

ANÁLISE PREDITIVA DE DESEMPENHO ACADÊMICO: UM ESTUDO SOBRE EVASÃO ESTUDANTIL



SUMÁRIO

- Introdução
- Objetivo
- Materiais e Métodos
- Resultados
- Discussão
- Conclusão

INTRODUÇÃO

- A evasão escolar, como apontado por **Torres Marques et al. (2022)**, representa um desafio significativo que exerce um impacto adverso sobre as instituições de ensino, acarretando consequências nas esferas social, acadêmica, econômica e ambiental. Esses impactos negativos, por sua vez, repercutem nas políticas de investimento e desenvolvimento educacional.
- A pesquisa de **Hasan, Rabby, Islam e Hossain (2019)** destaca a importância dos algoritmos de Machine Learning na esfera educacional, visando a previsão e otimização do desempenho dos estudantes.

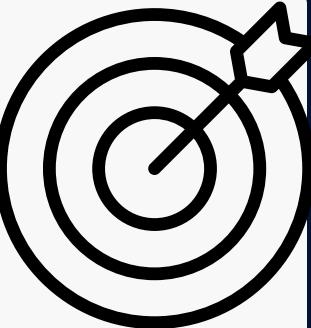
.

“—

| “As taxas de conclusão dos graduandos nos Estados Unidos dentro de seis anos após a matrícula são de apenas 62,3% em 2022.”

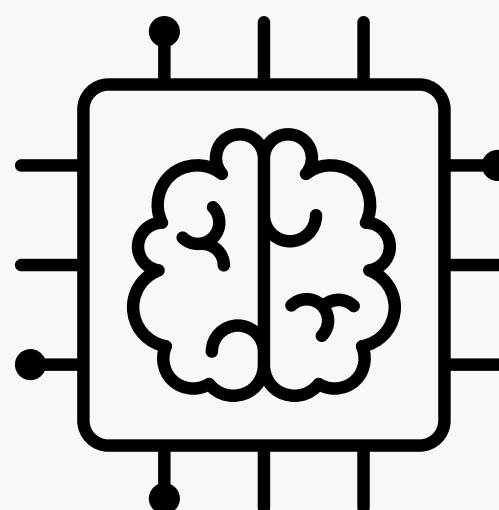
– Hanneh Bareham e Chelsea Wing (bankrate.com)





OBJETIVO

- O objetivo é criar um modelo viável que utilize técnicas de aprendizagem para identificar os estudantes que podem estar em risco de abandonar o ensino superior depois de se matricularem. Ao identificar estes estudantes numa fase inicial, podemos implementar estratégias para os apoiar e reduzir a probabilidade de abandonarem o ensino superior.



MATERIAIS E MÉTODOS

FERRAMENTAS

- **Linguagem de programação Python**
- **numpy:**
 - uma biblioteca fundamental que traz o poder computacional de linguagens como C para Python
- **matplotlib:**
 - uma biblioteca de visualização abrangente
- **seaborn:**
 - uma biblioteca de visualização de dados baseada na matplotlib
- **pandas:**
 - uma biblioteca para limpeza e manipulação dos dados
- **sklearn:**
 - uma biblioteca de aprendizagem automática para processamento e modelação de dados.

DESCRIÇÃO DOS DADOS



- Para esse projeto foi utilizado o conjunto de dados da UC Irvine, cujo nome é: “**Predict Students’ Dropout and Academic Success**”
- Os dados desse DataSet incluem informações que são conhecidas desde o momento em que um aluno se inscreve até ao segundo semestre de inscrição, incluindo:
 - **percurso académico (academic path)**
 - **dados demográficos (demographics)**
 - **factores sócio-económicos (social-economic factors)**
- O problema é formulado como uma tarefa de classificação em três categorias, em que cada estudante é classificado como evadido, matriculado ou graduado ao final do período normal do curso.

Dropout

Enrolled

Graduate

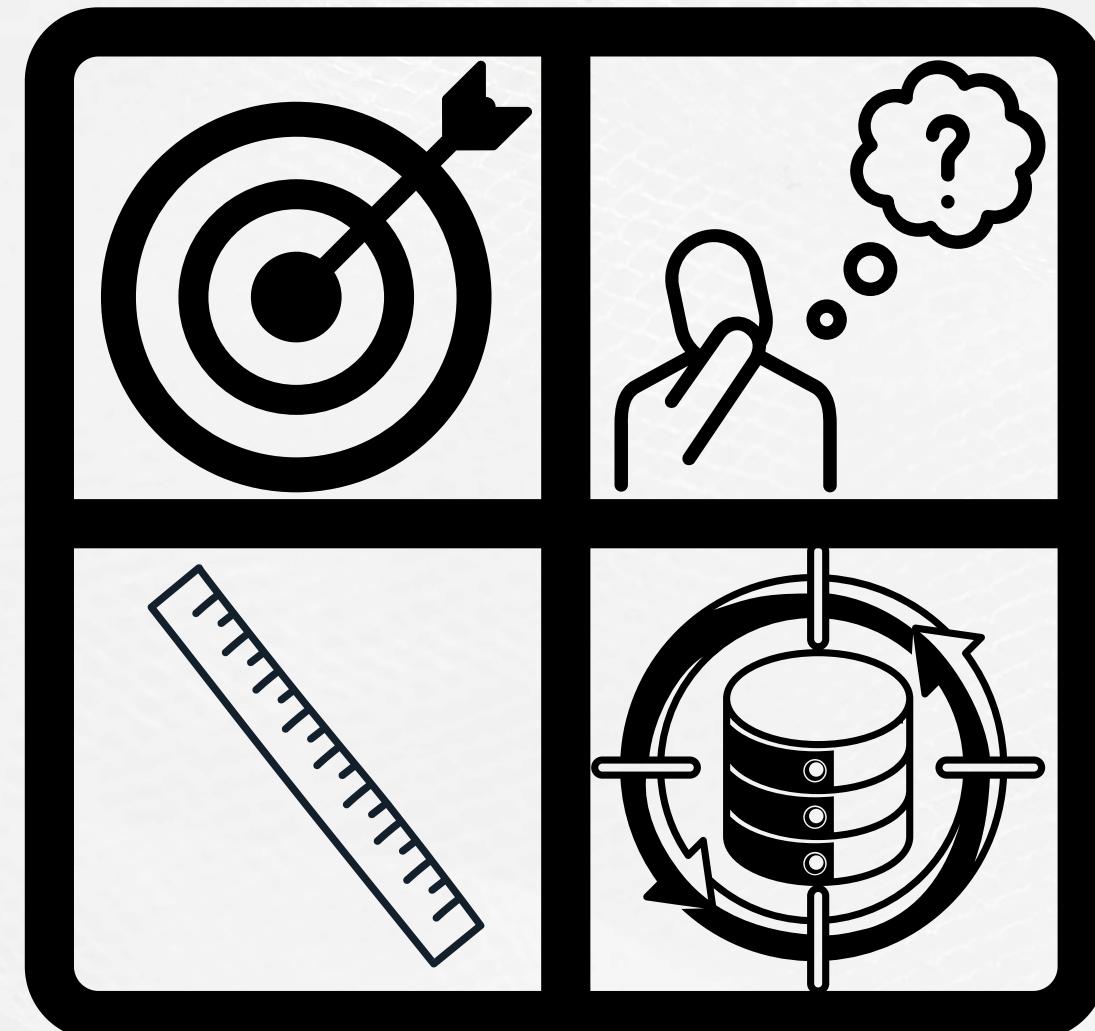
PARÂMETROS DE AVALIAÇÃO

1. Acurácia

Proporção de rótulos corretamente previstos em relação ao total de amostras no conjunto de teste.

