Análise Exploratória e Preditiva de Dados Acadêmicos Sobre Evasão Escolar em Cursos de Graduação do Instituto Federal da Bahia

Ana Ruth Nery Lima Bezerra[1],
Pablo Vieira Florentino[2]

[1]Análise e Desenvolvimento de Sistemas – Instituto Federal da Bahia (IFBA)
Salvador – BA – Brasil

an.bezerra@gmail.com

[2]Departamento de Computação – Instituto Federal da Bahia (IFBA)
Salvador – BA – Brasil

pablovf@ifba.edu.br

Abstract. *School dropout is a global socio-educational and economic problem affecting public and private institutions at all educational levels, with impacts ranging from the individual to the collective sphere. In Brazil, student retention measures span from national basic education programs, like* Pé-de-Meia*, to specific higher education policies, such as those implemented by the Federal Institute of Bahia (IFBA). However, undergraduate dropout rates remain high, demanding continuous analyses to support new strategies and outline profiles of students prone to leaving. Thus, this descriptive and quantitative study conducted an exploratory and predictive analysis of academic data, developing and training machine learning models to identify school dropout patterns.*

Resumo. *A evasão escolar é um problema socioeducacional e econômico global que afeta instituições públicas e privadas em todos os níveis de ensino. Seus impactos estendem-se do âmbito individual ao coletivo. No Brasil, ações de permanência estudantil abrangem desde programas nacionais na educação básica, como o* Pé-de-Meia*, até políticas específicas no ensino superior, como as adotadas pelo Instituto Federal da Bahia (IFBA). Contudo, os índices de evasão na graduação permanecem expressivos, exigindo análises contínuas para subsidiar novas estratégias e delinear perfis de estudantes propensos a evadir. Assim, este estudo descritivo e quantitativo realizou uma análise exploratória e preditiva de dados acadêmicos, desenvolvendo e treinando modelos de aprendizado de máquina para a identificação de padrões de evasão escolar.*

1. Introdução

A evasão escolar é um problema socioeducacional multifatorial persistente na sociedade mundialmente ao longo da história. Ela consiste no abandono dos estudos e pode ocorrer em qualquer nível de ensino, seja fundamental, médio ou superior, em instituições de ensino públicas ou privadas (sobretudo nas públicas) [1]. Se trata de um fenômeno que, ao mesmo tempo em que expõe problemas sociais complexos que o motivam, também impacta a sociedade ao reforçar um ciclo de desigualdade social [2].

As motivações para esse fenômeno são diversas: vão desde fatores demográficos e geográficos, até questões psicológicas, socioemocionais e pedagógicas. É um tema amplamente estudado por diversas áreas do conhecimento, como a Pedagogia, Psicologia, Economia, e Direito, sendo também contemplado pela área da Computação, especialmente na ciência de dados aplicada, com o objetivo de investigar suas motivações, por meio da descoberta de conhecimento em banco de dados. Por se tratar de um fenômeno multifatorial, as abordagens para sua melhor compreensão e intervenção também precisam ser diversas [3].

Para ilustrar esse cenário, é possível mencionar como exemplo os indicadores divulgados em 2021 pelo Instituto Nacional de Estudo e Pesquisa Educacional Anísio Teixeira (INEP), na Sinopse Estatística da Educação Superior de 2019, $43,5\%$ dos alunos matriculados em cursos de graduação presencial e à distância no Brasil abandonaram os estudos [4]. Outro dado mais recente foi trazido à luz no relatório *Education at a Glance 2025*, da Organização para a Cooperação e Desenvolvimento Econômico (OCDE), no qual foi constatado que 25% dos estudantes do ensino superior no Brasil abandonam os estudos logo no primeiro ano do curso. Ainda nesse relatório, a OCDE apurou que apenas 49% dos brasileiros que ingressaram em cursos superiores conseguiram se formar, mesmo após o prazo de três anos após a previsão de formatura [5].

O fenômeno da evasão escolar oferece muitos impactos, seja para o indivíduo que abandonou os estudos, seja para a sociedade de modo geral. No âmbito individual, os impactos permeiam áreas da vida da pessoa como socioemocional, renda, e até mesmo saúde e longevidade. É importante ressaltar que a decisão de evadir pode ser considerada um processo alimentado por muitos aspectos que afastam o estudante da instituição, como o mal estar escolar gerado por *bullying* e problemas em relacionamentos, ou dificuldades de aprendizado. Já outros aspectos o puxam para fora dela, como a necessidade imediata de gerar renda. De acordo com Rumberger, a consequência desse movimento pode ser considerado um motor para a evasão [6].

Já no âmbito coletivo, é possível mencionar a perpetuação da desigualdade social como um impacto significativo, visto que a evasão é um fator de estagnação da mobilidade social [2]. Sob uma ótica econômica, outro aspecto que impacta a sociedade é o custo da evasão. Seja no quesito de investimentos da gestão pública que são perdidos quando o aluno evade, seja na perda do capital humano, já que esse estudante deixa de se tornar um trabalhador qualificado capaz de construir conhecimento e cultura e gerar inovação e renda. Ambos aspectos mencionados corroboram o fato de que a evasão escolar é um fenômeno historicamente associado a subempregos e reforça ciclos de pobreza [7].

Em vista da complexidade da questão da evasão escolar e diante de sua persistência e expressivos impactos, é crucial identificar o perfil estudantil mais propenso à evasão e os principais fatores que influenciam esse fenômeno de forma antecipada. Isso é para que a tomada de decisões que envolvam a elaboração e implantação de políticas públicas de permanência estudantil sejam melhor embasadas [8] [9]. É importante que esse tipo de política pública seja aplicada de forma consciente e bem informada, visto que a sociedade como um todo perde com a evasão escolar.

Neste estudo descritivo e exploratório com abordagem quantitativa, foi dada ênfase aos cursos de graduação do Instituto Federal da Bahia (IFBA). O objetivo deste

trabalho, portanto, foi realizar um estudo sobre a relação entre atributos de potencial impacto na permanência estudantil, de modo a analisar contexto e perfil estudantis de alunos com maior propensão à evasão escolar. Para isso, foram desenvolvidos processos de ETL (*Extract, Transform, Load*) nas bases de dados de matrícula e de eficiência acadêmica extraídas da Plataforma Nilo Peçanha (PNP), referentes ao período entre os anos de 2011 a 2023, divulgados em 2024. Foi realizada uma análise exploratória de dados acadêmicos e elaboradas modelagens preditivas de *machine learning* para analisar contexto e perfil de alunos com propensão à evasão escolar.

A metodologia de trabalho adotada envolveu a revisão bibliográfica, com a exploração dos conceitos envolvidos e análise de trabalhos correlatos; coleta, preparação e remoção de duplicatas das bases de dados, consolidando os *DataFrames* a serem consumidos nas etapas de EDA (*Exploratory Data Analysis*) e modelagem; etapa de Análise Exploratória de Dados, que envolveu análise univariada e bivariada dos atributos com a visualização de gráficos e análise de coorte temporal; desenvolvimento do modelo de aprendizado de máquina de classificação (*Random Forest* e *XGBoost*), bem como a validação e avaliação dos modelos; além da discussão de escopo, limitações e apresentação dos resultados do estudo.

Assim, à luz da análise de dados, esse estudo aplicou um olhar investigativo sobre dados acadêmicos e sugeriu relações entre variáveis potencialmente influentes no fenômeno da evasão escolar que, a olho nu, não seriam perceptíveis. Além disso, aplicou técnicas de modelagem preditiva para analisar contextos e perfis estudantis mais propensos ao abandono dos estudos. Com isso, é esperado que este estudo sirva de auxílio para gestores e entidades pedagógicas no combate à evasão, como uma forma de embasar a tomada de decisão para a aplicação de políticas públicas de permanência estudantil.

2. Revisão Bibliográfica

A revisão bibliográfica deste estudo aprofundou os principais conceitos e pesquisas correlatas que sustentaram o desenvolvimento do trabalho. Foi adotada uma abordagem interdisciplinar, unindo Ciência de Dados, Inteligência Artificial e Gestão Educacional. Para isso, esta seção foi organizada de modo a dividir Conceitos Envolvidos e Trabalhos Correlatos. Essa estrutura ofereceu o suporte teórico necessário para a compreensão das etapas de tratamento de dados e para a discussão das dinâmicas de evasão escolar analisadas.

Esta seção foi construída com auxílio das bases de dados acadêmicos como SBC-OpenLib (SOL), SciELO e Google Acadêmico. A busca foi concentrada na intersecção entre Ciência de Dados e Educação, utilizando palavras-chave como “Evasão Escolar”, “Machine Learning na Educação” e “Análise Preditiva”. Assim, a partir de livros, artigos e dissertações, foram estabelecidos os principais pilares teóricos desta pesquisa.

Nas subseções seguintes, foram explorados os conceitos fundamentais que nortearam a realização do trabalho. O escopo macro abrange desde as premissas da Gestão Educacional e o desenho metodológico de coortes temporais, passando pelas técnicas de análise exploratória e tratamento de dados, até a consolidação dos algoritmos classificadores de aprendizado de máquina e suas respectivas métricas de avaliação. Dessa forma, foi estabelecida uma base conceitual unificada para as etapas subsequentes do projeto.

2.1. Gestão Educacional

No âmbito da gestão educacional, a compreensão dos fatores que impactam a eficiência institucional e o sucesso discente passa pela análise do fluxo escolar, que se apoia nos conceitos de evasão e retenção. A evasão é amplamente entendida como o abandono definitivo do curso por parte do estudante, motivado por uma complexidade de variáveis econômicas, geográficas, sociopedagógicas e psicológicas [9]. Essa ruptura prejudica a vida do estudante, as instituições de ensino e a sociedade como um todo [10].

Estreitamente associada à problemática da evasão, a retenção estudantil caracteriza a situação em que o discente permanece vinculado à instituição, mas não avança na matriz curricular dentro do tempo previsto, frequentemente em decorrência de reprovações sucessivas ou trancamentos. O estudo da retenção é crucial porque esse fenômeno atua como um estágio precursor da evasão [3].

O desgaste provocado pelo atraso na integralização do curso reduz a motivação do discente, agrava suas vulnerabilidades e culmina no seu desligamento definitivo. Desse modo, o monitoramento conjunto desses dois indicadores permite que a gestão educacional mapeie o risco de forma antecipada e promova políticas de permanência assertivas [10].

Na Rede Federal de Educação Profissional, Científica e Tecnológica, a flutuação desses fluxos de evasão e retenção impacta diretamente a avaliação de eficiência e a distribuição de recursos orçamentários e os dados que expressam essa dinâmica são mantidos pela Plataforma Nilo Peçanha (PNP). Para mensurar a complexidade da oferta institucional diante dessas variações, a gestão contemporânea utiliza indicadores gerenciais, dos quais se destacam para este estudo Fator de Esforço de Curso (FEC)[1] e Carga Horária.

Carga Horária é a duração estabelecida em horas no projeto pedagógico de um curso. O FEC, por sua vez, funciona como um regulador métrico que ajusta a contagem de matrículas com base na intensidade do trabalho docente exigido. É um valor que varia de 1.0 a 1.35 e é estabelecido e utilizado pelo MEC para o cálculo de outras métricas relevantes para a gestão educacional [11].

Essa ponderação considera a necessidade de atividades práticas, o uso de laboratórios ou situações que demandem a subdivisão de turmas. Permite, portanto, que a gestão relacione de forma precisa o investimento financeiro e a complexidade operacional dos cursos afetados pelo abandono escolar [11].

2.2. Análise de Coorte Temporal

Para compreender como a dinâmica da evasão escolar se manifestou ao longo do tempo, este estudo adotou a abordagem de Análise de Coorte Temporal. Na literatura educacional, uma coorte consiste em um grupo de estudantes que partilha de um evento acadêmico comum em um mesmo período, tipicamente o ingresso em um determinado curso e semestre letivo, portanto, turmas em um mesmo curso de graduação [12].

[1] Como o FEC possui valores indexados que variam conforme a natureza de cada habilitação, sua tabela completa com os indexadores oficiais encontra-se regulamentada no Anexo II da Portaria nº 146, de 25 de março de 2021, do Ministério da Educação, disponível neste link: https://www.in.gov.br/web/dou/-/portaria-n-146-de-25-de-marco-de-2021-310597431 [11].

Acompanhar essas coortes ao longo do tempo permite mapear o fluxo escolar e identificar em quais momentos críticos a retenção ou a evasão se concentraram. Diferente de análises que consideram um panorama estático, o estudo de coortes acompanha a trajetória dos indivíduos. Essa perspectiva temporal auxilia a compreensão de como a evasão se manifestou de acordo com a passagem dos anos de forma comparativa [13].

Sob a ótica da Ciência de Dados aplicada à educação, a análise das coortes enriquece o estudo ao fornecer um contexto temporal aos eventos em questão. Isso permite que a gestão educacional tome conhecimento de tendências e planeje políticas públicas ou pedagógicas direcionadas para as fases em que as turmas se mostram mais vulneráveis.

2.3. Processo de Descoberta do Conhecimento Em Bancos de Dados

Para que a extração de informações em grandes volumes de dados educacionais seja eficiente e metodologicamente rigorosa, a literatura adota processos estruturados como a Descoberta de Conhecimento em Banco de Dados, ou *Knowledge Discovery in Databases* (KDD). O KDD consiste em um fluxo macro que envolve diversas etapas com o objetivo de identificar padrões, realizar previsões e levantar correlações entre os atributos analisados [14].

Ao organizar o tratamento da base de dados, esse processo possibilita a obtenção de *insights*. Na literatura, esse conceito é traduzido como percepções profundas, ou seja, compreensões claras e acionáveis obtidas a partir da análise. São essas percepções que fundamentam a tomada de decisão, o que consolida o processo de KDD como uma base metodológica clássica para a mineração de dados [15].

Na prática contemporânea, a execução das primeiras etapas desse fluxo do KDD fica a cargo da Engenharia de Dados. Esta disciplina envolve o desenvolvimento, arquitetura e manutenção da infraestrutura tecnológica necessária para lidar com grandes volumes de informação. É por meio da engenharia de dados que os fluxos teóricos de manipulação são automatizados, garantindo que os dados brutos sejam movidos e tratados com eficiência e segurança antes de chegarem aos modelos de inteligência artificial [16].

Dentro do fluxo do KDD, a consolidação dos dados brutos ocorre por meio da Extração, Transformação e Carga, frequentemente referenciada pela sigla ETL (*Extract, Transform, Load*). Essa fase operacional é dividida em três momentos essenciais. Na etapa de extração, as informações são coletadas a partir de suas diversas fontes originais.

