Previsão de Evasão Acadêmica no Ensino Superior: Aplicação de Modelos Preditivos para Identificação de Riscos de Abandono

Nome dos autores: Guilherme Fontana, Henrique Laste Cansi, Lucas Klug Arndt

Resumo

Um dos desafios mais significativos enfrentados pelo ensino superior atualmente, especialmente nos cursos da área de Ciências Exatas, é a alta taxa de evasão acadêmica. Dados indicam que cerca de 50% dos estudantes matriculados em cursos de Exatas não conseguem concluir suas formações. Essa realidade não apenas compromete a sustentabilidade e a continuidade dos próprios cursos, mas também acarreta impactos negativos em uma escala mais ampla. A falta de profissionais qualificados nessas áreas essenciais reflete diretamente no mercado de trabalho, prejudicando o desenvolvimento econômico, científico e tecnológico do país e do mundo. Diante desse cenário preocupante, este estudo se propõe a abordar a questão da evasão no ensino superior, com foco específico na previsão de comportamentos e fatores que levam os alunos a abandonarem suas graduações.

Palavras-chave

Evasão acadêmica, inteligência artificial, classificação, análise de dados.

1. Introdução

A evasão no ensino superior é um dos desafios mais prementes enfrentados por instituições educacionais em todo o mundo. Este fenômeno, particularmente notável em cursos de Ciências Exatas e da Terra, afeta cerca de 50% dos estudantes matriculados, conforme apontam estudos recentes [1, 2]. Além de impactar a sustentabilidade desses cursos, a evasão compromete a formação de profissionais essenciais para o desenvolvimento científico e tecnológico. A redução desse índice é, portanto, uma questão de interesse não apenas acadêmico, mas também social e econômico.

Compreender as razões que levam os estudantes a abandonarem seus cursos é fundamental para propor estratégias eficazes de mitigação. A literatura sugere que fatores como dificuldades financeiras, lacunas na preparação acadêmica, falta de suporte institucional e desafios pessoais desempenham papéis críticos no processo de evasão [1]. Contudo, apesar dos avanços em estudos sobre o tema, ainda há lacunas significativas no entendimento sobre como esses fatores interagem e como prever, de forma eficiente, os casos de evasão.

Este trabalho busca contribuir para o campo de estudo da evasão no ensino superior ao explorar uma abordagem baseada em dados para identificar os principais fatores associados a esse problema. Com o uso de técnicas de análise de dados e modelos preditivos, o objetivo central não é apenas fornecer insights que subsidiem a criação de políticas educacionais mais eficazes, mas também prever quais alunos possuem perfil de evasão. Essa capacidade preditiva visa capacitar a coordenação acadêmica a tomar decisões mais assertivas e proativas, possibilitando intervenções personalizadas que ajudem a mitigar o risco de abandono. Essa abordagem inovadora reforça a relevância do tema no contexto acadêmico e oferece caminhos concretos para enfrentar o problema de maneira orientada por evidências e focada em resultados.

2. Material e Métodos

O trabalho foi conduzido em um ambiente de notebook Python, estruturado em quatro etapas principais: análise exploratória dos dados, tratamento e limpeza dos dados, análise das variáveis (features) e da variável alvo (target), e, finalmente, o treinamento e avaliação dos modelos preditivos. Para o desenvolvimento dessa análise, foram utilizadas diversas bibliotecas, como o Pandas para manipulação e análise de dados, NumPy para operações numéricas, Seaborn e Matplotlib para visualização de dados, Scikit-learn para a implementação dos modelos de machine learning e avaliação de desempenho, e Imbalanced-learn para lidar com desbalanceamento nas classes dos dados por meio de técnicas como SMOTE.

2.1 Análise da Base de Dados

A base de dados conta com 11807 amostras, contendo 28 classes.

A tabela a seguir apresenta as classes, com seus respectivos tipos de dados e número de variação de valores.