2. Precisão

Mede a exatidão do modelo ao prever verdadeiros positivos como uma proporção de todos os positivos.



3. Recall

Mede a capacidade do modelo de identificar corretamente todas as instâncias positivas.

4. F1 - Score

Média harmônica entre precisão e recall, usada quando queremos considerar tanto falsos positivos quanto falsos negativos.

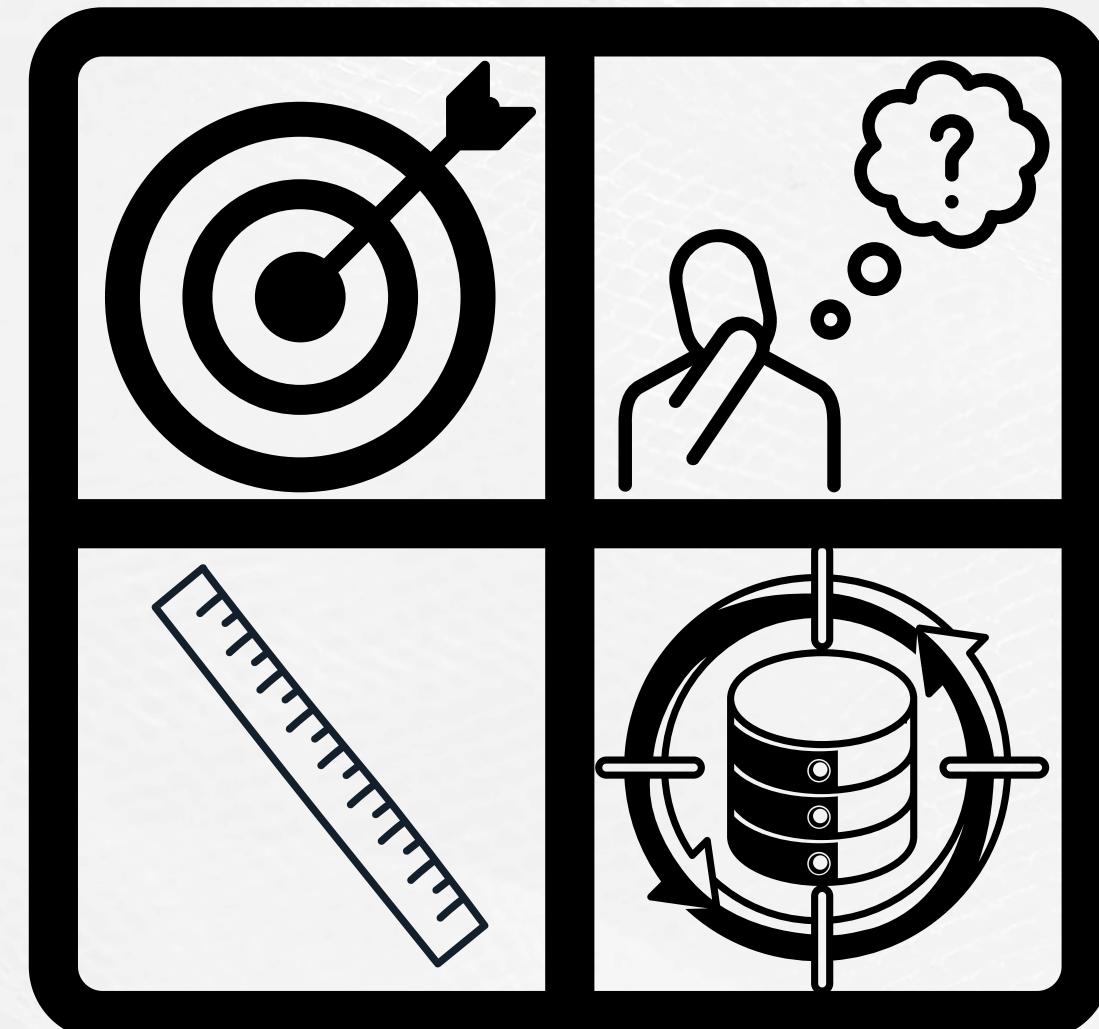
PRIORIDADE

1. Acurácia

proporção de rótulos corretamente previstos em relação ao total de amostras no conjunto de teste.

2. Precisão

mede a exatidão do modelo ao prever verdadeiros positivos como uma proporção de todos os positivos.



3. Recall

mede a capacidade do modelo de identificar corretamente todas as instâncias positivas.

4. F1 - Score

a média harmônica entre precisão e recall, usada quando queremos considerar tanto falsos positivos quanto falsos negativos.

POR QUE RECALL?