Em seguida, na transformação, os dados são submetidos a técnicas de limpeza e tratamento, cujas regras variam de acordo com o tipo de variável e o propósito específico do projeto. Por fim, a etapa de carga centraliza e persiste esses dados já padronizados [4], garantindo a integridade necessária para as etapas subsequentes.

Para esse propósito, uma prática recomendada é a utilização do formato Apache Parquet, um formato de arquivo de armazenamento colunar de código aberto que preserva nativamente o esquema e a tipagem das variáveis. Essa característica impede falhas de conversão, como a alteração involuntária de tipos numéricos para texto, algo frequente em arquivos baseados em texto plano como o CSV (*Comma-Separated Values*, ou Valores Separados por Vírgulas). Logo, a escolha pelo formato Parquet garante que a tipagem de dados definida na etapa de tratamento seja preservada para o consumo analítico [17].

2.3.1. Análise Exploratória de Dados

Dentro do processo de descoberta de conhecimento em banco de dados, uma vez estruturada a base de dados por meio do processo de ETL, uma etapa relevante é a de Análise Exploratória de Dados, também conhecida como EDA (*Exploratory Data Analysis*). Esta etapa compreende um conjunto de métodos voltados para a exploração inicial, descrição, visualização e interpretação do conjunto de informações [18]. Ao aplicar a análise exploratória, torna-se possível mapear a natureza dos dados, destacar tendências e revelar relações preliminares entre as variáveis.

Neste processo de análise, verifica-se o tipo das variáveis, se são numéricas ou categóricas. Variáveis numéricas podem representar valores contínuos ou discretos. Os contínuos são obtidos por medidas e podem assumir qualquer valor dentro de um intervalo numérico, incluindo frações. Os valores discretos são obtidos por contagem e podem assumir apenas valores inteiros [19].

Já as variáveis categóricas ou qualitativas agrupam as informações em rótulos ou categorias, podendo ser ordinais ou nominais. As ordinais agrupam os dados de modo a seguir uma ordem lógica ou hierárquica. As nominais, por sua vez, não representam hierarquia, apenas descrevem as características da categoria [19].

Uma vez compreendida a natureza das variáveis, é possível dar seguimento à investigação através de análises univariadas e/ou bivariadas[2]. A univariada examina uma variável por vez para descrever seu comportamento e identificar padrões, servindo para conhecer as características individuais do seu conjunto de dados. Ela não busca relações de causa e efeito, mas o resumo e a distribuição dos dados. Para este tipo de análise, aplica-se estatística descritiva e as visualizações são feitas através de histogramas, gráficos de barras e diagramas de caixa (*boxplots*) [19].

Nessa investigação de variáveis individuais é possível detectar valores atípicos, tecnicamente denominados *outliers*. Esses elementos consistem em observações ou pontos de dados que se distanciam drasticamente do padrão geral do restante do conjunto amostral. Sua ocorrência pode indicar tanto inconsistências no registro das informações quanto comportamentos de exceção altamente relevantes para a compreensão do fenômeno estudado [18].

Enquanto isso, a análise bivariada investiga a relação entre duas variáveis para identificar se a mudança de uma interfere no comportamento da outra, e como essa alteração se dá. Quando essa alteração acontece, dá-se o nome de **correlação** a este fenômeno, que, portanto, indica que duas variáveis modificam-se simultaneamente [19].

Na análise de correlação, são utilizados métodos para calcular o coeficiente de correlação, isto é, a medida que indica a força e a direção da relação entre duas variáveis. Para isso, as abordagens mais utilizadas são as de Pearson e Spearman. O coeficiente de correlação de Pearson (ρ) mensura a força e a direção de uma relação linear entre duas variáveis contínuas [19], sendo calculado por:

[2]Há, também, as análises multivariadas. Contudo, não estão no escopo deste estudo e, portanto, não serão abordadas neste tópico. Caso seja do desejo do leitor compreender com profundidade este tipo de análise e suas técnicas, vide [20].

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{1}$$

Contudo, essa medida é sensível a *outliers* e, diante de distribuições não lineares ou variáveis ordinais, o coeficiente de Spearman (r_s) mostra-se mais adequado. Essa abordagem baseia-se nos postos (*ranks*) das observações em vez de seus valores brutos, avaliando quão bem a relação entre duas variáveis pode ser descrita por uma função monótona[3] [19], conforme expresso na equação:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2}$$

Em que d_i representa a diferença entre os postos de cada observação e n o número total de instâncias [19].

No processo de análise exploratória, a identificação de associações entre as variáveis é fundamental para capturar possíveis percepções e tendências. Contudo, é importante diferenciar correlação de causalidade: o fato de duas variáveis se modificarem simultaneamente não significa necessariamente que uma seja a causa da outra. Esse movimento pode ser explicado, inclusive, por uma terceira variável que influencia um fenômeno em comum, ou até por uma coincidência estatística [19]. Com isso em vista, é importate que a análise tenha diversas camadas capazes de abarcar o problema em questão, e não se apoiar em apenas uma métrica isolada.

2.3.2. Inteligência Artificial e Aprendizado de Máquina

Ainda no processo de KDD, a Inteligência Artificial (IA) tem se destacado e expandido como ferramenta de automação analítica de dados. É possível defini-la como o campo da ciência da computação voltado para a compreensão, descrição e replicação de comportamentos análogos à inteligência humana por meio de agentes processadores capazes de simular o raciocínio humano [4]. Como um subcampo proeminente e prático da IA, o Aprendizado de Máquina, ou *Machine Learning* (ML), concentra-se no desenvolvimento de sistemas que aprendem de forma automática a partir do histórico de dados disponível.

Essa técnica é amplamente aplicada no reconhecimento de padrões e na realização de análises preditivas para apoiar a tomada de decisão [18]. No âmbito do Aprendizado de Máquina Supervisionado — paradigma em que os modelos são treinados a partir de dados rotulados, ou seja, informações que já contêm a resposta final que o algoritmo deve aprender a predizer —, as tarefas são divididas de acordo com a natureza do resultado que se deseja obter, destacando-se a classificação e a regressão [21].

A classificação consiste no processo de mapear uma instância de dados a uma categoria predefinida, com base no comportamento de seus atributos. Diferentemente da regressão, que se ocupa da predição de valores numéricos contínuos, a classificação busca

[3] Função que mantém um único sentido de variação em todo o seu domínio, isto é, só pode ser crescente ou decrescente [19].

rotular os dados em grupos específicos [21]. No contexto deste trabalho, o problema configura-se como uma classificação binária, uma vez que o objetivo é predizer se um estudante pertence à classe dos "evadidos" ou dos "não evadidos".

Para a resolução de problemas de classificação com essa natureza, é possível mencionar os modelos preditivos baseados em árvores de decisão. Uma árvore de decisão consiste em um modelo estruturado de forma hierárquica, que se assemelha a um fluxograma, no qual os dados são segmentados repetidamente por meio de regras condicionais aplicadas aos seus atributos. Esse processo de divisão lógica estende-se a partir de nós de decisão que se ramificam de acordo com as características dos dados, até atingir os nós folha, os quais determinam a classificação final do objeto analisado [22].

Quando múltiplas árvores são combinadas para atuar de forma conjunta, formam-se os chamados comitês de árvores ou "florestas", estratégia que visa contornar as limitações e falhas de uma única estrutura isolada. Um dos algoritmos mais consolidados nessa categoria é a Floresta Aleatória, ou *Random Forest* [22].

Esse método consiste em um conjunto de árvores de decisão estruturadas por meio da técnica de *bagging* (*Bootstrap Aggregating*), que se baseia na geração de múltiplos subconjuntos de dados independentes por meio de amostragem com reposição. Neste tipo de amostragem, cada elemento selecionado é devolvido ao conjunto original antes da próxima escolha, o que permite que uma mesma instância seja sorteada mais de uma vez para o mesmo subconjunto [22].

Cada árvore é formada isoladamente com uma dessas subamostras e com um escopo aleatório de variáveis. Ao final, as predições individuais são combinadas por meio de votação majoritária, isto é, os palpites de todas as árvores são contabilizados pelo algoritmo, e a classe que obteve maioria será escolhida como resultado final do modelo, definindo, assim, a classificação [22].

Esse processo reduz a variação dos resultados e evita o problema do superajuste (*overfitting*). Isto se trata da tendência de um algoritmo de aprendizado de máquina a aprender regras muito específicas com dados de treino e perder a capacidade de generalizar e gerar predições com novos conjuntos de dados. Adotar a abordagem da floresta contorna esse problema, portanto, ao compensar o erro de uma árvore de decisão isolada com o acerto de várias outras [22].

Além de ser robusto, o *Random Forest* apresenta alta eficiência computacional. O tempo de execução de cada árvore da floresta é delimitado pela complexidade de $O(m \cdot n \log n)$, em que n representa o número de amostras e m a quantidade de atributos. Já o tempo necessário para a predição é proporcional à altura da árvore gerada [22].

Como alternativa evolutiva aos métodos de amostragem por *bagging*, os algoritmos fundamentados no princípio de *boosting* (melhoria) oferecem uma abordagem focada na otimização sequencial de erros. Nesse escopo, destaca-se o algoritmo *XGBoost*, uma ferramenta de aprendizado de máquina que cria árvores de decisão de forma consecutiva, em que cada nova árvore atua corrigindo os erros residuais deixados pelos modelos anteriores [22].

Esse processo captura e trata os padrões de erro de forma iterativa até que os resíduos pareçam aleatórios. O principal diferencial do *XGBoost* é sua extrema eficiência

de execução, contando com suporte nativo para paralelismo e concorrência computacional, além de dispensar algumas etapas de pré-processamento de dados, já que lida nativamente com valores ausentes e com diferentes escalas numéricas das variáveis [22].

Para controlar a forma como os modelos de aprendizagem de máquina aprendem e otimizar os seus desempenhos, é importante ajustar os seus hiperparâmetros. Ao contrário dos parâmetros internos, que o algoritmo aprende e ajusta sozinho durante o treinamento, os hiperparâmetros são configurações externas definidas previamente pelo analista para controlar o comportamento do algoritmo e guiar o processo de aprendizado [21]. Além de aumentar a capacidade de o modelo fazer previsões corretas, esse ajuste fino evita o *overfitting* e o *underfitting* — quando o modelo é simples demais e não consegue aprender os padrões dos dados [22].

Outro aspecto relevante da etapa da modelagem preditiva é o cuidado que o fenômeno do desbalanceamento de classes carece. Ele ocorre quando o conjunto de dados apresenta uma disparidade expressiva na distribuição das categorias. Geralmente, existe uma classe majoritária e uma minoritária que, apesar de sua ocorrência ser rara, costuma ser o foco principal da análise. A ela se dá o nome **variável alvo**, pois representa o evento que se pretende prever. Já as variáveis preditoras são as utilizadas pelo modelo para fazer essa previsão [21].

Essa disparidade cria um viés que, se não tratado, prejudica o aprendizado do modelo. O algoritmo aprende muito sobre a classe abundante e quase nada sobre a rara, adivinhando sempre a classe majoritária e falhando em prever os casos raros, relevantes para a análise. É possível contornar esse comportamento com a reamostragem, seja reduzindo a quantidade de dados da classe majoritária para igualar o número de exemplos da classe minoritária (subamostragem), ou aumentando a quantidade de dados da minoritária até equilibrar com a majoritária (sobreamostragem) [21].

Dentre técnicas de reamostragem aplicadas para evitar esse viés, é possível mencionar o SMOTE (*Synthetic Minority Over-sampling Technique*), que faz o rebalanceamento criando novos dados sintéticos para a classe minoritária. Assim, ele iguala a proporção das classes e evita que o modelo de aprendizagem de máquina seja dominado pela classe majoritária [21].

No escopo da modelagem preditiva voltada para fenômenos educacionais, a interpretabilidade surge como um requisito metodológico crucial. A interpretabilidade refere-se ao grau em que um ser humano consegue compreender e explicar as razões por trás das decisões ou predições tomadas por um modelo computacional [23]. Um modelo altamente performático, mas que funciona como uma "caixa-preta" que é difícil de compreender como chegou aos resultados, pode se apresentar como uma barreira para a tomada de decisões para aplicação de políticas públicas.

Diante disso, a importância de atributos (*feature importance*) atua como um mecanismo direto para garantir essa transparência. Esse conceito consiste em um índice computado pelo algoritmo que mensura o nível de contribuição de cada variável individual na composição do resultado final do modelo [23]. Tanto o *Random Forest* quanto o *XGBoost* são amplamente valorizados por oferecerem suporte nativo a essa funcionalidade [22]. Esses modelos permitem aos gestores, além de gerar predições sobre quais estudantes estão em risco de evasão, também a identificação dos fatores subjacentes mais

críticos para orientar ações preventivas.

2.3.3. Métricas de Avaliação dos Modelos Preditivos

A avaliação do desempenho de modelos classificadores exige a escolha criteriosa de métricas que reflitam a realidade estatística dos dados analisados. Em estudos voltados a fenômenos educacionais, como a evasão escolar, é frequente deparar-se com o fenômeno do desbalanceamento de classes [21]. Diante disso, torna-se inapropriado o uso isolado de métricas globais de acerto, pois os algoritmos tendem a priorizar a classe majoritária para maximizar a performance geral e, assim, mascara a incapacidade do modelo em detectar os casos raros. Assim, utiliza-se a matriz de confusão como ferramenta analítica base.

A matriz de confusão consiste em uma tabela bidimensional que cruza os dados reais do esquema com as predições feitas pelo modelo, mapeando os resultados em quatro quadrantes fundamentais: Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN) [21]. A partir do mapeamento desses quadrantes, derivam-se as métricas detalhadas a seguir.

Acurácia) A acurácia reflete a proporção global de predições corretas realizadas pelo modelo em relação ao total de instâncias avaliadas. Apesar de sua interpretação intuitiva, essa métrica mostra-se altamente sensível ao desbalanceamento de classes, conforme discutido anteriormente [22]. Sua formulação matemática é expressa pela Equação 3:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{3}$$

Precisão) A precisão mensura a exatidão das predições positivas geradas pelo algoritmo, indicando a proporção de estudantes classificados sob risco de evasão que de fato evadiram [22]. Essa métrica é essencial para otimizar a alocação de recursos pedagógicos, mitigando a ocorrência de falsos alarmes, e é calculada de acordo com a Equação 4:

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{4}$$

Revocação (*Recall*) Também denominada sensibilidade, a revocação avalia a capacidade do modelo em identificar corretamente todas as amostras positivas [22]. No contexto da evasão, ela aponta o percentual de alunos evadidos que o algoritmo foi capaz de capturar com sucesso, sendo uma métrica crítica para evitar que estudantes em risco fiquem sem intervenção assistencial (falsos negativos). Sua determinação é descrita pela Equação 5:

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{5}$$

***F1-Score*)** O *F1-Score* atua como a média harmônica [4] entre a precisão e a revocação, fornecendo uma métrica única e balanceada para o desempenho do modelo. Trata-se do indicador mais robusto para cenários de forte desbalanceamento de classes, calculada conforme a Equação 6:

$$\textit{F1-Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{6}$$

Suporte) O suporte indica a quantidade absoluta de instâncias pertencentes a cada classe específica dentro do conjunto de dados de teste. Essa métrica fornece o contexto numérico para interpretar as demais métricas, permitindo identificar se a base sobre a qual a precisão e a revocação foram calculadas é estatisticamente representativa [24].

