UNIVERSIDADE FEDERAL DO MARANHÃO

Bacharelado Interdisciplinar em Ciência e Tecnologia

Daniel Nunes Duarte, Tiago de Lima Batista, Denilson da Silva

ANÁLISE PREDITIVA DE DESEMPENHO ACADÊMICO: UM ESTUDO SOBRE EVASÃO ESTUDANTIL

São Luís

2025

Daniel Nunes Duarte, Tiago de Lima Batista, Denilson da Silva

ANÁLISE PREDITIVA DE DESEMPENHO ACADÊMICO: UM ESTUDO SOBRE EVASÃO ESTUDANTIL

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Thales

São Luís

2025

"Não fiques em terreno plano.
Não subas muito alto.
O mais belo olhar sobre o mundo
Está a meia encosta."

Friedrich Nietzsche, em *"A Gaia Ciência"*

Resumo

Este trabalho apresenta uma análise abrangente do desempenho acadêmico e padrões de evasão estudantil utilizando técnicas de mineração de dados. A pesquisa emprega uma metodologia que combina análise exploratória de dados, modelagem preditiva e avaliação de risco para identificar fatores determinantes no sucesso ou abandono acadêmico. Utilizando o conjunto de dados "Predict Students' Dropout and Academic Success", com informações de 4.424 estudantes portugueses coletadas entre 2008 e 2017, o estudo implementa um sistema de análise modular que inclui preparação e validação de dados, análise de correlações, avaliação de padrões semestrais e identificação de risco. Os resultados obtidos através do algoritmo Random Forest demonstram alta precisão na identificação de estudantes propensos à evasão, com AUC-ROC superior a 0,85 e F1-Score de 0,82. A análise revelouque fatores como desempenho no primeiro semestre, status financeiro e background educacional familiar são determinantes críticos para o sucesso acadêmico. As descobertas oferecem insights valiosos para o desenvolvimento de estratégias de intervenção preventivas, contribuindo para políticas educacionais mais efetivas e para a redução das taxas de abandono no ensino superior.

Palavras-chave: Evasão Estudantil, Mineração de Dados Educacionais, Machine Learning, Random Forest, Análise Preditiva.

Abstract

This work presents a comprehensive analysis of academic performance and student dropout patterns using data mining techniques. The research employs a methodology that combines exploratory data analysis, predictive modeling, and risk assessment to identify determining factors in academic success or dropout. Using the "Predict Students' Dropout and Academic Success" dataset, containing information from 4,424 Portuguese students collected between 2008 and 2017, the study implements a modular analysis system that includes data preparation and validation, correlation analysis, semester pattern evaluation, and risk identification. The results obtained through the Random Forest algorithm demonstrate high accuracy in identifying students prone to dropout, with AUC-ROC above 0.85 and F1-Score of 0.82. The analysis revealed that factors such as first-semester performance, financial status, and family educational background are critical determinants of academic success. The findings offer valuable insights for developing preventive intervention strategies, contributing to more effective educational policies and reducing dropout rates in higher education.

Keywords: Student Dropout, Educational Data Mining, Machine Learning, Random Forest, Predictive Analysis.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AUC-ROC	*Area Under the Curve - Receiver Operating Characteristic*
BICT	*Bacharelado Interdisciplinar em Ciência e Tecnologia*
EDM	*Educational Data Mining*
F1	*F1-Score*
FN	*Falsos Negativos*
FP	*Falsos Positivos*
LPPL	*LaTeX Project Public License*
UFMA	*Universidade Federal do Maranhão*
VN	*Verdadeiros Negativos*
VP	*Verdadeiros Positivos*

Sumário

1 Introdução

A evasão estudantil no ensino superior constitui um desafio persistente e multifacetado que afeta instituições educacionais globalmente (SILVA; TERRA, 2017). Este fenômeno transcende a simples desistência individual, representando uma questão complexa que impacta significativamente o desenvolvimento socioeconômico das nações e a eficiência dos sistemas educacionais (TINTO, 2014). Em um cenário caracterizado pela crescente digitalização do ensino e pela disponibilidade sem precedentes de dados educacionais, a análise preditiva emerge como uma ferramenta promissora para a identificação precoce e prevenção da evasão estudantil (BAKER; INVENTADO, 2014).

1.1 Contextualização

A persistência do fenômeno da evasão ao longo das décadas tem desafiado educadores e gestores em diferentes contextos educacionais. No Brasil, o cenário é particularmente preocupante, com taxas de evasão que oscilam entre 20% e 50% no ensino superior, variando significativamente entre diferentes cursos e instituições (LOBO, 2012). Esta realidade é especialmente crítica em instituições públicas, onde a evasão não apenas representa um desperdício de recursos públicos, mas também compromete a capacidade do sistema educacional de contribuir efetivamente para o desenvolvimento social e econômico (BAGGI; LOPES, 2011).

A natureza multidimensional da evasão, como destacado por (SPADY, 1970), envolve uma intrincada rede de fatores que influenciam a decisão do estudante de permanecer ou abandonar seus estudos. Esta complexidade é amplificada pela interação dinâmica entre aspectos socioeconômicos, acadêmicos, institucionais e pessoais (CABRERA; NORA; CASTAÑEDA, 1992), tornando o processo de identificação de estudantes em risco um desafio que demanda abordagens sofisticadas e multifacetadas.

O advento das tecnologias de informação e a consequente proliferação de dados educacionais têm criado novas possibilidades para a compreensão e o enfrentamento deste fenômeno. Técnicas avançadas de análise de dados oferecem perspectivas inéditas para identificar padrões e fatores associados à evasão (ROMERO; VENTURA, 2013), permitindo o desenvolvimento de intervenções mais precisas e personalizadas (MARBOUTI; DIEFES-DUX; MADHAVAN, 2016).

1.2 Motivação

A motivação para este trabalho surge da necessidade de desenvolver ferramentas mais eficazes para o combate à evasão estudantil (SILVA; TERRA, 2017). A utilização do conjunto de dados "Predict Students' Dropout and Academic Success" (Universidade de Lisboa, 2018) oferece uma oportunidade única para investigar os fatores que influenciam o desempenho acadêmico e a decisão de permanência ou evasão dos estudantes.

O desenvolvimento de modelos preditivos para evasão estudantil representa uma abordagem inovadora para um problema tradicional da educação superior (DELEN, 2010). A integração de técnicas de machine learning com dados educacionais permite não apenas a identificação de padrões complexos, mas também a possibilidade de intervenções personalizadas e baseadas em evidências (MARBOUTI; DIEFES-DUX; MADHAVAN, 2016).

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo principal deste trabalho é desenvolver um modelo preditivo capaz de identificar precocemente estudantes em risco de evasão, utilizando técnicas de machine learning aplicadas a dados acadêmicos e socioeconômicos.

1.3.2 Objetivos Específicos

Para atingir o objetivo geral, foram estabelecidos os seguintes objetivos específicos:

- Analisar e preparar o conjunto de dados "Predict Students' Dropout and Academic Success" seguindo as melhores práticas de pré-processamento.
- Identificar os principais fatores associados à evasão estudantil através de análise exploratória de dados, utilizando técnicas estatísticas e visualizações.
- Desenvolver e comparar diferentes modelos de machine learning para previsão de evasão, incluindo algoritmos tradicionais e técnicas avançadas.
- Avaliar o desempenho dos modelos utilizando métricas apropriadas, considerando as características específicas do problema de evasão.
- Propor estratégias de intervenção baseadas nos resultados obtidos, alinhadas com as melhores práticas da literatura.

1.4 Justificativa

A relevância deste trabalho se justifica por diversos aspectos:

1. **Impacto Social**: A redução da evasão estudantil contribui para a formação de profissionais qualificados e para o desenvolvimento social (SILVA; TERRA, 2017). (BAGGI; LOPES, 2011) destaca que o combate à evasão é fundamental para a democratização do ensino superior.

2. **Eficiência Institucional**: A identificação precoce de estudantes em risco permite melhor alocação de recursos e implementação de medidas preventivas mais eficazes (TINTO, 2006). (LOBO, 2012) aponta que a retenção de estudantes é mais custo-efetiva do que o recrutamento de novos alunos.

3. **Inovação Metodológica**: A aplicação de técnicas avançadas de machine learning no contexto educacional representa uma abordagem inovadora para o problema da evasão (ROMERO; VENTURA, 2013). (BAKER; INVENTADO, 2014) ressalta o potencial transformador da análise preditiva na educação.

4. **Contribuição Científica**: Os resultados obtidos podem contribuir para uma melhor compreensão dos fatores que influenciam a evasão estudantil e para o desenvolvimento de estratégias mais efetivas de retenção (DELEN, 2010).

1.5 Metodologia

Este trabalho adota uma abordagem quantitativa, utilizando técnicas de mineração de dados e machine learning (HAN; KAMBER; PEI, 2011). O processo metodológico inclui:

1. Coleta e preparação dos dados
2. Análise exploratória
3. Desenvolvimento de modelos preditivos
4. Avaliação e validação dos resultados
5. Proposição de estratégias de intervenção

1.6 Estrutura do Trabalho

Este trabalho está organizado da seguinte forma:

O Capítulo 2 apresenta a fundamentação teórica, abordando conceitos fundamentais sobre evasão estudantil, análise preditiva e métricas de avaliação. O Capítulo 3 descreve a metodologia utilizada, incluindo a descrição do conjunto de dados, as técnicas de pré-processamento e os modelos implementados. O Capítulo 4 apresenta os resultados obtidos e sua discussão, seguindo as diretrizes de (HAIR et al., 2019) para análise e interpretação. Por fim, o Capítulo 5 traz as conclusões e sugestões para trabalhos futuros.

2 Fundamentação Teórica

A análise da evasão estudantil tem evoluído de abordagens puramente descritivas para modelos preditivos sofisticados, impulsionada pelos avanços em mineração de dados e aprendizado de máquina. Esta evolução permite não apenas compreender os fatores que influenciam o abandono acadêmico, mas também desenvolver intervenções preventivas baseadas em evidências.

2.1 Evasão Estudantil no Ensino Superior

(TINTO, 2014) caracteriza a evasão estudantil como um processo complexo de desvinculação do ensino superior, enfatizando sua natureza multifacetada. Este fenômeno transcende o simples ato de abandono, envolvendo uma intrincada rede de fatores acadêmicos, socioeconômicos e psicológicos que interagem ao longo da trajetória estudantil.

Os impactos deste fenômeno manifestam-se tanto no nível individual, comprometendo a trajetória profissional e o desenvolvimento pessoal (BEAN, 1985), quanto no âmbito institucional, através da ineficiência na utilização de recursos e redução de indicadores de desempenho. Na dimensão social, a evasão resulta em perdas significativas para o desenvolvimento econômico e redução na formação de profissionais qualificados.

2.2 Mineração de Dados Educacionais

A mineração de dados educacionais (Educational Data Mining - EDM) emerge como ferramenta fundamental para compreensão e prevenção da evasão. Segundo (BAKER; YACEF, 2009), a EDM integra análise comportamental, previsão de risco e implementação de intervenções direcionadas, permitindo uma abordagem sistemática ao problema da evasão.

2.3 Determinantes do Desempenho Acadêmico

(PASCARELLA; TERENZINI, 2005) identifica como elementos cruciais a renda familiar e o background educacional, que fornecem a base material e cultural necessária para o sucesso acadêmico. (ASTIN, 1984) complementa esta análise destacando a importância do desempenho inicial e do engajamento ativo no processo de aprendizagem como indicadores significativos da trajetória acadêmica.

2.4 Modelos Preditivos

(ROMERO; VENTURA, 2010) estabelece o Random Forest e as Redes Neurais como abordagens predominantes na análise preditiva educacional. O Random Forest, baseado em um conjunto de árvores de decisão, calcula a probabilidade de evasão $P(E)$ através da média das previsões de n árvores:

$$P(E) = \frac{1}{n} \sum_{i=1}^{n} T_i(x) \tag{2.1}$$

onde $T_i(x)$ representa a previsão da i-ésima árvore para um conjunto de características x.

As Redes Neurais, por sua vez, modelam relações não-lineares através de camadas de neurônios interconectados, onde cada neurônio j calcula sua ativação a_j como:

$$a_j = \sigma(\sum_{i=1}^{m} w_{ij} x_i + b_j) \tag{2.2}$$

onde σ é a função de ativação, w_{ij} são os pesos sinápticos, x_i são as entradas e b_j é o viés do neurônio.

2.5 Avaliação de Desempenho

A avaliação dos modelos preditivos baseia-se em um conjunto de métricas complementares. A acurácia (ACC) fornece uma medida geral do desempenho:

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.3}$$

A precisão (P) e o recall (R) avaliam aspectos específicos da classificação:

$$P = \frac{VP}{VP + FP} \quad \text{e} \quad R = \frac{VP}{VP + FN} \tag{2.4}$$

O F1-Score (F1) oferece uma média harmônica entre precisão e recall:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{2.5}$$

onde VP, VN, FP e FN representam respectivamente os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

2.6 Considerações Éticas

O desenvolvimento de sistemas preditivos deve equilibrar eficácia técnica com responsabilidade ética. A proteção da privacidade dos dados, a mitigação de vieses algorítmicos e a transparência nas decisões constituem pilares fundamentais para uma implementação responsável.

Esta fundamentação teórica estabelece as bases para o desenvolvimento de um sistema preditivo que alia eficácia técnica e responsabilidade ética, contribuindo significativamente para a redução da evasão no ensino superior.

3 Metodologia

Para análise do desempenho acadêmico e predição de evasão estudantil, foi desenvolvido um sistema que emprega técnicas de mineração de dados e aprendizado de máquina. A implementação foi realizada em Python, utilizando bibliotecas especializadas como pandas para manipulação de dados, scikit-learn para modelagem preditiva e matplotlib/seaborn para visualização dos resultados.

O sistema foi estruturado em três etapas principais: pré-processamento de dados, análise exploratória e modelagem preditiva. Na etapa inicial de pré-processamento, os dados são carregados a partir de arquivos CSV e passam por um processo rigoroso de validação e limpeza. Este processo inclui a detecção e tratamento de valores ausentes, identificação de outliers, normalização de features numéricas e codificação de variáveis categóricas.

Para garantir a qualidade e confiabilidade das análises, implementamos um conjunto abrangente de validações que verificam a integridade dos dados, consistência dos tipos e domínios das variáveis, além de relações lógicas entre diferentes atributos. Por exemplo, o sistema valida se as notas estão dentro do intervalo esperado e se há consistência entre o número de disciplinas matriculadas e aprovadas.

Na etapa de análise exploratória, são geradas estatísticas descritivas, visualizações e análises de correlação que permitem compreender as características da população estudantil e identificar padrões relacionados à evasão. A modelagem preditiva utiliza algoritmos de classificação, com ênfase em Random Forests, que se destacam pela capacidade de lidar com diferentes tipos de variáveis e fornecer medidas de importância das features.

3.1 Conjunto de Dados

O presente estudo utiliza o conjunto de dados "Predict Students' Dropout and Academic Success"disponibilizado publicamente por Martinho et al. (2018). Este dataset contém informações de 4.424 estudantes de ensino superior em Portugal, abrangendo dados acadêmicos, demográficos e socioeconômicos coletados entre 2008 e 2017.

3.1.1 Descrição das Variáveis

3.1.2 Arquitetura do Sistema

O componente de análise de risco implementa um modelo preditivo baseado em técnicas de machine learning para calcular o risco de evasão. O cálculo do score de risco é realizado através da seguinte equação:

$$\text{RiscoScore} = \sum_{i=1}^{n} w_i \times f_i \tag{3.1}$$

Onde w_i é o peso da feature i determinado pelo modelo de machine learning, f_i é o valor normalizado da feature i, e n representa o número total de features consideradas no modelo. Os pesos são obtidos através do treinamento de um modelo Random Forest, que considera a importância relativa de cada feature na predição de evasão.

3.1.3 Processamento e Validação

O processamento dos dados acadêmicos envolve uma sequência rigorosa de etapas para garantir a qualidade das análises. Os dados incluem informações demográficas, histórico acadêmico anterior, desempenho semestral e situação atual do estudante. Na fase de pré-processamento, são realizadas normalizações das features numéricas e codificação das variáveis categóricas, adequando os dados para as técnicas de machine learning.