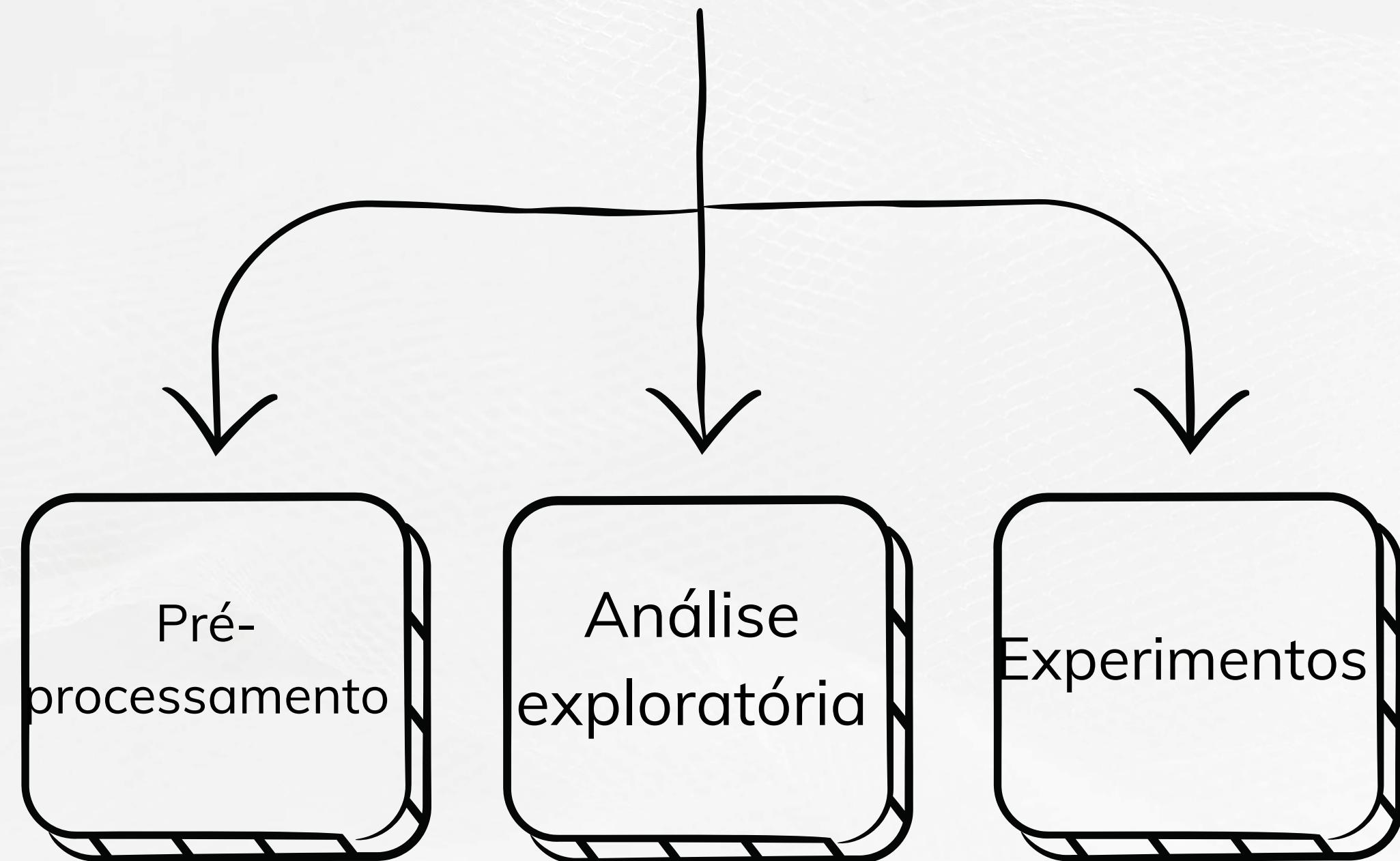
Ao priorizar recall, buscamos reduzir o número de falsos negativos e aumentar a capacidade do modelo de identificar corretamente os estudantes em risco de fracasso acadêmico.

Minimizar o número de falsos negativos ajudará a garantir que menos estudantes que precisam de assistência educacional passem despercebidos e não sejam devidamente identificados e apoiados.



$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

ETAPAS



ETAPAS

Dados são carregados a partir do arquivo CSV e passa por um processo rigoroso de validação e limpeza.

Pré-
processamento

Análise
exploratória

Modelagem
preditiva

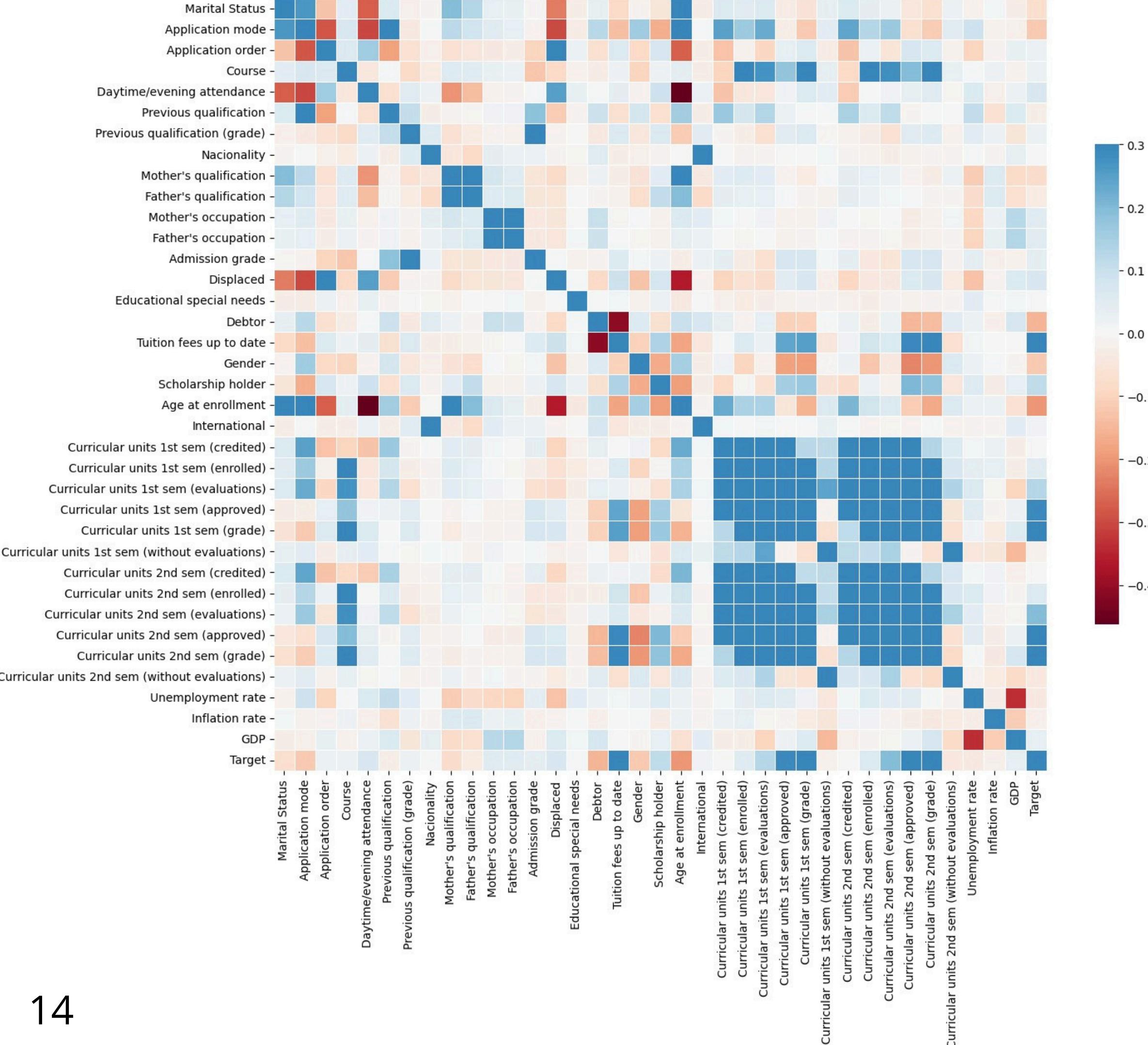
ETAPAS

São geradas estatísticas descritivas, visualizações e análises de correlação que permitem compreender as características da população estudantil e identificar padrões relacionados à evasão.

