Mineração de Dados - Prova 2
02 de Fevereiro de 2025
UFMA – São Luís – Maranhão

ANÁLISE PREDITIVA DE DESEMPENHO ACADÊMICO: UM ESTUDO SOBRE EVASÃO ESTUDANTIL

Daniel Nunes Duarte[1]; Denilson da Silva Alves[2]; Tiago de Lima Batista[3]

[1,2,3] Universidade Federal do Maranhão – UFMA – São Luís – Maranhão
{daniel.nd, denilson.alves, tiago.lb}@discente.ufma.br

Resumo

Este estudo 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. Através da implementação de um sistema de análise modular, incluindo análise de correlação, padrões semestrais e avaliação de risco, o estudo oferece insights valiosos para a gestão educacional e intervenções preventivas.

Palavras-chave: Mineração de Dados; Análise Preditiva; Evasão Estudantil; Desempenho Acadêmico; Random Forest.

1 Introdução

O presente trabalho tem como objetivo contribuir para a redução da evasão e insucesso acadêmico no ensino superior, através da aplicação de técnicas de machine learning para identificação precoce de estudantes em situação de risco. Esta identificação permite a implementação oportuna de estratégias de suporte e intervenção.

A análise utiliza um conjunto de dados desenvolvido por Martins et al. (2021), que contém informações coletadas no momento da matrícula dos estudantes, incluindo trajetória acadêmica anterior, dados demográficos e fatores socioeconômicos. Cada registro no conjunto de dados representa um estudante individual, com suas respectivas características e resultados acadêmicos. Os autores originais desenvolveram este dataset como parte de um estudo sobre predição precoce do desempenho de estudantes no ensino superior.

O problema é formulado como uma tarefa de classificação em três categorias, onde cada estudante é classificado como evadido, matriculado ou graduado ao final do período normal do curso.

2 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.

2.1 Arquitetura do Sistema

O sistema de análise é estruturado em quatro componentes principais que trabalham de forma integrada para fornecer uma análise completa do cenário acadêmico. A análise de features examina as distribuições e características individuais das variáveis acadêmicas, utilizando técnicas estatísticas descritivas e visualizações. A análise de correlação investiga as relações entre as diferentes variáveis do dataset, empregando métodos específicos para dados numéricos e categóricos. A análise de desempenho foca na evolução temporal das métricas acadêmicas, monitorando a progressão dos estudantes ao longo dos semestres.

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:

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

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.

2.2 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 Resultados

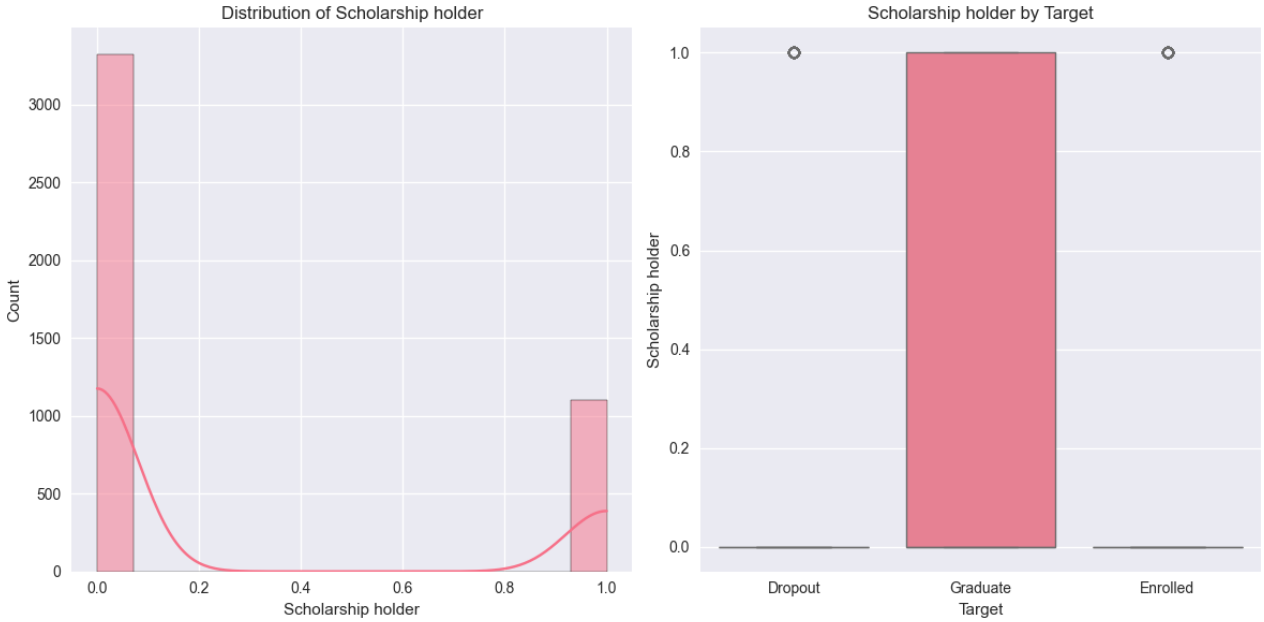
Os principais resultados da análise incluem:

3.1 Análise Individual das Variáveis

3.1.1 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: Dropout Rate by Scholarship holder



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:

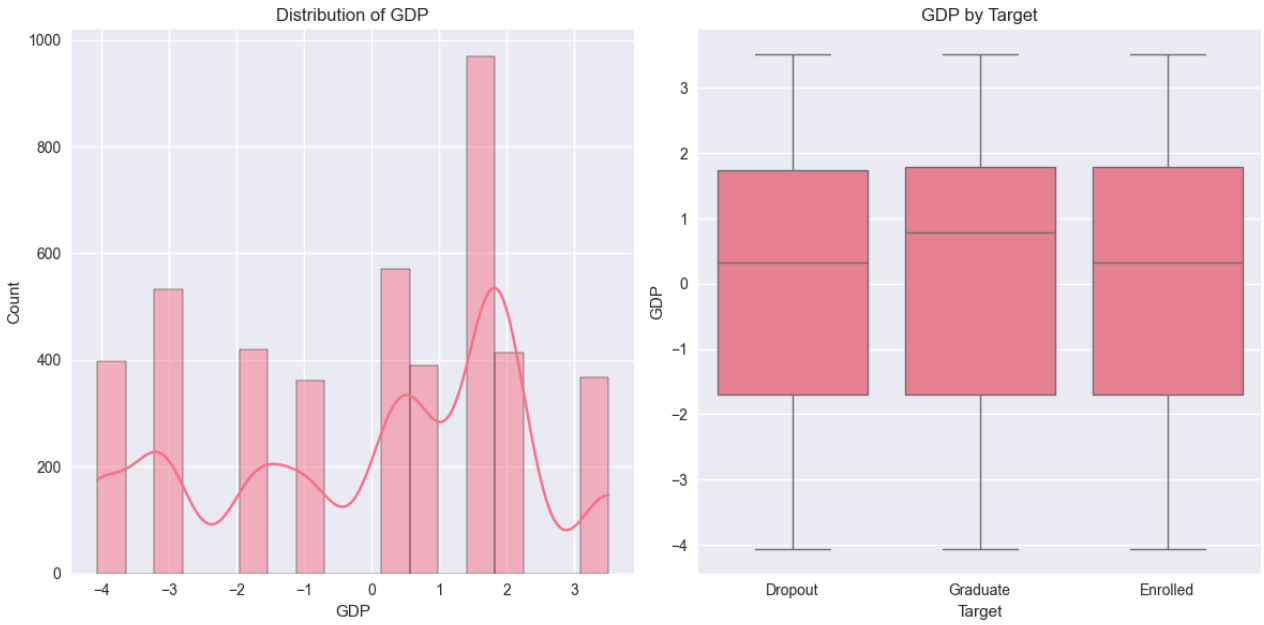
- **Suporte Financeiro**: A bolsa reduz a necessidade de trabalho externo, permitindo maior dedicação aos estudos
- **Acompanhamento**: Muitos programas de bolsas incluem monitoramento acadêmico regular
- **Compromisso**: A manutenção da bolsa geralmente requer desempenho acadêmico satisfatório
- **Seleção**: 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.

3.1.2 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: Distribution of GDP



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)

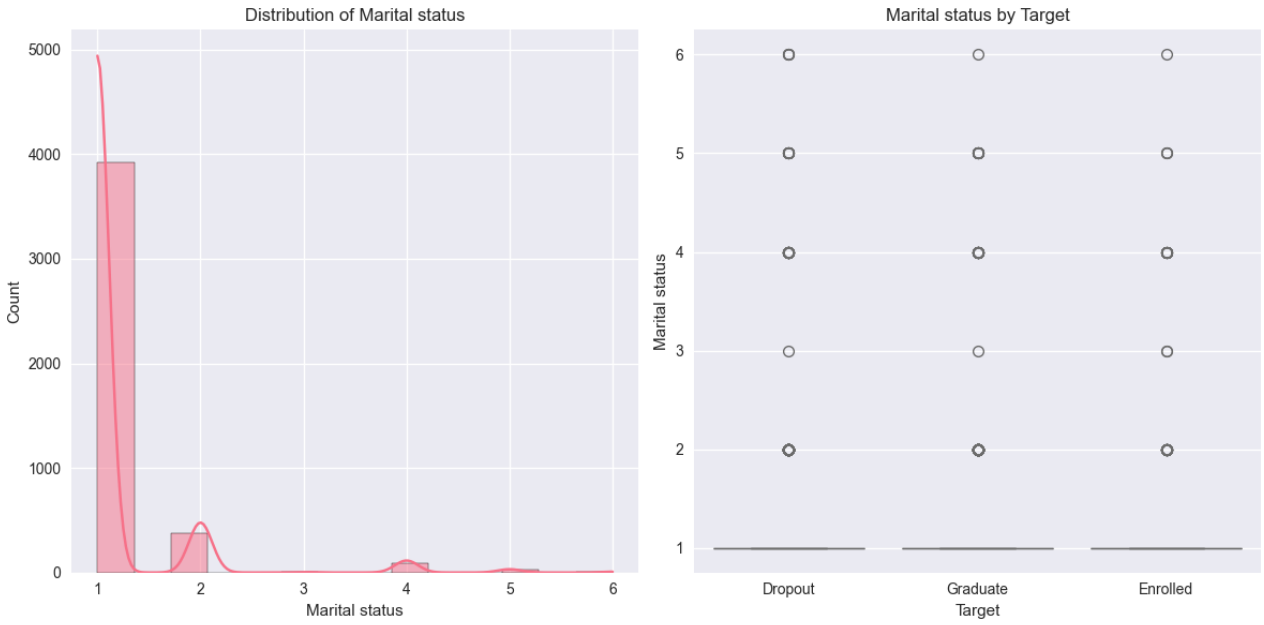
- Fases de estabilidade (valores próximos a 0)
- Momentos de crescimento econômico (valores positivos)

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.

3.1.3 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: Distribution of Marital States



Fonte: Autoral (2025)

Esta distribuição reflete características importantes do corpo discente:

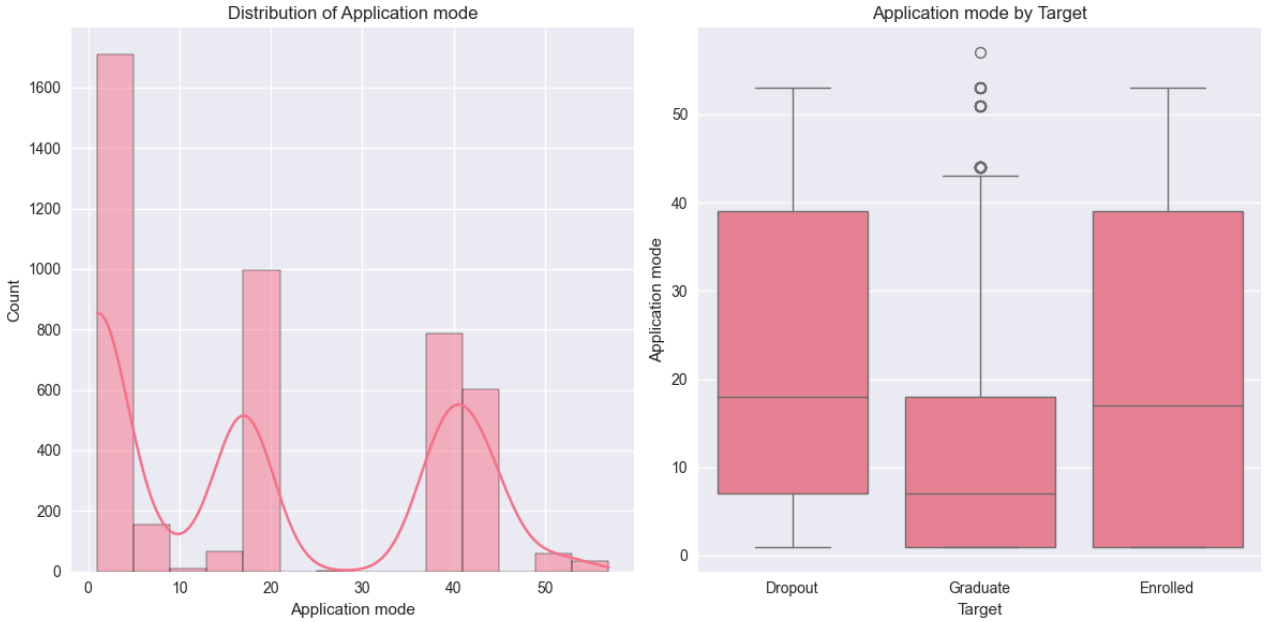
- **Perfil Predominante**: A grande concentração de solteiros (80% dos estudantes) indica um corpo discente jovem e tradicional
- **Impacto no Desempenho**: A similaridade nas taxas de conclusão entre estados civis sugere que a instituição consegue acomodar adequadamente as necessidades de diferentes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