Classes	Tipo de Dado	Variações
ANO_MATRICULA	Inteiro	2
PERIODO_MATRICULA	Inteiro	2
ANO_PROJETADO	Inteiro	1
PERIODO_PROJETADO	Inteiro	2
SIGLA	Texto	9
CADASTRO	Texto	7324
ALUNO_NOVO	Inteiro	2
ANO_INGRESSO	Inteiro	13
DT_AGENDAMENTO	Inteiro	2
DT_ORIENTACAO	Inteiro	8
DT_SOLICITACAO	Inteiro	2
DT_MATRICULA	Inteiro	2
DEBITO	Inteiro	3
SEMESTRE_ALUNO	Inteiro	31
NRO_DISC_MATRICULADAS	Inteiro	19
NRO_DISC_CANCELADAS	Inteiro	7
CANCELAMENTO_TOTAL	Inteiro	2
CH_MATRICULA	Inteiro	97
PERC_CH_CURSADA	Ponto Flutuante	1424
FALTAS_HA	Inteiro	189
DISC_REPROVADAS	Inteiro	10
NRO_NEGOCIACAO	Inteiro	14
MEDIA_10	Ponto Flutuante	4863
FINANCIAMENTO	Inteiro	2
TIPO_ESCOLA	Texto	5
PRESENCIAL	Inteiro	2
TIPO_MATRICULA	Texto	7
POSSIVEL_FORMANDO	Inteiro	2

Tabela 1. Classes, tipos de dados e número de variações de valores no conjunto de dados

Nota-se que os dados não variam muito, onde 19 das 28 classes apresentam uma variação inferior a 10.

Os gráficos a seguir mostram a comparação, da variação de valores entre as classes, e a distribuição dos valores nelas.

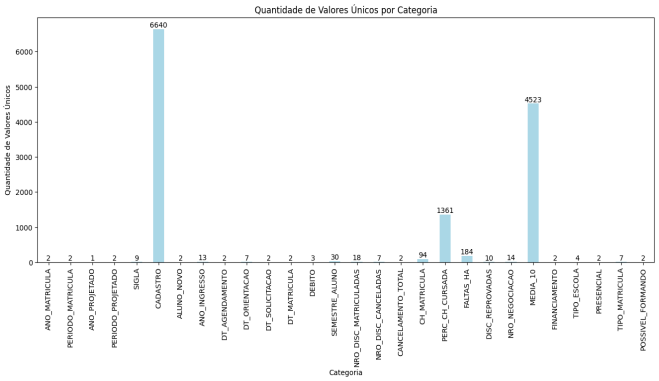

Fig. 1: Gráfico de barras da variação de dados entre classes.

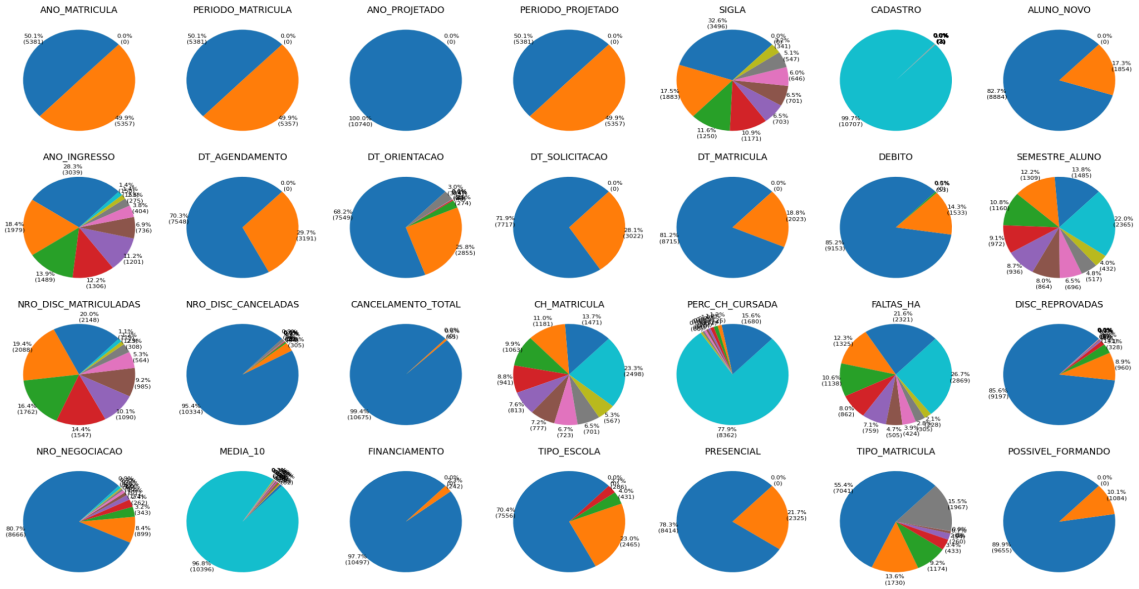

Fig. 2: Gráficos de pizza com a distribuição dos dados, dentro das classes.