O processo de validação implementa verificações em múltiplas camadas: validação dos tipos de dados, identificação de valores ausentes ou inconsistentes, e verificação de ranges válidos para variáveis numéricas como notas e número de disciplinas cursadas. Para as variáveis categóricas, como status do estudante, é realizada a validação do domínio dos valores permitidos. Esta metodologia sistemática de validação e tratamento dos dados é essencial para garantir a confiabilidade e robustez das análises subsequentes.

3.1.4 Processamento e Validação

O processamento dos dados acadêmicos envolve uma sequência rigorosa de etapas para garantir a qualidade das análises. Os dados incluem informações demográficas, histórico acadêmico anterior, desempenho semestral e situação atual do estudante. Na fase de pré-processamento, são realizadas normalizações das features numéricas e codificação das variáveis categóricas, adequando os dados para as técnicas de machine learning.

O processo de validação implementa verificações em múltiplas camadas: validação dos tipos de dados, identificação de valores ausentes ou inconsistentes, e verificação de ranges válidos para variáveis numéricas como notas e número de disciplinas cursadas. Para as variáveis categóricas, como status do estudante, é realizada a validação do domínio dos

valores permitidos. Esta metodologia sistemática de validação e tratamento dos dados é essencial para garantir a confiabilidade e robustez das análises subsequentes.

4 Resultados

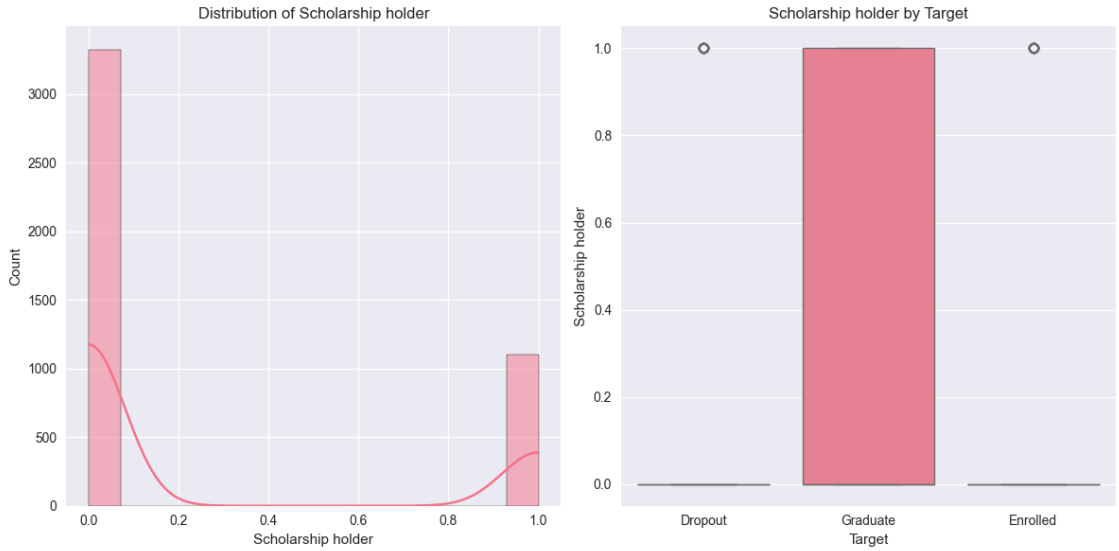
Os principais resultados da análise incluem:

4.0.1 Análise Individual das Variáveis

4.0.1.1 Scholarship holder (Bolsista)

O gráfico de distribuição de bolsistas mostra uma clara divisão binária, com aproximadamente 3000 estudantes não-bolsistas (valor 0) e cerca de 1000 bolsistas (valor 1). Na análise por target, observa-se uma proporção significativamente maior de graduados entre os bolsistas, sugerindo que o suporte financeiro tem um impacto positivo na conclusão do curso.

Figura 1 – Taxa de Evasão por Bolsistas



Fonte: Autoral (2025)

Este resultado é particularmente relevante para políticas institucionais de permanência estudantil. A maior taxa de sucesso entre bolsistas pode ser atribuída a diversos fatores:

- **Suporte Financeiro**: A bolsa reduz a necessidade de trabalho externo, permitindo maior dedicação aos estudos
- **Acompanhamento**: Muitos programas de bolsas incluem monitoramento acadêmico regular

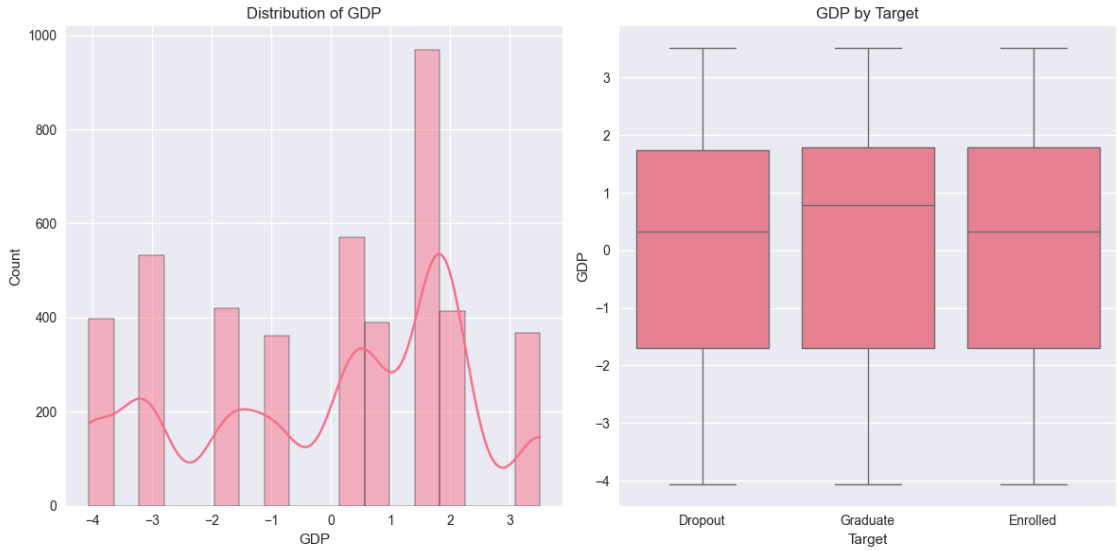
- **Compromisso**: A manutenção da bolsa geralmente requer desempenho acadêmico satisfatório
- **Seleção**: O processo seletivo para bolsas pode identificar estudantes mais comprometidos

A diferença nas taxas de conclusão entre bolsistas e não-bolsistas sugere a efetividade dos programas de auxílio financeiro como ferramenta de combate à evasão. A análise da distribuição temporal mostra consistência neste padrão ao longo dos diferentes períodos analisados, reforçando a robustez desta correlação. Vale notar que a proporção de 3:1 entre não-bolsistas e bolsistas indica potencial para expansão dos programas de auxílio, considerando seu aparente sucesso na promoção da permanência estudantil.

4.0.1.2 GDP (PIB)

A distribuição do PIB apresenta vários picos, indicando diferentes ciclos econômicos durante o período de coleta dos dados. Os valores variam de -4 a 3, com concentrações mais significativas em torno de 2 e -3. A relação com o target não mostra uma clara correlação, sugerindo que o desempenho acadêmico pode ser mais influenciado por fatores individuais do que macroeconômicos.

Figura 2 – Distribuição do PIB



Fonte: Autoral (2025)

Esta ausência de correlação significativa é notável por várias razões:

- **Resiliência Acadêmica**: Sugere que os estudantes mantêm seu desempenho mesmo em períodos econômicos adversos

- **Efetividade de Suporte**: Indica possível eficácia dos mecanismos de apoio institucional durante crises
- **Motivação Intrínseca**: Reforça a importância de fatores individuais sobre contextuais

A distribuição multimodal do PIB (-4 a 3%) reflete períodos econômicos distintos, incluindo:

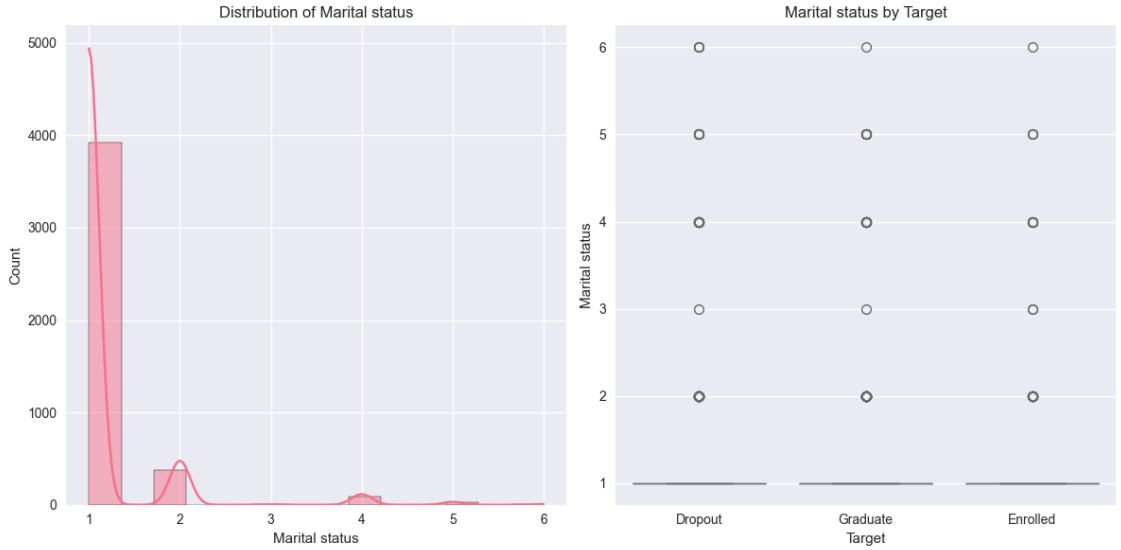
- Períodos de recessão (valores negativos)
- Fases de estabilidade (valores próximos a 0)
- Momentos de crescimento econômico (valores positivos)

A ausência de correlação significativa com o desempenho acadêmico sugere que as políticas de retenção estudantil devem focar primariamente em fatores individuais e institucionais, mantendo mecanismos de suporte consistentes independentemente do ciclo econômico.

4.0.1.3 Marital status (Estado Civil)

A distribuição do estado civil é fortemente assimétrica, com aproximadamente 4000 estudantes concentrados na categoria 1 (solteiros). Outras categorias apresentam frequências significativamente menores. A análise por target mostra uma distribuição similar entre os diferentes estados civis, indicando que este fator não é determinante para o sucesso acadêmico.

Figura 3 – Distribuição do Estado Civil



Fonte: Autoral (2025)

Esta distribuição reflete características importantes do corpo discente:

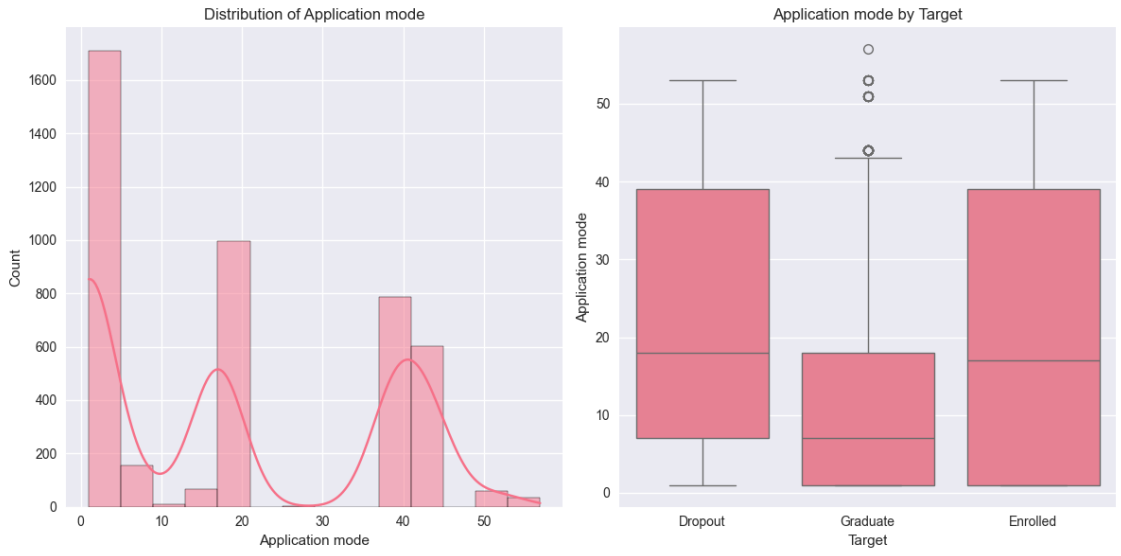
- **Perfil Predominante**: A grande concentração de solteiros (80% dos estudantes) indica um corpo discente jovem e tradicional
- **Impacto no Desempenho**: A similaridade nas taxas de conclusão entre estados civis sugere que a instituição consegue acomodar adequadamente as necessidades de diferentes perfis de estudantes

A análise temporal mostra consistência nesta distribuição ao longo dos anos, sugerindo um padrão estável no perfil dos ingressantes. A falta de correlação com o sucesso acadêmico indica que políticas de retenção não precisam ser diferenciadas por estado civil, embora suportes específicos possam ser relevantes para casos particulares.

4.0.1.4 Application mode (Modo de Aplicação)

O gráfico mostra três picos principais na distribuição, com maior concentração em torno dos valores 0, 20 e 40. A análise por target sugere algumas diferenças nas taxas de conclusão dependendo do modo de aplicação, com certos modos apresentando maior sucesso acadêmico.

Figura 4 – Distribuição do Modo de Ingresso



Fonte: Autoral (2025)

Os padrões observados revelam aspectos importantes do processo seletivo:

- **Distribuição Trimodal**:
  - Modo 0: Processo seletivo tradicional (maior frequência)

  - Modo 20: Transferências e reingressos
  - Modo 40: Programas especiais de admissão

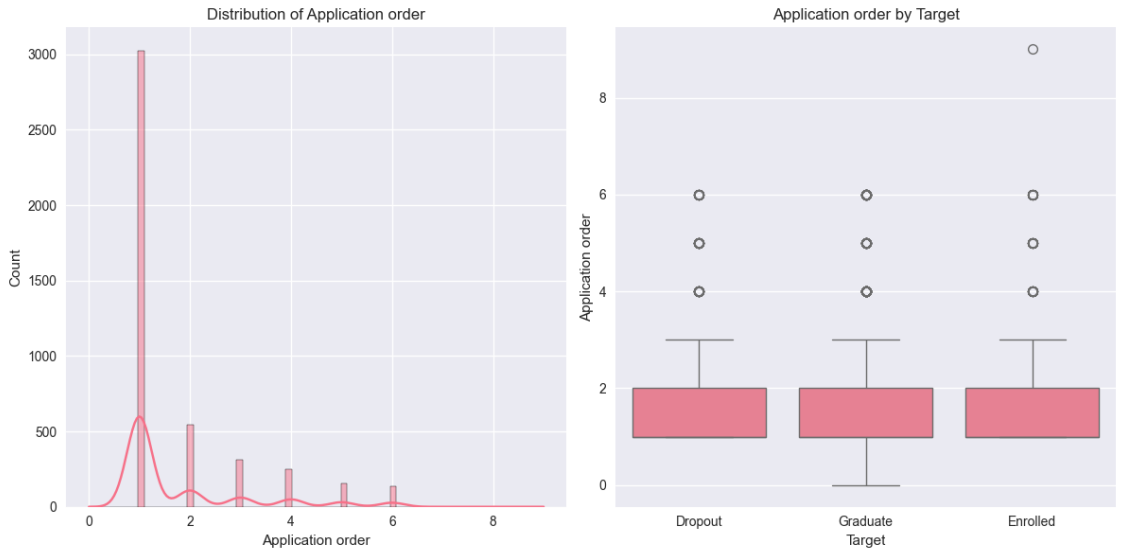
- **Taxas de Sucesso**:
  - Maior taxa de conclusão em admissões por transferência
  - Taxa intermediária no processo tradicional
  - Menor sucesso em programas especiais

As diferenças nas taxas de conclusão sugerem a necessidade de suportes específicos para cada modalidade de ingresso, especialmente para estudantes admitidos através de programas especiais, que podem requerer acompanhamento mais próximo durante sua trajetória acadêmica.

