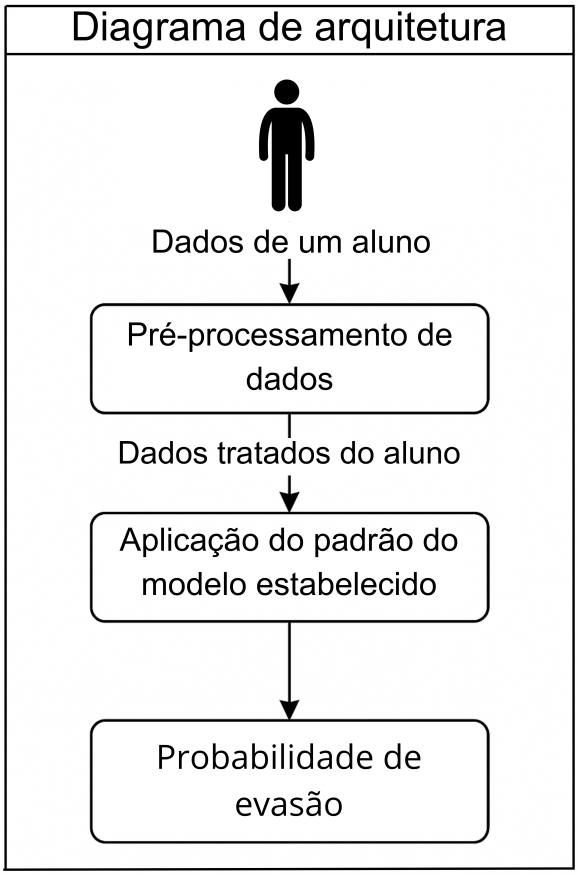

Figura 3.2 – Arquitetura de funcionamento da estratégia de predição proposta.

De acordo com a Figura 3.2, o fluxo operacional da estratégia de predição é composto por quatro etapas: (a) a primeira etapa, descrita na Subseção 3.2.1, consiste em receber como entrada os dados acadêmicos e sociodemográficos de um único aluno; (b) a segunda etapa, apresentada na Subseção 3.2.2, realiza o pré-processamento dessas informações, aplicando as mesmas transformações utilizadas durante o treinamento, de modo a gerar dados tratados consistentes com o modelo; (c) a terceira etapa, descrita na Subseção 3.2.3, aplica o padrão do modelo previamente estabelecido como o de melhor eficácia entre os avaliados; e (d) a quarta etapa, apresentada na Subseção 3.2.4, gera como saída a probabilidade de evasão do aluno, expressa em percentual, indicando que valores próximos a 100% representem alta chance de evasão e valores próximos a 0% representem baixa probabilidade.

3.2.1 Entrada de dados de um aluno

A primeira etapa do fluxo operacional da estratégia consiste em receber como entrada os dados acadêmicos e sociodemográficos de um único aluno. Essas informações incluem os mesmos sete atributos definidos durante o treinamento dos modelos: coeficiente de rendimento (CR), reprovações por falta, reprovações por nota, sexo, idade, ingresso por cotas e naturalidade, que devem ter o tipo indicado na Figura 3.3. A utilização dos mesmos atributos garante a consistência

entre as fases de treinamento e predição, permitindo que o modelo aplique os padrões estatísticos previamente aprendidos sobre a nova instância.

```
# predict.py
"""
Módulo: predict
Responsabilidade: Carregar o pipeline final (RandomForest) e, dado um aluno de entrada,
retornar a probabilidade (%) de evasão e uma classificação por limiar.

Compatível com o pipeline salvo em:
../outputs/modelos_salvos/randomforest_model.joblib

Entrada esperada (features):
- sexo: "M" ou "F" (string)
- idade: int
- tipo_de_cota: ex "AC" (string)
- coeficiente: float
- naturalidade_op: 0 ou 1 (int)  -> 1 se nasceu em Ouro Preto, 0 caso contrário
- total_reprovacao_nota: int
- total_reprovacao_falta: int
"""
```

Figura 3.3 – Tipos de dados dos atributos de entrada no módulo predict.py.