2.2 Análise Exploratória dos Dados

A análise exploratória inicial permitiu uma visão geral do dataset, com destaque para as colunas, tipos de dados e a distribuição das variáveis [4]. Utilizando gráficos de distribuição, foram analisadas as variáveis numéricas como ANO_MATRICULA, PERIODO_MATRICULA, ANO_PROJETADO, PERIODO_PROJETADO, ALUNO_NOVO, ANO_INGRESSO, DT_AGENDAMENTO, DT_ORIENTACAO, DT_SOLICITACAO, DT_MATRICULA, DEBITO, SEMESTRE_ALUNO, NRO_DISC_MATRICULADAS, NRO_DISC_CANCELADAS, CANCELAMENTO_TOTAL, CH_MATRICULA, FALTAS_HA, DISC_REPROVADAS, NRO_NEGOCIACAO, FINANCIAMENTO, PRESENCIAL e POSSIVEL_FORMANDO, permitindo observar a frequência de cada valor. Além disso, as variáveis categóricas, como SIGLA, TIPO_ESCOLA e TIPO_MATRICULA, foram analisadas para identificar quais valores predominam no conjunto de dados. A distribuição da variável alvo, CANCELAMENTO_TOTAL, que representa o cancelamento total da matrícula (evasão), revelou um desbalanceamento significativo: 11.717 casos com o valor zero (0 – sem cancelamento) e apenas 90 casos com o valor um (1 – cancelamento).

2.3 Tratamento e Limpeza dos Dados

A etapa de tratamento e limpeza dos dados envolveu várias ações. Primeiramente, foram descartadas colunas irrelevantes, como ANO_MATRICULA, ANO_PROJETADO, PERIODO_MATRICULA, PERIODO_PROJETADO, CADASTRO e ANO_INGRESSO, pois não agregam valor para o objetivo da análise. Para facilitar a interpretação do código, comentários explicativos foram adicionados em todas as células do notebook. Em seguida, foram feitas conversões de dados para garantir que todos os valores estivessem no formato numérico, substituindo as strings quando necessário.

Os valores nulos foram tratados utilizando o *SimpleImputer* [4]. Para a variável MEDIA_10, foi preenchida com a média dos valores presentes, enquanto para a variável TIPO_ESCOLA, foi utilizada a moda (valor mais frequente) para substituir os dados ausentes. Após essa etapa, a normalização das variáveis foi realizada. Para as variáveis com intervalo definido, como PERC_CH_CURSADA e MEDIA_10 (que variam entre 0-100 e 0-10, respectivamente), foi utilizado o *MinMaxScaler*, que normaliza os dados dentro de um intervalo entre 0 e 1. Para as variáveis sem intervalo definido, utilizou-se o *StandardScaler* para garantir que todas as variáveis estivessem na mesma escala [4].

2.4 Processamento das Variáveis Categóricas Treinamento e avaliação de modelos

Para adequar os dados aos modelos de classificação, foi necessário transformar variáveis categóricas em formato numérico, agrupando categorias com características similares e representando-as em colunas binárias. Para isso, foi utilizada a correlação de Pearson como técnica para identificar relações entre categorias que representavam conceitos semelhantes ou exerciam influência similar sobre a evasão. Essa análise permitiu consolidar variáveis de maneira mais informativa, reduzindo redundâncias e destacando os grupos mais relevantes.