Curva ROC e AUC (*Area Under the Curve*) A curva ROC (*Receiver Operating Characteristic*) é um gráfico que ilustra o desempenho de um classificador binário à medida que o seu limiar (*threshold*) de decisão varia, mapeando a taxa de verdadeiros positivos contra a taxa de falsos positivos. O eixo vertical do gráfico representa a Taxa de Verdadeiros Positivos (TVP), equivalente à revocação (*recall*), calculada por:

$$TVP = \frac{VP}{VP + FN} \tag{7}$$

O eixo horizontal representa a Taxa de Falsos Positivos (TFP), dada por:

$$TFP = \frac{FP}{FP + VN} \tag{8}$$

A análise visual do gráfico pode ser subjetiva. Para auxiliar esta tarefa, utiliza-se a AUC (Área Sob a Curva ROC) como uma métrica quantitativa [24]. Matematicamente, a AUC calcula a área sob o gráfico traçado por essas taxas. Considerando N limiares de decisão ordenados, a área pode ser obtida numericamente pela regra do trapézio:

$$AUC = \sum_{i=1}^{N-1} \frac{(TVP_{i+1} + TVP_i) \times (TFP_i - TFP_{i+1})}{2} \tag{9}$$

A AUC varia de 0 a 1, em que um valor igual a 1.0 representa um modelo preditivo perfeito e um valor menor ou igual a 0.5 indica um desempenho equivalente ao de um palpite aleatório. Trata-se de uma métrica robusta ao desbalanceamento de classes, pois avalia a capacidade do algoritmo de ordenar as instâncias corretamente, distinguindo os estudantes em risco de evasão daqueles que permanecerão frequentes [24].

[4] Medida estatística calculada como o número de elementos dividido pela soma dos inversos desses elementos. Ideal para situações que envolvem grandezas inversamente proporcionais [22].

2.4. Trabalhos Correlatos

Foram analisados trabalhos correlatos que utilizaram dados educacionais para identificar padrões de abandono, e que avaliaram algoritmos de classificação a partir de métricas de desempenho (Acurácia, *F1-Score*, *AUC-ROC*) e técnicas de interpretabilidade de modelos (como SHAP ou *Feature Importance*). Essa etapa serviu de base comparativa para esta pesquisa.

O KDD é amplamente utilizado na investigação dos fatores que potencialmente influenciam o fenômeno da evasão escolar, já que permite processos de predição, através de ferramentas como as técnicas de aprendizagem de máquina. Diversos modelos de aprendizado de máquina foram utilizados nesse processo, principalmente algoritmos supervisionados, isto é, aqueles que trabalham com dados rotulados [8].

Tal abordagem é vista no trabalho de Souza [25], em que foi utilizada a metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*) para a aplicação do processo de aprendizagem de máquina. Nessa pesquisa, diversos algoritmos de classificação foram testados, a fim de apoiar a tomada de decisão de gestores no combate à evasão escolar. Dentre os algoritmos utilizados, os melhor avaliados foram *Two-class Boosted Decision Tree*, e *Two-Class Neural Network*, apresentando, respectivamente, a métrica de Acurácia de 0,964 e 0,944.

Já em Trajano [26], foi feita uma análise preditiva a partir de dados da Plataforma Nilo Peçanha (PNP) sobre estudantes que evadiram de cursos superiores no Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB). Esse estudo visou a investigação de padrões nesses alunos evadidos. Para isso, foram utilizados os algoritmos de classificação *Support Vector Machine* (SVM), *Decision Tree, K-Nearest Neighbors* e *Logistic Regression*, dentre os quais foi escolhido o último, devido o seu desempenho diante dos critérios de análise do modelo *Recall* e área sob a curva (ROC).

Alinhado a essa abordagem, Oliveira [4] trabalhou com uma análise preditiva da evasão escolar com dados educacionais do IFPB. Utilizou as bases de dados do Sistema Unificado de Administração Pública (SUAP), bem como da PNP, relacionadas aos cursos de graduação do IFPB, a fim de investigar a interrelação entre características dos estudantes que abandonaram os estudos. Foram utilizados os algoritmos de classificação Árvore de Decisão, Floresta Aleatória, *Naive Bayes*, *Multilayer Perceptron* e SVM. A principal métrica para avaliar os modelos foi o *F1-Score*, que variou entre 0,84 e 0,98 para dados do SUAP, e 0,58 e 0,94 para os da PNP. Essa oscilação sugere a necessidade de abordagens específicas para diferentes contextos e bases de dados.

Barbosa et al. [8], por sua vez, conduziu a sua análise preditiva sobre evasão escolar com dados acadêmicos do Instituto Federal de Pernambuco - Campus Jaboatão dos Guararapes, extraídos da base de dados do Q-Acadêmico. Foram utilizados alguns modelos classificadores, dentre os quais o *Random Forest* apresentou melhor desempenho diante das métricas *Recall* e *F1-Score*.

Ainda que a abordagem da análise exploratória e preditiva de dados acadêmicos sobre evasão escolar seja aplicada de forma frequente ao longo dos anos, a investigação desse fenômeno a partir desse conjunto de técnicas não se faz menos necessária. A recorrência desse tipo de estudo é um sintoma da persistência da evasão.

Diante disso, o Aprendizado de Máquina é aplicado como uma ferramenta apropriada para analisar a evasão escolar, visto o caráter complexo e de fatores mutáveis desse problema. Os atributos que o influenciavam em 2020, no contexto da pandemia da COVID-19, por exemplo, podem não ser os mesmos que atuam na realidade atual.

3. Metodologia de Trabalho

Esta pesquisa se caracteriza como um estudo descritivo e exploratório, com abordagem quantitativa, fundamentada no processo de Descoberta de Conhecimento em Bases de Dados (KDD - *Knowledge Discovery in Databases*). A metodologia foi estruturada de modo a cobrir a revisão bibliográfica, os processos de ETL, EDA e modelagem dos dados, bem como a avaliação dos modelos preditivos de classificação. Ainda nesta seção foram discutidas as limitações e o escopo da pesquisa.

3.1. Coleta e Preparação dos Dados

O ambiente de dados deste projeto foi construído a partir de duas fontes públicas extraídas da Plataforma Nilo Peçanha (PNP), mantidas pelo Ministério da Educação (MEC): os *Microdados de Matrículas* e os *Microdados de Eficiência Acadêmica*, ambos contendo dados relativos ao período entre os anos de 2011 a 2023 e divulgados em 2024 [27]. O escopo geográfico e institucional foi delimitado aos cursos de graduação de todos os *campi* do Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA).

3.1.1. Preparação dos Microdados de Matrículas

A base de matrículas foi extraída e tratada a fim de fornecer atributos do contexto sociodemográfico e econômico dos estudantes do IFBA, além das informações dos seus cursos. Cada atributo foi selecionado de forma estratégica para contemplar o escopo e interesse do estudo nas etapas de EDA e modelagem preditiva.

Na etapa de preparação da base de matrículas, constatou-se que as colunas referentes à oferta de vagas dos cursos, como vagas de ampla concorrência ou de cotas, continham dados repetidos de forma redundante para todos os registros individuais de matrículas. Ou seja, para cada linha individual de matrícula do estudante, continham registros de todas as colunas de tipos de vaga, independente da que o aluno estava vinculado. Isso significa que a partir das colunas de tipo de vaga, não é possível saber o tipo de ingresso do estudante, mas elas contém a quantidade de alunos que ingressaram no mesmo ciclo que ele em cada tipo de vaga.

Para contornar esse problema, o arquivo original de matrículas foi decomposto em duas estruturas: uma tabela para as vagas, que manteve os registros de ofertas dos tipos de vaga para cada curso e uma tabela de matrículas, que conservou os registros individualizados do perfil sociodemográfico e econômico dos estudantes, sem as informações replicadas. Ambas preservaram o atributo `'codigo_ciclo_matricula'` como chave comum.

3.1.2. Preparação dos Microdados de Eficiência Acadêmica

A base de *Eficiência Acadêmica* foi tratada para fornecer variáveis de contexto macro do curso, como a Carga Horária e o Fator Esforço de Curso (FEC), e enriquecer a análise do conjunto de atributos preditores. Esses atributos numéricos foram selecionados por se tratarem de indicadores estruturais dos cursos que não estão presentes na primeira base mencionada. A natureza dessa base é de análise por ciclo/turma.

Na etapa de preparação desses dados, foi constatado que as linhas de cada ciclo constavam multiplicadas pela quantidade de estudantes matriculados em cada turma, ou seja, os registros apareciam replicados de forma redundante. Foi então aplicada uma contagem para registrar a quantidade total de alunos por turma e, posteriormente, essa duplicação foi removida do conjunto.

3.2. Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados foi conduzida com o uso da linguagem Python e as bibliotecas Pandas, Matplotlib e Seaborn. Esta etapa buscou traçar o perfil dos estudantes do IFBA e a distribuição de suas situações de vínculo. Foram aplicadas análises univariadas e bivariadas, cruzando as variáveis preditoras com a variável alvo (`categoria_situacao`, que distingue alunos em curso, concluintes e evadidos). Também foi realizada uma análise de coorte temporal baseada no ano de ingresso dos estudantes.

Ainda nesta etapa foram analisados aspectos institucionais e de distribuição de ingressos de turmas no IFBA no decorrer dos anos e atributos numéricos como FEC e carga horária. A análise bivariada também foi realizada aqui com o cruzamento dos dados para a compreensão de como se manifestaram as taxas de evasão para essas diferentes categorias.

3.3. Desenvolvimento do Modelo de Aprendizado de Máquina

O *pipeline* de modelagem preditiva foi desenvolvido utilizando a biblioteca Scikit-Learn e bibliotecas dos algoritmos de aprendizagem de máquina escolhidos. Foram selecionados os algoritmos de classificação baseados em árvores de decisão *Random Forest* e *XGBoost*. A escolha se justifica pela dominância de variáveis categóricas e de natureza não linear no conjunto de dados. Esses algoritmos lidam de forma eficiente com dados categóricos codificados, apresentando menor sensibilidade [21].

A não utilização de outras famílias de algoritmos de classificação se justificou pelas restrições da natureza da base de dados e pelo objetivo da pesquisa. Modelos lineares paramétricos, como a Regressão Logística, foram preteridos por exigirem o cumprimento de pressupostos estatísticos estritos e apresentarem limitações na captura automática de interações não lineares complexas.

Algoritmos baseados em cálculo de distância, como o *K-Nearest Neighbors* (KNN) e as Máquinas de Vetores de Suporte (SVM), mostraram-se inadequados frente à predominância de atributos categóricos na base de dados utilizada. A transformação matemática desses dados (codificação) poderia prejudicar o desempenho métrico desses estimadores.

Por fim, arquiteturas de Redes Neurais não foram inclusas por operarem frequentemente como modelos de "caixa-preta", já que é difícil para um ser humano monitorar

como exatamente o modelo chegou à tomada de decisão. Este aspecto poderia comprometer a interpretabilidade dos atributos, fator indispensável para a formulação de intervenções e políticas de permanência estudantil no contexto institucional. O fluxo de desenvolvimento da aprendizagem de máquina, portanto, compreendeu as seguintes sub-etapas:

(i) separação dos dados em conjuntos independentes de treino e teste;

(ii) aplicação de técnicas de transformação de variáveis categóricas em numéricas (codificação) e aplicação da técnica SMOTE para balancear as classes;

(iii) ajustes dos hiperparâmetros dos algoritmos preditivos de classificação; e

(iv) avaliação de desempenho a partir da validação do modelo com base na Matriz de Confusão e métricas baseadas nela (***Recall***, **Precisão** e ***F1-Score***). Adicionalmente, considerou-se **AUC-ROC**, por ser adequada diante do fenômeno de desbalanceamento de classes, que se manifestou nesta pesquisa.

3.4. Limitações/Escopo

O presente estudo se delimita à análise da situação dos estudantes em ciclos de matrículas registrados no IFBA entre os anos de 2011 a 2023, dados divulgados em 2024. A principal limitação metodológica enfrentada se dá devido as restrições de acesso aos dados educacionais públicos de nível individual no Brasil, intensificadas a partir de 2022.

Sob a vigência da Lei Geral de Proteção de Dados Pessoais (LGPD), o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) alterou a política de divulgação de seus dados, incluindo o Censo da Educação Superior e o Censo Escolar, omitindo informações demográficas e cruzamentos que pudessem permitir a reidentificação de indivíduos. Embora o acesso seja permitido por meio do Serviço de Acesso a Dados Protegidos (Sedap), a burocracia e o tempo de resposta institucional poderiam inviabilizar a execução desta pesquisa dentro do cronograma previsto.

A Plataforma Nilo Peçanha (PNP), por sua vez, por possuir um escopo focado na Rede Federal de Educação Profissional, Científica e Tecnológica, adota uma modelagem baseada em Ciclos de Matrícula vinculados ao Sistema Nacional de Informações da Educação Profissional e Tecnológica (Sistec). Essa arquitetura de dados permite o acesso público a microdados de estudantes anonimizados, o que viabiliza o cruzamento entre o perfil socioeconômico do aluno e os indicadores de eficiência da unidade de ensino, sem violar as diretrizes da privacidade individual.

Adicionalmente, é possível apontar como limitação para este estudo a ausência de algumas colunas indicadoras na tabela de matrículas que poderiam enriquecer a análise: pessoa com deficiência, tipo de ingresso, e data da última atividade acadêmica. Embora a tabela de vagas mapeie o conjunto de vagas ofertadas por curso ($L9$), a impossibilidade de identificar quais estudantes possuem deficiência e como ingressaram (se foi por canais alternativos de inclusão, por exemplo) impede que o algoritmo de *machine learning* aprenda o potencial impacto da condição de PCD, ou dos tipos de ingresso em relação à evasão escolar no IFBA.