No sistema desenvolvido, essa entrada é estruturada de forma que o usuário (professor ou gestor) forneça os dados brutos, que são então organizados em um objeto de dados (como um dicionário). Para exemplificar o funcionamento operacional, a inserção de um aluno segue o exemplo ilustrado na Figura 3.4.

```
(venv) marcellasilveiracampos@MacBook-Air-de-Marcella src % python predict.py
 --sexo F --idade 22 --tipo_de_cota AC --coeficiente 7.5 --naturalidade_op 1
--total_reprovacao_nota 0 --total_reprovacao_falta 0
```

Figura 3.4 – Demonstração da inserção de dados brutos de um aluno fictício no terminal do sistema.

Nessa etapa, o objetivo é padronizar o ponto de partida para cada aluno analisado, assegurando que todas as previsões sejam realizadas a partir de um conjunto de informações homogêneo e comparável ao que foi utilizado para a construção do modelo. Assim, é possível tratar o caso de cada discente de maneira individual, respeitando suas particularidades, mas mantendo a estrutura de análise rigorosa do sistema.

3.2.2 Pré-processamento dos dados do aluno

A segunda etapa consiste em aplicar o pré-processamento ao conjunto de informações do aluno no momento da consulta, garantindo que os dados de entrada estejam em conformidade

com o formato esperado pelo modelo final. Essa etapa é realizada de forma transparente, pois o modelo é carregado como um objeto serializado que já contém o *Pipeline* de transformações ajustado durante o treinamento.

Assim, os dados fornecidos pelo usuário são submetidos aos seguintes processos de tratamento imediato:

- **Alinhamento de Atributos**: o *script* organiza os dados de entrada (como idade, CR e totais de reprovação) em uma estrutura de dados (*DataFrame*) que respeita a ordem e o nome exato das colunas utilizadas na fase de aprendizado.
- **Transformação via Pipeline**: os dados nominais, como sexo e tipo de cota, são processados automaticamente pelo transformador de colunas (*ColumnTransformer*) embutido no modelo. Isso garante que a codificação *One-Hot Encoding* aplicada ao novo aluno seja idêntica à utilizada na base histórica.
- **Reescalonamento**: os valores numéricos de CR e idade passam pela normalização estatística (*StandardScaler*) contida no modelo salvo, assegurando que o novo dado seja comparável à escala de treinamento sem a necessidade de reprocessar toda a base de dados.

Dessa forma, a estratégia operacional garante que as previsões não sejam afetadas por inconsistências de formato na entrada individual, permitindo que o sistema responda de maneira ágil e precisa a cada nova consulta realizada.

3.2.3 Aplicação do padrão do modelo selecionado

Na terceira etapa, os dados tratados do aluno são inseridos no modelo previamente estabelecido como o de melhor eficácia durante a fase de comparação. No *script*, essa operação é iniciada pelo carregamento do arquivo serializado, que contém não apenas os pesos do algoritmo, mas todo o conhecimento estatístico extraído da base histórica da UFOP.

A aplicação prática consiste em submeter os dados do aluno ao método `predict_proba` do modelo selecionado. Diferente de uma classificação rígida, este método é responsável por processar as características do estudante e calcular a probabilidade estatística de ele pertencer à classe de evasão, baseando-se nas relações complexas identificadas durante o treinamento.

O uso de um padrão já validado assegura maior confiabilidade no processo preditivo, visto que a escolha do modelo foi fundamentada em métricas rigorosas de acurácia e precisão.

3.2.4 Predição da evasão do aluno

A quarta e última etapa da estratégia consiste na geração da saída do modelo, que corresponde à probabilidade de evasão associada ao aluno em questão. No sistema implementado,

o valor bruto de probabilidade gerado pelo algoritmo é convertido em uma escala centesimal. Esse resultado é expresso em percentual, permitindo uma interpretação direta e prática: valores próximos a 100% indicam elevada chance de evasão, enquanto valores próximos a 0% sugerem baixa probabilidade.