A variável TIPO_MATRICULA, por exemplo, apresentava originalmente sete categorias distintas: *matricula-parcelamento-fixo*, *matricula-parcelamento-fixo-personalizado*, *mensalidade-fixa*, *matricula-parcelamento-por-disciplina*, *credito*, *matricula-dias-aula-personalizado* e *matricula-dias-aula*. Essas categorias foram agrupadas de acordo com o tipo de pagamento associado. As categorias *matricula-parcelamento-fixo*, *matricula-parcelamento-fixo-personalizado* e *mensalidade-fixa* foram classificadas como pagamento fixo, representadas pelo valor 1 na nova variável binária criada, denominada PAGAMENTO_FIXO. Por outro lado, as categorias *matricula-parcelamento-por-disciplina*, *credito, matricula-dias-aula-personalizado* e *matricula-dias-aula* foram classificadas como pagamento variável, representadas pelo valor 0.

No caso da variável TIPO_ESCOLA, que abrangia categorias como *Particular Regular*, *Pública Regular*, *Particular Supletivo* e *Pública Supletivo*, foram criadas duas variáveis binárias para destacar as categorias de maior relevância. A primeira, ESCOLA_PARTICULAR_REGULAR, foi utilizada para identificar alunos oriundos de escolas particulares regulares, assumindo o valor 1 para esses casos e 0 para todos os outros. A segunda, ESCOLA_SUPLETIVO, foi criada para identificar alunos vindos de escolas supletivas, sejam elas públicas ou particulares, sendo representada pelo valor 1 para essas categorias e 0 para os demais tipos de escola.

Essas transformações foram realizadas com o objetivo de destacar padrões relevantes para o problema de estudo, permitindo que os modelos de classificação capturassem relações significativas entre as características dos alunos e a probabilidade de evasão. A abordagem binária simplifica a representação dos dados, ao mesmo tempo em que preserva as informações essenciais para a análise.

A variável SIGLA, que inicialmente estava representada em valores textuais, foi transformada utilizando o *OneHotEncoder* [4]. Esse processo gerou uma coluna binária para cada campus, permitindo que a variável categórica fosse representada numericamente de forma adequada para os modelos de classificação. Contudo, ao analisar a correlação de Pearson entre as colunas criadas e a variável alvo CANCELAMENTO_TOTAL, foi identificado que a correlação era muito baixa, indicando que SIGLA não exercia influência significativa sobre o cancelamento de matrícula. Com base nessa análise, a variável foi removida do modelo, uma vez que sua inclusão não contribuía para melhorar o desempenho preditivo.

2.5 Análise das Features e do Target

Após o tratamento e transformação das variáveis, foi realizada uma análise da correlação entre as variáveis e a variável alvo (CANCELAMENTO_TOTAL). Utilizando a correlação de Pearson [5], foi possível identificar quais variáveis estavam mais fortemente associadas ao target. Por exemplo, a variável SIGLA apresentou baixa correlação com o cancelamento total, o que indicou que essa variável poderia ser removida do modelo.

Além disso, foi realizada uma análise de correlação entre as variáveis independentes para identificar redundâncias e melhorar a eficiência do modelo. Variáveis com alta correlação entre si foram avaliadas, e apenas as mais relevantes foram mantidas. Por exemplo, as variáveis DT_ORIENTACAO e DT_AGENDAMENTO apresentaram uma correlação de 0,91, indicando que forneciam informações muito semelhantes; nesse caso, optou-se por manter DT_AGENDAMENTO. Situação semelhante ocorreu com as variáveis PERC_CH_CURSADA, SEMESTRE_ALUNO e ALUNO_NOVO, todas altamente correlacionadas, onde SEMESTRE_ALUNO foi mantida por ser a mais representativa. Outro caso envolveu DEBITO e NRO_NEGOCIACAO, sendo DEBITO a variável escolhida.