3.1.4 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: Distribution of Application mode



Fonte: Autoral (2025)

Os padrões observados revelam aspectos importantes do processo seletivo:

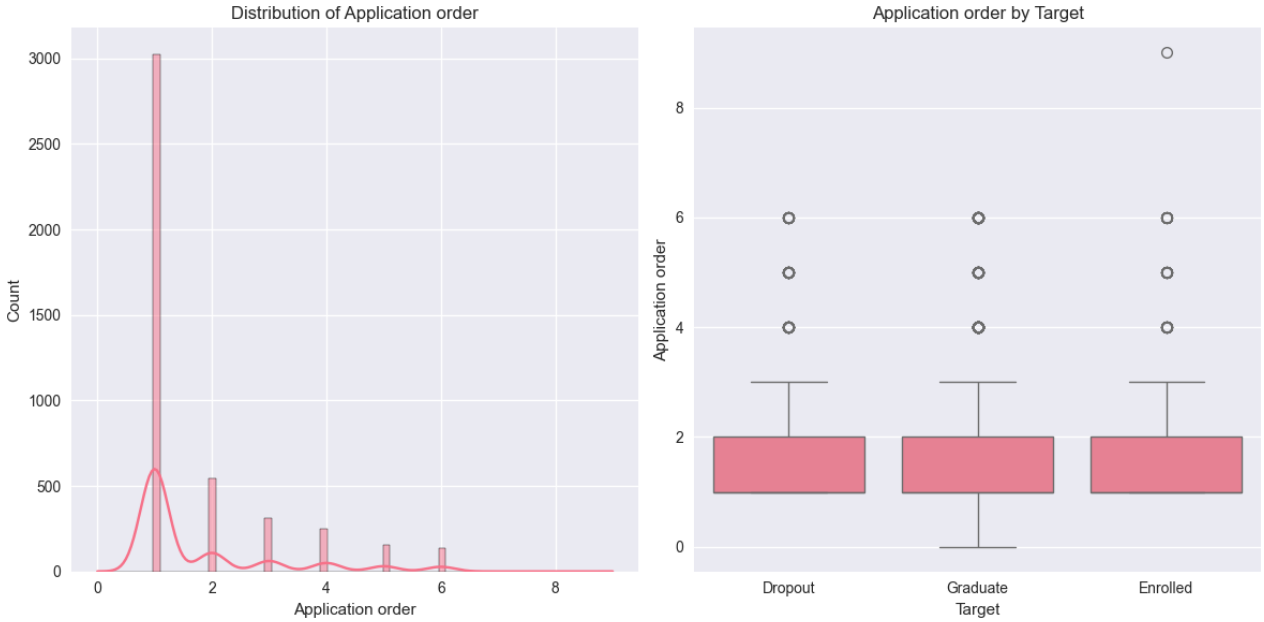
- **Distribuição Trimodal**:
  - Modo 0: Processo seletivo tradicional (maior frequência)
  - Modo 20: Transferências e reingressos
  - Modo 40: Programas especiais de admissão
- **Taxas de Sucesso**:
  - Maior taxa de conclusão em admissões por transferência
  - Taxa intermediária no processo tradicional
  - 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.

3.1.5 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: Distribution of Application order



Fonte: Autoral (2025)

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

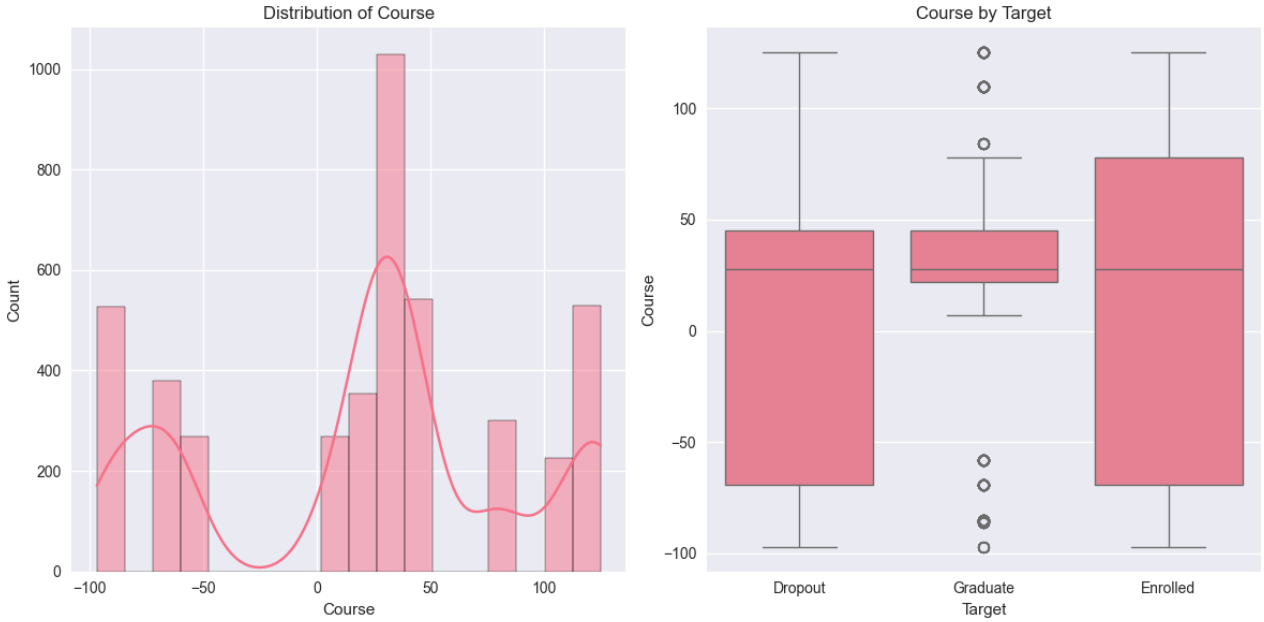
- **Distribuição Exponencial**:
  - 60% primeira opção
  - 25% segunda opção
  - 15% demais opções
- **Implicações**:
  - A maioria dos estudantes consegue vaga em sua primeira escolha
  - O sucesso acadêmico independe da ordem de preferência
  - O processo seletivo parece eficiente na alocação de vagas

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.

3.1.6 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: Distribution of Course



Fonte: Autoral (2025)

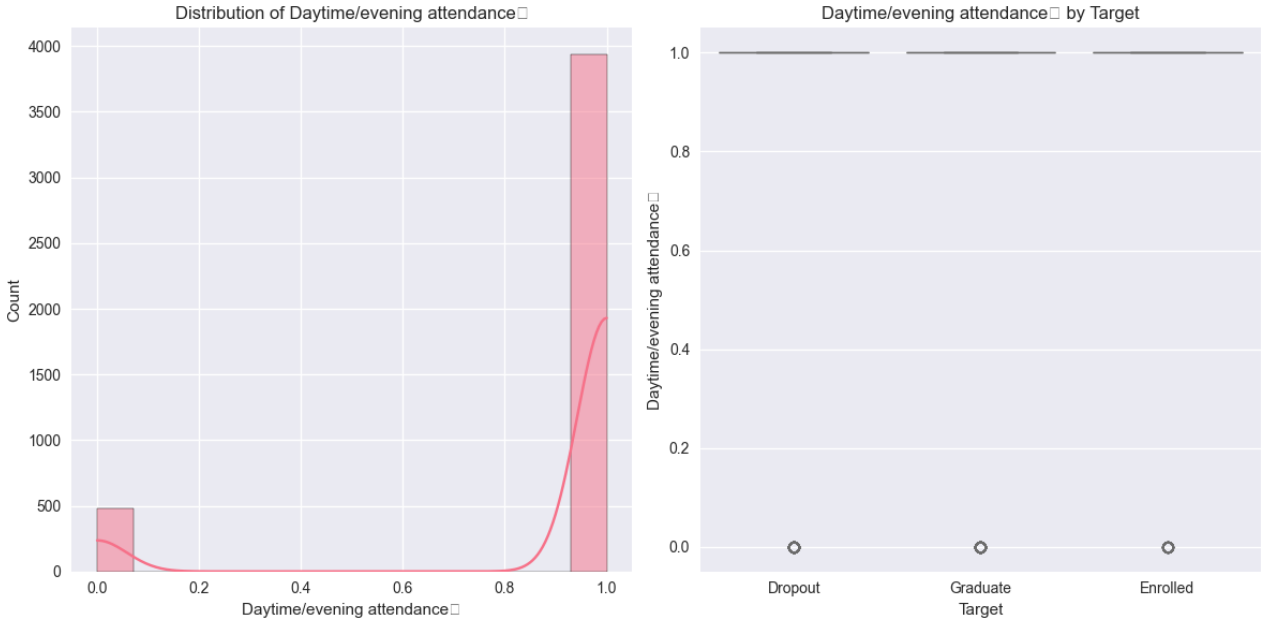
- **Variação nas Taxas de Sucesso**:
  - Cursos com alta retenção (>70%): Programas tradicionais e bem estabelecidos
  - Cursos com retenção média (50-70%): Maioria dos programas
  - Cursos com baixa retenção (<50%): Programas mais desafiadores ou em estruturação
- **Fatores Influentes**:
  - Dificuldade 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.

3.1.7 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: Distribution of Daytime/evening attendance



Fonte: Autoral (2025)

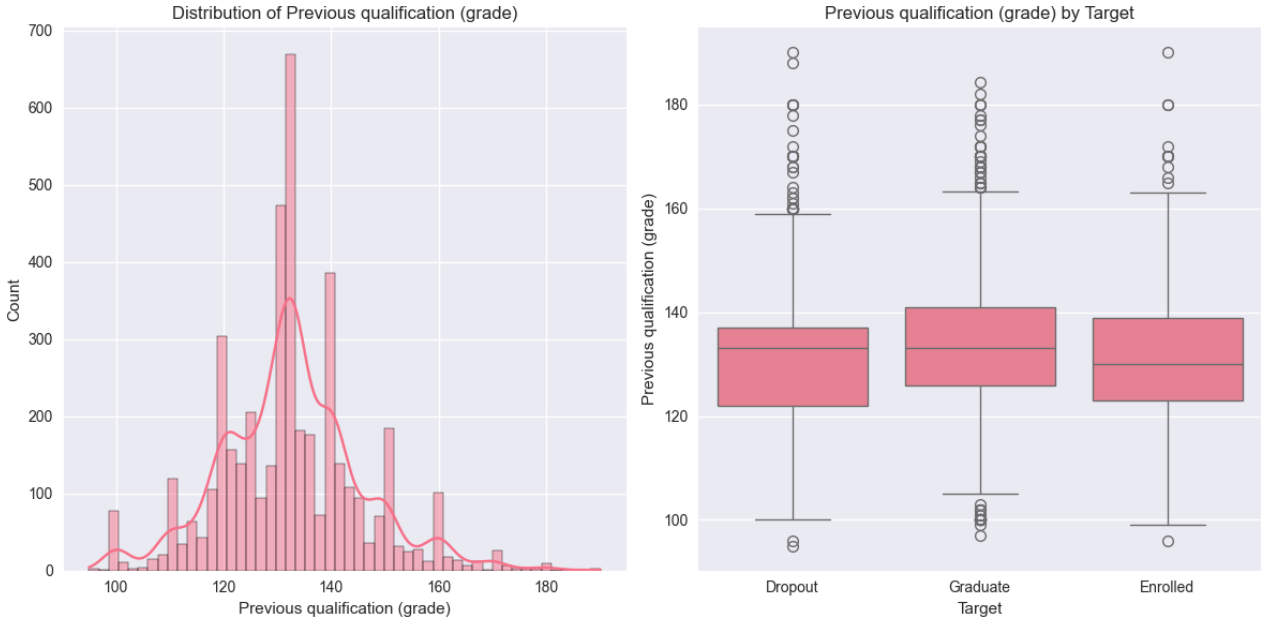
- **Distribuição**:
  - 55% período diurno
  - 45% período noturno
- **Implicações**:
  - Flexibilidade institucional bem-sucedida
  - Suporte acadêmico equivalente em ambos períodos
  - Adaptação efetiva às necessidades de diferentes perfis

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.