Além disso, as datas disponíveis se referem ao ingresso (início do ciclo) e à previsão de conclusão. Esse modelo impossibilita o mapeamento da situação efetiva de um

aluno que pode ainda estar registrado no sistema com o status de sua matrícula ativo, mesmo já tendo abandonado os estudos, o que configuraria uma evasão silenciosa.

4. Implementação e Arquitetura Tecnológica

A etapa de implementação deste projeto engloba o desenvolvimento do *pipeline* de tratamento dos dados, a execução da Análise Exploratória (EDA) e a construção do modelo preditivo de aprendizado de máquina. O projeto de software foi desenvolvido integralmente com o uso da linguagem de programação Python e suas principais bibliotecas de Ciência de Dados aplicada [5].

Para a etapa de engenharia de dados, foi utilizada a biblioteca Pandas, cuja estrutura de dados bidimensional tabular nativa (*DataFrame*) abstrai operações de ETL (*Extract, Transform, Load*). A etapa de persistência otimizada dos dados contou com a biblioteca PyArrow para manipulação de arquivos no formato Apache Parquet. Esse formato de arquivo de armazenamento colunar foi escolhido para salvar as etapas de tratamento aplicadas neste estudo pois preserva os tipos de dados, que foram transformados nessa ocasião.

4.1. Estrutura do Projeto e Repositório

O projeto foi estruturado seguindo os padrões de modularidade e reprodutibilidade recomendados para projetos de ciência de dados aplicada. O diretório `data` foi reservado de forma segmentada: dados brutos e imutáveis, em formato CSV, foram armazenados no subdiretório `raw`, enquanto as bases limpas e prontas para as etapas estatísticas e preditivas foram persistidas em formato *Parquet* no subdiretório `processed`.

Para a execução das etapas de tratamento, análise exploratória e modelagem dos dados, utilizou-se o ecossistema Jupyter Notebook. Essa ferramenta consiste em uma aplicação web de código aberto que permite a criação e o compartilhamento de documentos dinâmicos contendo código executável, equações e textos explicativos [28].

A estrutura fundamental do *Notebook* baseia-se em blocos independentes denominados "células". Essas células dividem-se em blocos de código isolados (onde fragmentos de código Python são compilados e executados separadamente de forma interativa e sequencial) e células de texto (utilizadas para documentação e contextualização das análises), garantindo a reprodutibilidade do experimento [28]. Todos os *Notebooks* foram isolados de forma sequencial no diretório `notebooks`, conforme ilustrado na Figura 1.

4.2. *Pipeline* de ETL (*Extract, Transform, Load*)

Para garantir a integridade e a qualidade dos dados antes das etapas de análise e modelagem, foi estruturado um *pipeline* de ETL nos *Notebooks 01* e *02*. O processo foi implementado seguindo uma sequência de filtragem, limpeza, padronização de tipos e remoção de redundâncias das bases de dados.

[5] A implementação completa deste projeto está disponível no repositório remoto que pode ser acessado através deste link: https://github.com/ana-ruth-bezerra/previsao_evasao

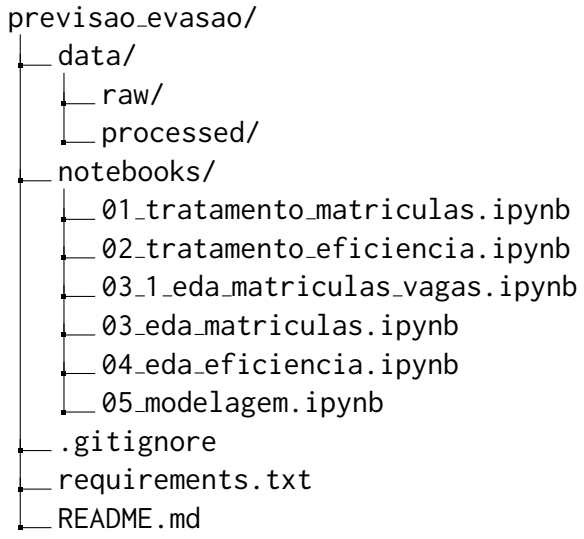

Figura 1. Estrutura de diretórios e organização modular do projeto.

4.2.1. Tratamento dos Microdados de Matrículas

O pipeline de tratamento dos microdados de matrículas foi integralmente desenvolvido no *Notebook 01*. O processo foi estruturado de forma cronológica, iniciando com a especificação das variáveis, extração filtrada, padronização sintática de cabeçalhos e transformação dos tipos de dados.

A etapa se iniciou com a seleção de 21 colunas de interesse contendo dados socioeconômicos, demográficos e institucionais. Para otimizar o processamento do arquivo bruto e o uso da memória RAM, a leitura foi feita em blocos (*chunksize=50000*). Durante a concatenação dos blocos, foi aplicado um filtro para manter apenas registros de instituições da Bahia que pertencessem a cursos de nível superior (Bacharelado, Licenciatura e Tecnologia). Isso feito, a coluna `'Instituição'` foi descartada por redundância.

Em seguida, as colunas foram renomeadas para o padrão sintático *snake_case*, com a remoção de acentuações e espaços em branco para uniformizar o código. Simultaneamente, realizou-se a adequação dos tipos de dados primitivos das variáveis — convertendo datas, indicadores quantitativos com suporte a nulos e chaves de identificação para suas respectivas tipagens nativas —, visando garantir a integridade estrutural e otimizar os índices categóricos para cruzamentos futuros.

Na sequência, foi isolada a tabela de vagas com o tratamento da anomalia do valor `-1` (substituído pelo nulo nativo do Python: `NaN` ou `None`) e posteriormente foram removidas das linhas duplicadas de cursos. O processo foi concluído com a remoção das colunas de vagas do *DataFrame* principal de matrículas, eliminando as duplicidades em nível de estudante.

A etapa de Carga (*Load*) encerra o *pipeline* persistindo de forma unificada as estruturas resultantes (Tabela de Matrículas e Tabela de Vagas). Foi utilizado o formato Apache Parquet, que garante a compressão e preserva as tipagens definidas nesta seção. A tabela de matrículas armazena os atributos granulares do estudante, se vinculando à tabela de vagas por meio do atributo `codigo_ciclo_matricula`. Este atua como Chave

Primária (PK) na estrutura de cursos e como Chave Estrangeira (FK) na estrutura de alunos, estabelecendo uma cardinalidade de um para muitos ($1 : N$).

O tratamento da base de dados de Eficiência Acadêmica foi realizado integralmente no *Notebook 02*, em processos similares ao que foi feito no *Notebook 01*. Assim como no processo de ETL visto anteriormente, foram selecionadas 7 colunas de interesse para o estudo.

É válido ressaltar que inicialmente também havia sido selecionada a coluna `'Tipo de Oferta'`, porém, na análise inicial da estrutura do *DataFrame* resultante, foi constatado que todos os registros constavam como `'Não se aplica'`. Por esse motivo, a coluna foi removida da seleção. Outra decisão tomada nessa etapa foi a de não incluir a coluna `'Modalidade de Ensino'`, que enriqueceria esta análise, entretanto a grande maioria dos cursos é da modalidade `'Educação Presencial'`. Constava apenas 1 registro de curso da modalidade `'Educação a Distância'` (Licenciatura em Computação, Campus Santo Amaro). Se trataria de uma amostra muito pequena para impactar este estudo.

4.2.2. Tratamento dos Microdados de Eficiência Acadêmica

Ainda na etapa de carga dos dados, processo semelhante ao que foi feito anteriormente de leitura do arquivo bruto em blocos, aplicação de filtro geográfico e institucional e remoção da coluna redundante `'Instituição'`, foi feito para essa base. A tabela resultante, portanto, ficou com um total de 2241 linhas. Ao analisar as primeiras linhas dessa estrutura, foi observado que haviam múltiplos registros para cada coluna, isto é, cada ciclo/turma (`'codigo_ciclo_matricula'`) apareceu replicado para cada registro de estudante em cada turma.

Foi aplicada uma contagem para capturar o tamanho de cada turma, o que revelou um total de 60 turmas (aproximadamente 10% do total das turmas vistas na base de matrículas). Também foram extraídas estatísticas sobre essas turmas, revelando uma média de 37.4 estudantes por turma, e valores atípicos (*outliers*) nas extremidades, com registro máximo de 260 estudantes e mínimo de 1 estudante. *Outliers* à parte, as maiores turmas apresentaram de 48 a 53 alunos, enquanto as menores de 16 a 24.

Outra inconsistência percebida foi a ocorrência de uma linha cujo código de matrícula não consta na base dos microdados de matrícula. Ela foi identificada posteriormente na etapa de *merge* com a segunda base mencionada, realizado mediante a chave comum `'codigo_ciclo_matricula'`. O *merge* e remoção das anomalias foram feitos apenas no *Notebook 05* de EDA de Eficiência Acadêmica, cujas impressões serão expostas no capítulo seguinte, mas vale registrar esse tratamento aplicado já nesta seção.

Mesmo que a base de microdados de Eficiência Acadêmica não tenha sido utilizada para a etapa de modelagem preditiva, esses valores anômalos poderiam distorcer a análise. No caso dos valores atípicos de quantidade de alunos por turma, esses dados gerariam distorções nas métricas de evasão, especialmente no caso extremo mínimo, já que esse único registro, caso constasse a situação de matrícula `'Evadido'`, a taxa de evasão seria de 100% e caso constasse `'Em Curso'`, seria 0%. Em qualquer um dos casos, a distorção da taxa de evasão seria expressiva.

Já em relação à linha cuja matrícula não foi localizada na base de matrículas, ela

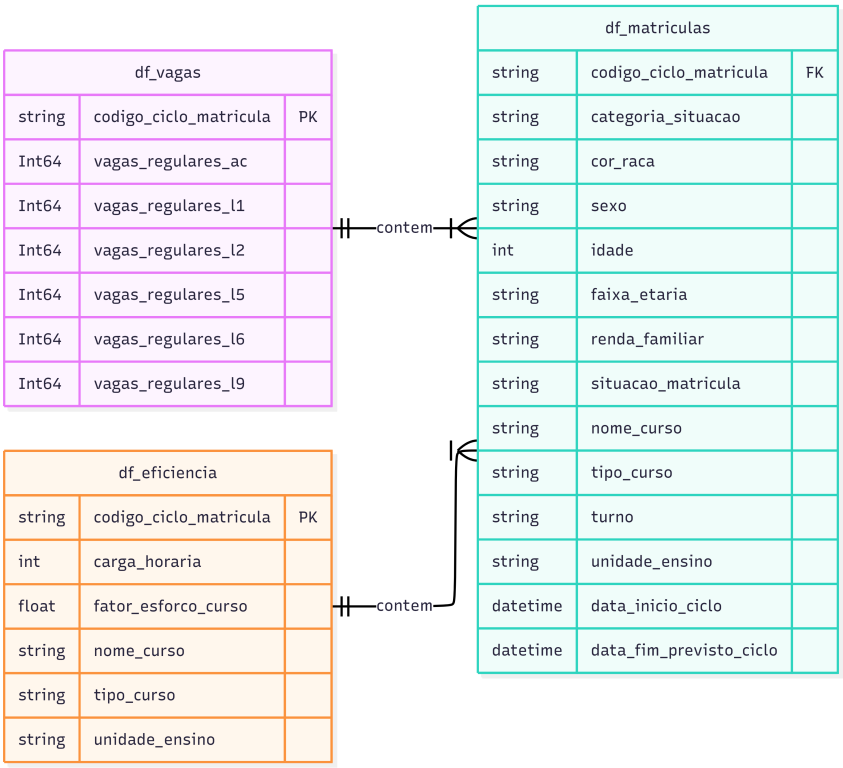

Figura 2. Modelo Lógico resultante da etapa de tratamento de dados

foi identificada no contexto do cruzamento dos dados entre duas bases, com o objetivo de extrair o percentual de evasão para cada ciclo. Seria, portanto, impossível localizar esse dado, já que `'situacao_matricula'` está na base das matrículas.

Com o cálculo dessas medidas, as duplicatas dessa base foram removidas e o seu índice foi redefinido, o que resultou numa tabela com apenas 60 ciclos únicos (turmas) para a EDA, e a etapa seguinte foi a de tratamento dos dados. Para essa etapa, os processos também foram similares ao que foi feito na base de matrículas, com variáveis renomeadas para o padrão sintático `snake_case`, e transformação dos tipos.

Para a persistência do *DataFrame* final tratado, foi aplicado processo similar ao feito no *Notebook* 01, salvando o arquivo em formato *Parquet*, para preservar a tipagem aqui definida. Assim, o *DataFrame* ficou pronto para ser consumido na etapa subsequente de EDA. Nessa etapa de análise exploratória, anomalias foram detectadas e removidas, o que resultou num *DataFrame* com 57 turmas/ciclos. As estruturas resultantes de todo o processo de tratamento de dados podem ser vistas no modelo lógico da Figura 2.

4.3. EDA (*Exploratory Data Analysis*) de Microdados de Matrículas e Eficiência Acadêmica

As etapas de Análise Exploratória dos Dados foram realizadas nos *Notebooks* 03, 04 e 05, organizados da seguinte forma:

- O primeiro (`03_eda_matriculas.ipynb`) foi reservado para a exploração e visualização dos dados da tabela de matrículas;
- O segundo (`03_1_eda_matriculas_vagas.ipynb`) para a tabela de vagas;

- E o terceiro (`04_eda_eficiencia.ipynb`) para eficiência acadêmica.

Em todos os *Notebooks* se seguiu um padrão da ordem lógica de execução das células, listada a seguir:

- Leitura dos arquivos Parquet tratados;
- Visualização dos tipos de dados e informações gerais dos *DataFrames* e na sequência, as primeiras linhas das estruturas para verificar o conteúdo;
- Extração das medidas estatísticas dos atributos numéricos;
- Investigação dos valores ausentes (*missing values*) no *DataFrame* e, caso necessário, o seu devido tratamento;
- Visualizações dos gráficos e engenharia de atributos para auxiliar essa etapa.

As bibliotecas do Python utilizadas para a exploração dos dados e visualização dos gráficos foram as duas principais para esse propósito, Matplotlib e Seaborn. Foram implementados gráficos de barras, gráficos de dispersão, gráficos de regressão, além das correlações de Pearson e Spearman, todos de acordo com as necessidades de cada visualização. Os gráficos mais relevantes para o estudo podem ser vistos de forma detalhada na seção 5.