Pré-
processamento

Análise
exploratória

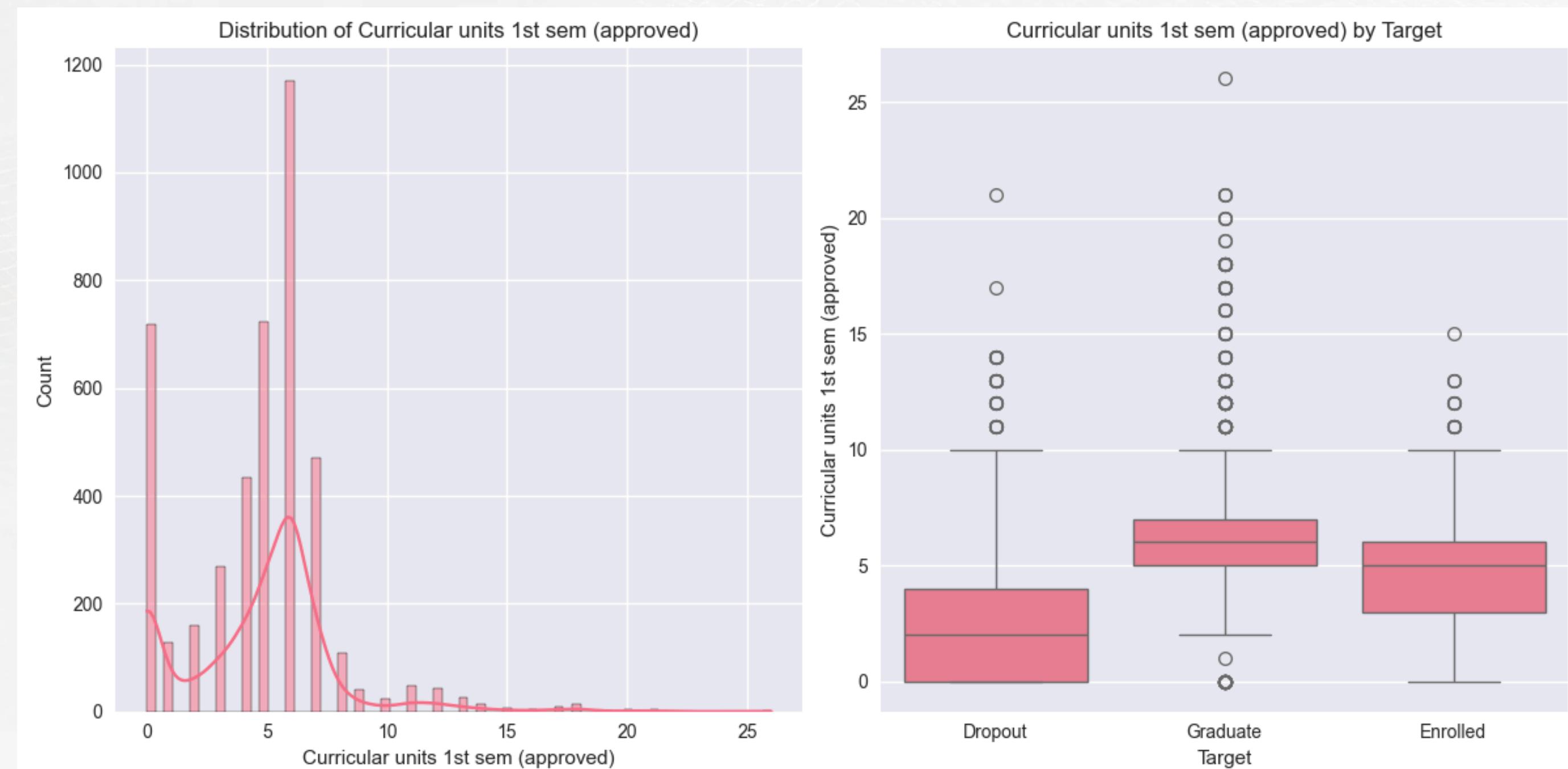
Modelagem
preditiva



Matriz de correlação entre features

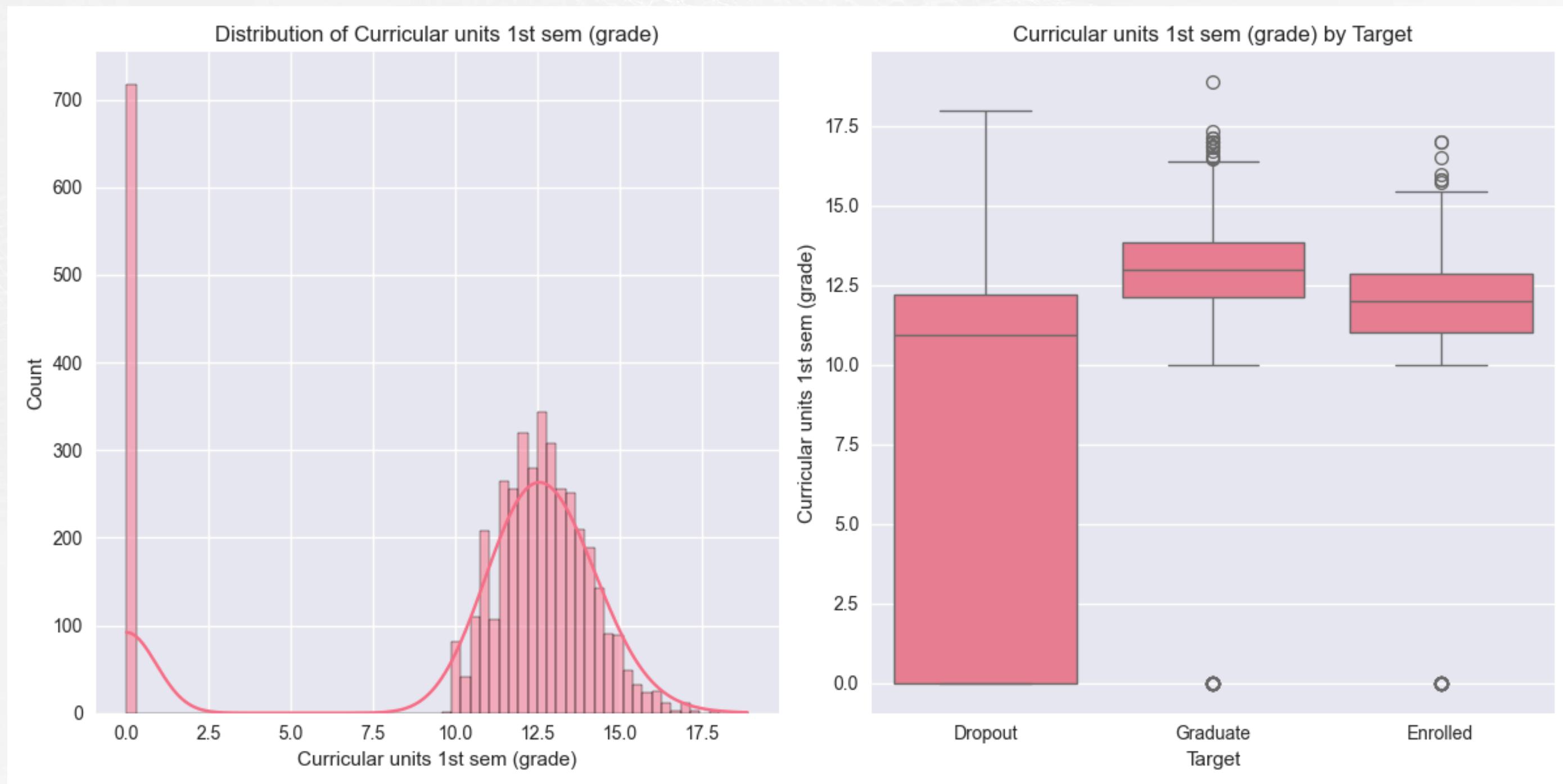
Unidades curriculares aprovadas no primeiro semestre

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.