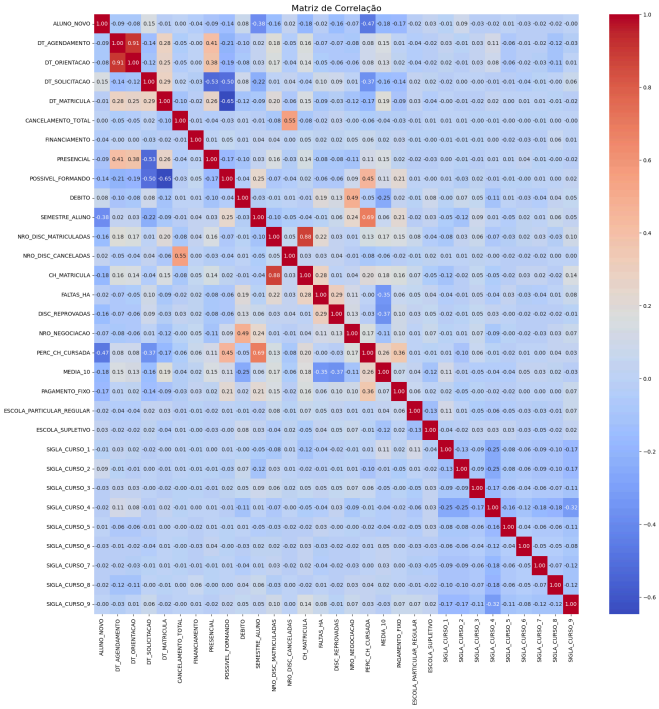

Fig. 3. Matriz de confusão da correlação de Pearson

3. Resultados

3.1 Treinamento do Modelo

O modelo foi treinado utilizando três técnicas diferentes: *Logistic Regression*, *Random Forest* e *Gradient Boosting*. Para isso, o conjunto de dados foi dividido seguindo a metodologia *Hold Out* [7], com 70% dos dados destinados ao treinamento e 30% para os testes. Contudo, um desafio importante enfrentado foi o desbalanceamento extremo da variável-alvo, que originalmente apresentava 11.717 ocorrências para o valor 0 (não cancelado) e apenas 90 ocorrências para o valor 1 (cancelado). Esse desbalanceamento acentuado poderia comprometer o desempenho dos modelos, levando-os a favorecer as previsões para a classe majoritária.

Para mitigar esse problema, foi aplicada a técnica SMOTE (Synthetic Minority Oversampling Technique) [3], que gera exemplos sintéticos para a classe minoritária, equilibrando o conjunto de dados. Após o balanceamento, o conjunto de treino resultante contou com 8.201 instâncias e o conjunto de teste com 3.606 instâncias, permitindo que os modelos tivessem uma representação mais equitativa das duas classes. Essa abordagem foi essencial para garantir que os modelos fossem capazes de identificar padrões associados tanto à evasão quanto à permanência, sem viés para a classe majoritária.

3.2 Avaliação dos Modelos

Indicador	Logistic Regression	Random Forest	Gradient Boosting
Precision (0)	1.00	1.00	1.00
Recall (0)	1.00	1.00	1.00
F1-Score (0)	1.00	1.00	1.00
Precision (1)	0.79	0.96	0.87
Recall (1)	1.00	0.89	0.96
F1-Score (1)	0.89	0.92	0.91
Accuracy	1.00	1.00	1.00
AUC	0.9997	0.9996	0.9999

Tab. 2: Classification Report

O modelo baseado em Regressão Logística obteve, para a classe 1, uma precisão de 0.79, um recall de 1 e um f1-score de 0.89 [6]. Para a classe 0, o desempenho foi perfeito, com 100% em todas as métricas. De modo geral, atingiu 0.9, 1 e 0.94 em média, para as três métricas, respectivamente, com uma AUC de 0.9997 [6].

O modelo treinado com Random Forest apresentou, para a classe 1, 0.89 de precisão, 0.811 de recall e um f1-score de 0.88. Para a classe 0, também atingiu 100% em todas as três métricas. Ao final, apresentou, na média, 0.98, 0.91 e 0.94 respectivamente para cada métrica, com uma AUC de 0.9997, igual ao Logistic Regression.

O modelo treinado utilizando Gradient Boosting apresentou para a classe 1 uma precisão de 0.87, recall de 0.96 e f1-score de 0.91. Para a classe 0, o desempenho foi impecável, com 100% em todas as métricas. No desempenho

geral, as médias ficaram em 0.93 para precisão,0.98 para recall e 0.96 para f1-score, enquanto a AUC atingiu 0.9999.

Quando geramos as matrizes de confusão das técnicas, a Regressão Logística resultou em 7 erros, todos falsos positivos. As outras duas técnicas apresentaram 1 falso negativo, cada, porém a Random Forest obteve 5 falsos positivos, enquanto a Gradient Boosting 4.