4.4. Modelagem Preditiva: *Random Forest* e *XGBoost*

A etapa de modelagem preditiva foi desenvolvida no *Notebook* 06, em que se seguiu um *pipeline* de carga de dados, filtragem, divisão de dados para treino e teste, codificação (transformar variáveis categóricas em numéricas), aplicação dos modelos de *machine learning*, e avaliação. Esse processo foi iniciado com a separação da base de treino e definição dos atributos que foram utilizados no modelo.

A respeito da consolidação do *DataFrame* utilizado para a modelagem, a seleção das variáveis preditoras seguiu critérios de representatividade amostral e capacidade de generalização do modelo. Variáveis com alta cardinalidade e distribuição assimétrica entre categorias, como nome do curso e unidade de ensino, não foram incluídas para evitar o sobreajuste. Também não se incluiu o atributo de idade e, em substituição, se manteve a faixa etária para reduzir sensibilidade a valores individuais. A estrutura resultade do *DataFrame* que se utilizou para o modelo pode ser visto na Figura 3.

Para dar continuidade à etapa de modelagem, definiu-se X e y para a base de treino, com o mapeamento da varável alvo `'categoria_situacao'`: 1 para "Evadidos" e 0 para "Concluídos" e "Concluintes". Alunos na situação "Em Curso" não foram inclídos aqui, já que se pretendeu trabalhar apenas com desfechos conhecidos nesta etapa. Na base de treino, então, foi feito o balanceamento de classes através da técnica SMOTE.

Na sequência, as variáveis categóricas foram transformadas em números (codificação) por meio da função `get_dummies`, e caracteres não suportados pelo *XGBoost* compatibilizados via `str.replace`. Isso feito, é aplicada a função `train_test_split` que é ferramenta padrão da biblioteca Scikit-Learn para particionar o *dataset* em *subsets* de treino e teste para os modelos de aprendizado de máquina.

Com os dados preparados, foram aplicados os modelos de aprendizagem de máquina *Random Forest* e *XGBoost*, ambos algoritmos de classificação, com ajustes dos seus hiperparâmetros. Eles se mostram apropriados para esse propósito por lidarem bem com dados granulares e heterogêneos, além de reduzirem o risco de *overfitting*.

df_modelo	
string	cor_raca
string	sexo
string	faixa_etaria
string	renda_familiar
string	tipo_curso
string	turno

Figura 3. Atributos definidos para a utilização no modelo

O *Random Forest* foi adotado em sua configuração padrão, e ajustes de hiperparâmetros foram aplicados ao *XGBoost*, de modo evitar o *overfitting* diante de dados qualitativos e de alta granularidade, que predominaram na base de dados usada para os modelos. Mesmo que esses modelos evitem o superajuste, não são imunes a ele e boas práticas para o evitar são recomendadas [21].

Foi, portanto, delimitada a profundidade máxima das árvores (`max_depth=4`) e configurado um número mínimo de amostras por folha (`min_child_weight=5`). Adicionalmente, foi incorporado o parâmetro `scale_pos_weight` para ajustar dinamicamente a penalização de erros entre as classes, assegurando que o modelo mantivesse alta sensibilidade na identificação do grupo de risco (alunos evadidos). Ambos foram alimentados pela mesma base de dados, para que a avaliação dos seus desempenhos fosse mais precisa, além da demonstração de como cada um deles mensura a importância de atributos.

Com o treinamento realizado, foi aplicado o método `predict()`, que gera previsões para a variável alvo (`categoria_situacao`), e a etapa que seguiu foi a de avaliação de desempenho dos modelos. Para essa etapa, foi utilizada a ferramenta de avaliação de performance `classification_report`, que gera como saída uma tabela com os valores das métricas de **Precisão**, **Revocação**, ***F1-score*** e **Suporte**.

Adicionalmente, foi utilizada a Matriz de Confusão, que compara as previsões do modelo com os valores reais dos dados, mostrando a sua taxa de erros e de acertos, além de *AUC-ROC* (área sob curva), que avalia a capacidade do modelo em distinguir categorias. Além disso, foi analisada a importância de variáveis, que indica quais atributos mais influenciaram as previsões do modelo. Por fim, foram implementados os gráficos da matriz de confusão e da importância de variáveis para melhor visualização e interpretabilidade, que podem ser vistos na seção seguinte.

5. Resultados e Discussão

Este capítulo apresenta e discute os resultados obtidos ao longo do desenvolvimento deste estudo, estruturado de modo a refletir o processo de exploração, tratamento e modelagem dos dados. A análise foi organizada em níveis, desde a escala estrutural dos dados até a interpretação do desempenho dos algoritmos.

Assim, o capítulo se inicia com os resultados do tratamento dos dados, apresentando detalhes sobre o *pipeline* de ETL e o processo adotado de remoção de dados redundantes. Na sequência, são apresentados aspectos do perfil dos estudantes do IFBA, em que se explora o seus perfis sociodemográficos e econômicos, além das características temporais de suas coortes de ingresso.

Isso visto, o foco muda para a perspectiva institucional, analisando o contexto de oferta de vagas dos cursos, bem como a relação de carga horária e fator de esforço de curso com as taxas de evasão. Por fim, o capítulo reúne os aspectos preditivos e analíticos, apresentando a avaliação do desempenho dos classificadores de aprendizado de máquina desenvolvidos e a interpretabilidade dos atributos de maior expressão para o risco de evasão escolar.

5.1. Resultados do Processamento e Tratamento das Bases de Dados

Antes de conduzir as visualizações gráficas e o treinamento dos algoritmos preditivos, o *pipeline* de ETL gerou os primeiros resultados estruturais do estudo ao consolidar a amostra a ser analisada referente ao IFBA. O arquivo bruto original de microdados de matrículas da Plataforma Nilo Peçanha (PNP), que contém o registro de toda a Rede Federal de Educação Profissional, Científica e Tecnológica do Brasil, foi filtrado e reestruturado, de modo a contemplar o interesse e escopo da pesquisa.

O processo de limpeza e a aplicação dos filtros de estado e tipo de curso tiveram como resultado uma base de dados com 10.034 registros de estudantes de graduação. Uma métrica relevante para a etapa de modelagem preditiva é a integridade da base de matrículas, que não apresentou valores ausentes nas suas colunas demográficas e socioeconômicas, o que elimina a necessidade da aplicação de técnicas de imputação de dados.

Além disso, o processo de separação da tabela de matrículas da tabela de vagas permitiu mapear a estrutura de oferta de vagas da instituição de forma isolada. A extração da tabela de vagas revelou que no contexto do IFBA foram registradas 599 turmas/ciclos de matrícula entre os anos de 2011 e 2023 na amostra fornecida pela PNP.

Durante a análise da consistência da tabela de vagas, foram identificados 6 valores ausentes, o que representa uma taxa de inconsistência de 1% da oferta total, que pode ser considerada muito pequena nesse contexto. Da mesma forma, anomalias, como o registro de vagas com valor negativo (-1), teve apenas uma ocorrência na coluna `vagas_regulares_l5`. Esse comportamento foi tratado por meio da transformação para nulo nativo do Python (`None`), o que evita distorções nas métricas durante as próximas análises.

Já o processo de tratamento de base de eficiência acadêmica, similar ao feito na base de matrículas, resultou numa estrutura de apenas 60 registros. Apesar dessa amostra ser pequena, foi possível extrair percepções relevantes para este estudo através da análise.

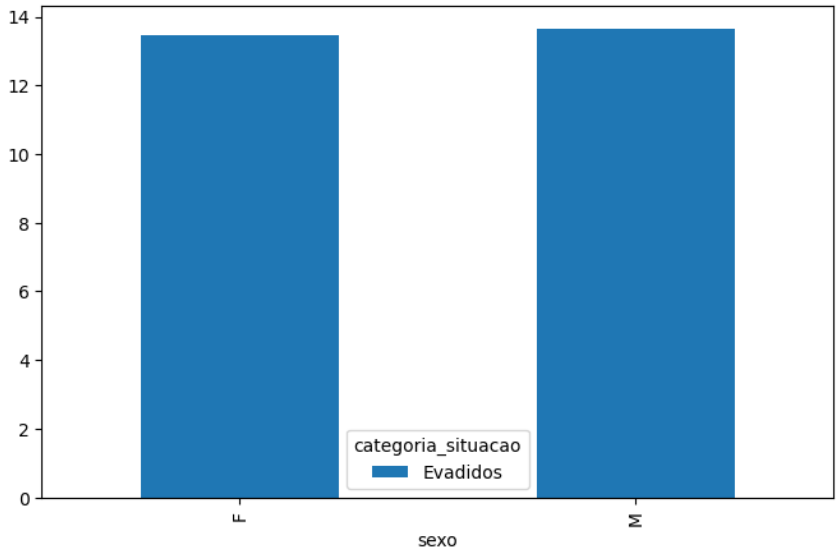

Figura 4. Evasão por Sexo

O tamanho dessa base também não impactou a etapa de aprendizado de máquina, já que ela não foi utilizada para fins preditivos.

5.2. Análise do Perfil Sociodemográfico e Econômico do Estudante

Esta subseção explora os microdados de 10.034 estudantes ativos ou evadidos do IFBA, mapeando a influência das características individuais sobre a evasão escolar por meio de análises bivariadas e representações visuais.

O quantitativo dos atributos demográficos demonstrou disparidade na distribuição de estudantes por gênero na comunidade do IFBA. Foi apresentado um total de 6.376 pessoas do sexo masculino (representando $63,5\%$ da amostra), enquanto o do sexo feminino foi de 3.658 pessoas ($36,5\%$). Apesar disso, o cruzamento bivariado indicou que a taxa de evasão se mantém muito próxima entre esses grupos, registrando $13,4\%$ para o sexo feminino e $13,6\%$ para o masculino, como visto na Figura 4.

A respeito da distribuição estudantil por autodeclaração de cor ou raça, foi visto que a maioria é de pardos. Eles contabilizaram um total de 4239 pessoas, o que representa $42,2\%$ da amostra, seguidos dos brancos com 2810 (28%), pretos com 1900 ($18,9\%$), indígenas com 135 ($1,3\%$) e amarelos, 69 ($0,7\%$). O total dos que não declararam foi de 881 pessoas, o que representa $8,8\%$ do total da amostra. Ao cruzar essas categorias com as taxas de evasão, foi observado comportamento similar ao visto entre os gêneros, apresentando índices homogêneos que variaram entre $13,2\%$ e $14,2\%$, exceto pelos indígenas que tiveram o menor índice de evasão, com $3,7\%$. É possível visualizar essa distribuição na Figura 5.

Já a análise da Renda Familiar Per Capita (RFP) demonstrou que o pico de evasão não se manifesta na faixa de extrema vulnerabilidade ($0 < RFP \leq 0,5$, com 14%), mas sim na faixa intermediária ($1 < RFP \leq 1,5$, com $15,6\%$) e na faixa superior ($RFP > 3,5$, com 16%), como é possível ver na Figura 6. Esse fenômeno sugere, por um lado, a eficiência das políticas institucionais de assistência estudantil (bolsas e auxílios) na retenção de alunos de baixíssima renda e, por outro, um vetor de evasão por