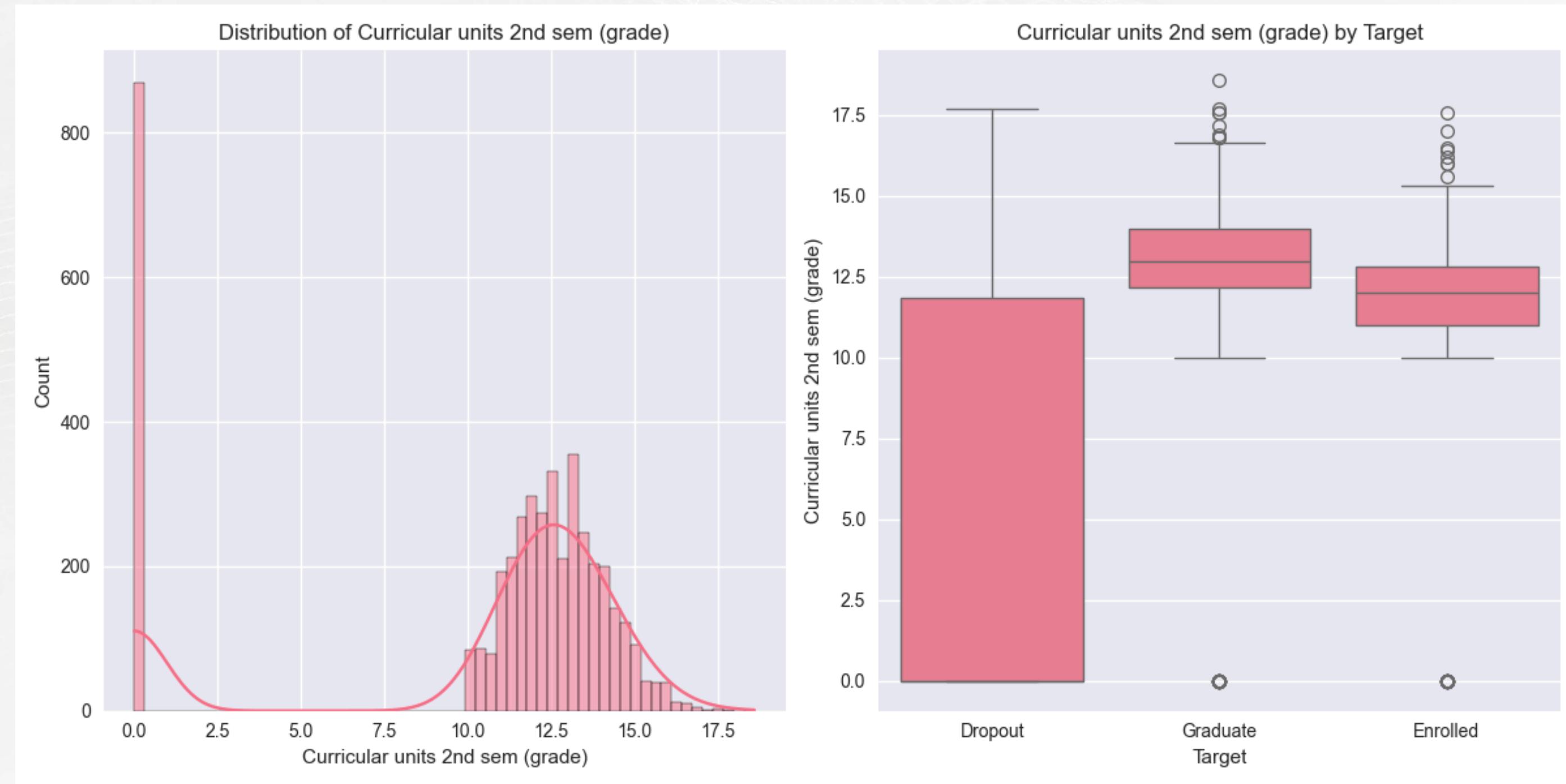
Notas do primeiro semestre

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.