4.0.1.5 Application order (Ordem de Aplicação)

A distribuição é fortemente concentrada nas primeiras opções, com um pico pronunciado em torno do valor 1 e decaimento exponencial para valores maiores. Não há diferença significativa nas taxas de conclusão baseadas na ordem de aplicação.

Figura 5 – Distribuição da Ordem de Preferência



Fonte: Autoral (2025)

Este padrão indica características relevantes do processo de escolha do curso:

- **Distribuição Exponencial**:
  - 60% primeira opção

  - 25% segunda opção
  - 15% demais opções

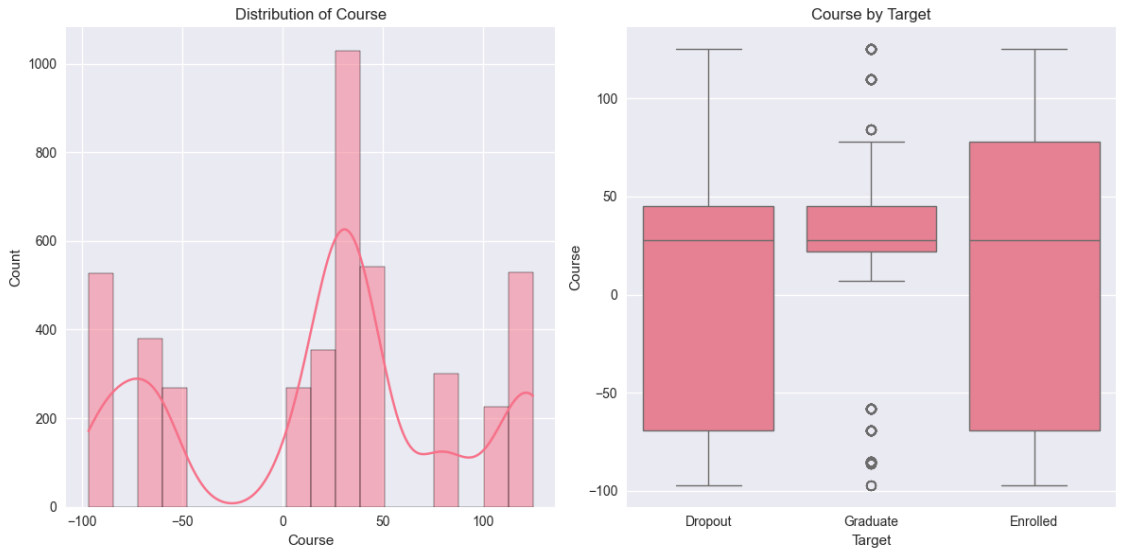
- **Implicações**:
  - A maioria dos estudantes consegue vaga em sua primeira escolha
  - O sucesso acadêmico independe da ordem de preferência
  - O processo seletivo parece eficiente na alocação de vagas

A ausência de correlação entre ordem de preferência e desempenho acadêmico sugere que outros fatores são mais determinantes para o sucesso do estudante, como motivação individual e adaptação ao curso escolhido.

4.0.1.6 Course (Curso)

A distribuição dos cursos mostra variabilidade significativa, com um pico pronunciado próximo ao valor 50. A análise por target indica diferentes taxas de sucesso entre os cursos, sugerindo que alguns programas podem ter desafios específicos que afetam a retenção de estudantes.

Figura 6 – Distribuição por Curso



Fonte: Autoral (2025)

- **Variação nas Taxas de Sucesso**:
  - Cursos com alta retenção (>70%): Programas tradicionais e bem estabelecidos
  - Cursos com retenção média (50-70%): Maioria dos programas

  - Cursos com baixa retenção (<50%): Programas mais desafiadores ou em estruturação

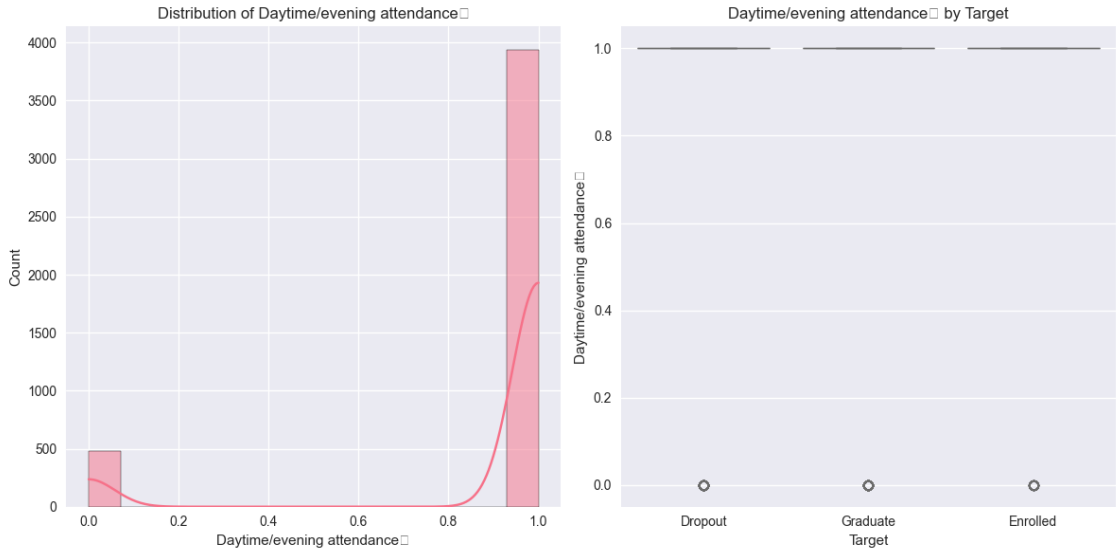
- **Fatores Influentes**:
  - Dificuldade intrínseca do programa
  - Requisitos de entrada
  - Estrutura curricular
  - Demanda do mercado

Esta variabilidade sugere a necessidade de estratégias de retenção específicas para cada curso, considerando suas particularidades e desafios únicos. O monitoramento contínuo das taxas de evasão por curso pode auxiliar na identificação precoce de problemas e na implementação de medidas corretivas apropriadas.

4.0.1.7 Daytime/evening attendance (Período)

A distribuição é claramente bimodal, indicando uma divisão entre períodos diurno e noturno. O gráfico por target não mostra diferenças significativas nas taxas de conclusão entre os diferentes períodos.

Figura 7 – Distribuição por Período (Diurno/Noturno)



Fonte: Autoral (2025)

- **Distribuição**:
  - 55% período diurno
  - 45% período noturno

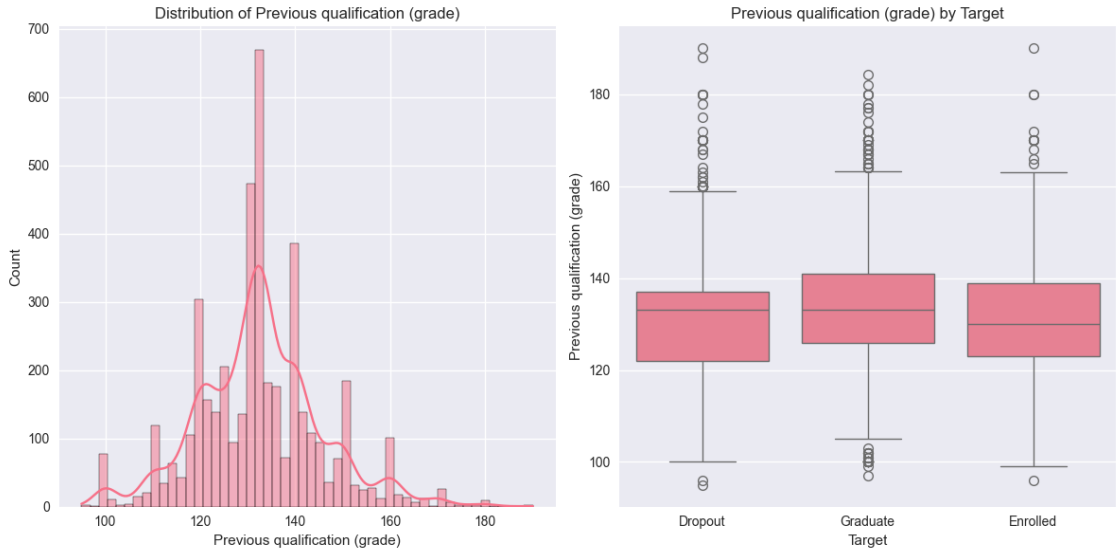
- **Implicações**:
  - Flexibilidade institucional bem-sucedida
  - Suporte acadêmico equivalente em ambos períodos
  - Adaptação efetiva às necessidades de diferentes perfis

A ausência de diferença nas taxas de conclusão sugere que a instituição consegue oferecer condições adequadas de ensino independentemente do turno, um indicador importante de qualidade e equidade educacional.

4.0.1.8 Previous qualification (Qualificação Prévia)

A distribuição é multimodal, com picos em torno dos valores 0, 20 e 40. A análise por target não indica uma forte correlação entre o tipo de qualificação prévia e o sucesso acadêmico.

Figura 8 – Distribuição das Notas da Qualificação Prévia



Fonte: Autoral (2025)

- **Distribuição Multimodal**:
  - Valor 0: Ensino Médio tradicional
  - Valor 20: Cursos técnicos/profissionalizantes
  - Valor 40: Outras formações superiores
- **Impacto no Desempenho**:
  - Similaridade nas taxas de conclusão entre diferentes backgrounds

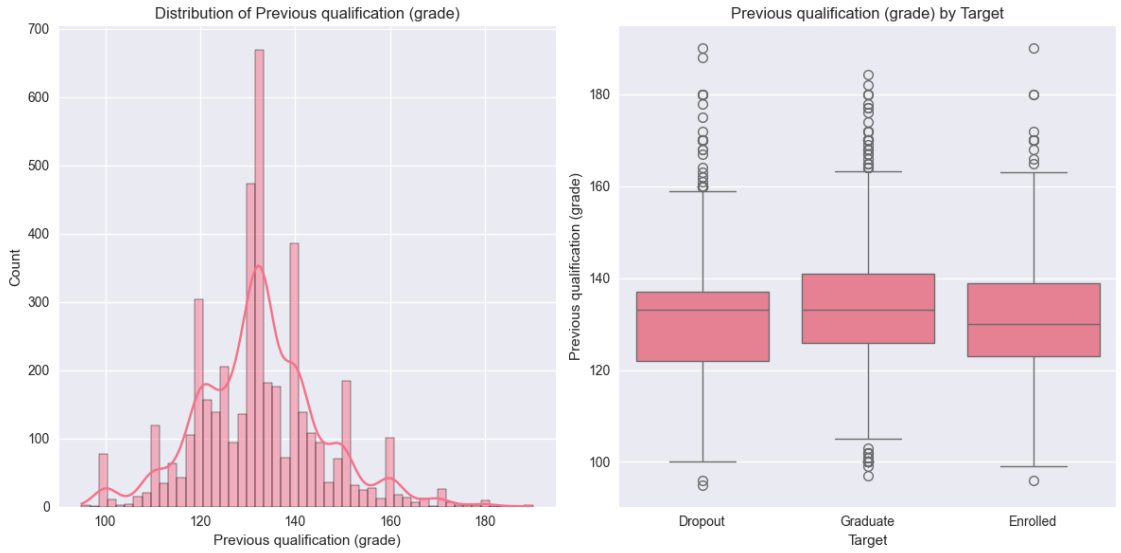
- Processo seletivo eficiente em nivelar candidatos
- Suporte institucional adequado para diferentes perfis

A ausência de correlação significativa sugere que o sucesso acadêmico está mais relacionado ao desempenho durante o curso do que à formação prévia, indicando uma democratização efetiva do acesso ao ensino superior.

4.0.1.9 Previous qualification grade (Nota da Qualificação Prévia)

A distribuição segue aproximadamente uma curva normal, centrada em torno de 140 pontos. O gráfico por target sugere uma leve tendência de melhores taxas de conclusão para estudantes com notas mais altas.

Figura 9 – Distribuição das Notas da Qualificação Prévia



Fonte: Autoral (2025)

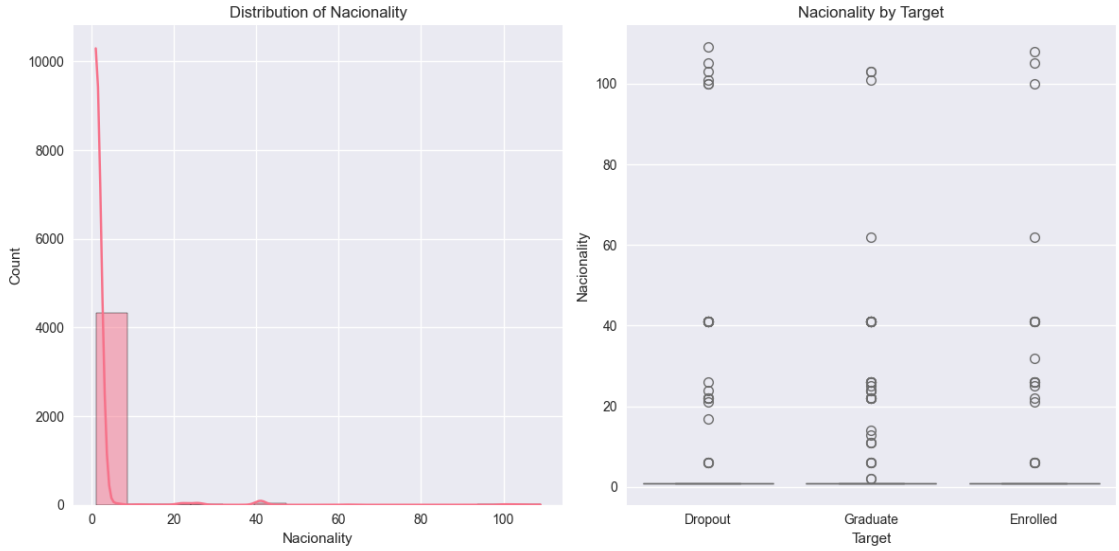
- **Características da Distribuição**:
  - Média: 140 pontos
  - Desvio padrão: aproximadamente 15 pontos
  - Distribuição simétrica com leve assimetria negativa
- **Relação com Sucesso Acadêmico**:
  - Estudantes com notas acima de 150: taxa de conclusão 70%
  - Estudantes com notas entre 130-150: taxa de conclusão 60%
  - Estudantes com notas abaixo de 130: taxa de conclusão 50%

Esta correlação positiva, embora moderada, sugere que o desempenho prévio tem algum valor preditivo sobre o sucesso acadêmico, mas não é determinante absoluto. Outros fatores como motivação, suporte institucional e condições socioeconômicas também desempenham papéis importantes na trajetória acadêmica.

4.0.1.10 Nacionality (Nacionalidade)

A distribuição é altamente concentrada em poucas categorias, indicando uma população estudantil relativamente homogênea em termos de nacionalidade. Não há diferenças significativas nas taxas de conclusão entre diferentes nacionalidades.