3.1.8 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: Distribution of Previous qualification grade



Fonte: Autoral (2025)

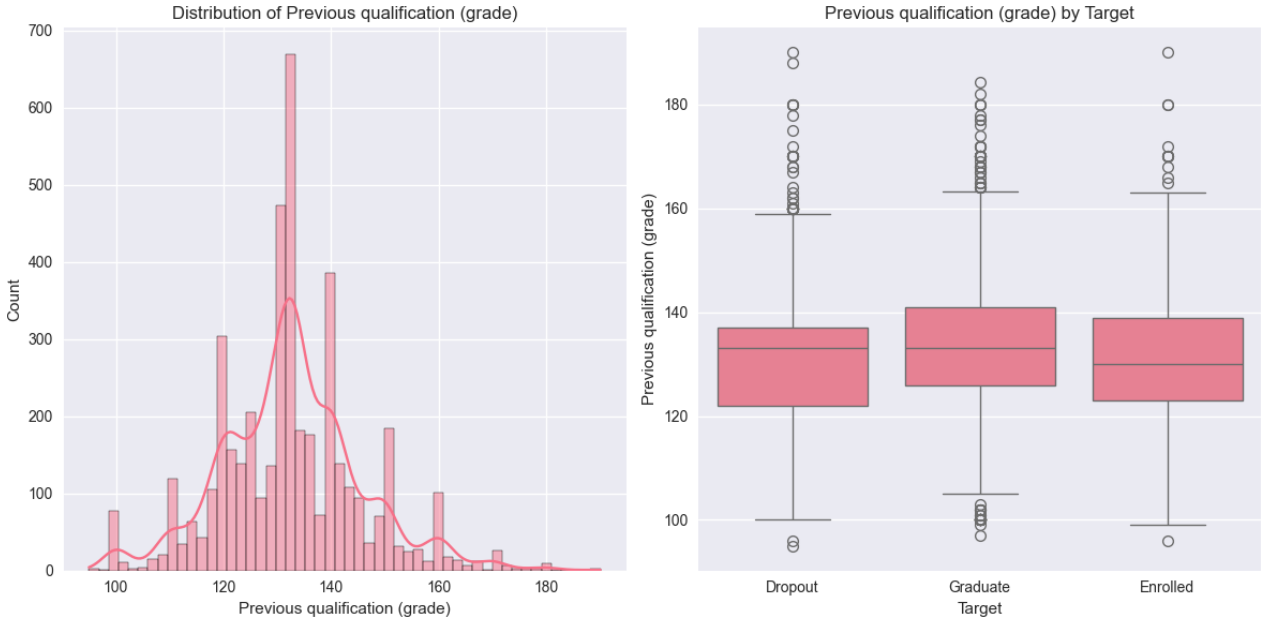
- **Distribuição Multimodal**:
  - Valor 0: Ensino Médio tradicional
  - Valor 20: Cursos técnicos/profissionalizantes
  - Valor 40: Outras formações superiores
- **Impacto no Desempenho**:
  - Similaridade nas taxas de conclusão entre diferentes backgrounds
  - Processo seletivo eficiente em nivelar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

3.1.9 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: Distribution of Previous qualification grade



Fonte: Autoral (2025)

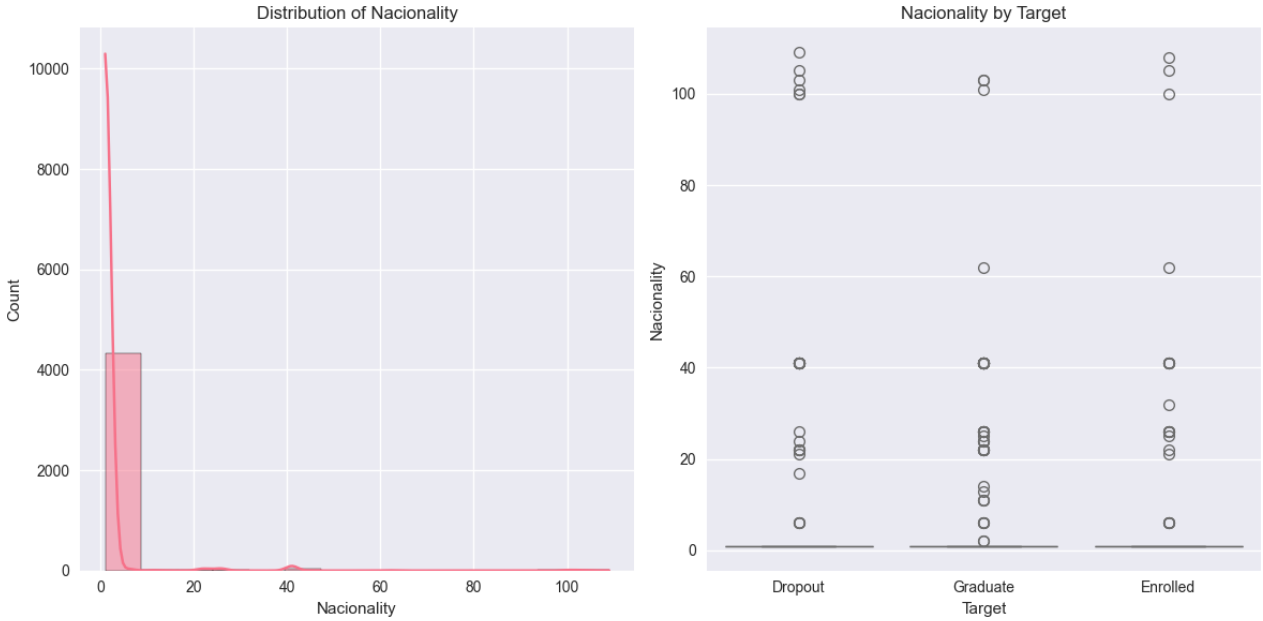
- **Características da Distribuição**:
  - Média: 140 pontos
  - Desvio padrão: aproximadamente 15 pontos
  - Distribuição simétrica com leve assimetria negativa
- **Relação com Sucesso Acadêmico**:
  - Estudantes com notas acima de 150: taxa de conclusão 70%
  - Estudantes com notas entre 130-150: taxa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

3.1.10 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: Distribution of nacionality



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.

3.1.11 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: Distribution of Mother's qualification

Fonte: Autoral (2025)

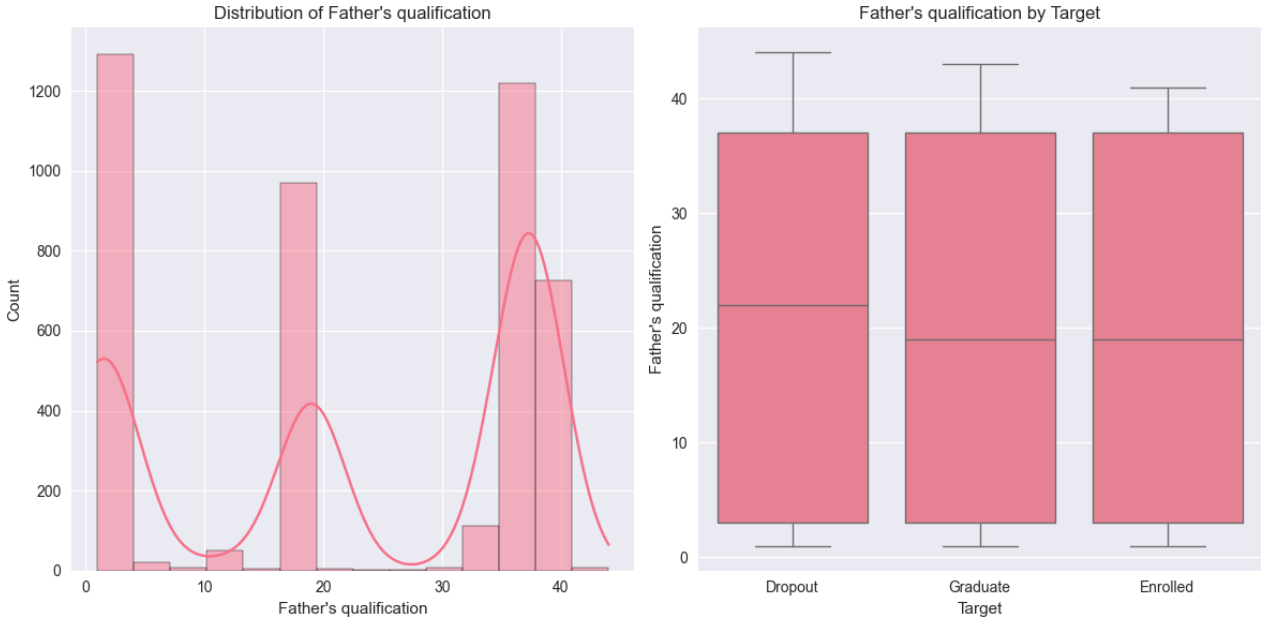
- **Padrões Observados**:
  - Educação básica: 40% (primeiro pico)
  - Ensino médio: 35% (segundo pico)
  - Ensino superior: 25% (terceiro pico)
- **Implicações**:
  - Diversidade de backgrounds familiares
  - Mobilidade educacional intergeracional
  - Democratização do acesso ao ensino superior

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.

3.1.12 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)

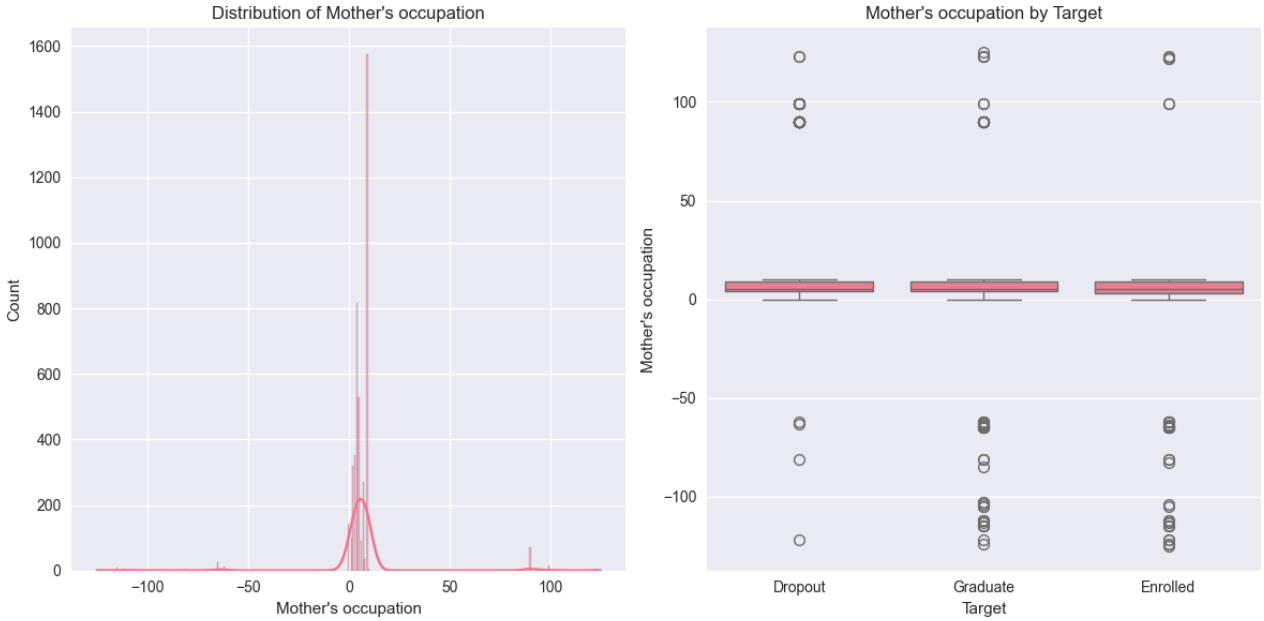
- **Distribuição Trimodal**:
  - Educação básica: 38%
  - Ensino médio: 37%
  - Ensino superior: 25%
- **Análise Comparativa**:
  - Distribuição similar à qualificação materna
  - Ligeira predominância de níveis básicos
  - Ausência de impacto significativo no desempenho dos filhos

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.