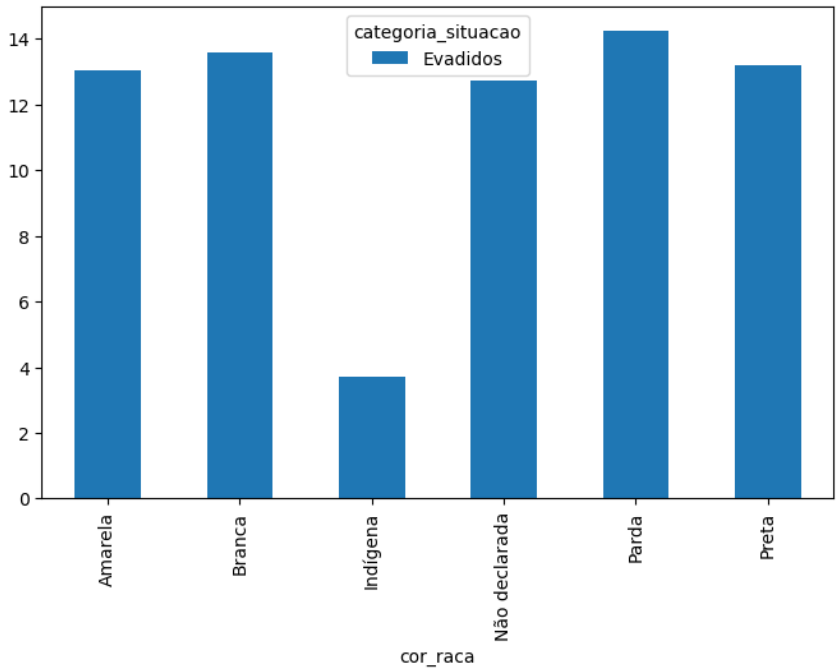

Figura 5. Evasão por Cor/Raça

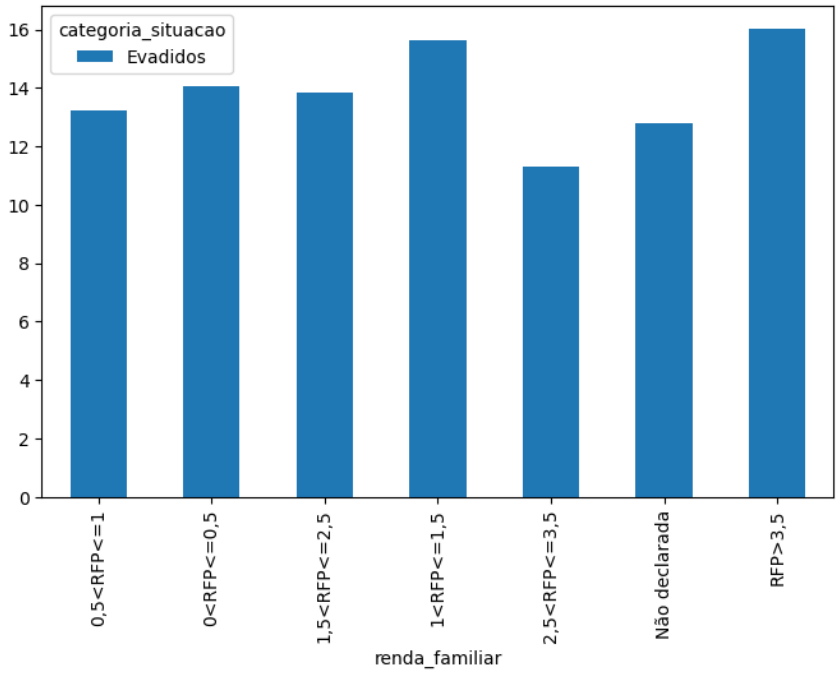

Figura 6. Evasão por Renda Familiar

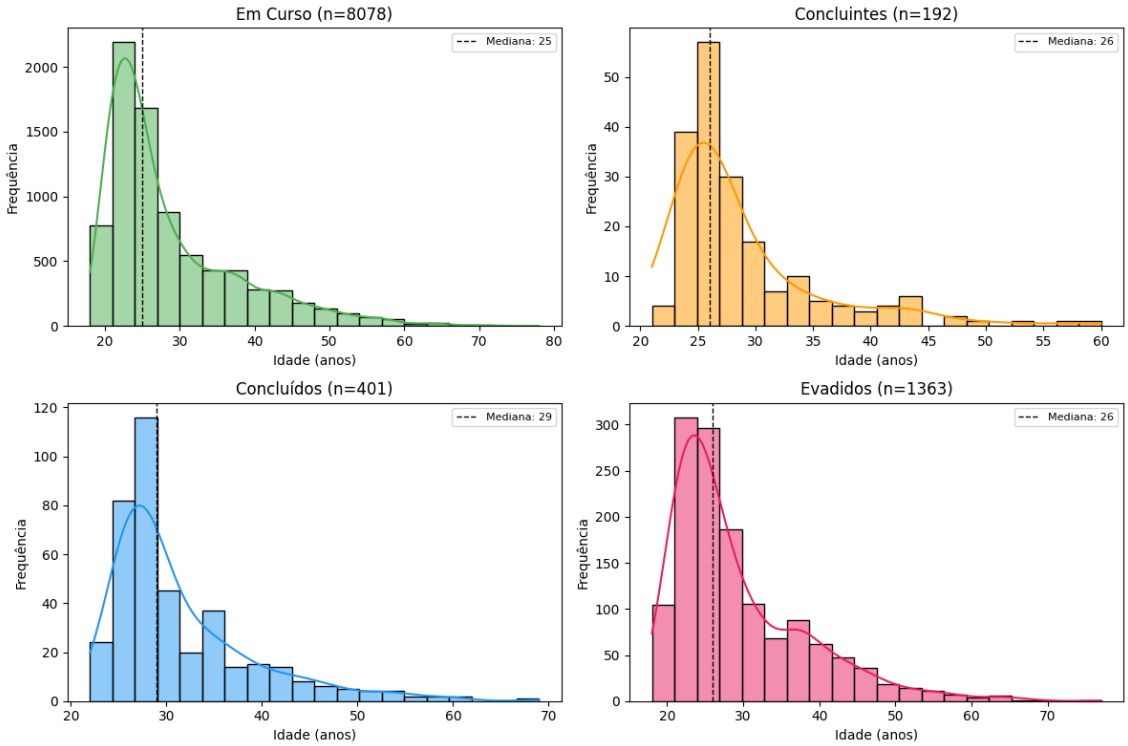

Figura 7. Distribuição de Idade por Situação de Matrícula no IFBA

transferência interna ou externa nos grupos de maior poder aquisitivo, ou seja, que têm a possibilidade de mudar de curso.

A análise da distribuição etária por situação de matrícula, apresentada na Figura 7, revelou perfis semelhantes entre os grupos, com mediana concentrada na faixa dos 25 a 29 anos em todas as situações. Chama atenção a presença de estudantes em idades avançadas em todos os grupos: a situação "Em Curso" registrou a maior idade máxima (78 anos), seguida de "Evadidos" (77 anos) e "Concluídos" (69 anos). A permanência de estudantes seniores na situação "Em Curso" por tempo indeterminado pode indicar retenção prolongada — estudantes que integralizaram disciplinas mas não concluíram formalmente o curso. A dinâmica da evasão por faixa etária pode ser vista na Figura 8, que apresenta a distribuição proporcional de abandonos ao longo das diferentes faixas.

A respeito das características dos cursos, é importante deixar explícito que as graduações do IFBA são majoritariamente bacharelados e de turno noturno. Dentre as 10.034 matrículas registradas, 4868 estavam vinculadas a cursos de bacharelado, 3122 a licenciaturas, e 2044 a cursos de tecnologia. Já sobre os quantitativos dos turnos, 6401 estudantes cursaram cursos noturnos, 1036 foram vespertinos, 975 matutinos, 779 de turno integral e 843 'Não se aplica'.

Tendo esses aspectos em vista, os turnos dos cursos se mostraram potenciais indicadores de risco. O Noturno registrou o maior índice de interrupção de vínculo com 14%, o que pode refletir as dificuldades geradas pelo cansaço diário e logística enfrentadas pelo estudante trabalhador. O turno Vespertino, por outro lado, disparou nos índices de Concluídos, com 5, 8%, como visto na Figura 9.

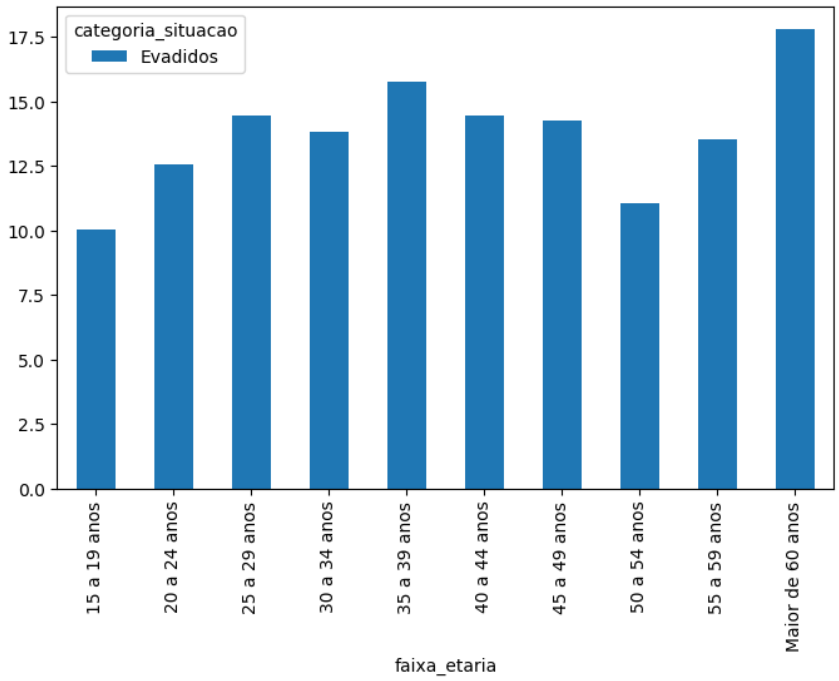

Figura 8. Evadidos por Faixa Etária

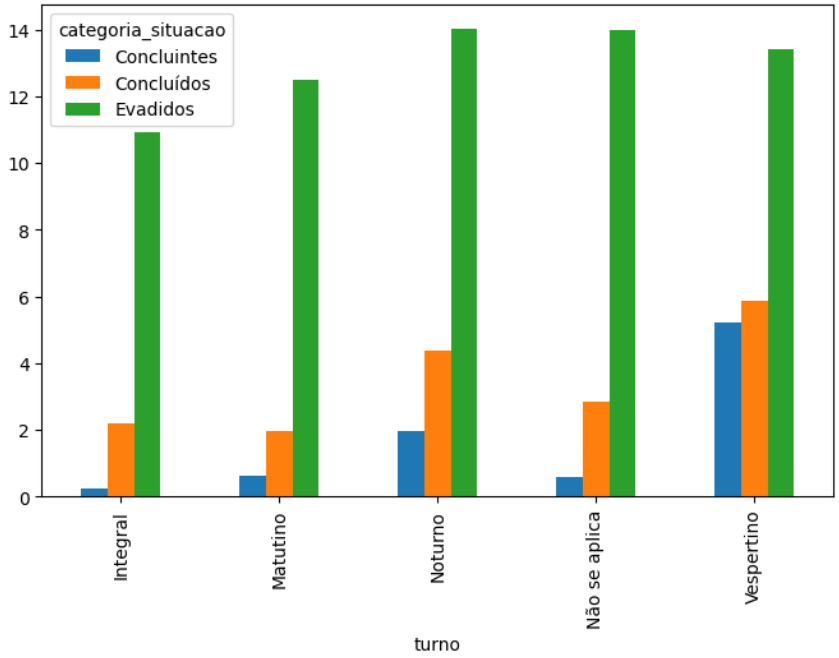

Figura 9. Evasão por Turno

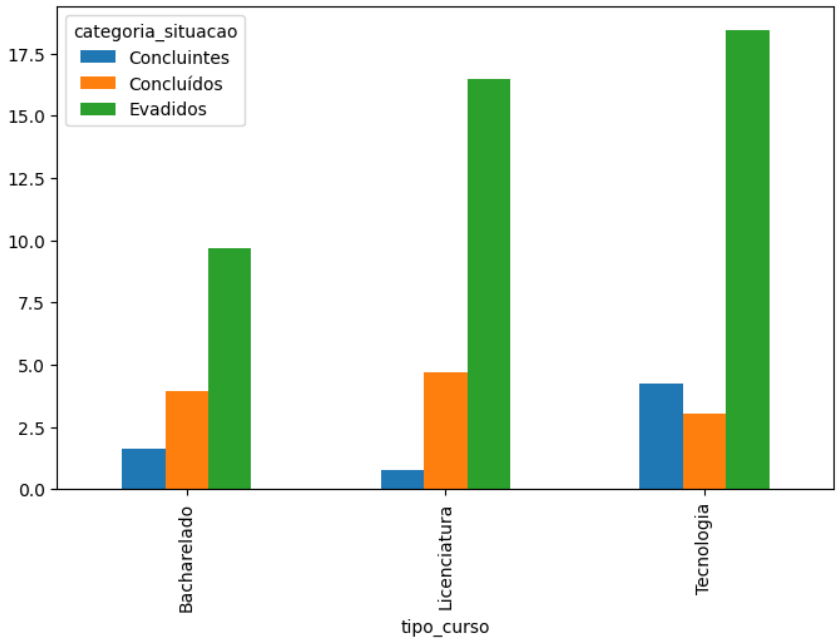

Figura 10. Evasão por Tipo de Graduação

Na análise dos tipos de graduação, os dados revelaram que os cursos de Tecnologia lideraram a evasão com $18,4\%$, superando os Bacharelados e as Licenciaturas (esta última registrando a melhor taxa de conclusão, $4,7\%$, enquanto os Bacharelados apresentaram a menor taxa de evadidos, com $9,6\%$). É possível visualizar essa dinâmica na Figura 10.

5.3. Análise de Coorte Temporal e Dinâmica da Evasão

Para compreender como o fenômeno da evasão se manifestou no decorrer dos anos, foi realizada a análise de coorte temporal baseada nos ciclos de matrículas ativos entre os anos de 2013 e 2023. Nessa subseção foi aplicado um recorte temporal a partir do ano de 2013, mesmo diante do acesso a dados desde 2011, reduzindo o escopo desta etapa da análise para exatamente uma década. É importante ressaltar, contudo, que as demais etapas do trabalho analisaram todos os dados disponíveis, de 2011 a 2023.

A análise dos dados temporais, aliada à visualização dos gráficos, apontou para dois contextos históricos em que se destacou o comportamento da evasão no IFBA. O primeiro se refere à coorte de 2020, no contexto da pandemia da COVID-19, em que a taxa de alunos na situação "Em Curso"apresentou o registro máximo de $89,61\%$, enquanto os de evasão despencaram para o mínimo histórico de $9,78\%$, como visto na Figura 11.

Esse cenário reflete o impacto das políticas institucionais de flexibilização acadêmica e suspensão temporária de jubilamentos durante o Ensino Remoto Emergencial, o que viabilizou a permanência dos estudantes. O segundo fenômeno se manifestou logo em seguida nas coortes de 2021 e 2022, num cenário pós pandêmico. Com a retomada compulsória das atividades presenciais, as deserções acumuladas vieram à tona, fazendo o indicador de evadidos saltar para $14,03\%$ em 2021 (um aumento relativo de quase 44%), e $14,7\%$ em 2022.

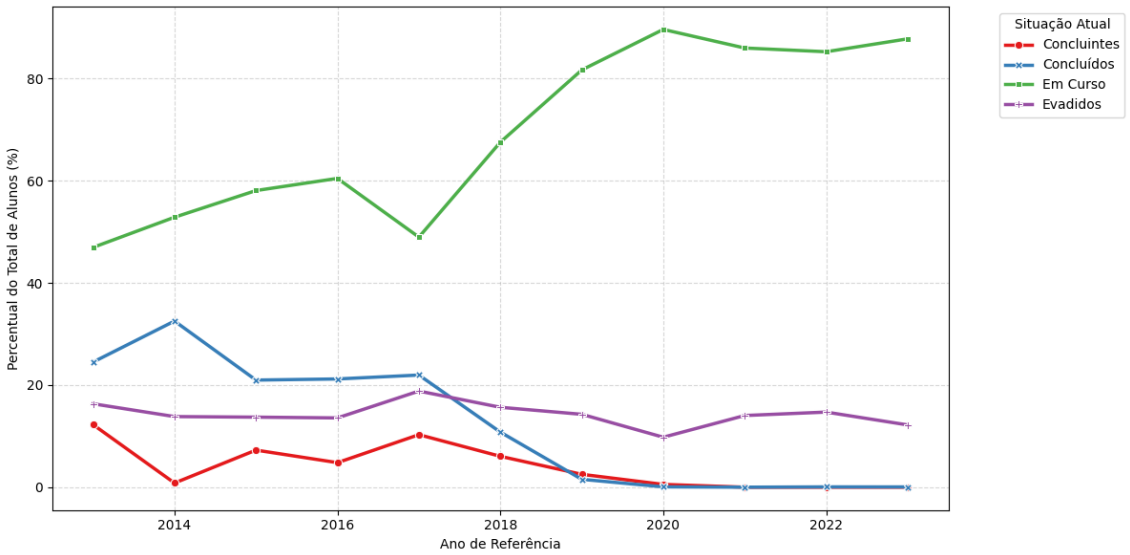

Figura 11. Evolução do Status de Matrícula por Coorte de Ingresso no IFBA

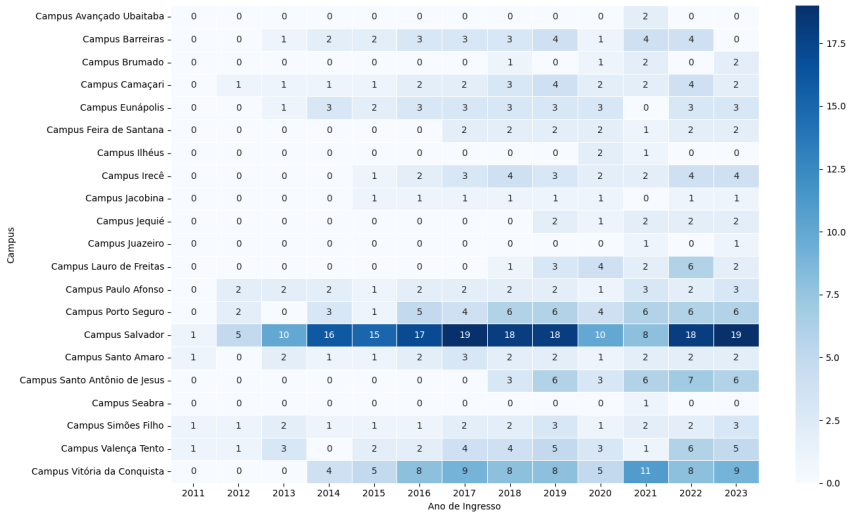

Figura 12. Ciclos de Matrícula Abertos por Campus e Ano de Ingresso

5.4. Panorama Institucional

Vistas as características e os aspectos das coortes e dos perfis estudantis, esta subseção move o foco analítico para o nível institucional. Foi investigada a estrutura de ofertas de vagas nos 599 ciclos registrados no IFBA. É importante considerar aspectos temporais nesse recorte também, já que a implantação de cursos superiores em alguns dos *campi* do IFBA aconteceu posteriormente a outros. Assim, na Figura 12 observa-se a distribuição de turmas consolidadas em cada *campus* com o passar dos anos.