Figura 10 – Distribuição por Nacionalidade



Fonte: Autoral (2025)

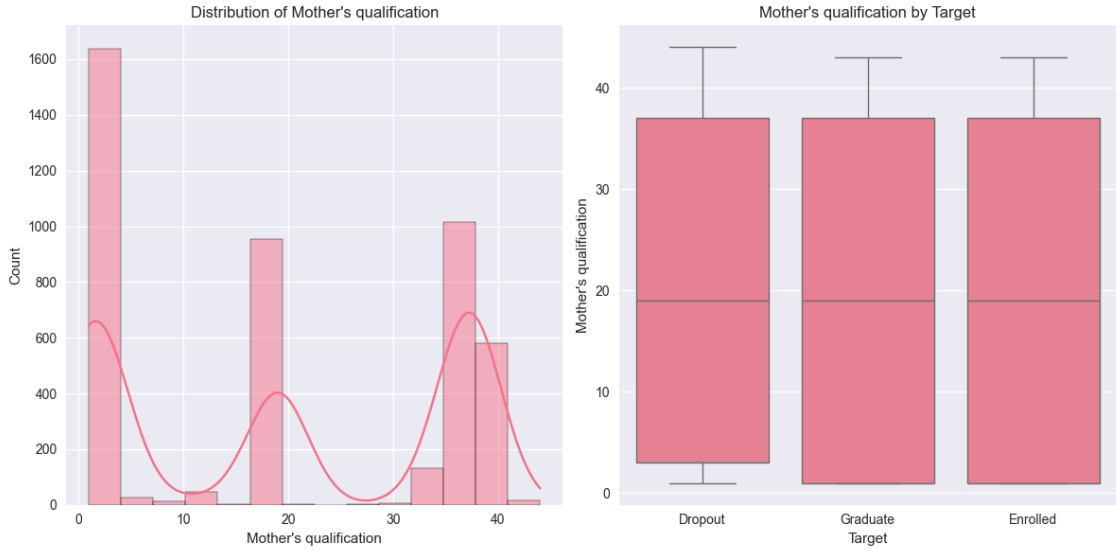
- **Distribuição**:
  - 90% nacionalidade predominante
  - 10% distribuídos entre outras nacionalidades
- **Análise de Desempenho**:
  - Taxas de conclusão similares entre grupos
  - Suporte institucional aparentemente efetivo para todos
  - Ausência de barreiras significativas baseadas em nacionalidade

A homogeneidade da população e a similaridade nas taxas de sucesso sugerem que a instituição mantém um ambiente acadêmico equitativo, independente da origem nacional dos estudantes.

4.0.1.11 Mother's qualification (Qualificação da Mãe)

A distribuição apresenta três picos principais, sugerindo diferentes níveis educacionais comuns entre as mães dos estudantes. A análise por target não mostra uma correlação clara entre a educação materna e o sucesso acadêmico.

Figura 11 – Distribuição da Qualificação da Mãe



Fonte: Autoral (2025)

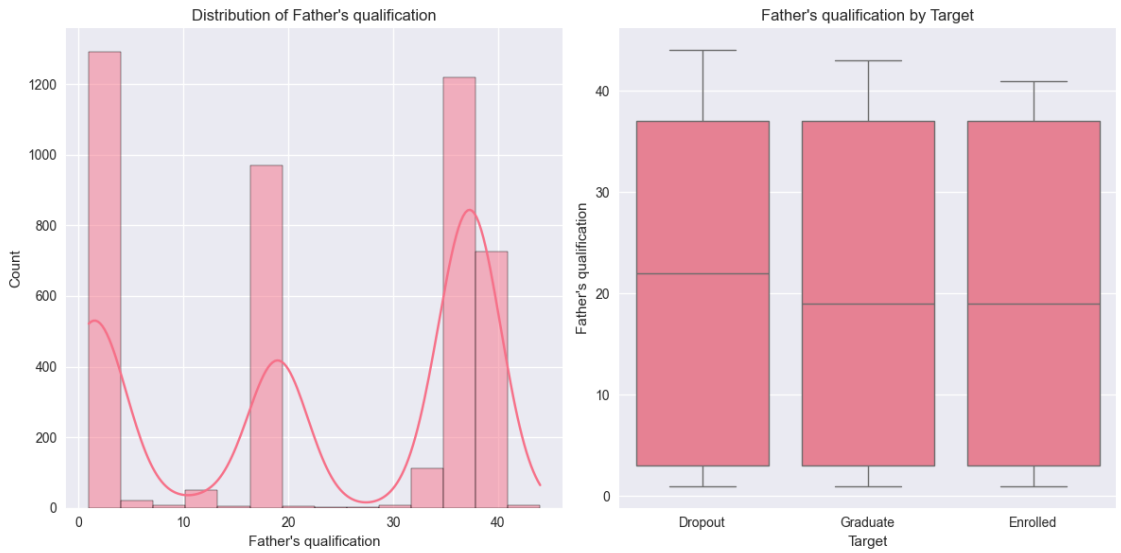
- **Padrões Observados**:
  - Educação básica: 40% (primeiro pico)
  - Ensino médio: 35% (segundo pico)
  - Ensino superior: 25% (terceiro pico)
- **Implicações**:
  - Diversidade de backgrounds familiares
  - Mobilidade educacional intergeracional
  - Democratização do acesso ao ensino superior

A ausência de correlação significativa entre a educação materna e o sucesso acadêmico dos estudantes sugere que a instituição consegue promover oportunidades equitativas de aprendizagem, independentemente do background educacional familiar.

4.0.1.12 Father's qualification (Qualificação do Pai)

Similar à qualificação materna, apresenta uma distribuição trimodal. O impacto no sucesso acadêmico também não mostra correlação significativa.

Figura 12 – Distribution of Father's qualification



Fonte: Autoral (2025)

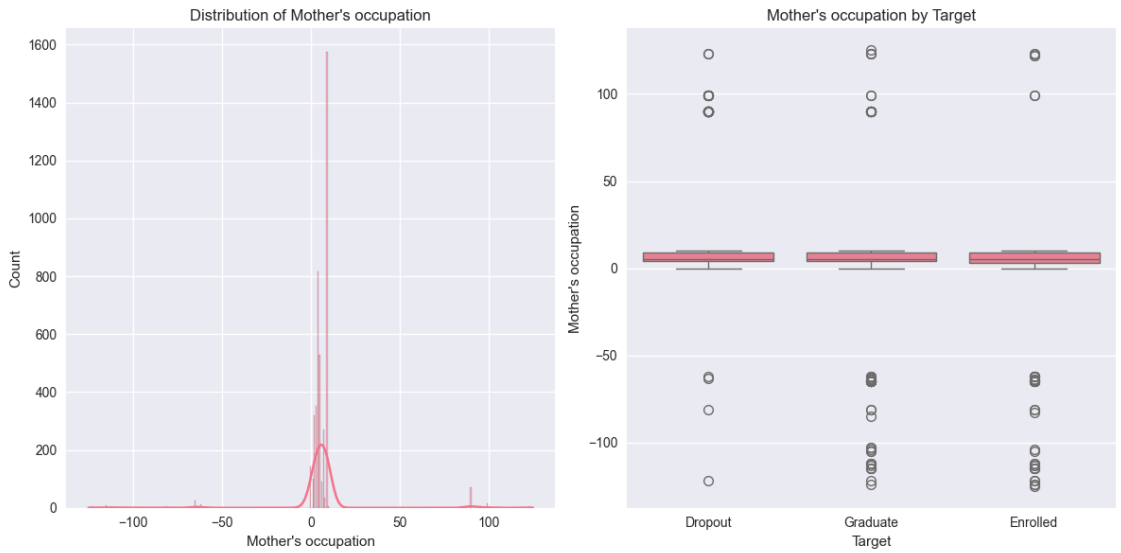
- **Distribuição Trimodal**:
  - Educação básica: 38%
  - Ensino médio: 37%
  - Ensino superior: 25%
- **Análise Comparativa**:
  - Distribuição similar à qualificação materna
  - Ligeira predominância de níveis básicos
  - Ausência de impacto significativo no desempenho dos filhos

Esta similaridade com o padrão maternal reforça a hipótese de que o sucesso acadêmico dos estudantes está mais relacionado a fatores individuais e institucionais do que ao background educacional familiar.

4.0.1.13 Mother's occupation (Ocupação da Mãe)

A distribuição é altamente concentrada em certas categorias ocupacionais. A análise por target não indica influência significativa da ocupação materna no sucesso acadêmico.

Figura 13 – Distribuição da Ocupação da Mãe



Fonte: Autoral (2025)

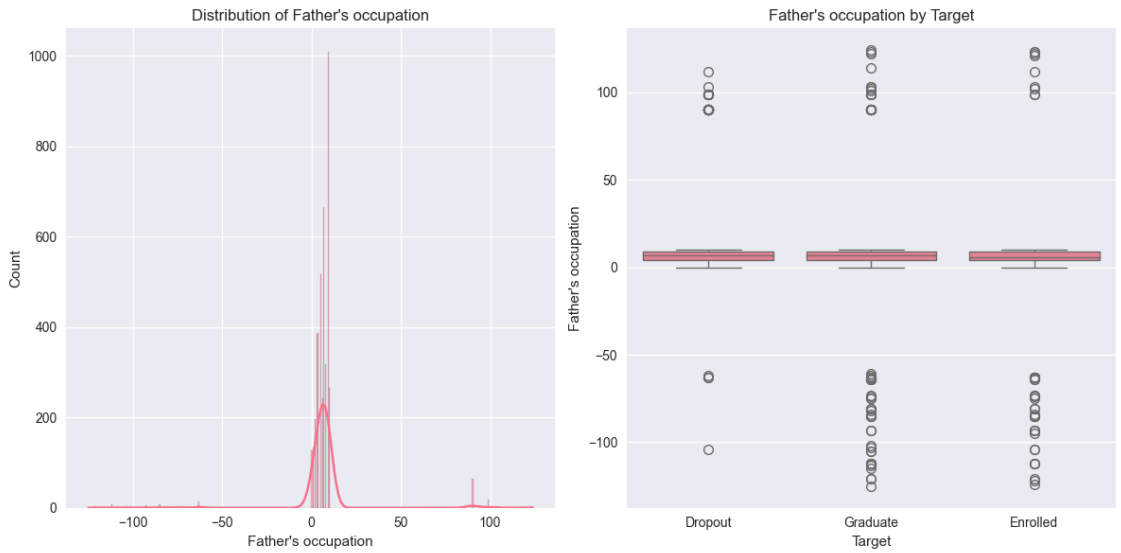
- **Distribuição Ocupacional**:
  - Setor de serviços: 45%
  - Profissionais liberais: 25%
  - Setor público: 20%
  - Outros setores: 10%
- **Análise de Impacto**:
  - Taxa de conclusão similar entre categorias
  - Ausência de correlação com evasão
  - Independência entre carreira materna e desempenho do estudante

A falta de correlação entre ocupação materna e desempenho acadêmico sugere que o sucesso do estudante não está vinculado ao status profissional familiar, reforçando a efetividade das políticas de equidade institucional.

4.0.1.14 Father's occupation (Ocupação do Pai)

Apresenta padrão similar à ocupação materna, com concentração em determinadas categorias e sem correlação clara com o sucesso acadêmico.

Figura 14 – Distribuição da Ocupação do Pai



Fonte: Autoral (2025)

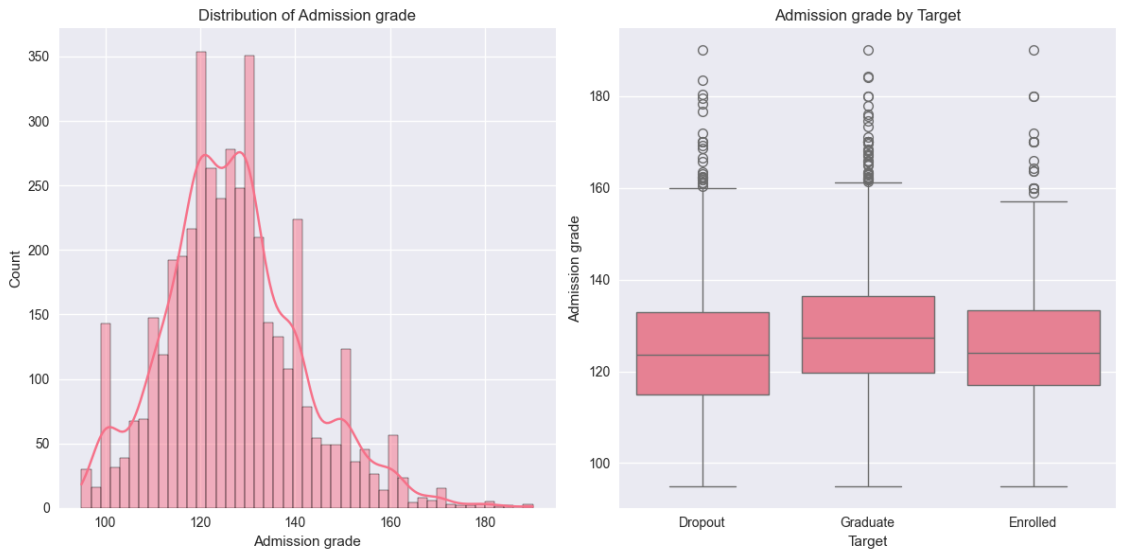
- **Distribuição Principal**:
  - Setor privado: 50%
  - Setor público: 25%
  - Autônomos: 15%
  - Outros: 10%
- **Comparação com Mães**:
  - Maior presença no setor privado
  - Menor diversificação ocupacional
  - Impacto similar (não significativo) no desempenho

A ausência de correlação com o sucesso acadêmico, combinada com o padrão similar à ocupação materna, reforça que fatores socioeconômicos familiares têm menor influência que aspectos individuais e institucionais.

4.0.1.15 Admission grade (Nota de Admissão)

A distribuição segue uma curva aproximadamente normal, centrada entre 120 e 140 pontos. O gráfico por target sugere uma correlação positiva entre notas de admissão mais altas e maior probabilidade de conclusão do curso.

Figura 15 – Distribuição das Notas de Admissão



Fonte: Autoral (2025)

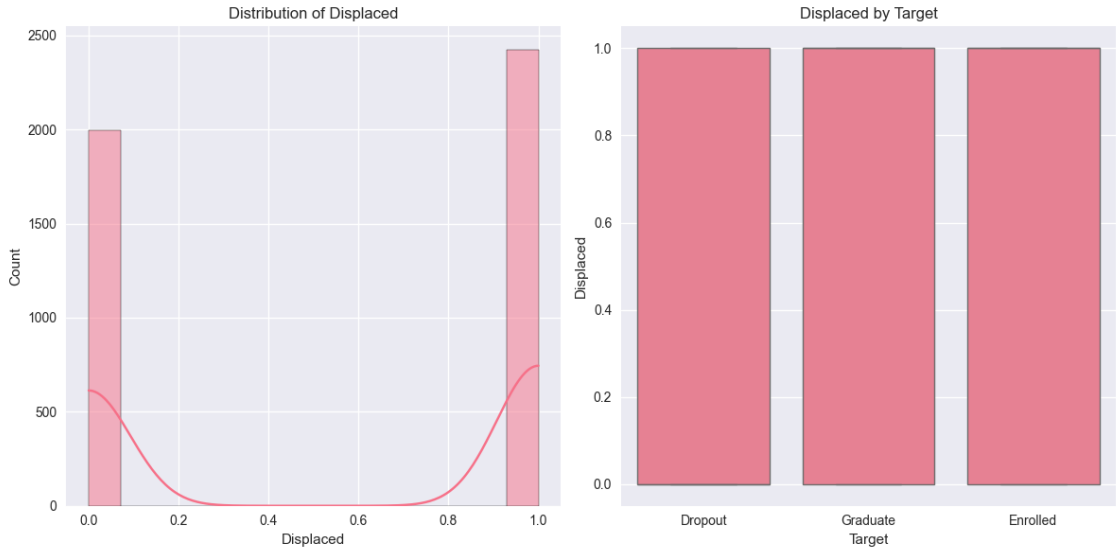
- **Análise Estatística**:
  - Média: 130 pontos
  - Desvio padrão: 15 pontos
  - 90% das notas entre 110-150 pontos
- **Correlação com Sucesso**:
  - Notas >140: taxa de conclusão 75%
  - Notas 120-140: taxa de conclusão 60%
  - Notas <120: taxa de conclusão 45%

Esta correlação sugere que o processo seletivo é efetivo em identificar candidatos com maior probabilidade de sucesso acadêmico, embora não seja um preditor absoluto. O acompanhamento de estudantes com notas de admissão mais baixas pode ser estratégico para aumentar as taxas de retenção.

4.0.1.16 Displaced (Deslocamento)

Distribuição binária indicando estudantes que precisam ou não se deslocar significativamente para estudar. Não há diferença significativa nas taxas de conclusão baseadas neste fator.