Esse percentual funciona como um indicador objetivo para orientar a tomada de decisão por parte da instituição de ensino, possibilitando a implementação de estratégias preventivas e personalizadas de intervenção. A interpretação clara e intuitiva dos resultados favorece sua aplicação em cenários reais, oferecendo aos gestores e professores um recurso adicional no acompanhamento da permanência estudantil e na priorização de suporte pedagógico de acordo com o nível de risco indicado pelo modelo.

4 Experimentação Prática

Este capítulo apresenta a experimentação prática e a análise dos resultados obtidos pela estratégia proposta. O objetivo central é validar a eficácia da estratégia proposta de predição de evasão, utilizando o conjunto de dados reais fornecido pela UFOP, consolidando as etapas de treinamento e comparação discutidas anteriormente. Para tanto, este capítulo está organizado em quatro partes principais: (i) o detalhamento do ambiente de desenvolvimento e das configurações experimentais (vide Seção 4.1); (ii) a caracterização do conjunto de dados reais e o processo de particionamento de dados (vide Seção 4.2); e (iii) a apresentação comparativa do desempenho dos modelos Floresta Aleatória, *XGBoost* e KNN (vide Seção 4.3).

4.1 Ambiente de Desenvolvimento

A implementação da estratégia foi realizada utilizando a linguagem de programação *Python* (versão 3.10), escolhida por sua robustez e vasta disponibilidade de bibliotecas voltadas ao aprendizado de máquina. Para assegurar a reprodutibilidade dos experimentos e evitar conflitos entre dependências, utilizou-se um ambiente virtual para o isolamento dos pacotes.

Dentre as principais ferramentas utilizadas, destacam-se: (i) *Pandas* e *NumPy* para a manipulação e tratamento das matrizes de dados; (ii) *Scikit-Learn* para a implementação do *pipeline* de pré-processamento dos algoritmos Floresta Aleatória e KNN; e (iii) a biblioteca *XGBoost* para a execução do algoritmo de *gradient boosting*. Todo o desenvolvimento foi conduzido em ambiente *Linux*, utilizando o editor *Visual Studio Code*.

4.2 Configuração do Experimento

Para realização do experimento de validação da estratégia proposta, foi gerada uma base de dados reais composta por registros anônimos de históricos de discentes do curso de Ciência da Computação da UFOP. Tais registros de históricos foram obtidos por meio de uma solicitação formal ao FALABR, disponível no site do Governo Federal (*gov.br*). A base de dados final, após as etapas de limpeza e engenharia de características, totalizou 1.962 instâncias, sendo 1.098 pertencentes à classe de não evasão e 864 pertencentes à classe de evasão. A Figura 4.1 ilustra algumas dessas instâncias.

id_aluno	sexo	idade	tipo_de_cota	coeficiente	naturalidade_op	total_reprovacao_nota	total_reprovacao_falta	target_evasao
1	M	47	AC	5.6	0	0	0	0
2	F	21	AC	0.0	0	0	0	0
3	F	19	LB_PPI	6.0	0	0	0	0
4	M	29	LB_EP	0.0	0	0	0	0
5	M	20	LB_PPI	0.0	0	0	0	0
6	M	19	AC	0.0	0	0	0	0
7	F	20	LB_EP	0.0	0	0	0	0
8	F	19	LI_PPI	0.0	0	0	0	0
9	M	23	LI_EP	0.0	0	0	0	0
10	M	20	LI_EP	0.0	0	0	0	0
11	F	19	AC	0.0	0	0	0	0
12	M	28	LI_PPI	0.0	0	0	0	0
13	M	21	LI_EP	0.0	0	0	0	0
14	M	23	AC	0.0	0	0	0	0
15	M	19	AC	0.0	0	0	0	0
16	M	21	AC	0.0	0	0	0	0
17	M	20	LB_EP	0.0	0	0	0	0
18	F	20	LB_EP	0.0	0	0	0	0
19	M	22	LB_EP	0.0	0	0	0	0
20	M	23	LB_PPI	0.0	0	0	0	0

Figura 4.1 – Exemplos de instâncias após o pré-processamento dos dados reais.