Notou-se, novamente, o impacto da crise sanitária da COVID-19 a partir da redução do número de turmas nos anos de 2020 e 2021, ao mesmo tempo em que reforçou-se a eficiência das políticas emergenciais de permaência estudantil, como o auxílio financeiro emergencial [29] e as ações de inclusão digital [30], bem como a adoção do Ensino Remoto Emergencial, além de outras flexibilizações, para garantir a continuidade das ati-

vidades acadêmicas.

Foi possível perceber também que o histórico da distribuição de turmas do Ensino Superior do IFBA é desigual: enquanto alguns *campi* concentraram uma maior quantidade de ciclos, outros apresentaram dificuldade para fechar turmas. Por esse motivo, comparações entre quantitativos de ofertas de vagas entre *campi* seriam injustas neste estudo.

5.5. Resultados da Análise Exploratória dos Microdados de Eficiência Acadêmica

Ainda sob a perspectiva institucional, antes do levantamento das percepções extraídas a partir da análise exploratória dos microdados de eficiência acadêmica, os processos de ETL resultaram na estrutura consolidada que foi analisada. O arquivo bruto passou por uma seleção filtrada, remoção de duplicatas e anomalias, o que resultou numa base de dados com 57 registros, livre de valores ausentes.

Mesmo ao se tratar de uma amostra muito restrita, representando cerca de 10% do total dos ciclos registrados na base de microdados de matrículas, foi possível realizar análises que enriqueceram este estudo. Isso é porque os microdados de eficiência acadêmica fornecem atributos numéricos de panorama institucional inéditos em relação aos de matrículas, principalmente Fator de Esforço de Curso (FEC) e Carga Horária.

Ao iniciar a análise dessas variáveis, foi constatado que a correlação de Spearman entre elas foi de $0,33$. Isso indica que a associação entre elas é fraca e sugere que essas variáveis capturam dimensões complementares da estrutura curricular, justificando a inclusão independente de ambas na análise.

Foi realizado um *merge* com a base de matrículas para resgatar a taxa de evasão de cada ciclo da base de eficiência e fazer uma análise conjunta, mediante a chave comum `codigo_ciclo_matricula`, o que resultou em 57 ciclos. A partir dessa integração foi possível visualizar como a evasão se comporta junto aos indicadores de eficiência.

A correlação de Pearson entre a taxa de evasão e ambos atributos numéricos de eficiência é negativa, sendo $-0,17$ para FEC e $-0,31$ para Carga Horária. Isto sugere que estes indicadores de eficiência e a taxa de evasão se movimentaram em direções opostas nessa amostra de dados: à medida em que FEC e carga horária cresceram, a taxa de evasão caiu. É possível visualizar essa dinâmica na Figura 13, que apresenta essa tendência negativa nas retas de regressão linear e amplitude nos intervalos de confiança, que refletem as altas variabilidades entre os ciclos.

5.6. Avaliação e Desempenho dos Modelos Preditivos

Ao executar os modelos preditivos de classificação *Random Forest* e *XGBoost*, uma das métricas a se levar em consideração foi a AUC-ROC, já que é a mais adequada para dados desbalanceados. O XGBoost obteve AUC-ROC de $0,72$, superando o *Random Forest* $(0,70)$, indicando maior capacidade de discriminação entre as classes.

Valores entre 0,70 e 0,80 podem ser considerados aceitáveis na literatura de classificação binária, a depender do setor de aplicação, da complexidade do problema e do custo do erro [31]. No domínio do problema desta pesquisa, verdadeiros positivos ignorados podem ter um custo considerável, especialmente social e econômico. Este desempenho, portanto, mostrou-se insuficiente.

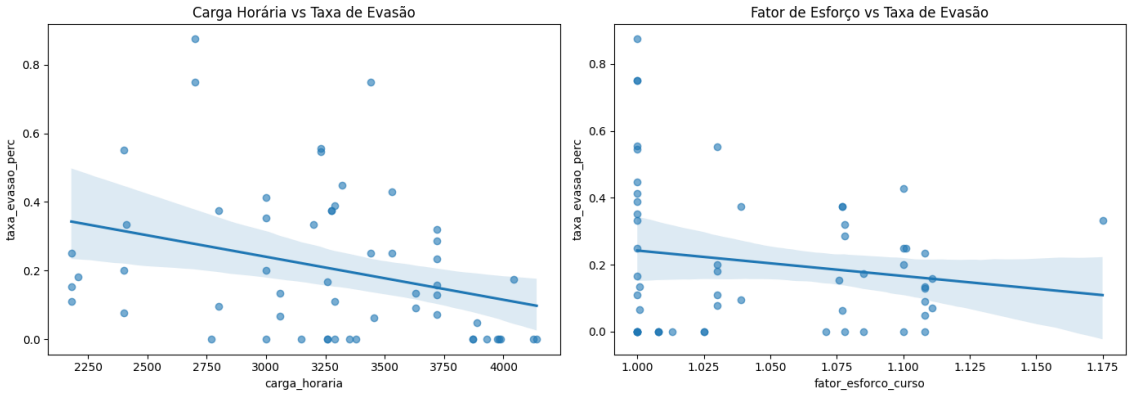

Figura 13. Dispersão entre carga horária, fator de esforço do curso e taxa de evasão por ciclo (N = 57), com reta de regressão linear

O *recall*, portanto, foi utilizado como uma métrica complementar, já que avalia como os modelos combrem esses casos em que o aluno de fato evadiu. Adicionalmente, em função do desbalanceamento entre as classes — 273 evadidos contra 119 concluintes e concluídos no conjunto de teste — a acurácia geral (70%) isoladamente não é um indicador confiável de desempenho. Por isso, o *F1-score* também foi adotado como métrica complementar, que pondera igualmente ambas as classes.

Para a classe Evadidos (classe positiva), o *XGBoost* alcançou precisão de $0,80$ e *recall* de $0,73$, com *F1-score* de $0,76$. No entanto, a identificação de alunos concluintes e concluídos (classe 0) apresentou desempenho mais modesto (F1 = $0,53$), o que reflete a limitação do conjunto de atributos disponíveis — restrito a variáveis socioeconômicas e de perfil de matrícula, sem dados de desempenho acadêmico.

O *Random Forest* apresentou *recall* ligeiramente superior para a classe de evasão ($0,78$ vs $0,73$). O *XGBoost*, contudo, demonstrou maior equilíbrio entre as classes e melhor AUC-ROC, sendo considerado o modelo de melhor desempenho geral para este estudo. Na Tabela 1 compara-se as métricas de avaliação dos dois modelos.

Tabela 1. Comparação de métricas entre os modelos Random Forest e XGBoost.

Métrica	**Random Forest**	**XGBoost**
AUC-ROC	0,70	**0,72**
F1-Score	0,65	0,65
Recall (Evadidos)	**0,78**	0,73
Recall (Concluídos + Concluintes)	0,52	**0,58**

Esses resultados são consistentes com estudos que utilizaram exclusivamente dados de perfil socioeconômico para predição de evasão, nos quais AUC-ROC entre $0,65$ e $0,75$ é reportado como teto esperado na ausência de variáveis de desempenho acadêmico [32]. A incorporação de indicadores como coeficiente de rendimento e histórico de reprovações tende a elevar substancialmente a capacidade preditiva dos modelos.

Adicionalmente, foram analisadas as matrizes de confusão dos modelos, dispostas na Figura 14, para compreender a quantidade de erros e de acerto deles e elas revelaram

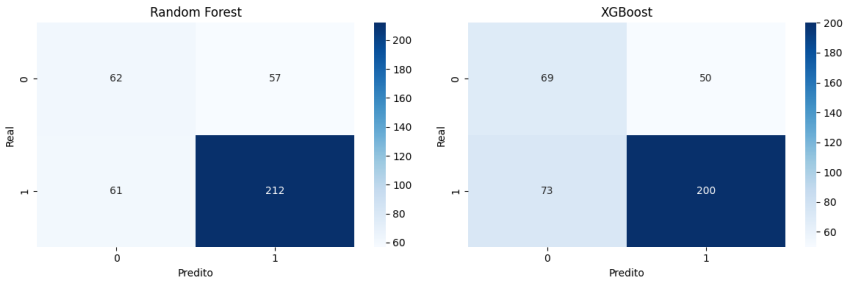

Figura 14. Matrizes de Confusão dos modelos Random Forest e XGBoost

padrões complementares entre os dois modelos. O *Random Forest* identificou corretamente 212 dos 273 alunos evadidos (*recall* de $77,7\%$), gerou 61 falsos negativos e 57 falsos positivos.

O *XGBoost*, por sua vez, apresentou comportamento mais conservador: acertou 200 dos 273 evadidos (*recall* de $73,3\%$), com 73 falsos negativos, porém com menor volume de falsos positivos (50). Essa diferença evidencia um *trade-off* entre os modelos: o *Random Forest* é mais sensível à detecção de evasão, ao custo de mais alertas falsos, enquanto o *XGBoost* é mais preciso na triagem, mas deixa de detectar um maior número de alunos em risco.

No contexto de políticas de intervenção institucional, a escolha entre os dois depende da capacidade de atendimento disponível: se o objetivo é maximizar a cobertura de alunos vulneráveis, o Random Forest se mostra mais adequado; se a prioridade é direcionar recursos de forma mais eficiente, o XGBoost é preferível.

5.7. Análise de Importância de Atributos (Interpretabilidade)

A análise de importância de atributos revelou concordância entre os dois modelos no fator mais relevante para as predições: a faixa etária de 20 a 24 anos foi o preditor de maior peso tanto no *Random Forest* ($11,2\%$) quanto no *XGBoost* ($25,8\%$). Esse resultado é coerente com o perfil predominante de ingressantes no ensino superior público, e com o dado de que as maiores taxas de evasão se concentram no primeiro ano da graduação [5].

Em segundo plano, o tipo de curso, especialmente Bacharelado e Tecnologia, aparece com destaque em ambos os modelos, sugerindo que a natureza e a duração do curso exercem influência sobre o risco de evasão. O turno noturno também figura entre os atributos relevantes, o que é consistente com a literatura: alunos noturnos frequentemente conciliam estudo e trabalho. Variáveis de cor/raça e renda familiar completam o conjunto de atributos mais informativos, reforçando o papel dos aspectos socioeconômicos no fenômeno da evasão.

É importante observar que o *Random Forest* apresentou distribuição de importância mais uniforme entre os atributos (característica típica do método, que divide o peso entre todas as variáveis relevantes), como é possível visualizar na Figura 15. O *XGBoost*, por sua vez, concentrou maior parte da importância no atributo dominante, reflexo da natureza sequencial do algoritmo de *boosting*, como visto na Figura 16.

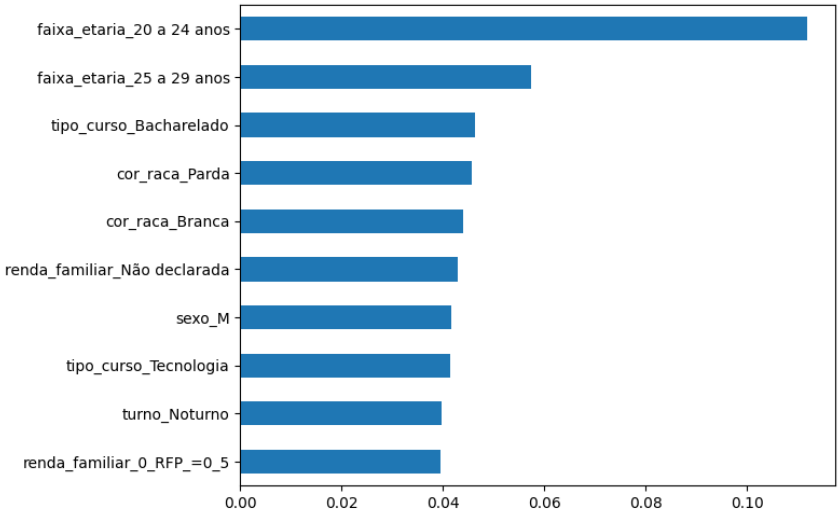

Figura 15. Top 10 Atributos — Random Forest

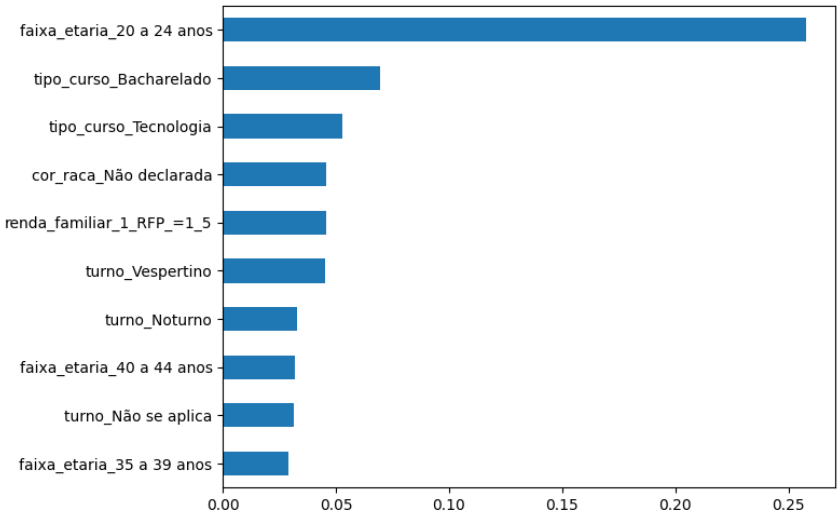

Figura 16. Top 10 Atributos — XGBoost

6. Conclusão

A evasão escolar é uma problema que historicamente impacta a sociedade através de múltiplas frentes e a ciência de dados aplicada tem oferecido técnicas promissoras para o estudo desse fenômeno. A investigação da evasão escolar através da análise de dados e algoritmos de aprendizado de máquina pode auxiliar gestores educacionais ao apoiar a tomada de decisão.