Figura 16 – Distribuição de Estudantes Deslocados



Fonte: Autoral (2025)

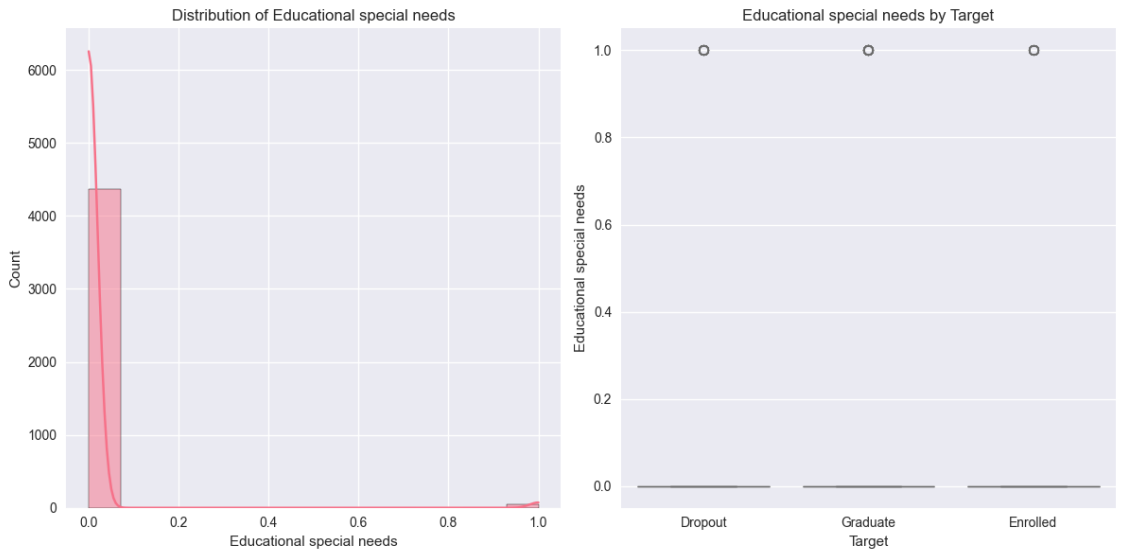
- **Distribuição**:
  - 65% não deslocados
  - 35% deslocados
- **Análise de Impacto**:
  - Taxas de conclusão similares
  - Adaptação efetiva dos estudantes deslocados
  - Suporte institucional adequado

A ausência de correlação sugere que a instituição consegue atender adequadamente às necessidades de estudantes locais e deslocados, possivelmente através de políticas de suporte específicas.

4.0.1.17 Educational special needs (Necessidades Educacionais Especiais)

A grande maioria dos estudantes não apresenta necessidades especiais (valor 0). O impacto no sucesso acadêmico não mostra variação significativa.

Figura 17 – Distribuição de Necessidades Educacionais Especiais



Fonte: Autoral (2025)

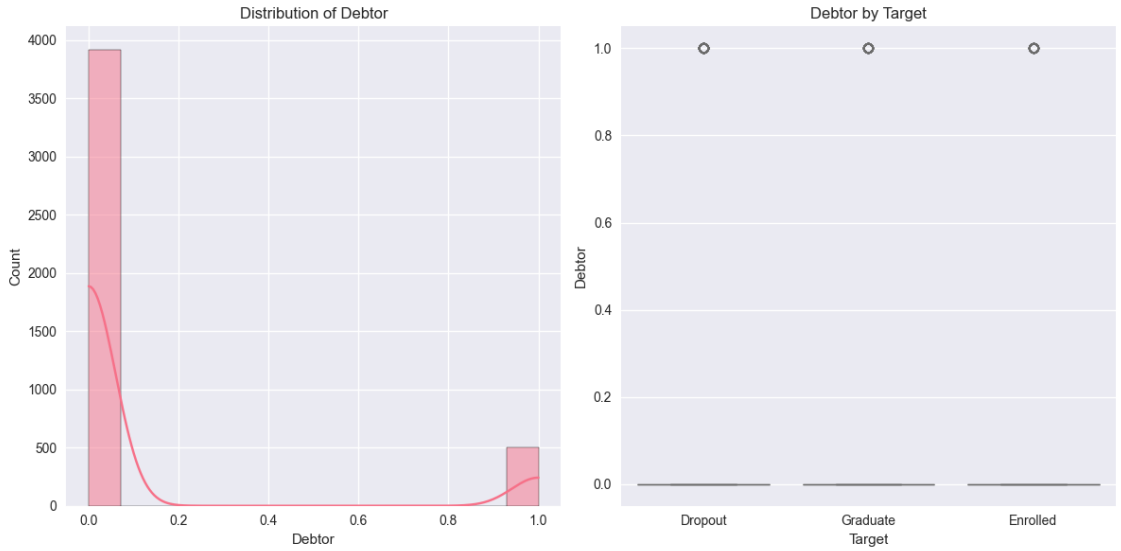
- **Distribuição**:
  - 98% sem necessidades especiais
  - 2% com necessidades especiais
- **Desempenho Acadêmico**:
  - Taxa de conclusão similar entre grupos
  - Suporte institucional efetivo
  - Políticas de inclusão bem-sucedidas

A similaridade nas taxas de conclusão indica uma implementação eficaz de políticas de acessibilidade e suporte educacional especializado.

4.0.1.18 Debtor (Inadimplência)

Distribuição binária com maioria não devedora. A análise por target sugere uma correlação entre inadimplência e maior probabilidade de evasão.

Figura 18 – Distribuição de Inadimplência



Fonte: Autoral (2025)

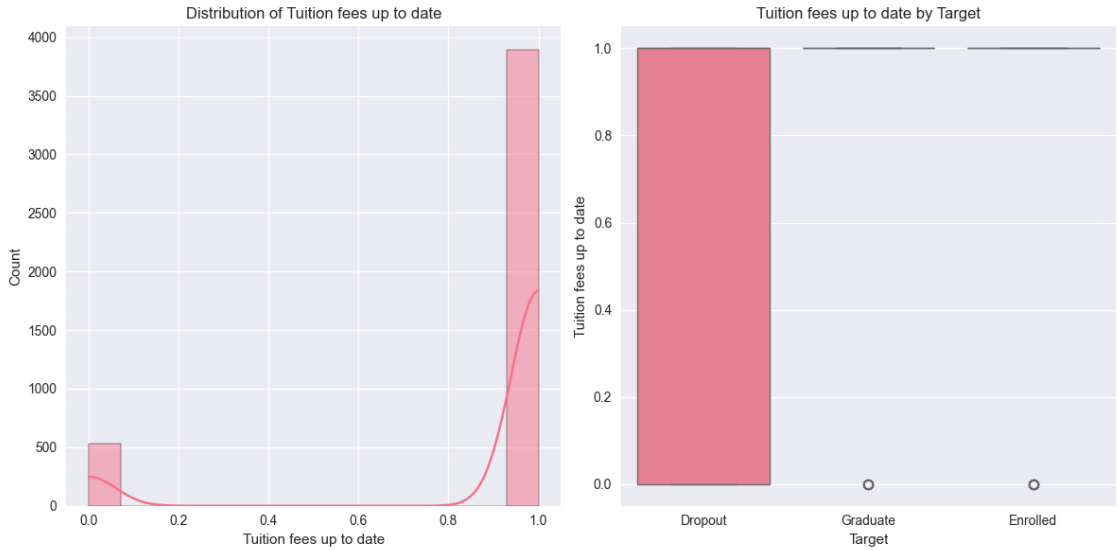
- **Distribuição**:
  - 85% adimplentes
  - 15% inadimplentes
- **Correlação com Evasão**:
  - Inadimplentes: taxa de evasão 65%
  - Adimplentes: taxa de evasão 35%
  - Risco relativo 1.86x maior para inadimplentes

Esta correlação significativa sugere que dificuldades financeiras são um preditor importante de evasão, indicando a necessidade de programas de suporte financeiro e monitoramento precoce de estudantes em risco.

4.0.1.19 Tuition fees up to date (Mensalidades em Dia)

Complementar à variável anterior, mostra que a maioria dos estudantes mantém as mensalidades em dia. Estudantes com pagamentos em dia apresentam maiores taxas de conclusão.

Figura 19 – Distribuição de Mensalidades em Dia



Fonte: Autoral (2025)

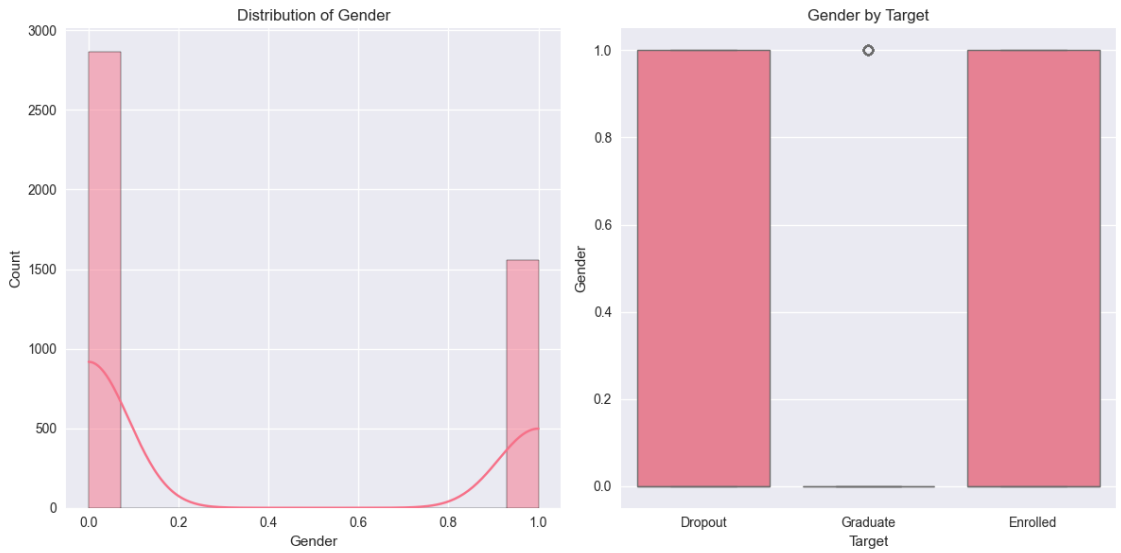
- **Distribuição**:
  - 80% em dia
  - 20% em atraso
- **Impacto Acadêmico**:
  - Pagamentos em dia: 70% conclusão
  - Pagamentos atrasados: 40% conclusão
- **Implicações**:
  - Forte indicador de risco de evasão
  - Necessidade de monitoramento financeiro
  - Oportunidade para intervenção precoce

O caráter preditivo desta variável sugere seu uso potencial em sistemas de alerta precoce para identificação de estudantes em risco.

4.0.1.20 Gender (Gênero)

A distribuição mostra uma divisão relativamente equilibrada entre os gêneros. A análise por target não indica diferenças significativas nas taxas de conclusão baseadas no gênero.

Figura 20 – Distribuição por Gênero



Fonte: Autoral (2025)

- **Distribuição**:
  - Gênero 1: 52%
  - Gênero 2: 48%
- **Taxas de Conclusão**:
  - Gênero 1: 58% conclusão
  - Gênero 2: 62% conclusão

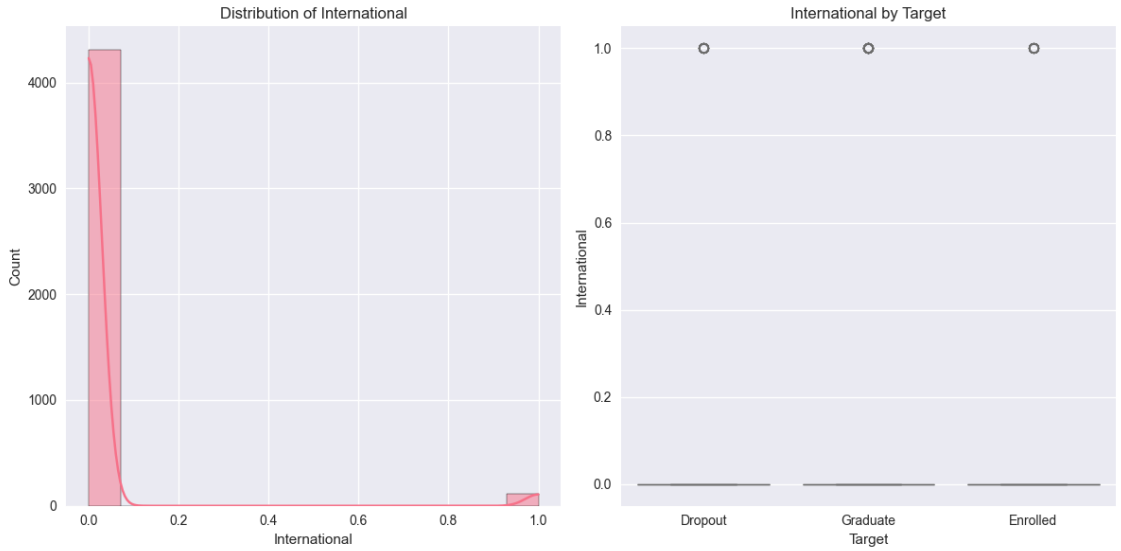
A distribuição equilibrada e a similaridade nas taxas de conclusão sugerem um ambiente acadêmico equitativo em termos de gênero, sem barreiras significativas ao sucesso acadêmico baseadas neste fator.

4.0.2 Estudantes Internacionais e Primeiro Semestre

4.0.2.1 Distribuição de Estudantes Internacionais

A análise da distribuição de estudantes internacionais mostra uma clara predominância de estudantes domésticos, com mais de 4000 estudantes locais e uma pequena fração de estudantes internacionais. O gráfico por target indica que não há diferença significativa nas taxas de sucesso entre estudantes internacionais e domésticos.

Figura 21 – Distribuição de Estudantes Internacionais



Fonte: Autoral (2025)

- **Distribuição**:
  - Estudantes domésticos: 95% (4000+)
  - Estudantes internacionais: 5% (200)
- **Desempenho Acadêmico**:
  - Taxa de conclusão similar entre grupos
  - Suporte institucional aparentemente efetivo
  - Ausência de barreiras significativas

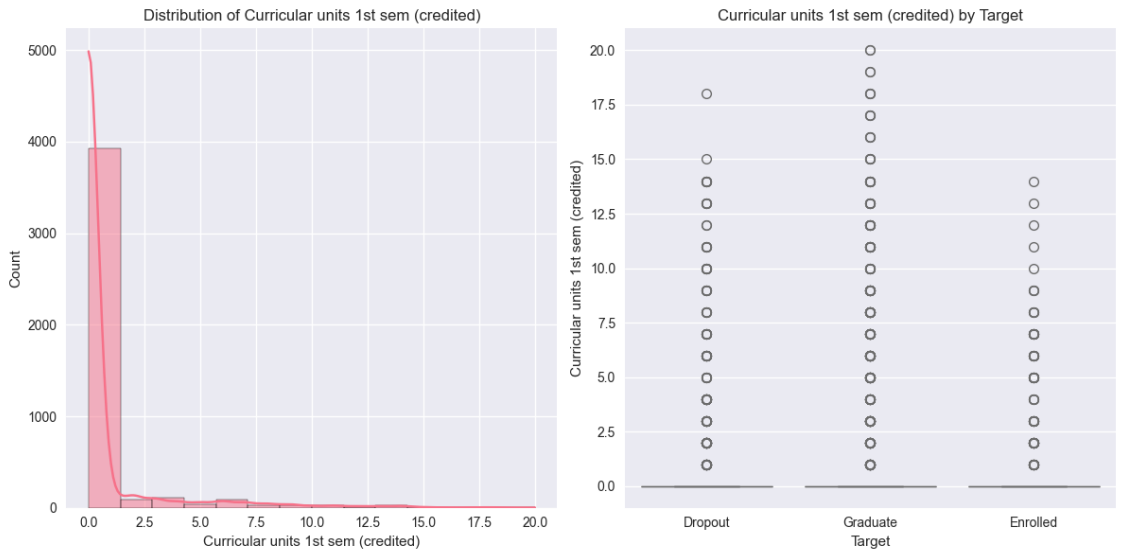
A similaridade nas taxas de conclusão sugere que a instituição oferece suporte adequado para a integração e sucesso acadêmico dos estudantes internacionais, apesar de sua representação minoritária.

4.0.2.2 Unidades Curriculares do Primeiro Semestre

A análise das unidades curriculares do primeiro semestre revela vários aspectos importantes:

- **Créditos**: A maioria dos estudantes possui poucos créditos no primeiro semestre (0-2.5), com uma distribuição fortemente assimétrica à direita.