Para a avaliação dos modelos, aplicou-se a técnica de *hold-out*, particionando o conjunto original em 80% para o treinamento dos algoritmos e 20% para o teste definitivo, como visto em Kohavi et al. (1995) que tal divisão atua na prevenção do superajuste (*overfitting*). Ao reservar uma fatia robusta de 80% para o ajuste dos parâmetros dos classificadores, preserva-se uma quantidade adequada de registros (20%) totalmente isolada do processo de modelagem, simulando com maior fidelidade o comportamento do sistema diante de novos discentes inseridos na instituição. Adicionalmente, durante a fase de ajuste de hiperparâmetros, empregou-se a validação cruzada estratificada (*Stratified K-Fold*) com $k = 5$. Esta abordagem é fundamental devido ao desbalanceamento inerente ao problema da evasão, garantindo que cada dobra (*fold*) mantivesse a proporção original entre alunos evadidos e não evadidos, evitando vieses na métrica de *F1-score*.

4.3 Resultados

O desempenho dos modelos foi avaliado com base em quatro métricas fundamentais: Acurácia, Precisão, Revocação e *F1-Score*, conforme descritas na Subseção 2.1.3. As tabelas 4.1 e 4.2 apresentam os resultados obtidos nos dados de teste. Os melhores resultados obtidos encontram-se em negrito.

Tabela 4.1 – Desempenho comparativo da Classe 0 (não evasão) dos modelos com dados reais da UFOP.

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
KNN	0,8697	0,8576	0,9200	0,8877
XGBoost	0,8737	0,8873	0,8873	0,8873
Floresta Aleatória	**0,8900**	**0,8932**	**0,9127**	**0,9029**

Tabela 4.2 – Desempenho comparativo da Classe 1 (evasão) dos modelos com dados reais da UFOP.

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
KNN	0,8697	0,8878	0,8056	0,8447
XGBoost	0,8737	0,8565	0,8565	0,8565
Floresta Aleatória	**0,8900**	**0,8857**	**0,8611**	**0,8732**

Os resultados experimentais obtidos com dados reais da UFOP, consolidados nas Tabelas 4.1 e 4.2, revelam aspectos fundamentais sobre a capacidade preditiva dos modelos e a natureza do fenômeno da evasão no curso de Ciência da Computação. A análise comparativa demonstra que todos os algoritmos alcançaram patamares elevados de assertividade, com acurácias globais variando entre 86,97% e 89,00%, o que valida a relevância do conjunto de variáveis acadêmicas e sociodemográficas selecionadas no Capítulo 3.

O algoritmo KNN apresentou uma acurácia global de 86,97%. Ao analisar seu comportamento, observa-se que o modelo tende a apresentar um desempenho ligeiramente superior na identificação da classe de risco (Classe 1), na qual obteve uma precisão de 0,8878 e um *F1-score* de 0,8447, em comparação com o *F1-score* de 0,8877 obtido na Classe 0. Contudo, o KNN registrou a menor revocação do experimento para a classe de evasão (0,8056). Como o KNN baseia-se na similaridade geométrica calculada a partir da distância euclidiana, esse comportamento indica que, embora o reescalonamento efetuado pelo *StandardScaler* tenha aproximado os perfis dos discentes, a fronteira de decisão do algoritmo foi influenciada pela densidade dos dados.

O modelo *XGBoost* exibiu um comportamento perfeitamente simétrico entre as classes, registrando métricas idênticas de precisão, revocação e *F1-score* (0,8873 na Classe 0 e 0,8565 na Classe 1) para uma acurácia global de 87,37%. Essa estabilidade decorre da sua arquitetura de impulsionamento sequencial (*gradient boosting*), configurada com 300 estimadores e uma taxa de aprendizado de 0,05, que força o otimizador a minimizar os resíduos de classificação de forma equilibrada. No entanto, o *XGBoost* demandou um tempo de processamento massivamente superior durante a etapa de busca exaustiva por parâmetros via *GridSearchCV*, sem conseguir superar a eficiência geral da Floresta Aleatória.