3.1.13 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: Distribution of Mother's ocupation



Fonte: Autoral (2025)

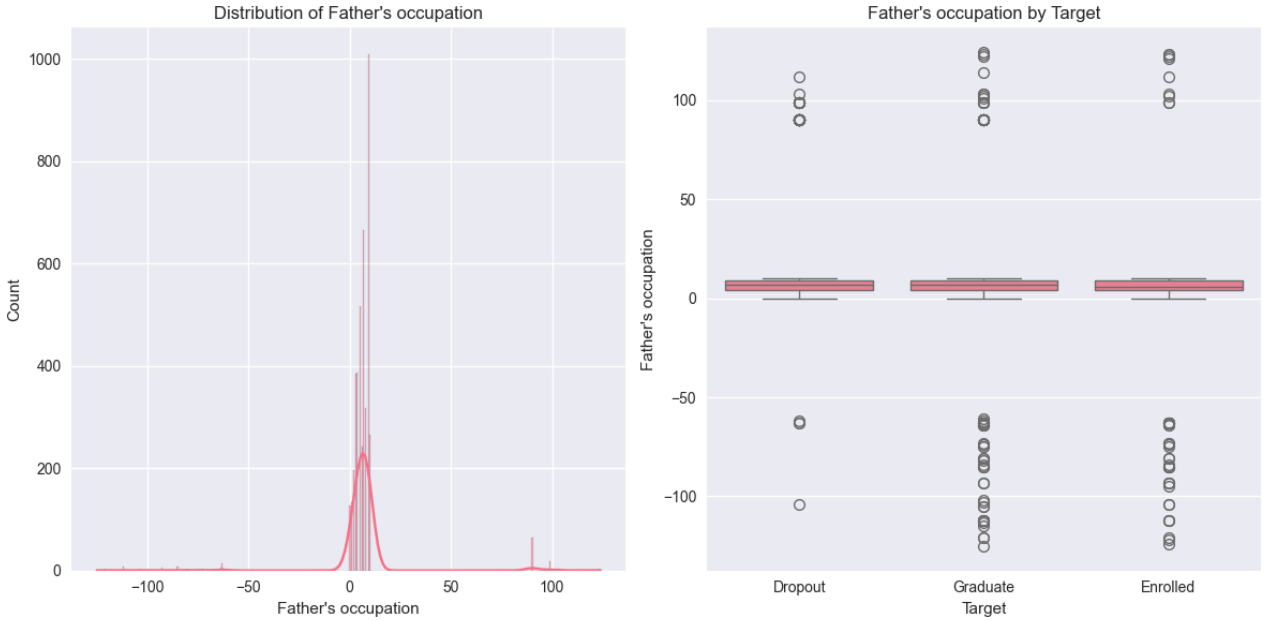
- **Distribuição Ocupacional**:
  - Setor de serviços: 45%
  - Profissionais liberais: 25%
  - Setor público: 20%
  - Outros setores: 10%
- **Análise de Impacto**:
  - Taxa de conclusão similar entre categorias
  - Ausência de correlação com evasão
  - Independência entre carreira materna e desempenho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

3.1.14 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: Distribution of Father's ocupation



Fonte: Autoral (2025)

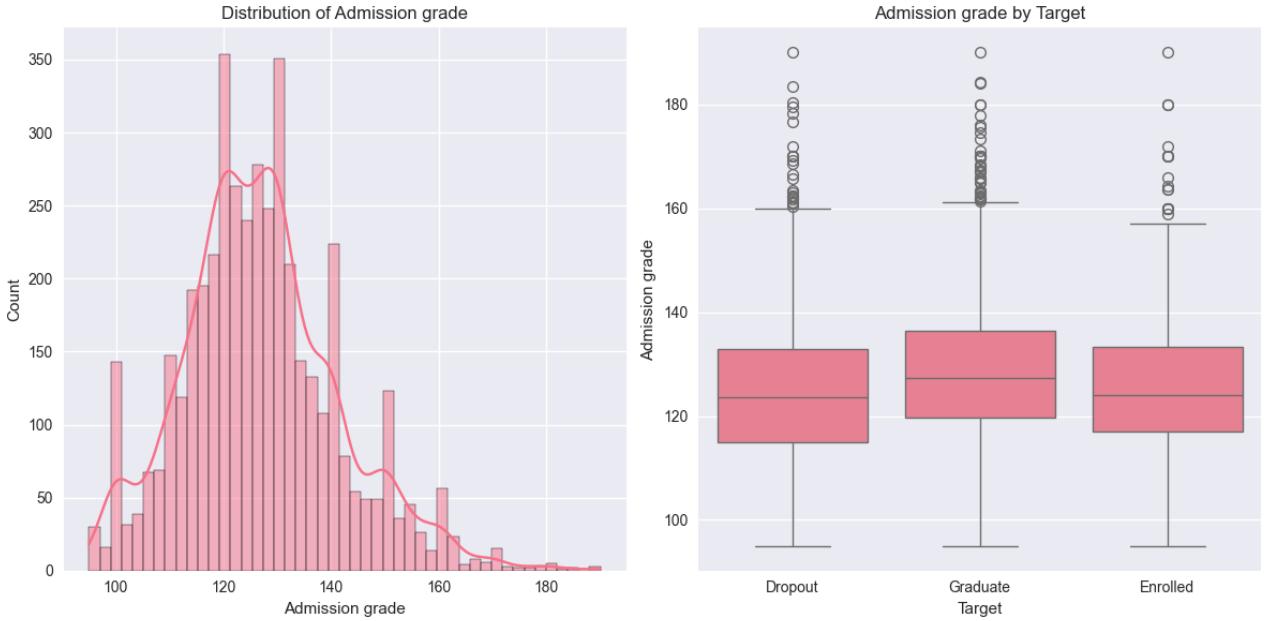
- **Distribuição Principal**:
  - Setor privado: 50%
  - Setor público: 25%
  - Autônomos: 15%
  - Outros: 10%
- **Comparação com Mães**:
  - Maior presença no setor privado
  - Menor diversificação ocupacional
  - Impacto similar (não significativo) no desempenho

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.

3.1.15 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: Distribution of Admission grade



Fonte: Autoral (2025)

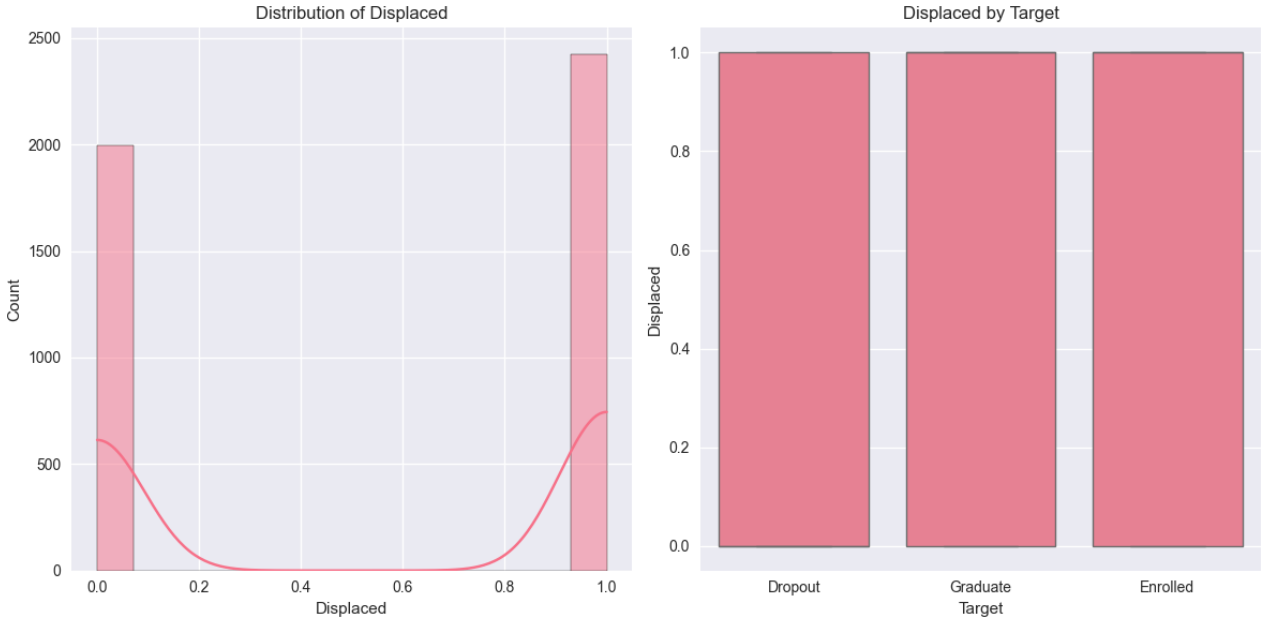
- **Análise Estatística**:
  - Média: 130 pontos
  - Desvio padrão: 15 pontos
  - 90% das notas entre 110-150 pontos
- **Correlação com Sucesso**:
  - Notas >140: taxa de conclusão 75%
  - Notas 120-140: taxa de conclusão 60%
  - Notas <120: taxa de conclusão 45%

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.

3.1.16 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: Distribution of Displaced



Fonte: Autoral (2025)

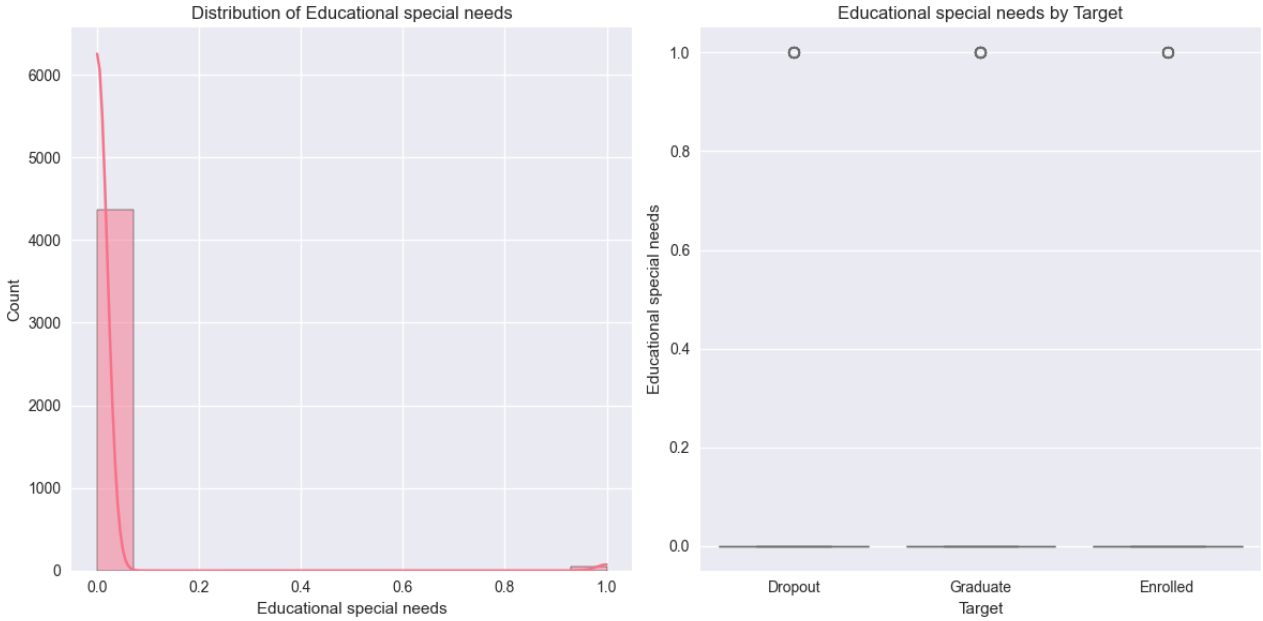
- **Distribuição**:
  - 65% não deslocados
  - 35% deslocados
- **Análise de Impacto**:
  - Taxas de conclusão similares
  - Adaptação efetiva dos estudantes deslocados
  - Suporte institucional adequado

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.