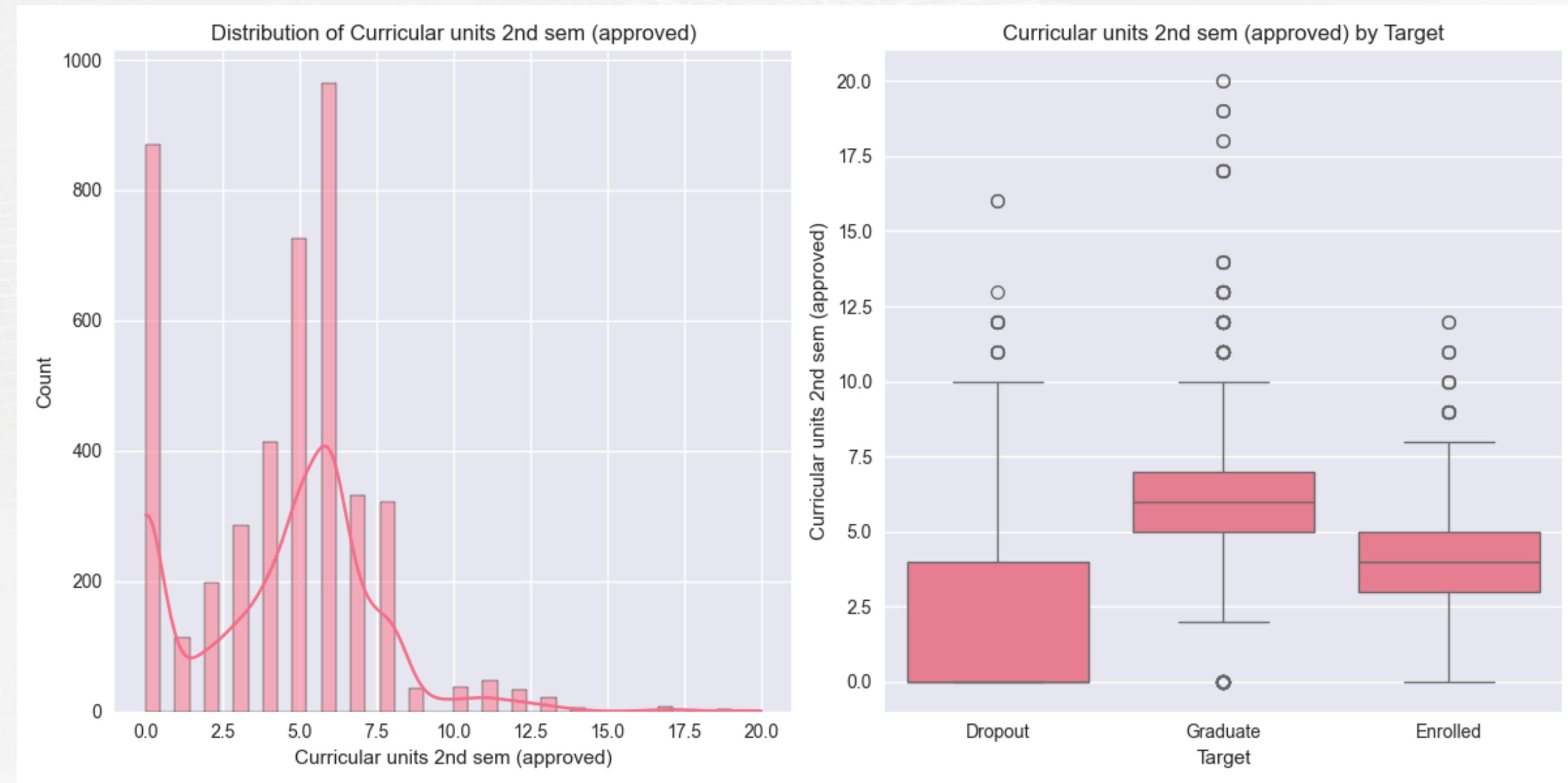
Notas do segundo semestre

As notas apresentam uma distribuição bimodal similar a do primeiro semestre, mas com uma ligeira melhora nas médias