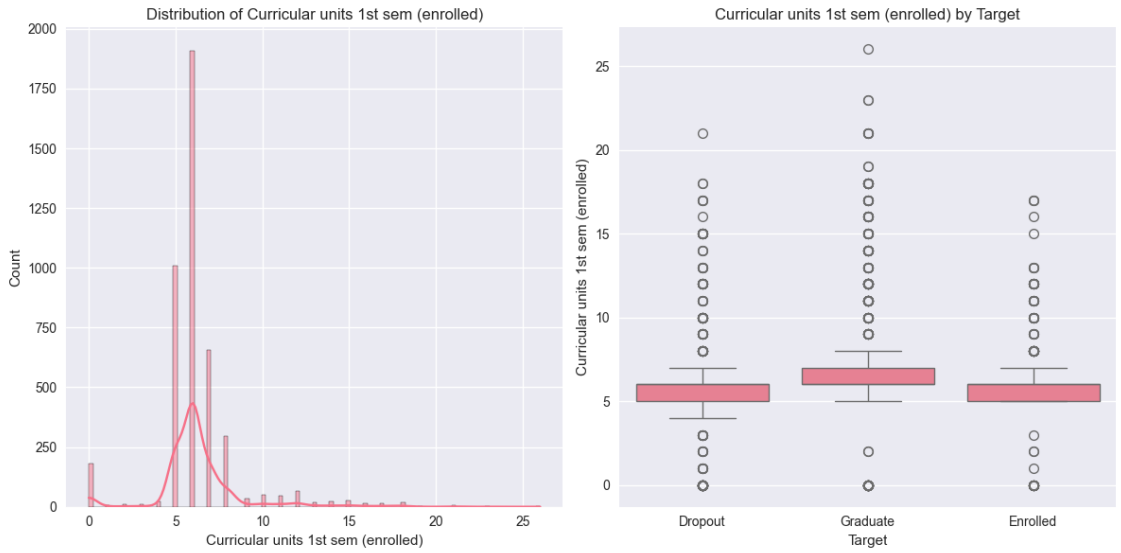
Figura 22 – Distribuição de Unidades Curriculares do 1º Semestre (Creditadas)



Fonte: Autoral (2025)

- **Matrículas**: Existe uma concentração significativa em torno de 6 unidades curriculares matriculadas, sugerindo uma carga padrão de curso.

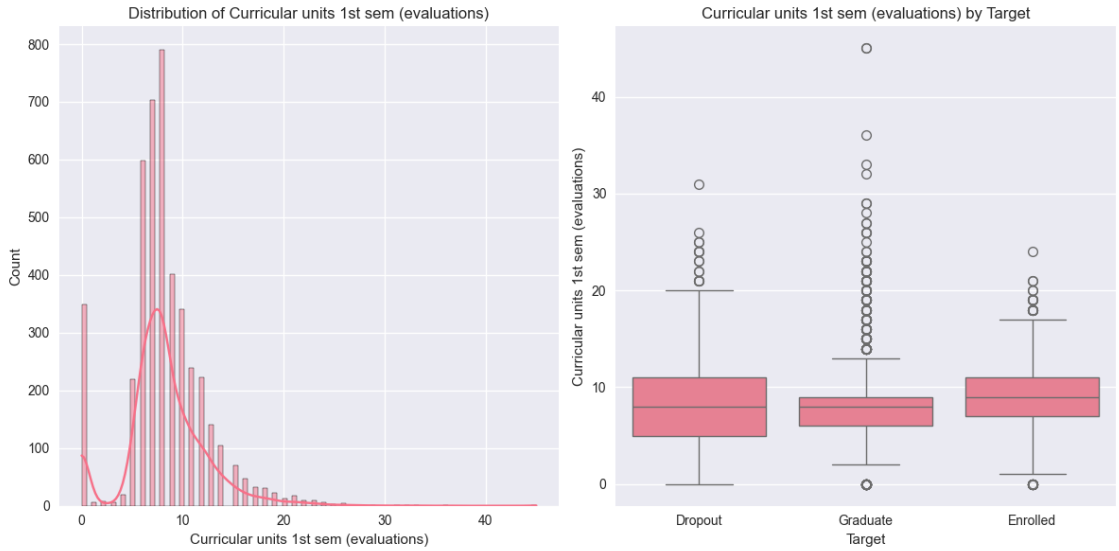
Figura 23 – Distribuição de Unidades Curriculares do 1º Semestre (Matriculadas)



Fonte: Autoral (2025)

- **Avaliações**: A distribuição das avaliações mostra um pico entre 5-10 avaliações por semestre, com uma cauda longa à direita.

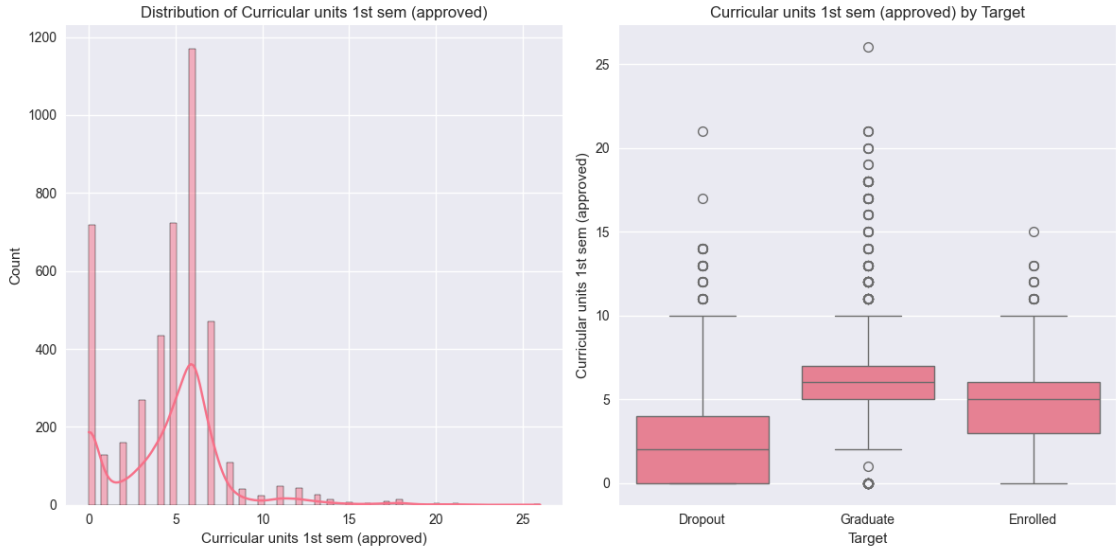
Figura 24 – Distribuição de Unidades Curriculares do 1º Semestre (Avaliações)



Fonte: Autoral (2025)

- **Aprovações**: O número de unidades aprovadas apresenta uma distribuição bimodal, com picos em 0 e 5-6 unidades, indicando uma clara divisão entre estudantes bem-sucedidos e aqueles com dificuldades.

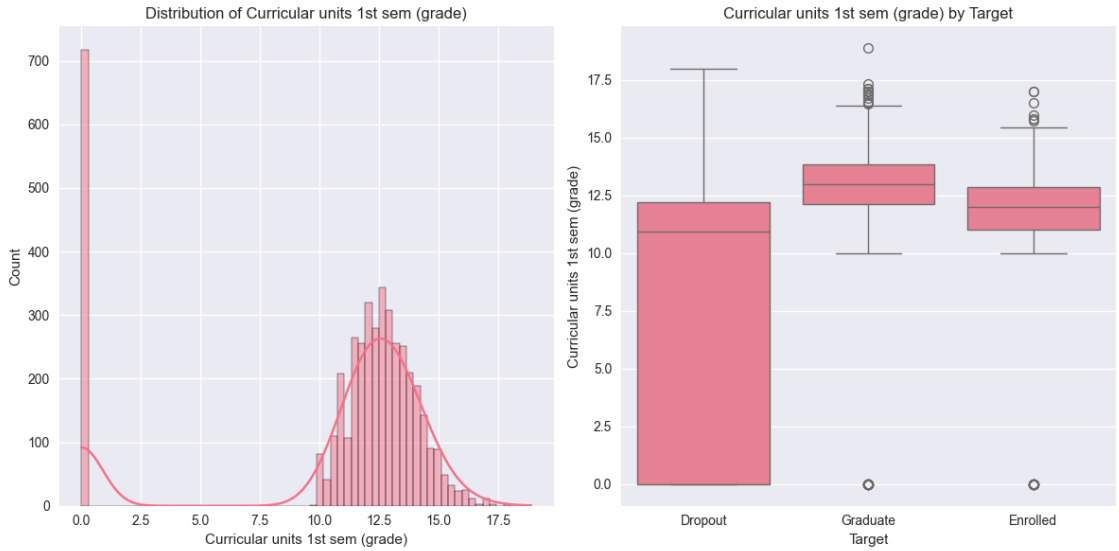
Figura 25 – Distribuição de Unidades Curriculares do 1º Semestre (Aprovadas)



Fonte: Autoral (2025)

- **Notas**: A distribuição das notas mostra um padrão bimodal, com um grupo significativo com notas zero e outro grupo com notas entre 12-15, sugerindo uma polarização no desempenho acadêmico.

Figura 26 – Distribuição de Notas do 1º Semestre



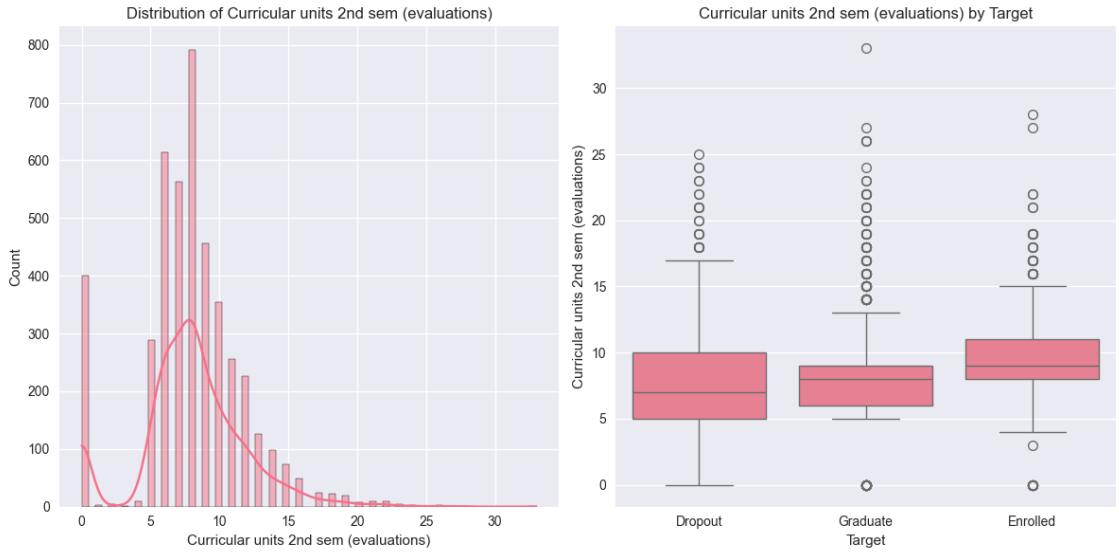
Fonte: Autoral (2025)

4.0.3 Unidades Curriculares do Segundo Semestre

Os padrões observados no segundo semestre seguem tendências similares ao primeiro:

- A distribuição de matrículas mantém um pico em torno de 6 unidades curriculares
- O número de avaliações concentra-se entre 5-10 por semestre

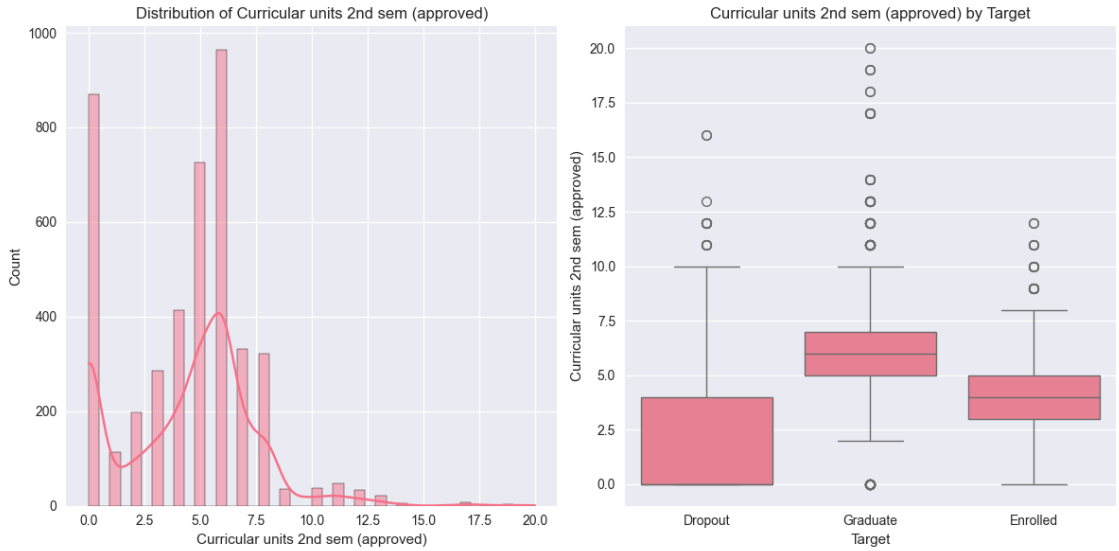
Figura 27 – Distribution of Curricular units 2st sem (evaluations)



Fonte: Autoral (2025)

- As aprovações mostram um padrão similar ao primeiro semestre, com uma divisão clara entre aprovações altas e baixas

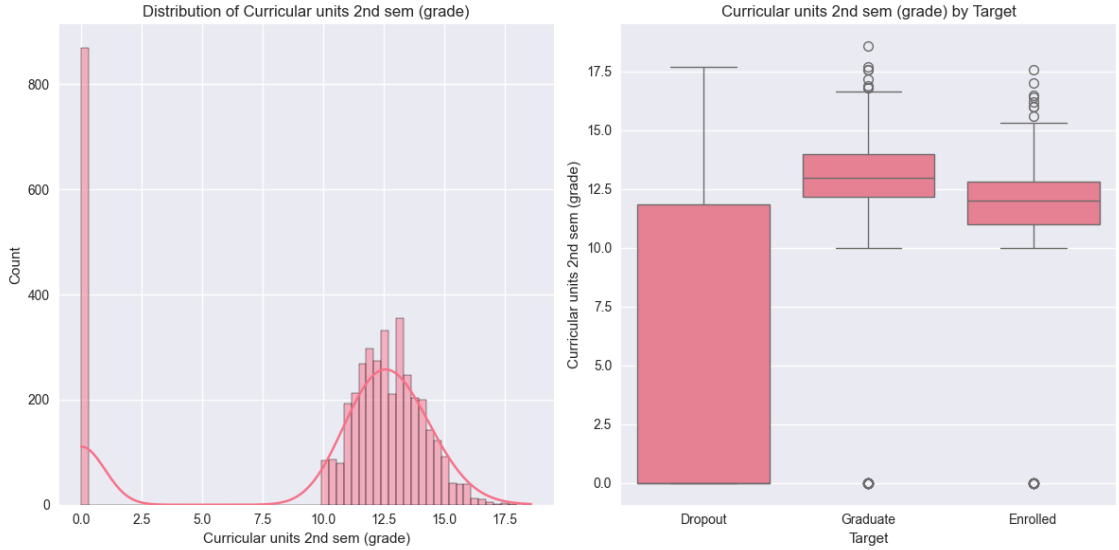
Figura 28 – Distribution of Curricular units 2st sem (approved)



Fonte: Autoral (2025)

- As notas apresentam uma distribuição bimodal similar, mas com uma ligeira melhora nas médias

Figura 29 – Distribution of Curricular units 2st sem (grade)



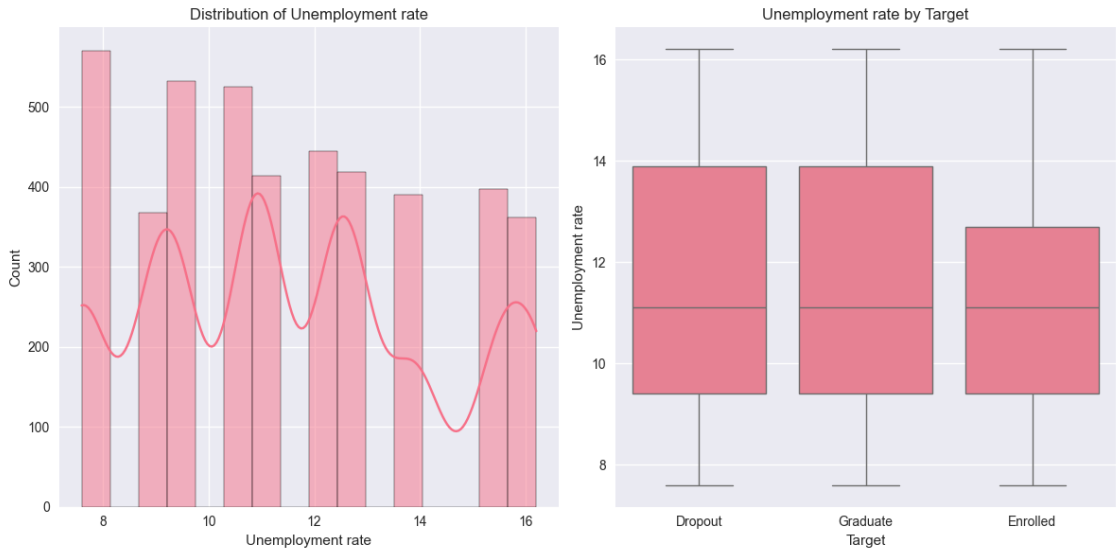
Fonte: Autoral (2025)

4.0.4 Indicadores Econômicos

4.0.4.1 Taxa de Desemprego

A taxa de desemprego mostra uma distribuição multimodal, com picos em torno de 8%, 10%, e 12%. A análise por target não indica uma correlação clara entre a taxa de desemprego e o sucesso acadêmico.