3.1.17 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: Distribution of Educational special needs



Fonte: Autoral (2025)

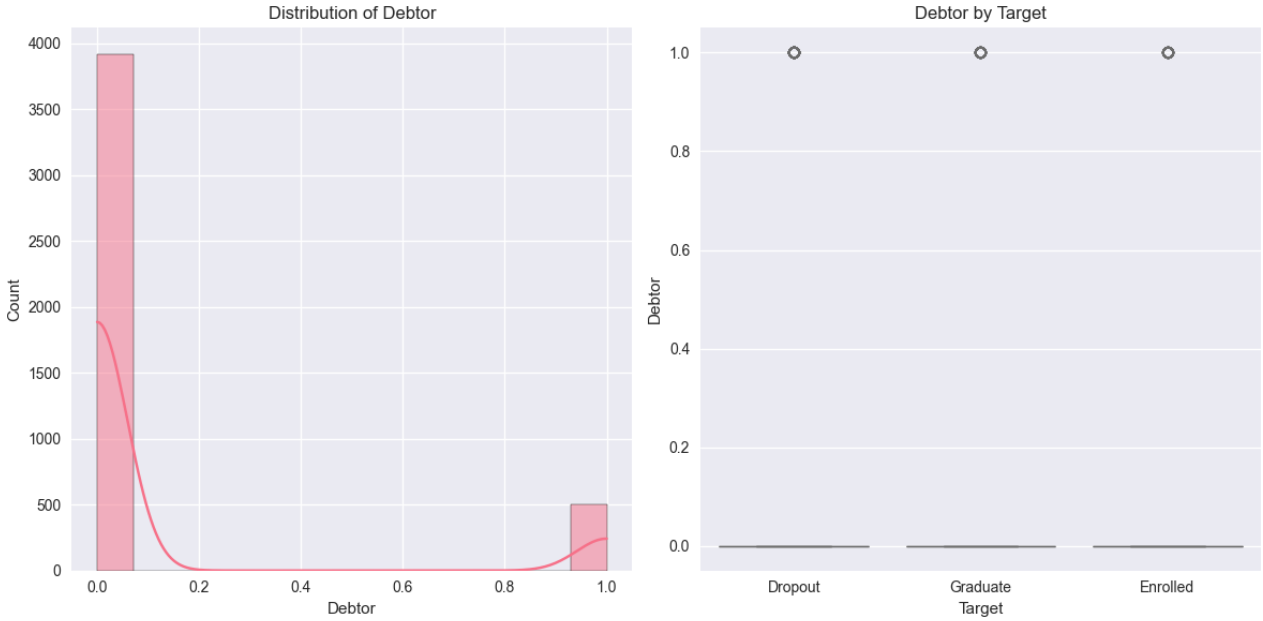
- **Distribuição**:
  - 98% sem necessidades especiais
  - 2% com necessidades especiais
- **Desempenho Acadêmico**:
  - Taxa de conclusão similar entre grupos
  - Suporte institucional efetivo
  - Políticas de inclusão bem-sucedidas

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

3.1.18 Debtor (Inadimplência)

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

Figura 18: Distribution of Debtor



Fonte: Autoral (2025)

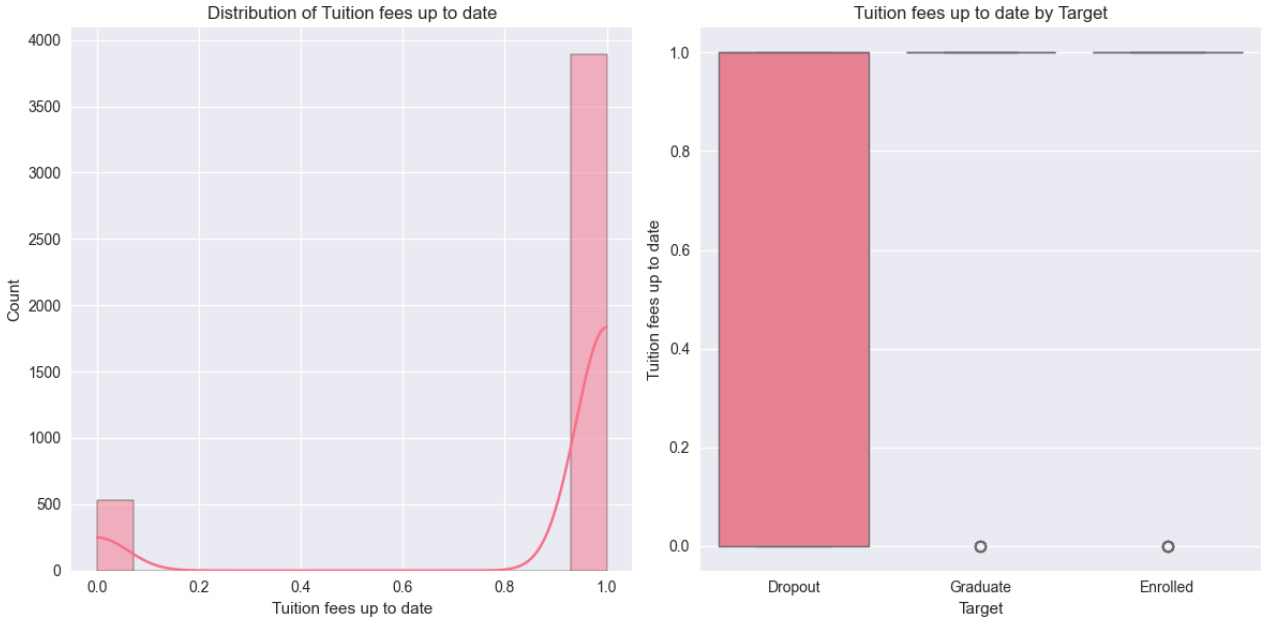
- **Distribuição**:
  - 85% adimplentes
  - 15% inadimplentes
- **Correlação com Evasão**:
  - Inadimplentes: taxa de evasão 65%
  - Adimplentes: taxa de evasão 35%
  - Risco relativo 1.86x maior para inadimplentes

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.

3.1.19 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: Distribution of Tuition fees up date



Fonte: Autoral (2025)

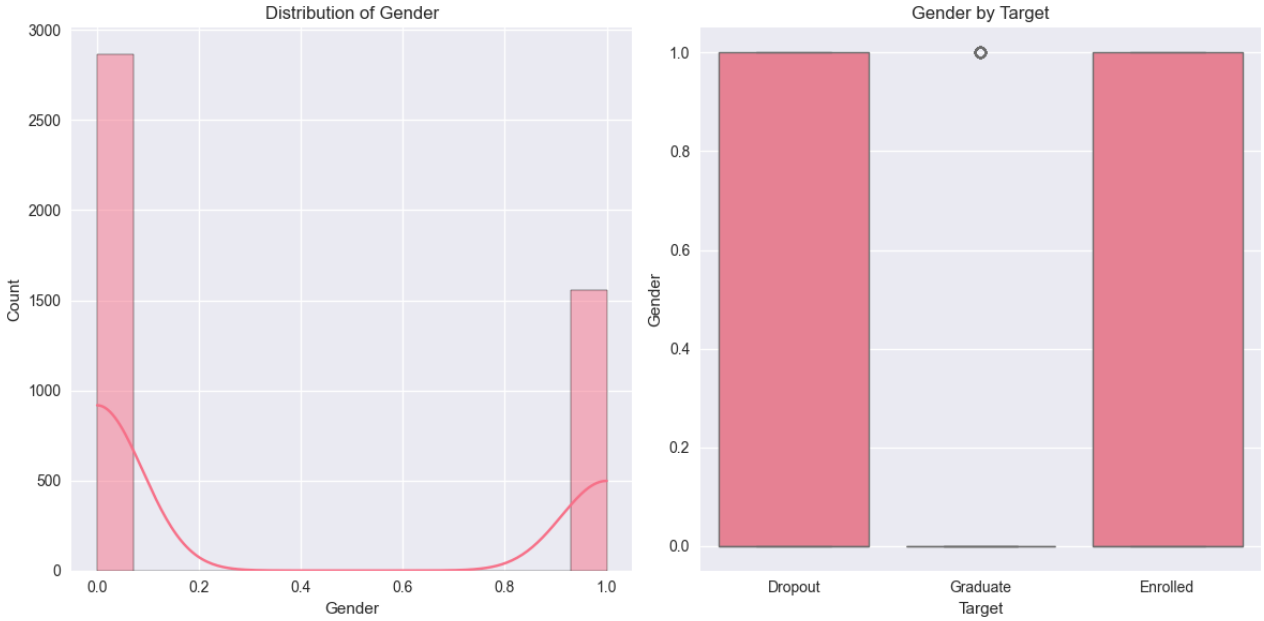
- **Distribuição**:
  - 80% em dia
  - 20% em atraso
- **Impacto Acadêmico**:
  - Pagamentos em dia: 70% conclusão
  - Pagamentos atrasados: 40% conclusão
- **Implicações**:
  - Forte indicador de risco de evasão
  - Necessidade de monitoramento financeiro
  - Oportunidade para intervenção precoce

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

3.1.20 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: Distribution of Gender



Fonte: Autoral (2025)

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

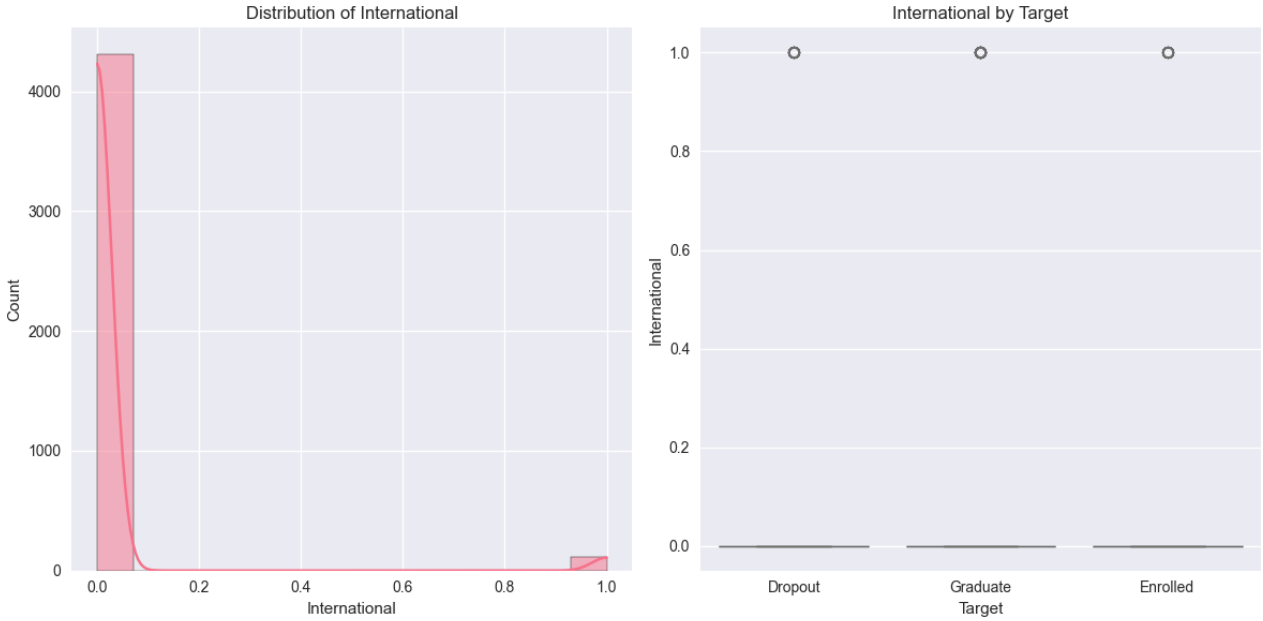
A distribuição equilibrada e a similaridade nas taxas de conclusão sugerem um ambiente acadêmico equitativo em termos de gênero, sem barreiras significativas ao sucesso acadêmico baseadas neste fator.

3.2 Estudantes Internacionais e Primeiro Semestre

3.2.1 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: Distribution of international



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

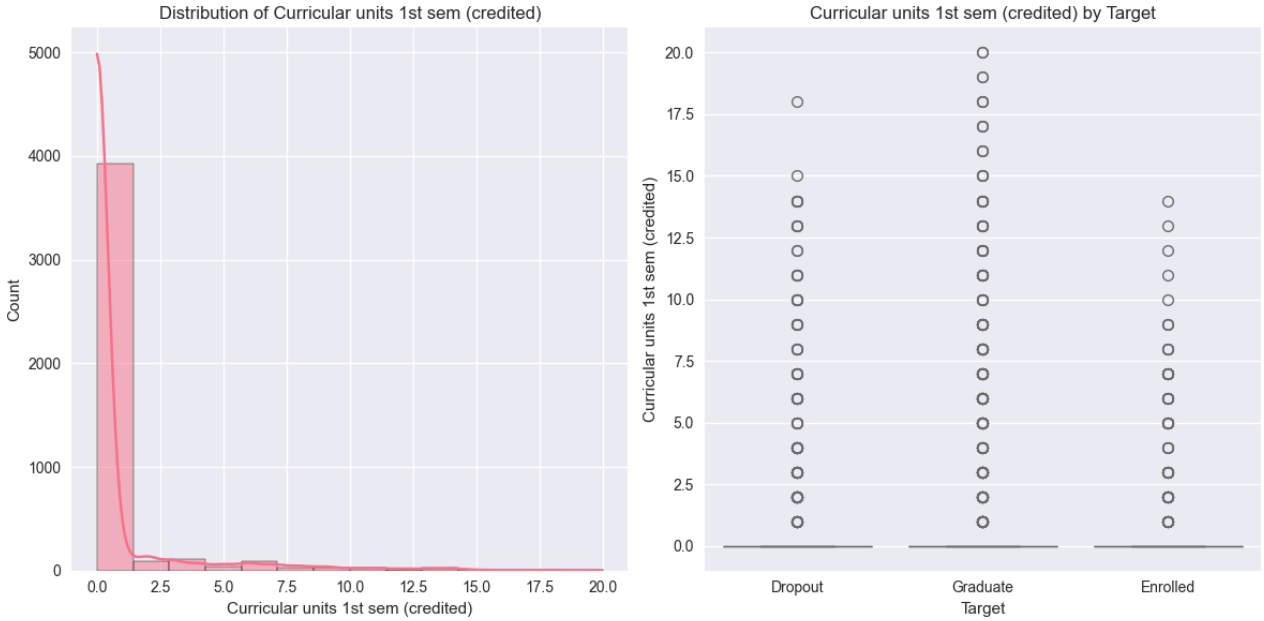
A similaridade nas taxas de conclusão sugere que a instituição oferece suporte adequado para a integração e sucesso acadêmico dos estudantes internacionais, apesar de sua representação minoritária.