A Floresta Aleatória consolidou-se como o classificador de melhor desempenho geral do estudo, atingindo a acurácia máxima de 89,00%. O modelo demonstrou alta robustez ao registrar o maior *F1-score* para a Classe 0 (0,9029), impulsionado por uma revocação de 0,9127. Na Classe 1, o algoritmo garantiu um *F1-score* de 0,8732 e uma revocação de 0,8611. Essa

capacidade de generalização justifica-se por sua estrutura baseada na combinação de 100 árvores de decisão independentes cultivadas com amostragem por *bootstrap*. Ao restringir a profundidade máxima das árvores em 10 níveis, o modelo conseguiu mitigar os riscos de sobreajuste comuns a bases de dados institucionais complexas. Adicionalmente, o mecanismo de seleção aleatória de subconjuntos de atributos a cada divisão de nó permitiu capturar o impacto combinado e não-linear de características heterogêneas, como a correlação entre as taxas de reprovação por nota, o coeficiente de rendimento (CR) e fatores sociodemográficos (como o tipo de cota e a naturalidade do discente). No contexto de políticas de permanência estudantil, o valor de 0,8611 na revocação é particularmente expressivo, pois indica que a estratégia é capaz de identificar corretamente a grande maioria dos alunos que efetivamente estão em risco de abandonar o curso.

Sob a perspectiva da aplicabilidade prática na gestão universitária, os resultados detalhados por classe nas Tabelas 4.1 e 4.2 revelam um balanço seguro entre precisão e revocação para a Classe 1 (Evasão), considerando todos os modelos avaliados. Em sistemas de alerta precoce educacionais, minimizar os falsos negativos (alunos em risco de abandono que o sistema não consegue detectar) é tradicionalmente mais crítico do que tolerar pequenos índices de falsos positivos. No entanto, para que uma estratégia institucional seja sustentável, é indispensável analisar o *F1-score*, métrica que sintetiza o equilíbrio harmônico entre a sensibilidade de detecção e a assertividade das indicações. Nesse cenário, o algoritmo Floresta Aleatória consolidou-se como a escolha ideal ao atingir um *F1-score* de 87,32% para a classe de evasão. Esse índice expressivo demonstra que, além de assegurar uma alta capacidade de identificação ativa (revocação de 86,11%), o modelo mitiga o desperdício de recursos administrativos e pedagógicos com alarmes falsos. Dessa forma, a estratégia operacional proposta valida-se como altamente confiável e balanceada, fornecendo aos coordenadores e docentes da UFOP dados precisos para disparar intervenções pedagógicas e políticas de acolhimento direcionadas antes que o desligamento definitivo do estudante se concretize.

5 Considerações Finais

Este capítulo apresenta as considerações finais sobre o trabalho desenvolvido. Para tanto, este capítulo está organizado em duas partes principais: (i) as conclusões obtidas a partir do desenvolvimento da pesquisa e da avaliação comparativa dos modelos preditivos (vide Seção 5.1); e (ii) as perspectivas de trabalho futuro, com foco no delineamento de ações institucionais e pedagógicas (vide Seção 5.2).

5.1 Conclusões

Este trabalho apresenta o desenvolvimento e a avaliação de uma estratégia operacional para a predição precoce da evasão estudantil no curso de Ciência da Computação da Universidade Federal de Ouro Preto (UFOP). O fenômeno do abandono acadêmico gera severos impactos sociais, institucionais e econômicos, demandando ferramentas analíticas robustas que auxiliem os gestores públicos na tomada de decisões preventivas. A abordagem proposta combinou técnicas de engenharia de características com algoritmos de aprendizado de máquina supervisionado, validando sua eficácia diretamente em uma base de dados reais composta por registros históricos da instituição.