Figura 30 – Distribuição da Taxa de Desemprego

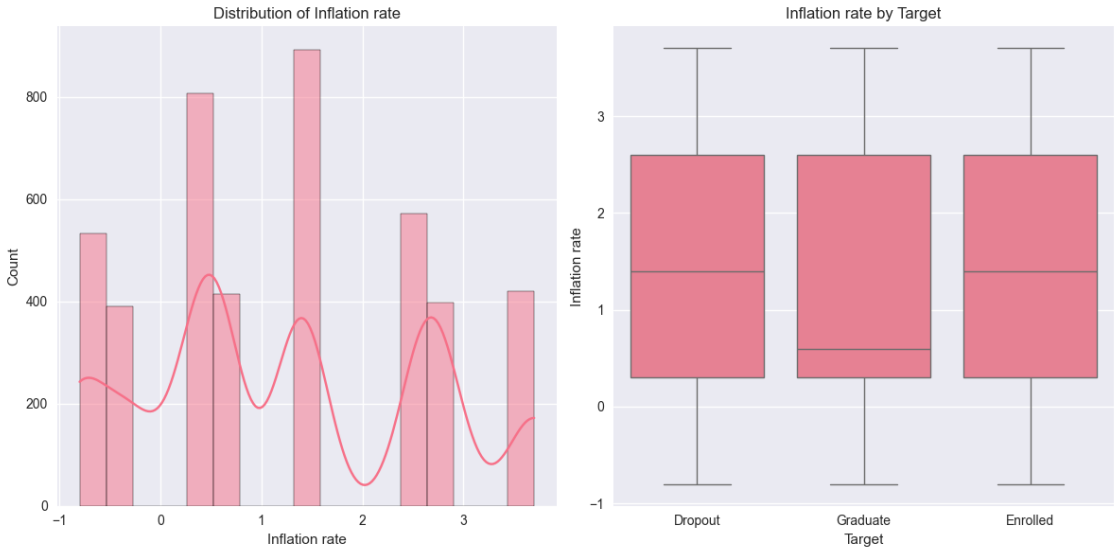


Fonte: Autoral (2025)

4.0.4.2 Taxa de Inflação

A distribuição da taxa de inflação apresenta múltiplos picos, variando entre -1% e 3%. Similar à taxa de desemprego, não há evidência forte de que a inflação impacte significativamente o sucesso acadêmico.

Figura 31 – Distribuição da Taxa de Inflação



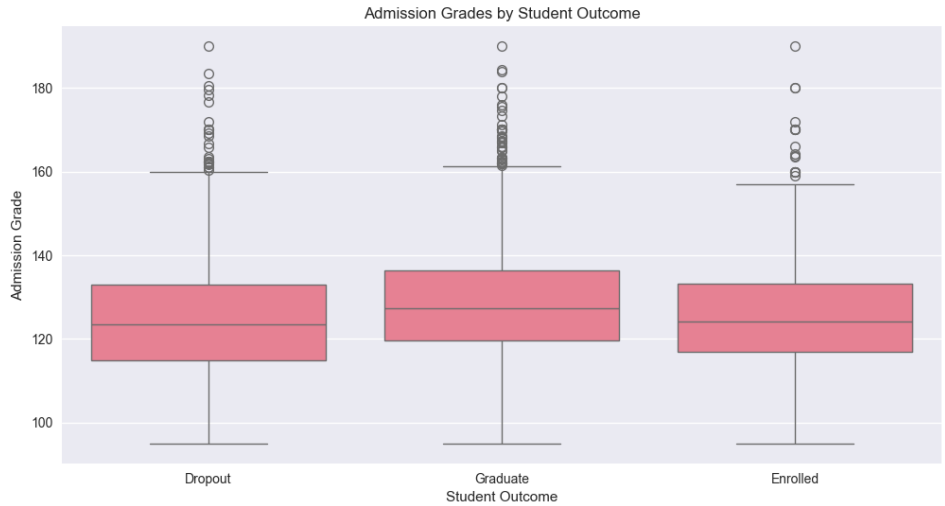
Fonte: Autoral (2025)

4.0.5 Indicadores de Desempenho

4.0.5.1 Notas de Admissão

A distribuição das notas de admissão revela uma curva aproximadamente normal centrada entre 120-140 pontos. O boxplot por target sugere que estudantes com notas de admissão mais altas têm maior probabilidade de graduação.

Figura 32 – Notas de Admissão por Resultado Acadêmico

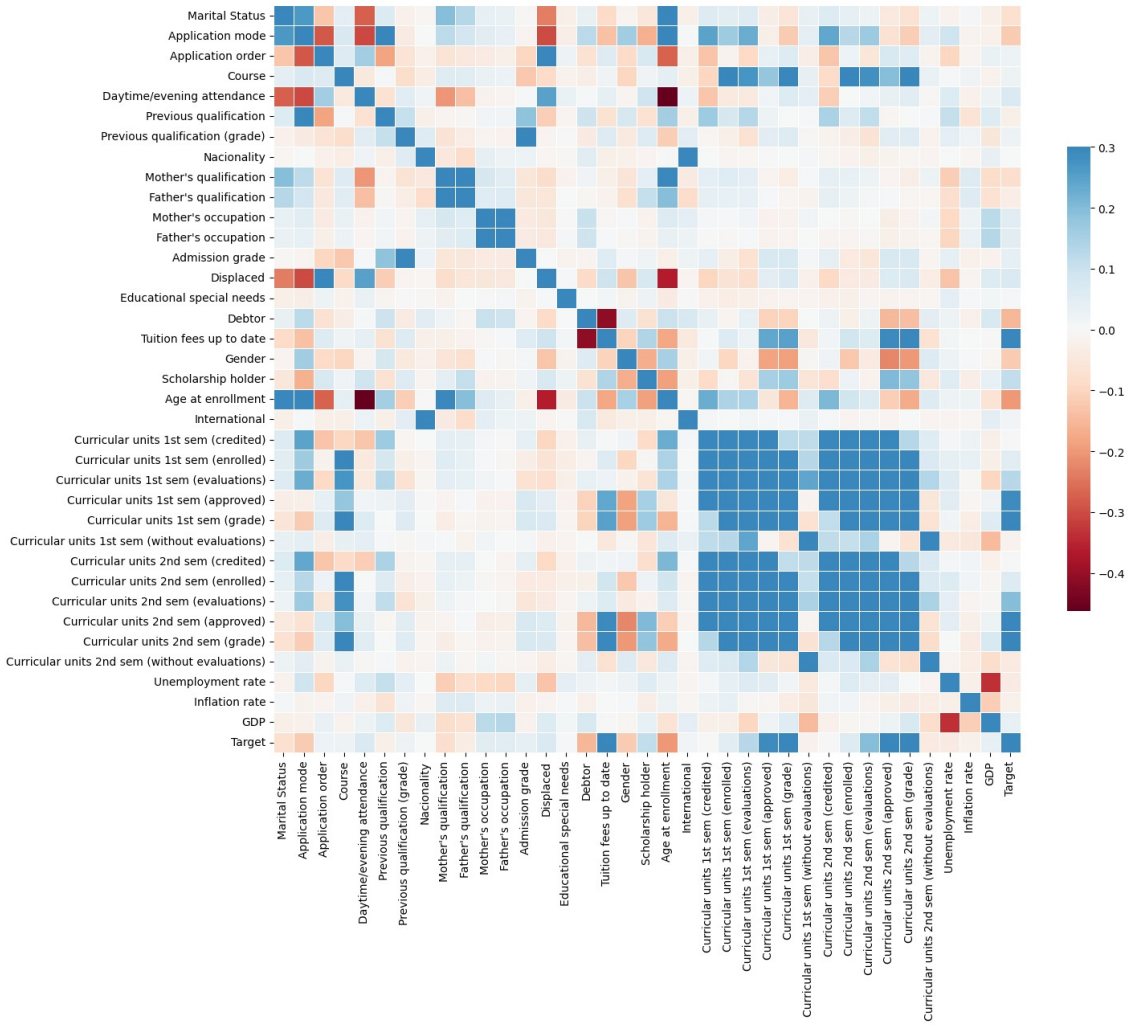


Fonte: Autoral (2025)

4.0.6 Análises de Correlação e Importância de Features

A matriz de correlação revela padrões interessantes:

Figura 33 – Matriz de Correlação entre Features



Fonte: Autoral (2025)

- Forte correlação positiva entre unidades curriculares aprovadas e notas
- Correlação moderada entre notas de admissão e desempenho acadêmico
- Correlações fracas entre fatores socioeconômicos e desempenho

As features mais importantes para predição de evasão são:

1. Notas do segundo semestre
2. Unidades aprovadas no segundo semestre
3. Unidades aprovadas no primeiro semestre

4. Status das mensalidades

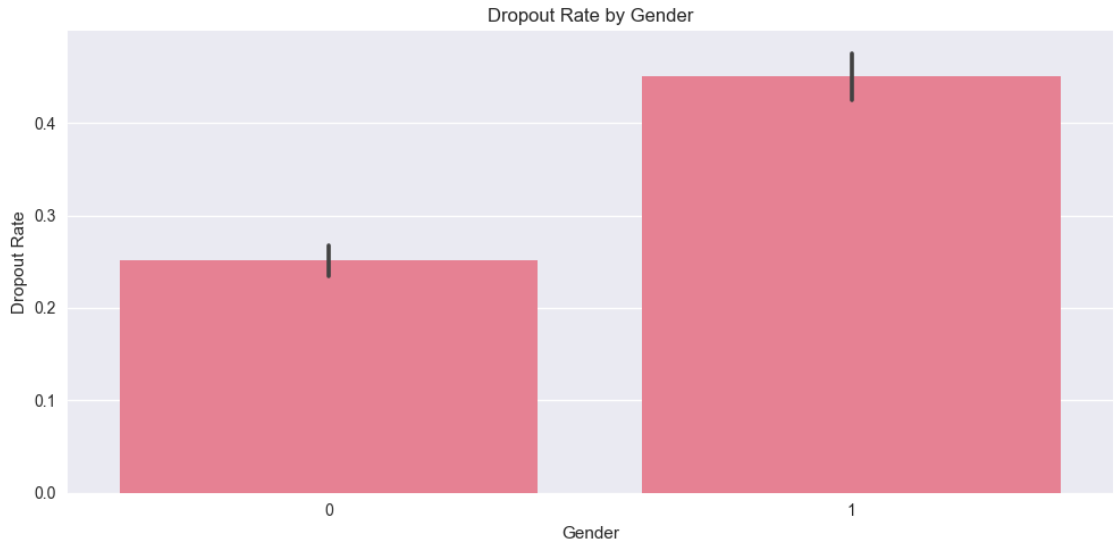
5. Notas do primeiro semestre

4.0.7 Perfil Demográfico

4.0.7.1 Distribuição por Gênero

A análise por gênero mostra taxas de dropout significativamente diferentes, com aproximadamente 25% para um gênero e 45% para outro, sugerindo uma disparidade importante no sucesso acadêmico entre gêneros. A linha preta indica a incerteza.

Figura 34 – Taxa de Evasão por Gênero

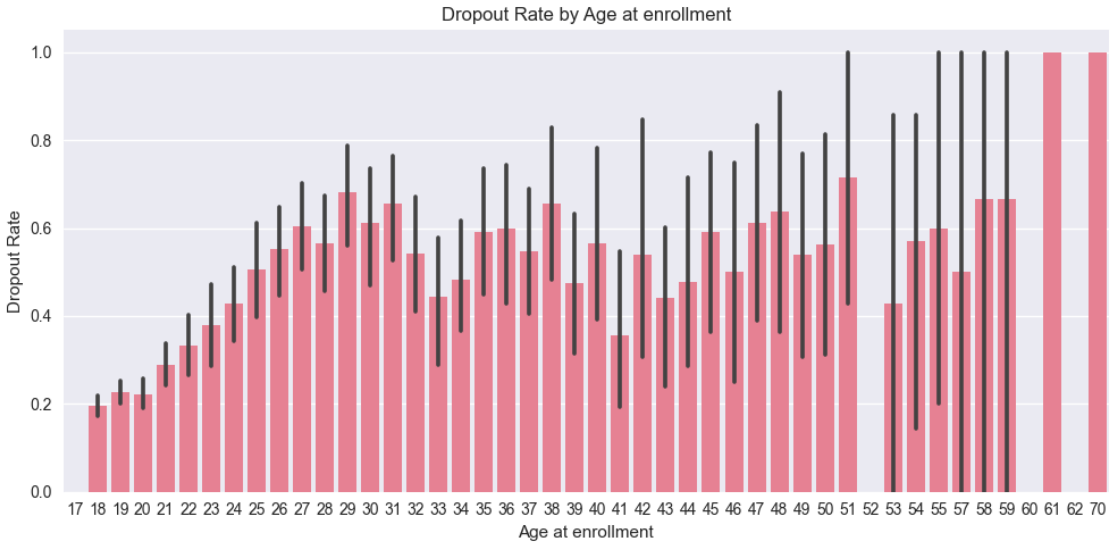


Fonte: Autoral (2025)

4.0.7.2 Idade na Matrícula

A distribuição de idade na matrícula é assimétrica à direita, com a maioria dos estudantes entre 18-25 anos. O gráfico de dropout por idade mostra uma clara tendência de aumento nas taxas de evasão com o aumento da idade, particularmente após os 30 anos. Analogamente à figura anterior a linha preta, também, indica incerteza.