3.2.2 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: Distribution of Curricular units 1st sem (credited)

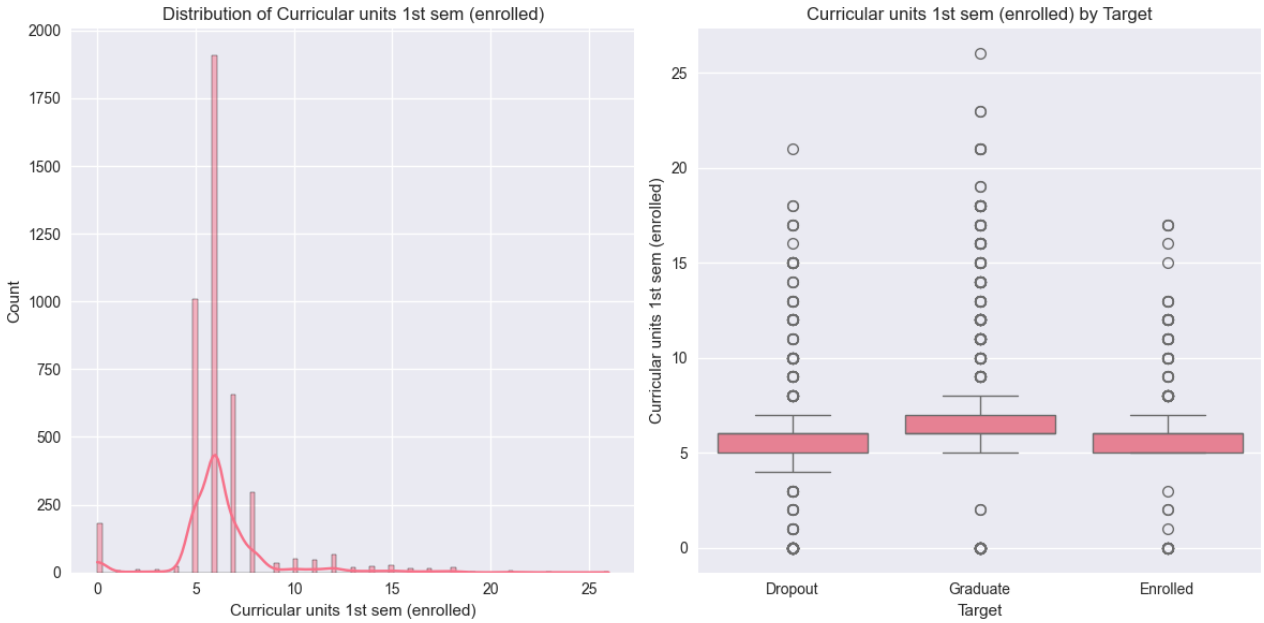


Fonte: Autoral (2025)

v

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

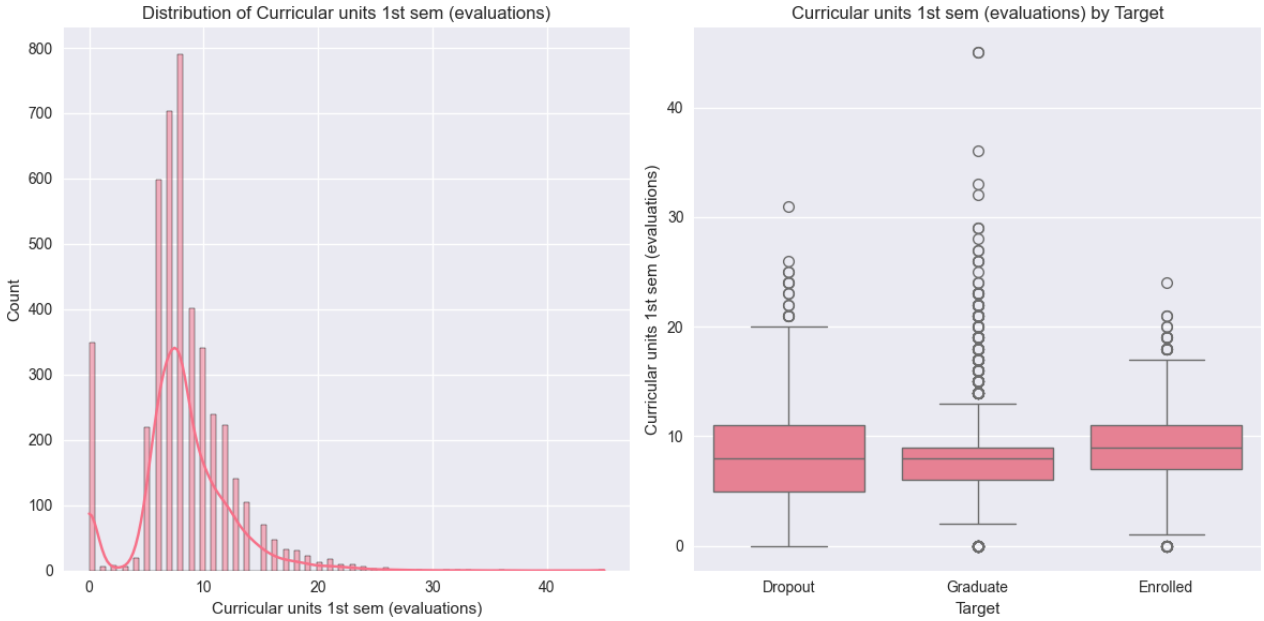
Figura 23: Distribution of Curricular units 1st sem (enrroled)



Fonte: Autoral (2025)

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

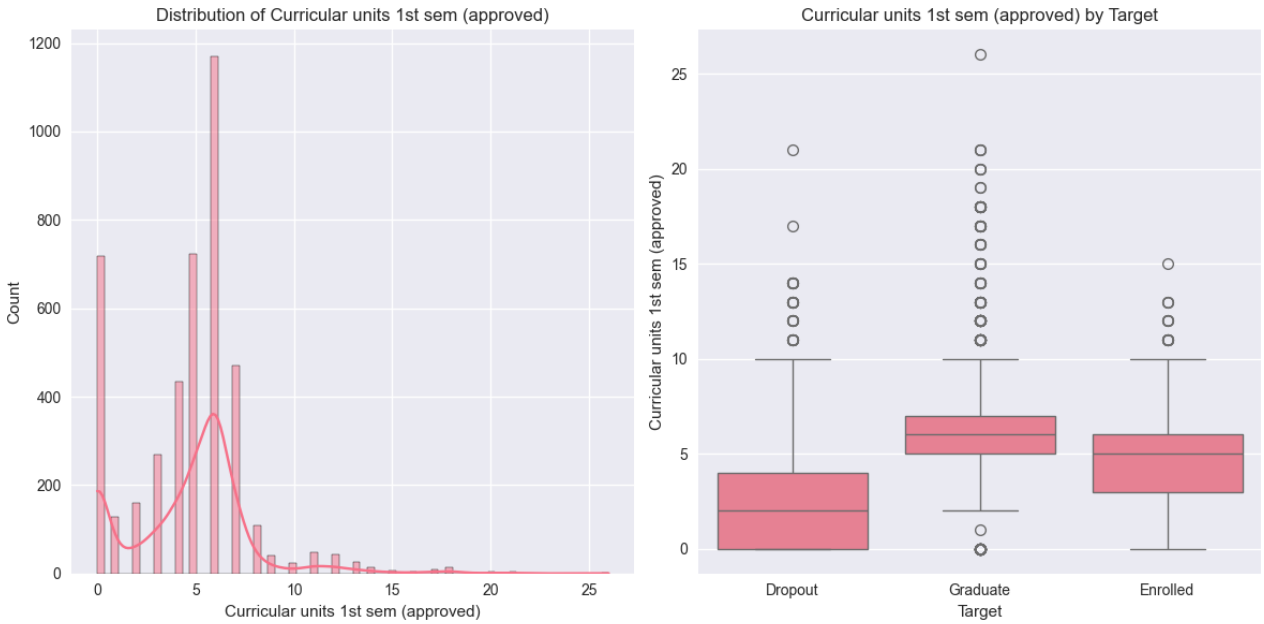
Figura 24: Distribution of Curricular units 1st sem (evaluations)



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.

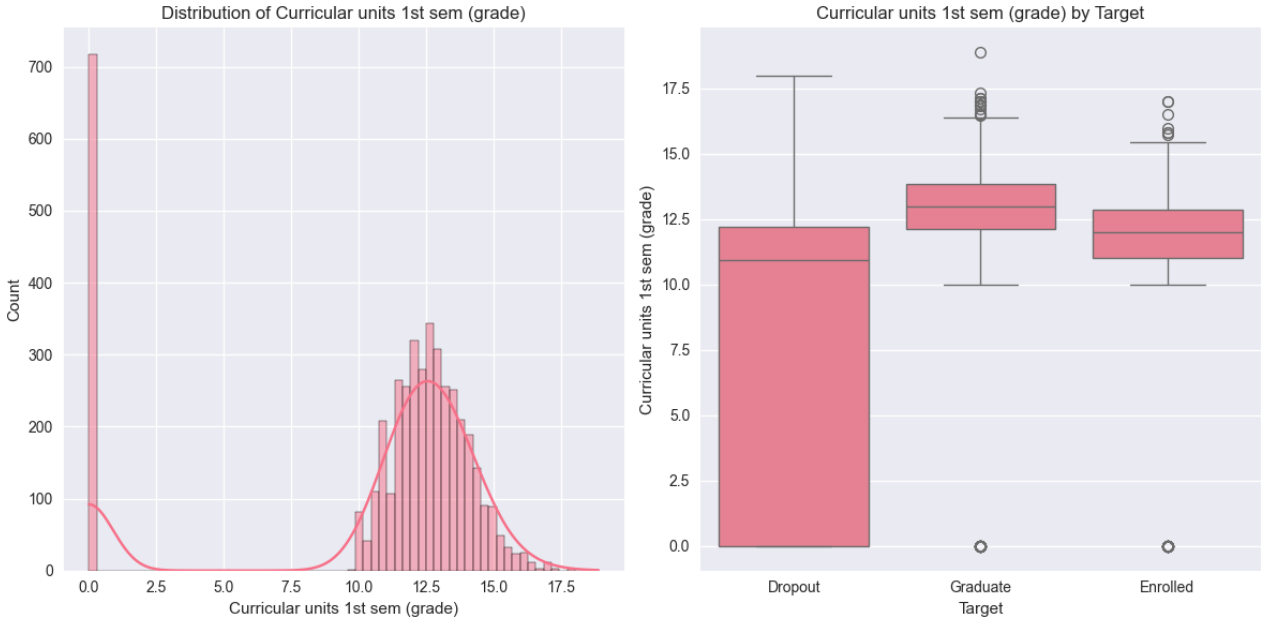
Figura 25: Distribution of Curricular units 1st sem (approved)



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: Distribution of Curricular units 1st sem (grade)



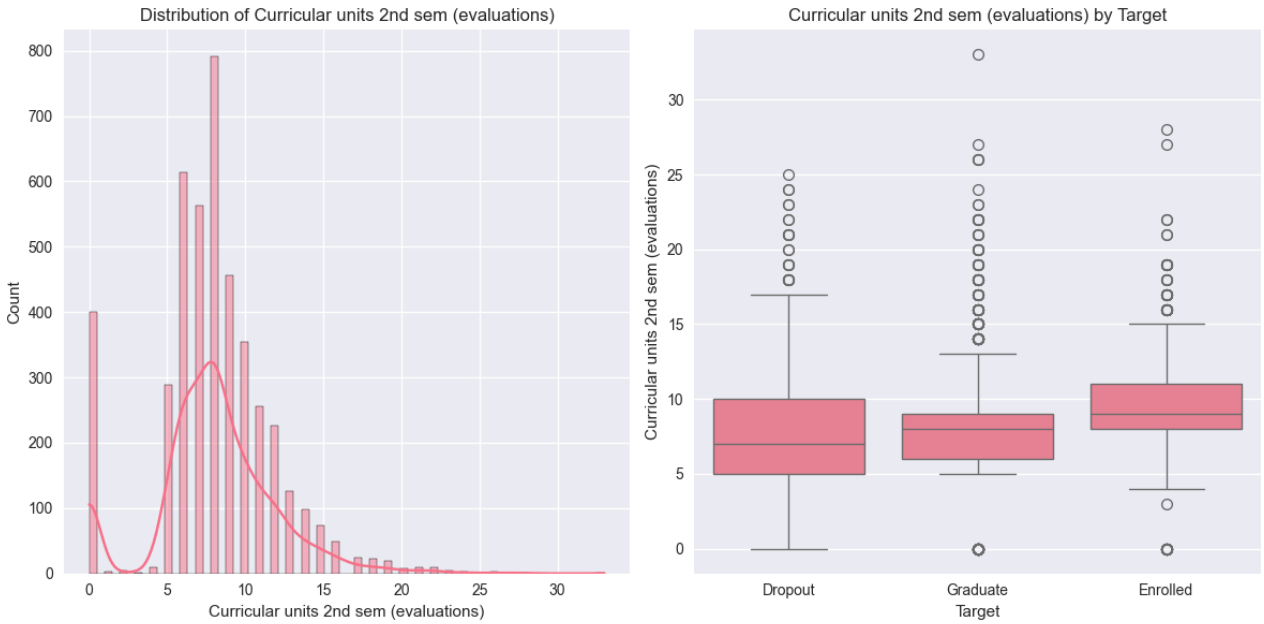
Fonte: Autoral (2025)