Nesta pesquisa, dados acadêmicos públicos da Plataforma Nilo Peçanha (PNP) foram extraídos e tratados de modo a formar uma base para o processo de descoberta de conhecimento em bancos de dados, cobrindo o escopo de cursos de ensino superior do Instituto Federal da Bahia. A partir dessa base de dados, foi realizado o processo de Análise Exploratória de Dados, no qual foram investigadas características institucionais e estudantis dos alunos de graduação de todos os *campi* do IFBA, além do desenvolvimento de modelos de aprendizagem de máquina para prever padrões de evasão a partir desses dados.

O objetivo deste trabalho foi realizar um estudo sobre a relação entre atributos de potencial impacto na permanência estudantil, de modo a analisar o contexto e perfil dos estudantes com maior propensão à evasão escolar. Para isso, foram desenvolvidos processos de extração, tratamento e carga de dados nas bases de matrícula e de eficiência acadêmica; foi realizada uma análise exploratória com esses dados acadêmicos e, para analisar os maiores fatores de risco para a evasão, foram utilizados dois modelos preditivos de aprendizado de máquina: *Random Forest* e *XGBoost*.

As características centrais da implementação do projeto envolveram o uso da linguagem de programação Python, suas bibliotecas principais para ciência de dados aplicada (Matplotlib, Scikit-learn, Pandas, etc.), e desenvolvimento integral no ambiente interativo Jupyter Notebook. Além desses aspectos é relevante mencionar a organização modular da arquitetura do projeto: dados (brutos e tratados) e *Notebooks* ficaram reservados em diretórios próprios, de modo a seguir boas práticas em análise de dados e garantir a reprodutibilidade do estudo.

Durante o desenvolvimento do projeto, algumas limitações se manifestaram. Dentre elas se destaca a dificuldade de acesso a dados de desempenho acadêmico, que poderiam se integrar aos dados demográficos dos estudantes e enriquecer o estudo, especialmente na etapa de aprendizagem de máquina. No entanto, o seu caráter é mais sensível e o seu acesso requer etapas burocráticas como a aprovação do Conselho de Ética em Pesquisa, que poderiam representar um gargalo para este estudo. Optou-se, portanto, por manter apenas os dados públicos da PNP, que ofereceram diversas nuances para a pesquisa.

Outras limitações relevantes de mencionar referem-se a alguns aspectos da fonte de dados utilizada. A ausência de atributos que expressassem o tipo de ingresso do aluno, sua situação efetiva no sistema acadêmico utilizado pela instituição, bem como informações sobre pessoa com deficiência reduziram o potencial explicativo do trabalho.

Esta pesquisa foi consoante a trabalhos correlatos existentes na literatura, ao adotar abordagens similares como o uso de algoritmos preditivos de classificação para analisar o fenômeno da evasão. Análises exploratórias de dados também são recorrentes no estudo desse problema. Mesmo que esse tipo de pesquisa seja amplamente desenvolvido

na literatura atual, é importante que essa abordagem continue sendo adotada, já que se trata de um problema persistente e de características voláteis, isto é, suas motivações podem mudar com o passar do tempo. Além disso, a realização dessa análise por meio de diferentes fontes de dados pode gerar novas percepções sobre o mesmo fenômeno e enriquecer muitas áreas do conhecimento, como a Educação e a Computação.

Tendo isso em vista, este estudo contribuiu com essas análises ao oferecer uma perspectiva em diferentes níveis, desde o estudantil, até o institucional e temporal. Os resultados obtidos a partir da etapa de análise exploratória desta pesquisa reforçaram o perfil predominante de estudantes trabalhadores no IFBA, com características demográficas e socioeconômicas diversas. Essas características isoladamente não expressaram grande potencial de risco para o fenômeno da evasão, diferente de alguns aspectos institucionais, como turno e os tipos de graduação. Já a análise de aspectos temporais foi relevante para mostrar como foram eficientes as políticas de permanência durante a crise sanitária da COVID-19.

A respeito da etapa de aprendizagem de máquina, notou-se que as métricas de avaliação dos modelos apresentaram desempenho modesto. Isso pode se justificar pela predominância de atributos de natureza demográfica e socioeconômica na base de dados usada para o treino dos modelos. A literatura em ciência de dados aplicada à área da gestão escolar sugere que esse tipo de atributo desempenha um papel secundário para este tipo de análise e atributos que expressam histórico e desempenho acadêmicos têm poder preditivo mais forte [33].

Mesmo diante disso, é relevante analisar variáveis de caráter demográfico, pois eles contextualizam a situação do estudante, enriquecem o estudo e, em conjunto com dados de desempenho acadêmico, tornam a pesquisa mais completa. Ressalta-se, ainda, que, na etapa de análise de importância de atributos, dados demográficos tiveram destaque para ambos os modelos. Isto, contudo, não contradiz o que foi exposto sobre o poder preditivo dos modelos, uma vez que esse tipo de dado predominou a base de treino e foi a partir desses dados disponíveis que os modelos aprenderam.

Quanto à utilização do projeto implementado, é voltada a gestores educacionais e/ou entidades envolvidas em processos de tomada de decisão para a aplicação de políticas públicas de permanência estudantil. A pesquisa pode auxiliar esse público ao indicar os fatores e contextos em que a evasão mais se manifestou. Além das percepções aqui levantadas, é possível que, futuramente, novas sejam trazidas à tona por aqueles que se envolverem nesses processos, seja a partir dos dados analisados, seja a partir de dados divulgados futuramente. É possível que o modelo de aprendizagem de máquina seja atualizado com dados novos, desde que as colunas da base de dados sejam compatíveis e que se siga o mesmo *pipeline* aqui consolidado.

Há, contudo, lacunas a serem cobertas em futuros trabalhos. Mesmo diante do acesso a uma estrutura tabular, fruto do processo de tratamento de dados, contendo informações sobre distribuição de vagas de diversas modalidades, este estudo não pretendeu lançar um olhar focado para essa dimensão. Essa tabela foi usada apenas para observar como se deu o ingresso de estudantes com o passar dos anos, considerando o tamanho de cada turma.

Entretanto, não se analisou como foi a distribuição dessas turmas por tipo de vaga,

nem foi feita a análise bivariada com o cruzamento de dados de evasão e modalidades de vagas. Recomenda-se, ainda, que em trabalhos futuros se acrescente, junto aos atributos preditores dos modelos de aprendizagem de máquina, fatores de desempenho acadêmico, como quantidade de reprovações, trancamentos e coeficientes de rendimento.

Outras possibilidades para a continuidade deste estudo incluem integrar modelos adicionais de aprendizagem de máquina para fazer uma análise comparativa. Além disso, outras técnicas de interpretabilidade além da Importância de Atributos podem ser acrescentadas, como LIME ou SHAP, que podem oferecer explicações mais detalhadas das previsões e aumentar a transparência para gestores.

Por fim, outra recomendação de trabalhos futuros se baseia numa limitação metodológica desta pesquisa. Do ponto de vista estatístico, a avaliação dos classificadores baseou-se apenas na separação de treino e teste dos dados, o que gera estimativas pontuais de desempenho. Para trabalhos futuros e para o aprimoramento do rigor metodológico, portanto, recomenda-se a implementação da Validação Cruzada (*k-fold Cross-Validation*).

A execução de múltiplos testes permitirá a aplicação de estatística intervalar (como o cálculo de Intervalos de Confiança para as métricas de avaliação). Essa abordagem permitirá mitigar possíveis vieses de particionamento e fornecer uma estimativa mais segura e robusta da capacidade de generalização dos algoritmos frente a novas coortes de estudantes.

Referências

[1] Maurício Thiago Ferreira de Lima. Uma análise baseada em dados sobre o desempenho ea evasão escolar do estudante da rede pública estadual do rn. 2023. 1

[2] Marcelo Cortes Neri. O tempo de permanência na escola e a desigualdade de renda no brasil. Technical report, Centro de Políticas Sociais, Fundação Getulio Vargas (FGV/CPS), Rio de Janeiro, 2009. 1, 2

[3] Felippe Flain Pires Santos, Landara Maitê Simon, and Nelson Guilherme Machado Pinto. Retenção e evasão escolar em um instituto federal de educação, ciência e tecnologia. *Revista Científica da Ajes*, 9(18), 2020. 2, 4

[4] Isleimar de Souza Oliveira. Análise de dados aplicada à evasão escolar: um estudo de caso do ifpb. Master's thesis, 2023. 2, 5, 7, 12

[5] Organização para a Cooperação e Desenvolvimento Econômico. Education at a glance 2025: OECD indicators. Technical report, OECD Publishing, Paris, 2025. Acesso em: 24 jan. 2026. 2, 31

[6] Russell W. Rumberger. *Dropping Out: Why Students Drop Out of High School and What Can Be Done About It*. Harvard University Press, Cambridge, MA, 2011. 2

[7] Ricardo Paes de Barros, Samuel Franco, and Hudson Mueller. Consequências econômicas da evasão escolar no brasil. Technical report, Insper, Fundação Roberto Marinho e Instituto Unibanco, São Paulo, 2020. 2

[8] Dyego Barbosa, Luciano Cabral, Filipe Dwan, Elyda Feitas, and Rafael Ferreira Mello. Previsão da evasão escolar através da análise de dados e aprendizagem de máquina: Um estudo de caso. In *Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil (WAPLA)*, pages 42–50. SBC, 2023. 2, 12

[9] Leonardo de Almeida Teodoro and Marco André Abud Kappel. Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no brasil. *Revista Brasileira de Informática na Educação*, 28:838–863, 2020. 2, 4

[10] Laís Bássame Rodrigues, Ricardo Kagimura, Brenda Gabrielly da Silva Cardoso, Alessandra Riposati Arantes, and Marili Peres Junqueira. Evasão e retenção no ensino superior: abordagem baseada em taxas quantitativas. *Revista Contemporânea de Educação*, 36(16):4–21, 2021. 4

[11] BRASIL. Ministério da Educação. *Portaria nº 146, de 11 de março de 2021. Dispõe sobre as diretrizes para o cálculo dos indicadores de gestão das instituições integrantes da Rede Federal de Educação Profissional, Científica e Tecnológica*. Secretaria de Educação Profissional e Tecnológica, Brasília, DF, 2021. 4

[12] INEP. Metodologia de cálculo dos indicadores de fluxo da educação superior. Technical report, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília, 2017. 4

[13] Roberto Leal Lobo Silva Filho, Paulo Roberto Motejunas, Oscar Hipólito, and Maria Beatrice de Carvalho Melo Lobo. A evasão no ensino superior brasileiro. *Cadernos de Pesquisa*, 37(132):641–659, 2007. 5

[14] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–54, 1996. 5

[15] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan Kaufmann, 3 edition, 2011. 5

[16] Joe Reis and Matt Housley. *Fundamentals of Data Engineering*. O'Reilly Media, 2022. 5

[17] Deepak Vohra. Practical hadoop ecosystem. *Chapter in Apache Parquet*, 177:178, 2016. 5

[18] Geraldo Mendes Batista Neto. Análise e previsão da evasão escolar no ensino médio em instituições federais brasileiras. B.S. thesis, 2024. 6, 7

[19] Wilton de O Bussab and Pedro A Morettin. *Estatística Básica*. Saraiva Educação S.A., São Paulo, 9 edition, 2017. 6, 7

[20] Joseph F Hair Jr, William C Black, Barry J Babin, Rolph E Anderson, and Ronald L Tatham. *Análise multivariada de dados*. Bookman Editora, Porto Alegre, 7 edition, 2014. 6

[21] Katti Faceli, Ana Carolina Lorena, João Gama, Tiago Agostinho de Almeida, and André Carlos Ponce de Leon Ferreira de Carvalho. Inteligência artificial: uma abordagem de aprendizado de máquina. 2021. 7, 8, 9, 10, 14, 21

[22] Matt Harrison. *Machine Learning: Guia de Referência Rápida: Trabalhando com Dados Estruturados em Python*. Novatec Editora, São Paulo, 2020. Tradução: Lúcia A. Kinoshita. 8, 9, 10, 11

[23] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2 edition, 2022. 9

[24] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., Sebastopol, CA, 3 edition, 2022. 11

[25] Alex Marques de Souza. *Machine learning e a evasão escolar: análise preditiva no suporte à tomada de decisão*. PhD thesis, Mestrado em Sistemas de Informação e Gestão do Conhecimento, 2020. 12

[26] Francisco Matheus Valença Trajano. Análise preditiva e exploratória de dados para auxiliar no combate à evasão estudantil nos cursos superiores do ifpb. B.S. thesis, 2023. 12

[27] BRASIL. Ministério da Educação. Secretaria de Educação Profissional e Tecnológica. Microdados da plataforma nilo peçanha: Ano de referência 2023. Dados Abertos Governamentais, 2023. Acesso em: 19 maio 2026. 13

[28] Wes McKinney. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter*. O'Reilly Media, Inc., Sebastopol, CA, 3 edition, 2022. 16

[29] INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA BAHIA. Resolução nº 08, de 17 de abril de 2020. Conselho Superior (CONSUP), 2020. Dispõe sobre as ações emergenciais de assistência estudantil durante a pandemia de Covid-19. Disponível em: ¡`https://ifba.edu.br`¿. Acesso em: 10 jul. 2026. 28

[30] INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA BAHIA. Resolução nº 23, de 17 de setembro de 2020. Conselho Superior (CONSUP), 2020. Regulamenta o Auxílio de Inclusão Digital Emergencial no âmbito do IFBA. Disponível em: ¡https://portal.ifba.edu.br/dmc/inclusao-digital¿. Acesso em: 10 jul. 2026. 28

[31] David W Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, New York, 2nd edition, 2000. 29

[32] Neema Mduma, Khamisi Kalegele, and Dina Machuve. A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal*, 18(1):14–14, 2019. 30

[33] Elaf Alyahyan and Dilek Düştegör. Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(1):1–21, 2020. 34