Os experimentos práticos permitiram realizar uma análise comparativa criteriosa entre os algoritmos KNN, *XGBoost* e Floresta Aleatória. Diante dos resultados, a Floresta Aleatória consolidou-se como o modelo preditivo mais eficiente para a estratégia, alcançando a acurácia máxima de 89,00%. Ao avaliar o comportamento segregado por classes, o modelo campeão demonstrou alta capacidade de generalização ao registrar um *F1-score* de 0,9029 para a classe de permanência (Classe 0) e 0,8732 para a classe de evasão (Classe 1). No contexto da gestão universitária, a revocação de 86,11% alcançada na Classe 1 valida o modelo como um sistema de alerta precoce seguro, capaz de interceptar a grande maioria dos discentes em situação de risco iminente antes do desligamento definitivo.

A engenharia de características desempenhou um papel central no sucesso da estratégia. O mapeamento automatizado de interações multifatoriais e não-lineares, tais como o Coeficiente de Rendimento (CR), as taxas específicas de reprovação por nota e por falta e as variáveis de perfil sociodemográfico, como a modalidade de ingresso por cotas e a naturalidade, provou ser um indicativo consistente. A discrepância observada nos modelos de linha de base, como o KNN, evidenciou que fronteiras geométricas simples são insuficientes para lidar com as nuances e o desbalanceamento dos dados institucionais, consolidando a arquitetura de árvores independentes da Floresta Aleatória como a solução ideal.

A pesquisa entrega à comunidade acadêmica da UFOP não apenas uma análise compara-

tiva rigorosa, mas também uma estratégia operacional computacionalmente viável e validada, apta a subsidiar políticas institucionais de acolhimento e retenção de alunos com embasamento científico de dados.

5.2 Trabalhos Futuros

Como desdobramento natural desta pesquisa, identificam-se oportunidades relevantes para a continuidade e expansão do estudo. Como principal trabalho futuro, propõe-se o estudo e o desenvolvimento de ferramentas para intervenção pedagógica direcionadas especificamente a alunos identificados com alta probabilidade de evasão, de acordo com a estratégia aqui estabelecida. Uma vez que o modelo matemático cumpre o papel de diagnosticar o risco de forma precoce, torna-se imperativo estruturar planos de ação institucionais que traduzam esses alertas em ações práticas de retenção, tais como:

- desenhar programas de monitoria e nivelamento acadêmico customizados para os perfis críticos identificados (por exemplo, focando em disciplinas que historicamente causam o acúmulo de reprovações por nota);
- o auxílio aos professores em identificar e atuar preventivamente com os alunos em risco de evasão, objetivando o aumento da formação em cursos de graduação nas áreas de tecnologia;
- propor fluxos de acolhimento psicopedagógico automatizados integrados aos sistemas de gestão da UFOP, disparando alertas confidenciais para as coordenações de curso sempre que o limiar de risco de um discente for ultrapassado;
- avaliar o impacto a longo prazo das intervenções realizadas, criando um ciclo de retroalimentação em que as ações pedagógicas bem-sucedidas ajudem a atualizar e recalibrar os pesos das variáveis do modelo preditivo.

Adicionalmente, quanto à estratégia proposta para predição de evasão propriamente dita, sugere-se a expansão da base de dados com a inclusão de novas dimensões informacionais, como o acompanhamento da frequência em tempo real ao longo do semestre letivo e indicadores de engajamento em plataformas de apoio ao ensino (Ambientes Virtuais de Aprendizagem, como o TôSabendo[1]). Por fim, vislumbra-se a aplicação e validação desta mesma metodologia em outros cursos de graduação da universidade, avaliando a capacidade de transferência e generalização do modelo para diferentes áreas do conhecimento técnico e humanístico.

[1] De acordo com França et al. (2021), o TôSabendo é uma plataforma baseada em *quizzes* desenvolvida para criar experiências envolventes de ensino e aprendizagem no Ensino Superior, utilizando elementos de gamificação para aumentar o engajamento dos estudantes.