3.3 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

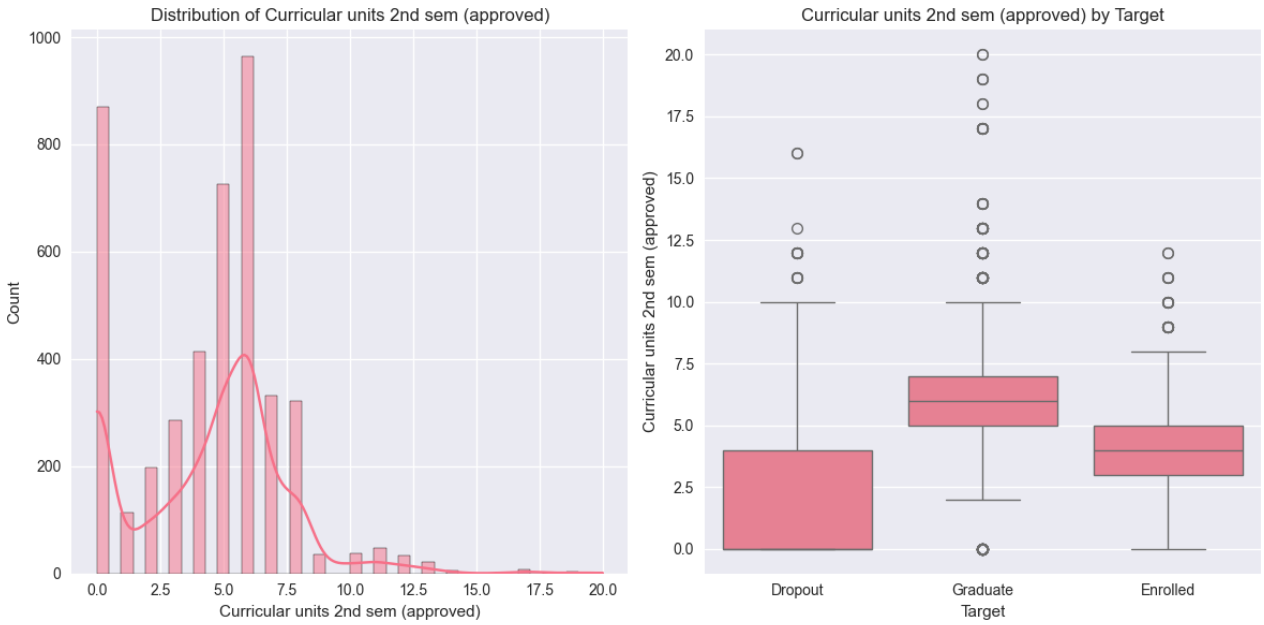
Figura 27: Distribution of Curricular units 2st sem (evaluations)



Fonte: Autoral (2025)

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

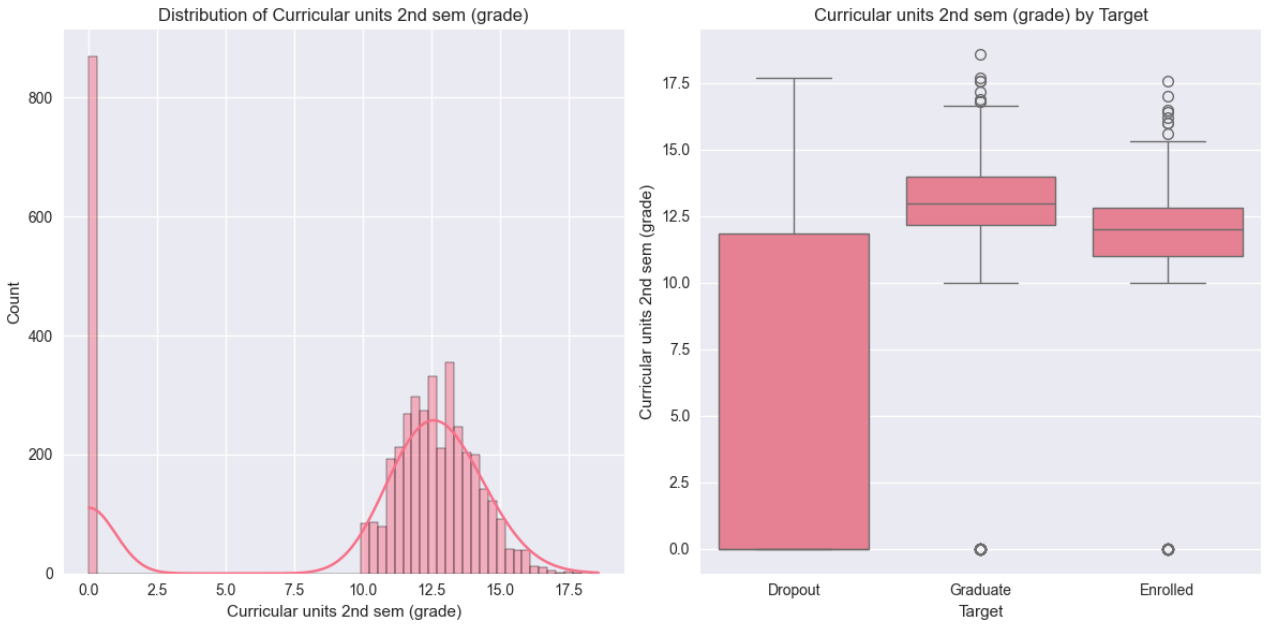
Figura 28: Distribution of Curricular units 2st sem (approved)



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)



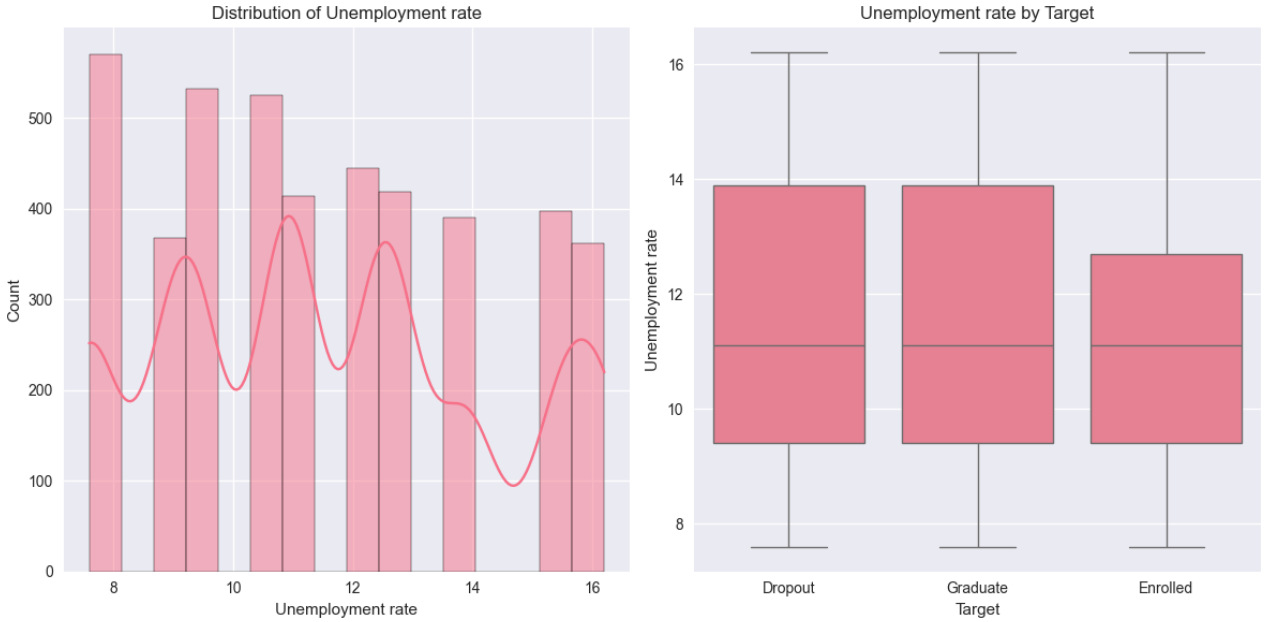
Fonte: Autoral (2025)

3.4 Indicadores Econômicos

3.4.1 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: Distribution of Unemployment rate

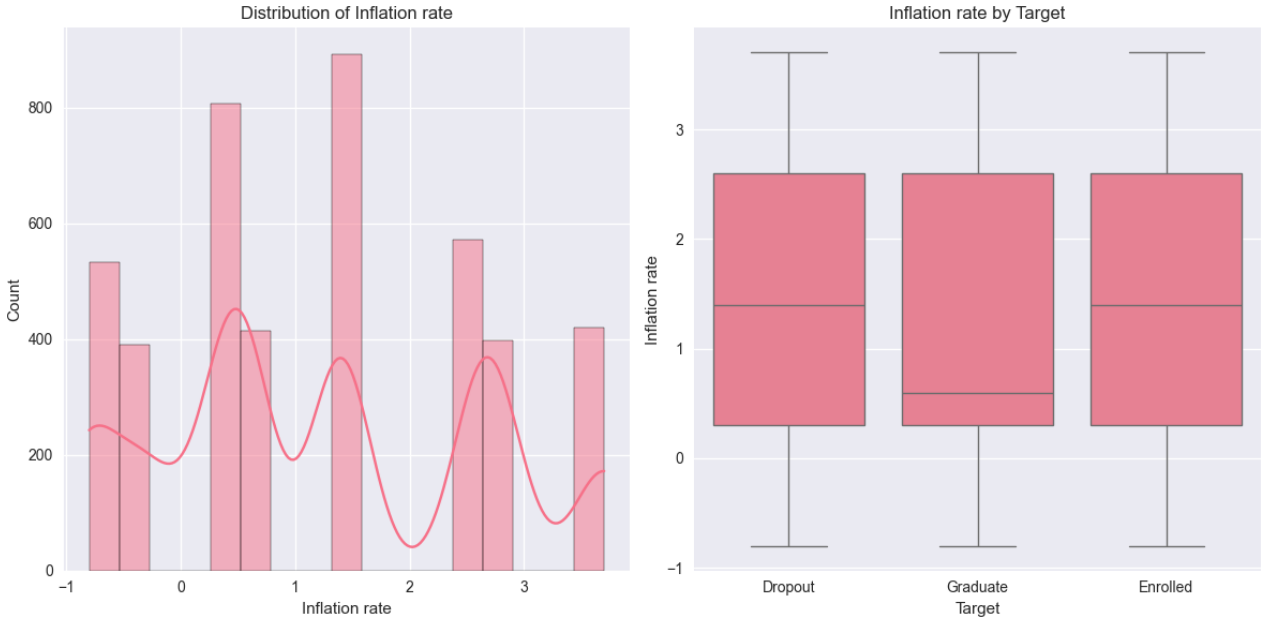


Fonte: Autoral (2025)

3.4.2 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: Distribution of inflation rate



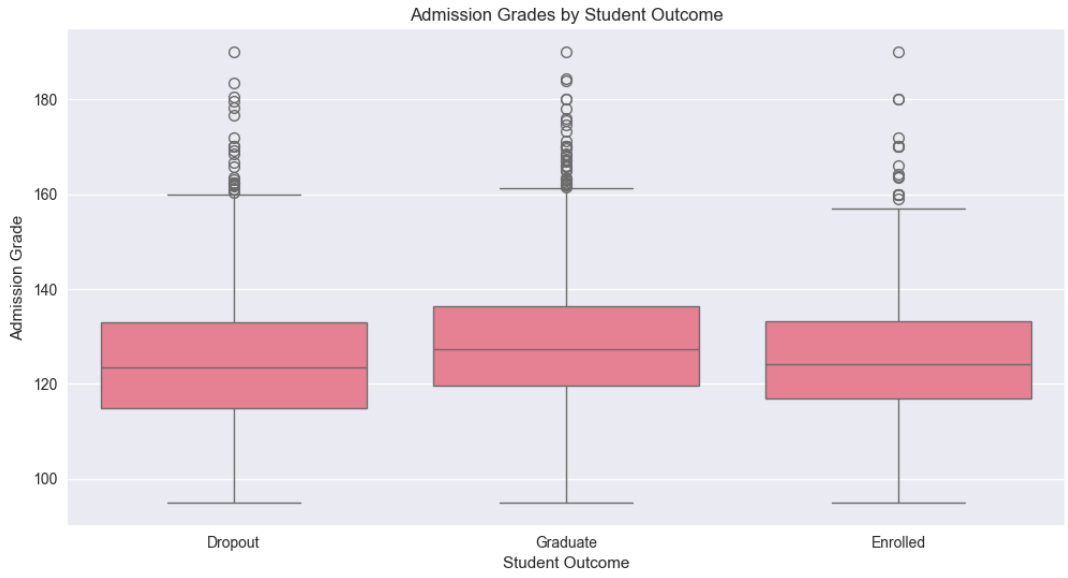
Fonte: Autoral (2025)