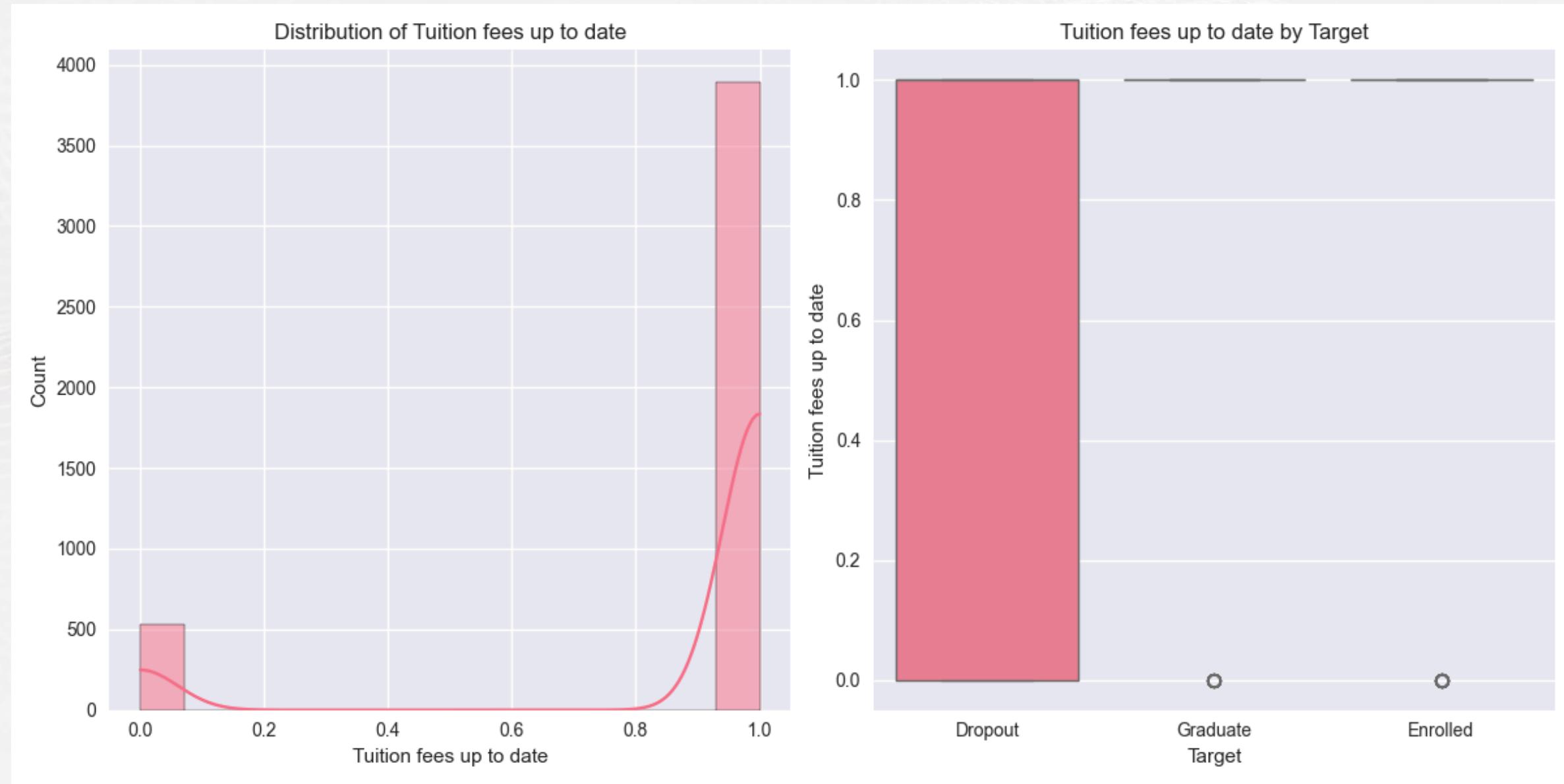


Unidades curriculares aprovadas no segundo semestre

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

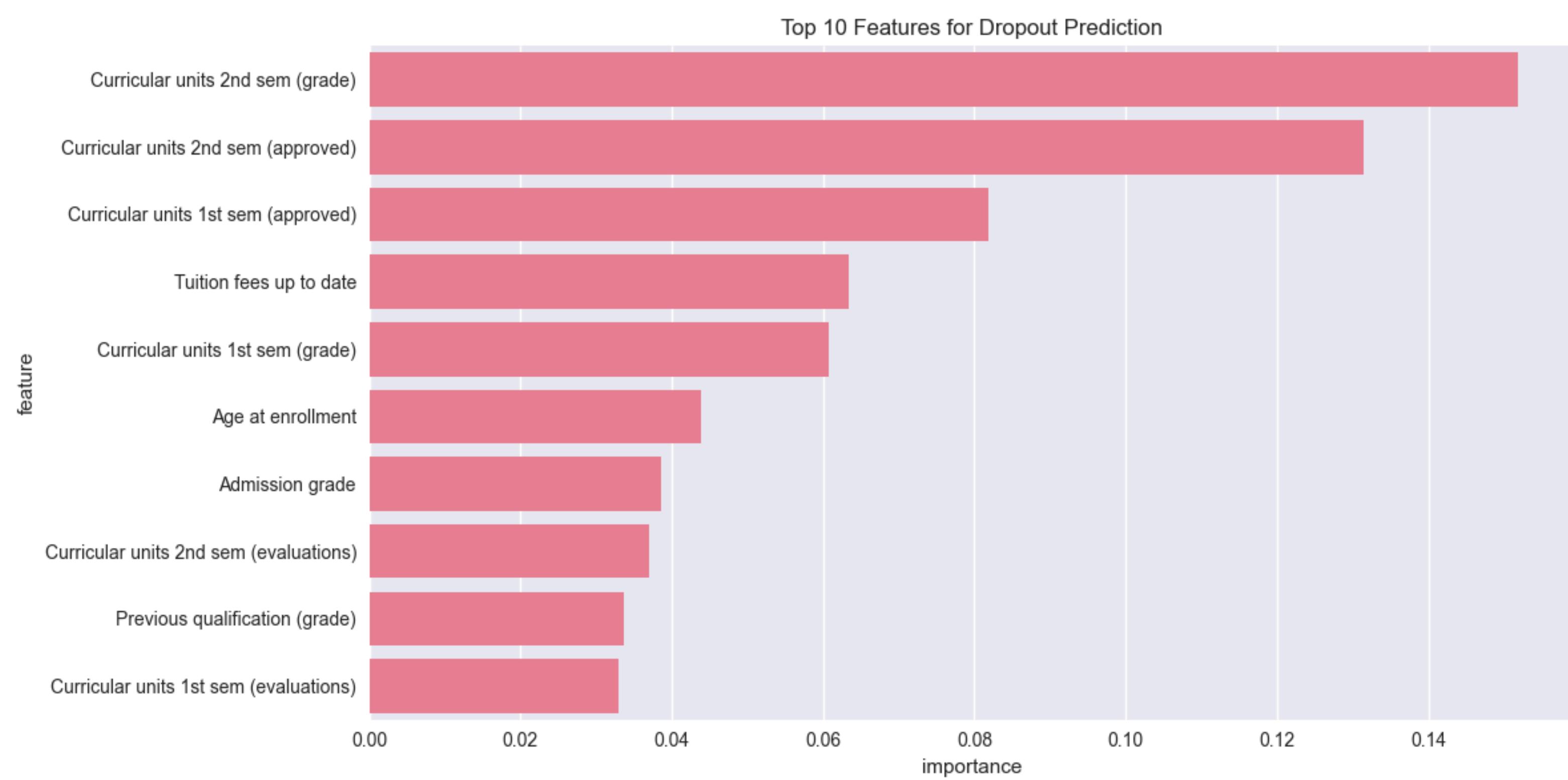


Mensalidades em dia



- **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

Resultados



Resultados

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

ETAPAS

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.

Pré-
processamento

Análise
exploratória

Modelagem
preditiva

MODELO DE MACHINE LEARNING

RANDOM FOREST



PRECISÃO

Combina múltiplas árvores de decisão para obter previsões mais estáveis e confiáveis. Diferente de modelos simples, essa abordagem reduz o overfitting, garantindo que o modelo generalize bem para novos dados.

INTERPRETAÇÃO

Permite identificar fatores que mais influenciam a evasão ou a permanência dos alunos. Além de prever os resultados, o modelo também auxilia na formulação de estratégias para reduzir a evasão.

RESISTÊNCIA

Como cada árvore no modelo aprende com diferentes subconjuntos dos dados, o impacto de inconsistências individuais é reduzido, tornando as previsões mais confiáveis.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC

random_forest	0.8687	0.8377	0.7501	0.7909	0.9158
gradient_boosting	0.8689	0.8609	0.7230	0.7848	0.9089
adaboost	0.8554	0.8219	0.7201	0.7671	0.9048

Dataset:

Total students: **4424**

Number of features analyzed: **42**

Numeric features: **42**

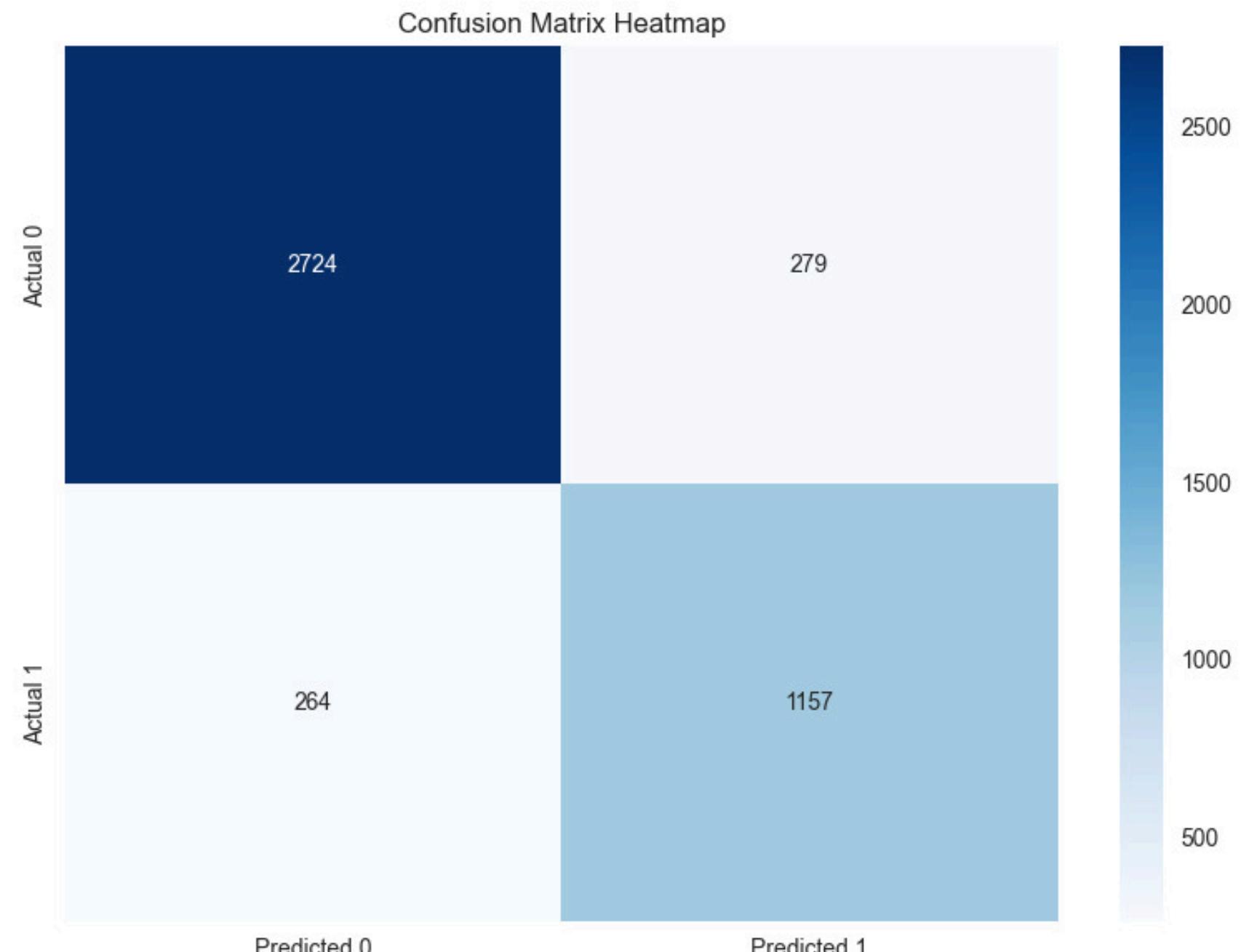
Categorical features: **0**

Principais Achados:

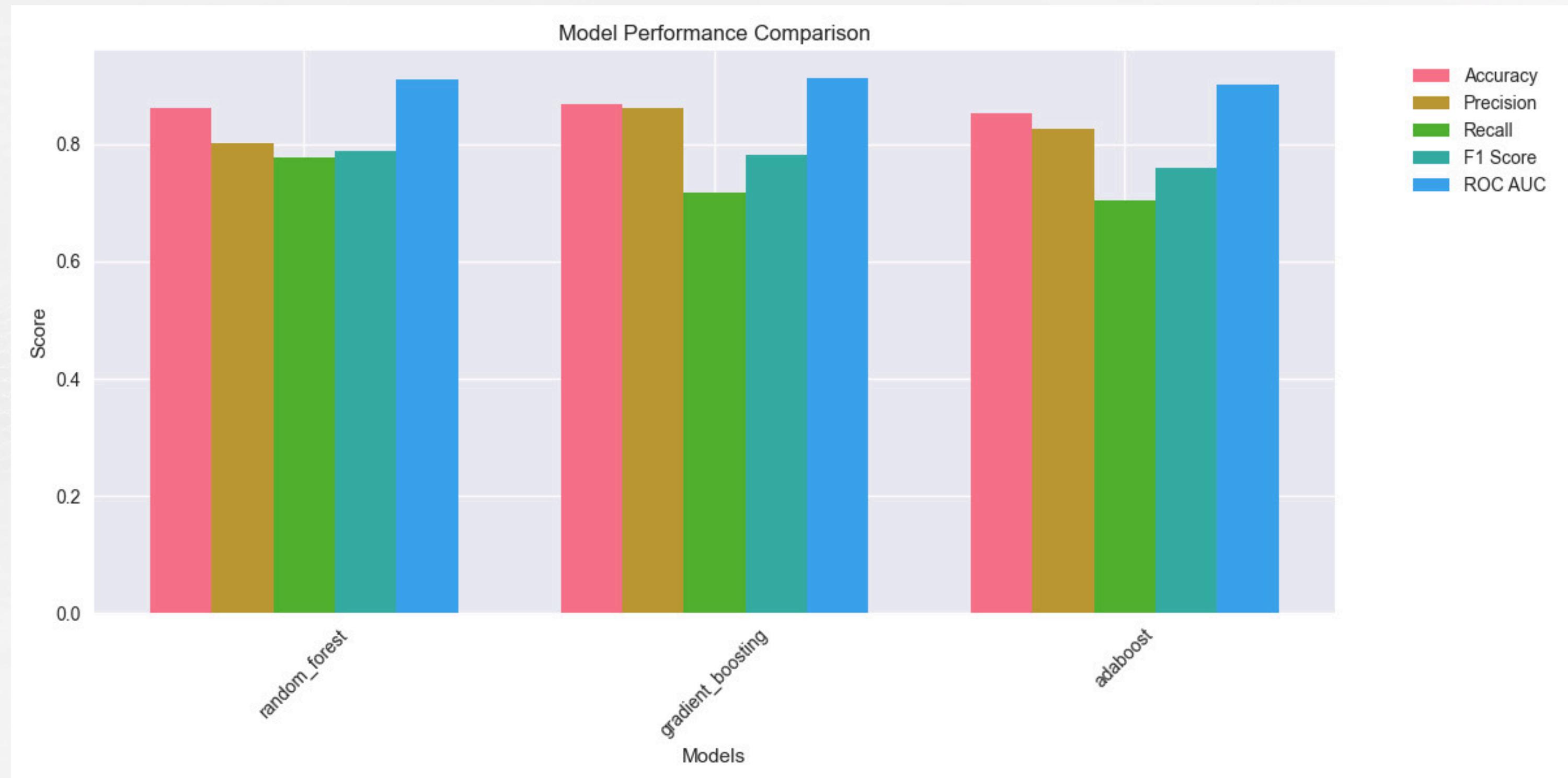
Overall dropout rate: **32.1%**

Dropout rates by semester:

Semester_1: **4.1%**



Comparação de performance do modelo



Os princípios achados revelam que o sucesso acadêmico está intimamente relacionado a aspectos com desempenho nos primeiros semestre, suporte financeiro e regularidade no pagamento das mensalidades.

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.

Discussão

Conclusão

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.

RECONHECIMENTOS E DIREITOS AUTORAIS

@autor: [DENILSON DA SILVA ALVES, TIAGO DE LIMA BATISTA, DANIEL NUNES DUARTE] @data última versão: [02/02/2025]
@versão: 1.0 @Agradecimentos: Universidade Federal do Maranhão (UFMA), Professor Doutor Thales Levi Azevedo Valente, e colegas de curso. Copyright/License Este material é resultado de um trabalho acadêmico para a disciplina “MINERAÇÃO DE DADOS E APLICAÇÕES NA ENGENHARIA”, sob a orientação do professor Dr. THALES LEVI AZEVEDO VALENTE, semestre letivo 2024.2, curso Engenharia da Computação, na Universidade Federal do Maranhão (UFMA). Todo o material sob esta licença é software livre: pode ser usado para fins acadêmicos e comerciais sem nenhum custo. Não há papelada, nem royalties, nem restrições de “copyleft” do tipo GNU. Ele é licenciado sob os termos da Licença MIT, conforme descrito abaixo, e, portanto, é compatível com a GPL e também se qualifica como software de código aberto. É de domínio público. Os detalhes legais estão abaixo. O espírito desta licença é que você é livre para usar este material para qualquer finalidade, sem nenhum custo. O único requisito é que, se você usá-lo, nos dê crédito.

Este projeto está licenciado sob os termos da [MIT License](#). Esta Permissão é concedida, gratuitamente, a qualquer pessoa que obtenha uma cópia deste software e dos arquivos de documentação associados (o “Software”), para lidar no Software sem restrição, incluindo sem limitação os direitos de usar, copiar, modificar, mesclar, publicar, distribuir, sublicenciar e/ou vender cópias do Software, e permitir pessoas a quem o Software é fornecido a fazê-lo, sujeito às seguintes condições:

REFERÊNCIAS

M.V. Martins, D. Tolledo, J. Machado, L. M.T. Baptista, V. Realinho. (2021) “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. DOI: [10.1007/978-3-030-72657-7_16](https://doi.org/10.1007/978-3-030-72657-7_16)