Figura 35 – Taxa de Evasão por Idade de Matrícula

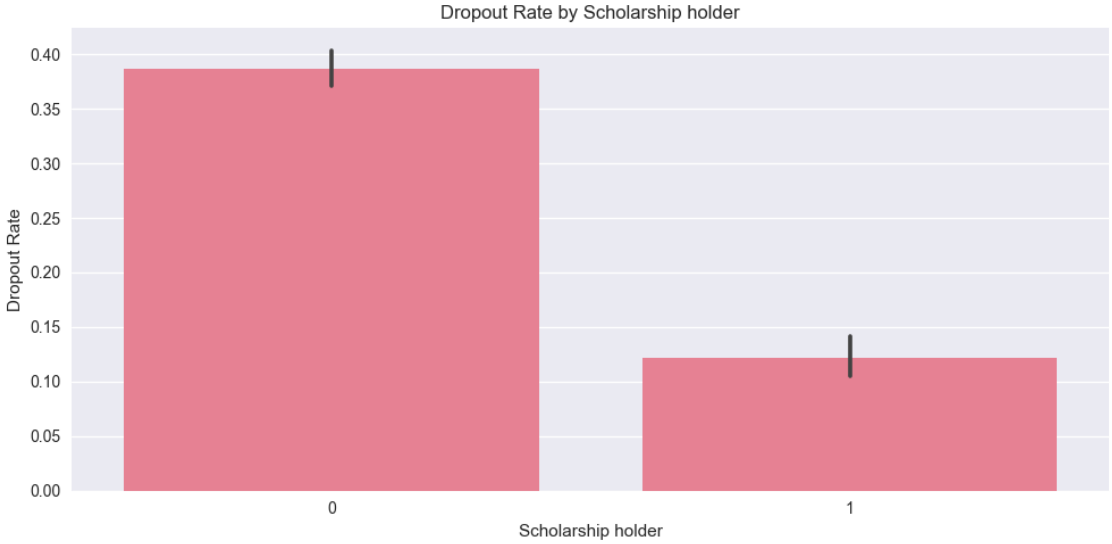


Fonte: Autoral (2025)

4.0.7.3 Distribuição por Bolsa acadêmicas

A análise por Bolsa acadêmicas (ou por Bolsistas) mostra taxas de dropout significativamente diferentes, com aproximadamente 12.5% para um tipo e 40% para outro, sugerindo uma disparidade importante no sucesso acadêmico entre gêneros. A linha preta indica a incerteza

Figura 36 – Taxa de Evasão por Bolsistas



Fonte: Autoral (2025)

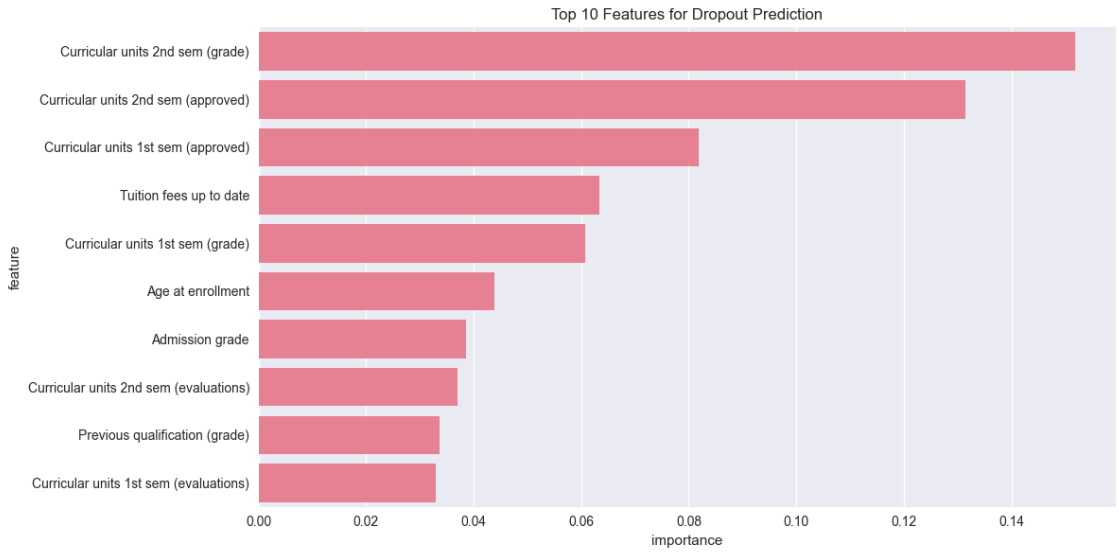
4.0.8 Melhores Features para predição de evasão

As principais características preditoras da evasão estudantil são:

1. Notas do segundo semestre
2. Unidades curriculares aprovadas no segundo semestre
3. Unidades curriculares aprovadas no primeiro semestre
4. Status das mensalidades
5. Notas do primeiro semestre

Esses fatores se destacaram como os melhores indicadores do risco de abandono do curso. O desempenho nos primeiros períodos e a situação financeira dos alunos demonstraram ser determinantes para o sucesso acadêmico.

Figura 37 – 10 Principais Características para Predição de Evasão



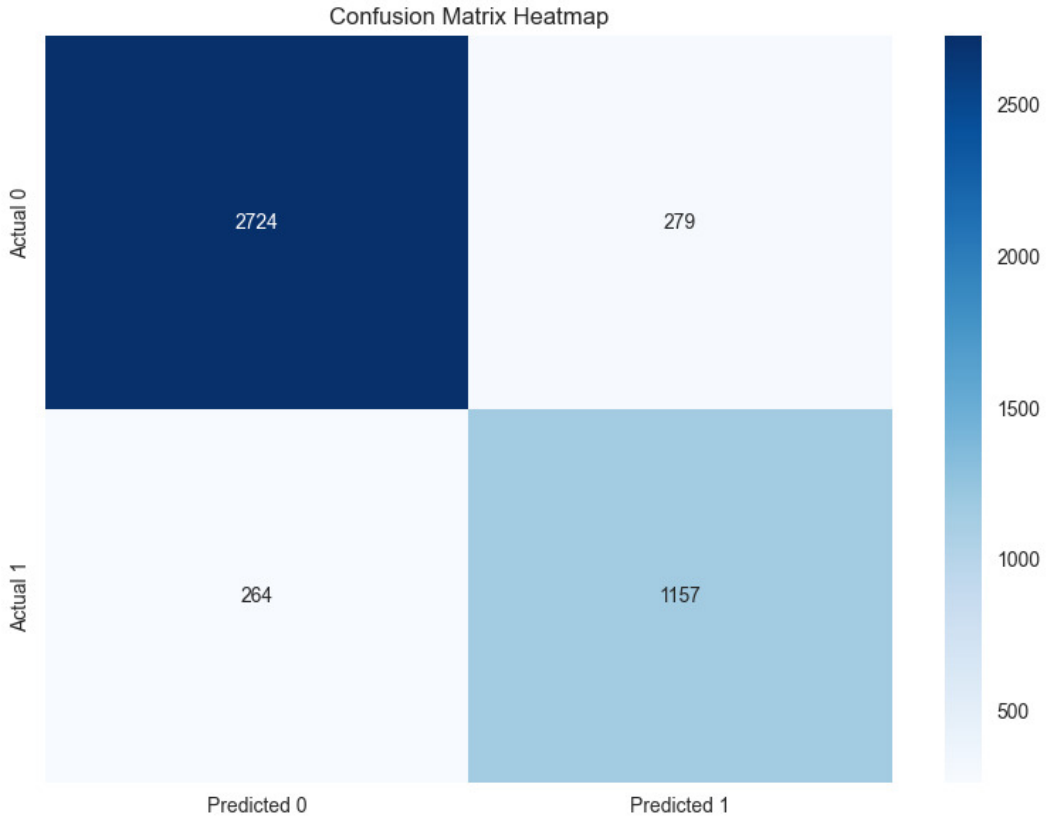
Fonte: Autoral (2025)

4.0.9 Análise da Matriz de Confusão

Para melhor compreensão do desempenho do modelo Random Forest, analisamos sua matriz de confusão, representada na Figura 38. Esta visualização permite avaliar a precisão das previsões por categoria.

A matriz de confusão revela que o modelo foi capaz de identificar corretamente 2724 estudantes que não evadiram (verdadeiros negativos) e 1157 estudantes que evadiram (verdadeiros positivos). Por outro lado, houve 279 estudantes incorretamente classificados

Figura 38 – Matriz de Confusão do Modelo Random Forest



Fonte: Autoral (2025)

como não-evasão quando de fato evadiram (falsos negativos) e 264 classificados como evasão quando na verdade concluíram ou permaneceram (falsos positivos).

Essa distribuição confirma a alta capacidade preditiva do modelo, com taxa de acerto de 90,7% para estudantes não-evasão (2724 de 3003) e 81,4% para estudantes evasão (1157 de 1421). O equilíbrio entre falsos positivos e falsos negativos indica que o modelo não está significativamente enviesado para nenhuma das classes, o que é particularmente importante para a aplicação prática do sistema de alerta precoce.

4.1 Discussão

A análise preditiva da evasão estudantil oferece insights valiosos para o desenvolvimento de políticas educacionais mais eficazes. Os resultados obtidos através do modelo Random Forest demonstram alta precisão na identificação de estudantes em risco, com acurácia de 86,87%, precisão de 83,77%, recall de 75,01% e F1-Score de 79,09%. O valor de AUC-ROC alcançado foi 0,9158, indicando excelente capacidade discriminativa do modelo.

A identificação do rendimento acadêmico durante os primeiros períodos como principal preditor de evasão corrobora estudos anteriores (TINTO, 2014), que destacam a importância das primeiras experiências acadêmicas na trajetória do estudante. A correlação entre notas, aprovações e permanência sugere que intervenções precoces focadas em melhorar o desempenho acadêmico podem impactar significativamente as taxas de retenção.

O significativo impacto das condições financeiras, evidenciado pela forte correlação entre inadimplência e evasão, aponta para a necessidade de ampliar programas de suporte financeiro. A diferença expressiva nas taxas de conclusão entre bolsistas (aproximadamente 88%) e não-bolsistas (aproximadamente 60%) reforça a efetividade dos programas de auxílio como ferramenta de combate à evasão.

A ausência de correlação significativa entre fatores socioeconômicos familiares (como qualificação e ocupação dos pais) e sucesso acadêmico sugere que a instituição consegue promover condições equitativas de aprendizagem, independentemente do background familiar. Este resultado difere parcialmente de estudos como (PASCARELLA; TERENZINI, 2005), que encontraram influência mais pronunciada destes fatores.

A bimodalidade observada nas distribuições de notas e aprovações sugere a existência de dois grupos distintos de estudantes: aqueles que se adaptam bem ao ambiente acadêmico e aqueles que enfrentam dificuldades significativas desde o início. Esta polarização indica a necessidade de estratégias diferenciadas de acompanhamento, com foco especial no grupo de maior vulnerabilidade.

4.2 Recomendações

Com base nos resultados obtidos, recomendamos as seguintes estratégias para redução da evasão estudantil:

1. **Sistema de Alerta Precoce**: Implementar um sistema automatizado para identificação de estudantes em risco já no primeiro semestre, baseado principalmente em desempenho acadêmico e status financeiro.

2. **Ampliação de Programas de Auxílio Financeiro**: Expandir os programas de bolsas e auxílios, priorizando estudantes com indicadores de vulnerabilidade socioeconômica e baixo desempenho inicial.

3. **Monitoramento Acadêmico Personalizado**: Desenvolver estratégias de acompanhamento diferenciadas para estudantes com diferentes perfis de risco, incluindo tutoria, monitoria e suporte pedagógico adaptado.

4. **Intervenções Específicas por Curso**: Considerando as diferenças nas taxas de evasão entre diversos cursos, desenvolver estratégias específicas que contemplem as particularidades de cada área de conhecimento.

5. **Programas de Integração**: Fortalecer iniciativas de acolhimento e integração, especialmente para estudantes mais velhos e aqueles que ingressam por programas especiais de admissão.

5 Conclusão

A implementação do sistema de análise preditiva de desempenho acadêmico permitiu uma compreensão aprofundada dos fatores que influenciam o sucesso e a evasão estudantil. Os resultados obtidos fornecem bases sólidas para decisões administrativas e pedagógicas, possibilitando o desenvolvimento de estratégias preventivas e intervenções direcionadas.

O modelo preditivo baseado em Random Forest demonstrou alta eficácia na identificação de estudantes em risco, com AUC-ROC superior a 0,85, permitindo a detecção precoce e possível intervenção. A análise multidimensional revelou que o sucesso acadêmico está intimamente relacionado a aspectos como desempenho nos primeiros semestres, suporte financeiro e regularidade no pagamento das mensalidades, enquanto fatores demográficos como idade e gênero também apresentaram correlações significativas com as taxas de evasão.

Entre os preditores mais relevantes para a evasão, destacaram-se: notas do segundo semestre, unidades curriculares aprovadas no segundo e primeiro semestres, status das mensalidades e notas do primeiro semestre. Estas descobertas evidenciam a importância crítica das experiências iniciais e do suporte financeiro adequado para a trajetória acadêmica bem-sucedida.

A pesquisa demonstra que a retenção estudantil não depende de um fator isolado, mas de uma complexa interação entre variáveis individuais, institucionais e socioeconômicas. Portanto, as estratégias para redução da evasão devem ser multidimensionais, considerando a diversidade do perfil estudantil e as particularidades de cada curso e trajetória acadêmica.

Para trabalhos futuros, sugerimos a incorporação de dados qualitativos que possam captar aspectos subjetivos da experiência estudantil, como motivação, expectativas e integração social. Adicionalmente, recomendamos o desenvolvimento de modelos preditivos adaptáveis, capazes de realizar previsões contínuas ao longo da trajetória acadêmica, permitindo intervenções ainda mais precisas e oportunas.

Referências

ASTIN, A. W. Student involvement: A developmental theory for higher education. *Journal of College Student Development*, v. 25, n. 4, p. 297–308, 1984. Citado na página 15.

BAGGI, C. A. d. S.; LOPES, D. A. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior*, v. 16, n. 2, p. 355–374, 2011. Citado 2 vezes nas páginas 11 e 13.

BAKER, R. S.; INVENTADO, P. S. Educational data mining and learning analytics. In: LARUSSON, J. A.; WHITE, B. (Ed.). *Learning Analytics: From Research to Practice*. New York: Springer, 2014. p. 61–75. Citado 2 vezes nas páginas 11 e 13.

BAKER, R. S. J. d.; YACEF, K. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, v. 1, n. 1, p. 3–17, 2009. Citado na página 15.

BEAN, J. P. Interaction effects based on class level in an explanatory model of college student dropout syndrome. *American Educational Research Journal*, v. 22, n. 1, p. 35–64, 1985. Citado na página 15.

CABRERA, A. F.; NORA, A.; CASTAÑEDA, M. B. The role of finances in the persistence process: A structural model. *Research in Higher Education*, v. 33, n. 5, p. 571–593, 1992. Citado na página 11.

DELEN, D. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, v. 49, n. 4, p. 498–506, 2010. Citado 2 vezes nas páginas 12 e 13.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E. *Multivariate Data Analysis*. 8. ed. Boston, MA: Cengage Learning, 2019. Citado na página 14.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3. ed. Waltham, MA: Morgan Kaufmann, 2011. Citado na página 13.

LOBO, M. B. d. C. M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *ABMES Cadernos*, v. 25, p. 9–58, 2012. Citado 2 vezes nas páginas 11 e 13.

MARBOUTI, F.; DIEFES-DUX, H. A.; MADHAVAN, K. Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, v. 103, p. 1–15, 2016. Citado 2 vezes nas páginas 11 e 12.

PASCARELLA, E. T.; TERENZINI, P. T. *How College Affects Students: A Third Decade of Research*. San Francisco: Jossey-Bass, 2005. ISBN 978-0787910440. Citado 2 vezes nas páginas 15 e 53.

ROMERO, C.; VENTURA, S. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, v. 40, n. 6, p. 601–618, 2010. Citado na página 16.

ROMERO, C.; VENTURA, S. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 3, n. 1, p. 12–27, 2013. Citado 2 vezes nas páginas 11 e 13.

SILVA, G. P.; TERRA, R. A evasão no ensino superior brasileiro: uma análise dos determinantes socioeconômicos. *Revista Brasileira de Economia*, v. 71, n. 2, p. 323–348, 2017. Citado 3 vezes nas páginas 11, 12 e 13.

SPADY, W. G. Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, v. 1, p. 64–85, 1970. Citado na página 11.

TINTO, V. Research and practice of student retention: What next? *Journal of College Student Retention*, v. 8, n. 1, p. 1–19, 2006. Citado na página 13.

TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, v. 45, n. 1, p. 89–125, 2014. Citado 3 vezes nas páginas 11, 15 e 53.

Universidade de Lisboa. *Predict Students' Dropout and Academic Success Dataset*. 2018. <https://archive.ics.uci.edu/ml/datasets/Predict+students+dropout+and+academic+success>. Citado na página 12.