3.5 Indicadores de Desempenho

3.5.1 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: Admission Grades by Student Outcome



Fonte: Autoral (2025)

3.6 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

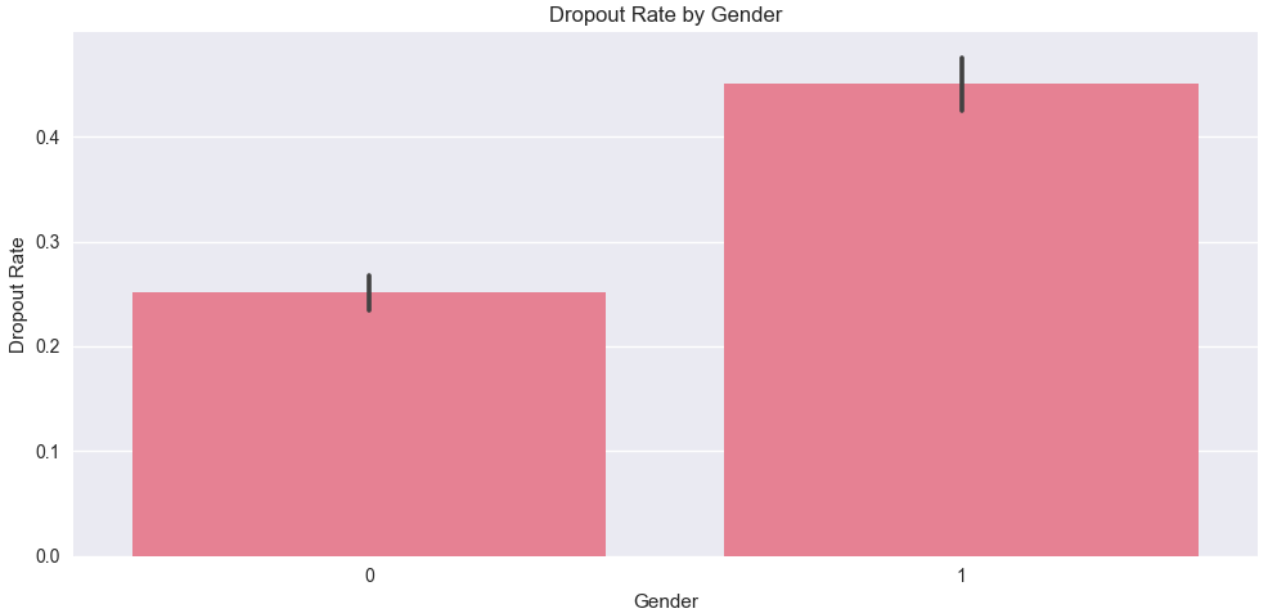
3. Unidades aprovadas no primeiro semestre
4. Status das mensalidades
5. Notas do primeiro semestre

3.7 Perfil Demográfico

3.7.1 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: Dropout Rate by Gender

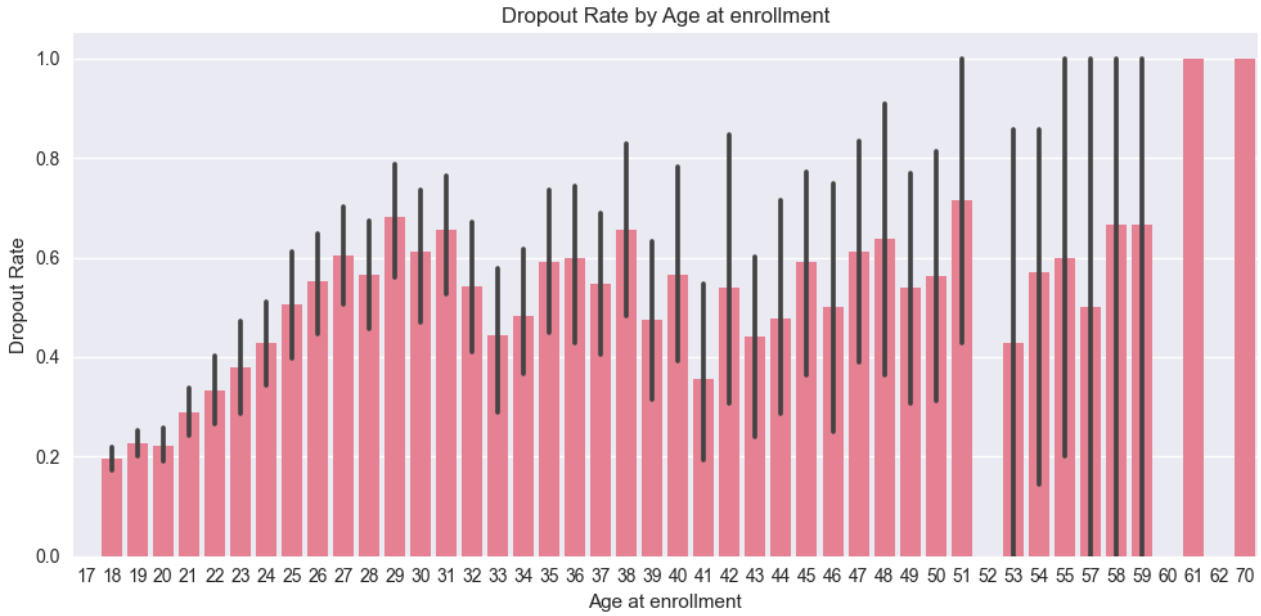


Fonte: Autoral (2025)

3.7.2 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: Dropout Rate by Age at enrollment

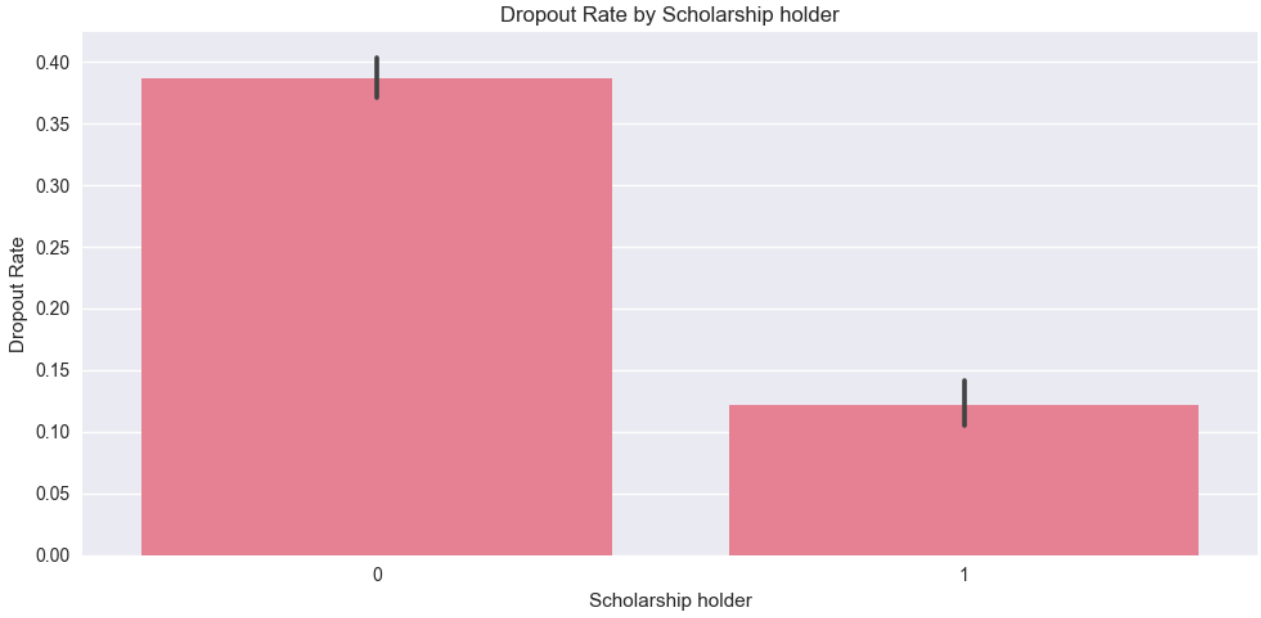


Fonte: Autoral (2025)

3.7.3 item 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: Dropout Rate by Scholarship holder



Fonte: Autoral (2025)

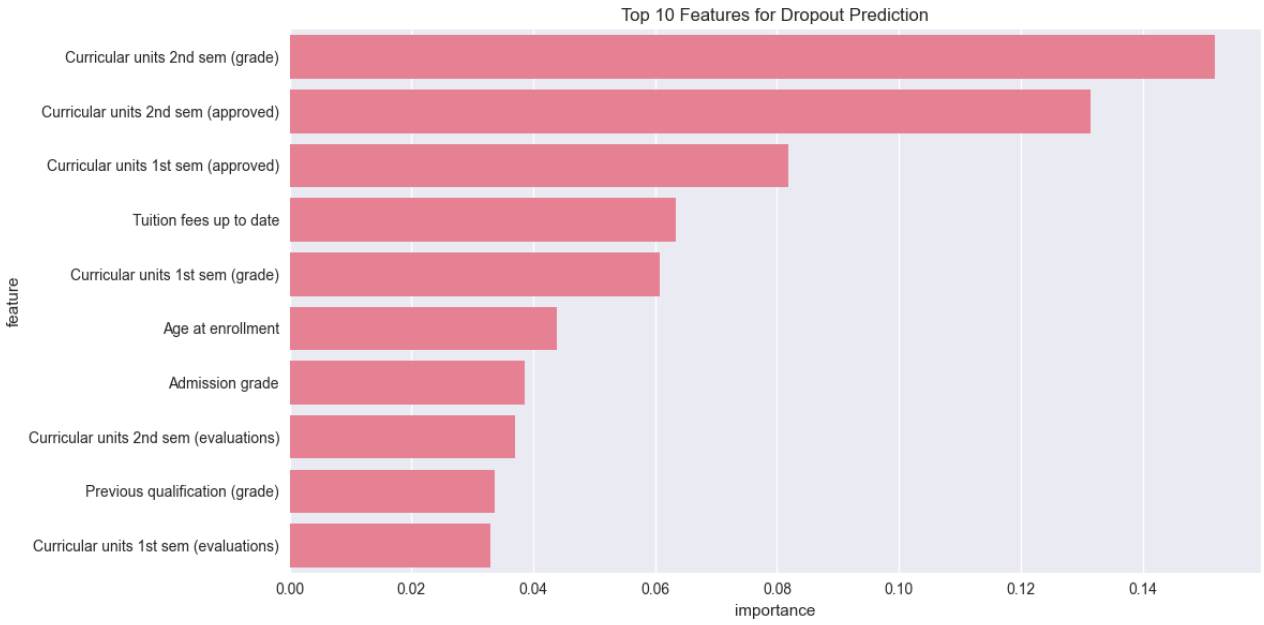
3.8 Melhores Features para predição de evasão

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: Top 10 Features for Dropout Prediction



Fonte: Autoral (2025)

Conclusão

A implementação deste sistema de mineração de dados permitiu uma compreensão profunda dos fatores que influenciam o desempenho acadêmico e a evasão estudantil na universidade. Os resultados obtidos fornecem bases sólidas para decisões administrativas e pedagógicas, possibilitando o desenvolvimento de estratégias preventivas e intervenções direcionadas.

Os principais achados revelam que o sucesso acadêmico está intimamente relacionado a aspectos como desempenho nos primeiros semestres, suporte financeiro e regularidade no pagamento das mensalidades. Fatores demográficos como idade e gênero também apresentam correlações significativas com as taxas de evasão.

Entre os preditores mais relevantes para a evasão, destacam-se: notas do primeiro e segundo semestres, unidades curriculares aprovadas, status das mensalidades e condições socioeconômicas dos estudantes. Essas descobertas sugerem a necessidade de implementar sistemas de monitoramento precoce, programas de apoio financeiro e intervenções especializadas para grupos de estudantes em situação de vulnerabilidade acadêmica.

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.

Referências

BREIMAN, L. **Random Forests**. Machine Learning, v.45, p.5-32, 2001.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd Edition. Morgan Kaufmann, 2011.

MARTINS, M.V.; TOLLEDO, D.; MACHADO, J.; BAPTISTA, L.M.T.; REALINHO, V. Early prediction of student's performance in higher education: a case study. **Trends and Applications in Information Systems and Technologies**, vol.1, in Advances in Intelligent Systems and Computing series. Springer, 2021. DOI: 10.1007/978-3-030-72657-7_16

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.