Referências

ABBAD, G.; CARVALHO, R. S.; ZERBINI, T. Evasão em curso via internet: explorando variáveis explicativas. *RAE eletrônica*, SciELO Brasil, v. 5, 2007.

ALMEIDA, L. S.; SOARES, A. P.; FERREIRA, J. A. G. Questionário de vivências acadêmicas (qva-r): Avaliação do ajustamento dos estudantes universitários. Instituto Brasileiro de Avaliação Psicológica (IBAP), 2002.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.]: ACM, 2016. p. 785–794.

FIX, E.; JR., J. L. H. Discriminatory analysis: Nonparametric discrimination: Consistency properties. *International Statistical Review*, v. 57, n. 3, p. 238–247, 1989.

FRANÇA, T. F.; FERREIRA, C. O.; OLIVEIRA, D. L. D.; ASSIS, G. T. D.; FERREIRA, A. A.; SILVA, E. J. D. Tôsabendo: a platform to create engaging teaching and learning experiences. In: IEEE. *2021 XVI Latin American Conference on Learning Technologies (LACLO)*. [S.l.], 2021. p. 275–281.

GIRAFFA, L.; KOHLS-SANTOS, P. Inteligência artificial e educação: conceitos, aplicações e implicações no fazer docente. *Educação em análise*, v. 8, n. 1, p. 116–134, 2023.

JESUS, H. O. D.; RODRIGUEZ, L. C.; JUNIOR, A. d. O. C. Predição de evasão escolar na licenciatura em computação. *Revista Brasileira de Informática na Educação*, v. 29, p. 255–272, 2021.

KANTORSKI, G.; FLORES, E. G.; SCHMITT, J.; HOFFMANN, I.; BARBOSA, F. Predição da evasão em cursos de graduação em instituições públicas. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2016. v. 27, n. 1, p. 906.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

MAIA, M. d. C.; MEIRELLES, F. d. S. Tecnologias de informação e comunicação e os índices de evasão nos cursos a distância. In: SN. *Proceedings of 12th International Congress of Distance Education*. [S.l.], 2005.

NOETZOLD, E.; PERTILE, S. d. L. Análise e predição de evasão dos alunos de um curso de graduação em sistemas de informação por meio da mineração de dados educacionais. *Revista Novas Tecnologias na Educação*, v. 19, n. 1, p. 351–360, 2021.

OLIVEIRA, R. d. S. *Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação*. Dissertação (Mestrado) — a, 2023.

REIS, J. N. de A.; SOUZA, W. D. S.; SANTANA, S. A. Proposta de ferramenta para predição de evasão no ensino superior. *e-RAC*, v. 8, n. 1, 2018.

ROLL, I.; WYLIE, R. Evolution and revolution in artificial intelligence in education. *International journal of artificial intelligence in education*, Springer, v. 26, n. 2, p. 582–599, 2016.

SACCARO, A.; FRANÇA, M. T. A.; JACINTO, P. d. A. Fatores associados à evasão no ensino superior brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de ciência, matemática e computação e de engenharia, produção e construção em instituições públicas e privadas. *Estudos Econômicos (São Paulo)*, SciELO Brasil, v. 49, n. 2, p. 337–373, 2019.

SANTOS, N. D. dos; MARCZAK, S. Fatores de atração, evasão e permanência de mulheres nas áreas da computação. In: SBC. *Women in Information Technology (WIT)*. [S.l.], 2023. p. 136–147.

UTIYAMA, F.; BORBA, S. d. F. P. Uma ferramenta de apoio ao controle da evasão de alunos em cursos a distância via internet. In: *III Congresso internacional de Computação*. [S.l.: s.n.], 2003.

ZELAYA, C. V. G. Towards explaining the effects of data preprocessing on machine learning. In: IEEE. *2019 IEEE 35th international conference on data engineering (ICDE)*. [S.l.], 2019. p. 2086–2090.