3.3 Análise das Métricas

Todos os modelos apresentam desempenho quase perfeito, com *precision*, *recall* e *F1-score* de 1.00 ou muito próximos.

O *Gradient Boosting* apresenta a melhor *macro avg* para *recall* (0.98) e *F1-score* (0.96), destacando um desempenho equilibrado entre as classes.

O *Gradient Boosting* tem menos falsos negativos que o *Random Forest* e menos falsos positivos que o *Logistic Regression*.

O *Gradient Boosting* destaca-se como o modelo mais adequado pelos seguintes motivos:

1. Equilíbrio superior entre precisão e recall: Apresenta excelente desempenho na identificação da classe minoritária (classe 1), reduzindo erros sem comprometer a precisão.
2. Maior AUC: Com um valor de 0.9999, demonstra a melhor capacidade de discriminação entre as classes, mesmo em cenários desbalanceados.
3. Menor taxa de erros críticos: Gera apenas 1 falso negativo, significativamente melhor que o *Random Forest*, e reduz falsos positivos em comparação ao *Logistic Regression*.

Dado que a prioridade é minimizar falsos negativos para a classe minoritária, o modelo *Gradient Boosting* é a escolha mais apropriada para este problema, equilibrando eficácia e confiabilidade.

4. Conclusões

Este estudo abordou um dos desafios mais críticos do ensino superior, a evasão acadêmica, com foco em cursos da área de Ciências Exatas. Utilizando técnicas de análise de dados e modelos preditivos, foi possível identificar padrões e fatores que influenciam o abandono de curso. Os modelos de classificação, incluindo Regressão Logística, Random Forest e Gradient Boosting, mostraram-se eficazes, com todos atingindo desempenho quase perfeito na previsão da evasão. Contudo, o Gradient Boosting destacou-se como a melhor escolha devido ao seu equilíbrio superior entre precisão e recall, menor taxa de falsos negativos e maior AUC, indicando uma excelente capacidade de discriminação, mesmo em cenários de desbalanceamento extremo.

Essa abordagem preditiva, além de fornecer insights valiosos para a implementação de políticas educacionais mais eficazes, oferece uma base sólida para ações proativas na redução da evasão. A identificação precoce de alunos em risco de evasão permitirá que instituições acadêmicas realizem intervenções mais personalizadas e direcionadas, contribuindo para o aumento das taxas de conclusão e formando profissionais mais qualificados para o mercado de trabalho. A aplicabilidade de modelos preditivos no contexto educacional abre novas possibilidades para a melhoria contínua da qualidade do ensino e para a promoção da equidade no acesso e permanência dos estudantes no ensino superior.

5. Bibliografia

[1] Almeida, P. F., & Souza, R. T. (2021). Causas da evasão em cursos de ciências exatas: uma revisão da produção acadêmica. Lume - Repositório Institucional da UFRGS. Disponível em: https://lume.ufrgs.br/bitstream/handle/10183/263019/001165072.pdf?sequence=1.

[2] Oliveira, A. L., & Almeida, P. F. (2020). Indicadores de evasão acadêmica no curso de Licenciatura em Matemática: números que apontam vulnerabilidades para permanência e êxito no Ensino Superior. Cocar - Revista Eletrônica de Ciências Sociais Aplicadas. Disponível em: https://periodicos.uepa.br/index.php/cocar/article/view/6334.

[3] Elor, G. & Averbuch-Elor, R. (2022). To SMOTE, or not to SMOTE. [Não disponível para acesso direto].

[4] Corso, L. (2024). Materiais da disciplina de Computação Aplicada I. Preparação de Dados. Universidade de Caxias do Sul.

[5] Corso, L. (2024). Materiais da disciplina de Computação Aplicada I. Preparação de Dados - Parte 2. Universidade de Caxias do Sul.

[6] Corso, L. (2024). Materiais da disciplina de Computação Aplicada I. Medidas de Desempenho em Inteligência Artificial. Universidade de Caxias do Sul.

[7] Corso, L. (2024). Materiais da disciplina de Computação Aplicada I. Métodos de Avaliação de Desempenho em Inteligência Artificial. Universidade de Caxias do Sul.