Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: dd/Mmm/yyyy; 1st round notif.: dd/Mmm/yyyy; New version: dd/Mmm/yyyy; 2nd round notif.: dd/Mmm/yyyy;
Camera ready: dd/Mmm/yyyy; Edition review: dd/Mmm/yyyy; Available online: dd/Mmm/yyyy; Published: dd/Mmm/yyyy;

Análise Preditiva da Permanência Estudantil: Um Estudo Comparativo de Fatores em Instituições de Ensino Superior Públicas e Privadas no Brasil

Title: Predictive Analysis of Student Retention: A Comparative Study of Factors in Public and Private Higher Education Institutions in Brazil

Arthur Yang Tung
Universidade de São Paulo
arthur.tung@usp.brdg

Kevin Tamayose
Universidade de São Paulo
kevin.tamayose@usp.br

Kawe Gomes
Universidade de São Paulo
kawe.gomes@usp.br

Filipe Valeriano Batista Oliveira
Universidade de São Paulo
fvaleriano@usp.br

Resumo

A permanência estudantil no ensino superior é um desafio complexo e multifatorial no Brasil, com dinâmicas distintas entre instituições públicas e privadas. Este plano de trabalho propõe uma investigação quantitativa para identificar e comparar os principais fatores correlacionados ao tempo de permanência dos estudantes. Utilizando técnicas de Mineração de Dados Educacionais (MDE) sobre microdados do Censo da Educação Superior (CES) e do Exame Nacional do Ensino Médio (ENEM), o estudo buscará construir e avaliar modelos preditivos de classificação. O objetivo é determinar quais variáveis demográficas, socioeconômicas e de desempenho acadêmico prévio são mais influentes para prever a permanência ou evasão. Espera-se que os resultados forneçam insights valiosos para a formulação de políticas institucionais mais eficazes, visando a redução das taxas de evasão e a promoção do sucesso acadêmico em ambos os setores do ensino superior brasileiro.
Palavras-chave: *Permanência Estudantil; Evasão Universitária; Mineração de Dados Educacionais; Análise Preditiva; Ensino Superior.*

Abstract

Student retention in higher education is a complex and multifactorial challenge in Brazil, with distinct dynamics between public and private institutions. This work plan proposes a quantitative investigation to identify and compare the main factors correlated with students' length of stay. Using Educational Data Mining (EDM) techniques on microdata from the Higher Education Census (CES) and the National High School Exam (ENEM), the study will seek to build and evaluate predictive classification models. The objective is to determine which demographic, socioeconomic, and prior academic performance variables are most influential in predicting retention or dropout. The results are expected to provide valuable insights for the formulation of more effective institutional policies aimed at reducing dropout rates and promoting academic success in both sectors of Brazilian higher education.
Keywords: *Student Retention; University Dropout; Educational Data Mining; Predictive Analysis; Higher Education.*

Cite as: Tung, A. Y., Tamayose, K., Gomes, K., & Oliveira, F. V. B. (Ano). Título do artigo. Revista Brasileira de Informática na Educação, vol, pp-pp. https://doi.org/10.5753/rbie.yyyy.id.

1 Introdução

A evasão no ensino superior representa um desafio persistente e de grande impacto para o sistema educacional brasileiro, acarretando custos significativos tanto para o Estado quanto para os indivíduos. As dinâmicas que influenciam a permanência dos estudantes, no entanto, variam consideravelmente entre as Instituições de Ensino Superior (IES) públicas e privadas. Nas IES públicas, a gratuidade do ensino é contrabalançada por fatores como o alto custo de oportunidade e a maior heterogeneidade do corpo discente. Já nas IES privadas, o encargo financeiro das mensalidades emerge como uma variável crítica. A ausência de uma compreensão aprofundada e, principalmente, comparativa sobre os fatores determinantes para a permanência em cada um desses contextos dificulta a formulação de políticas de apoio que sejam eficazes e direcionadas. Diante desse cenário, o objetivo desta pesquisa é aplicar técnicas de Mineração de Dados Educacionais (MDE) para construir e avaliar modelos preditivos que identifiquem os fatores mais relevantes associados ao tempo de permanência de estudantes de graduação, realizando uma análise comparativa entre os setores público e privado. A relevância deste estudo reside no seu potencial de subsidiar gestores educacionais com evidências empíricas para o desenvolvimento de programas de retenção mais eficientes, a otimização na alocação de recursos e a personalização do suporte oferecido a estudantes em potencial risco de evasão.

2 Fundamentos Teóricos

Esta pesquisa fundamenta-se em duas áreas principais: os estudos sobre evasão e permanência no ensino superior e a Mineração de Dados Educacionais (MDE). Primeiramente, abordam-se os conceitos de evasão, definida como o abandono definitivo do curso, e permanência, entendida como a trajetória bem-sucedida do estudante até a conclusão. Modelos teóricos consolidados, como o de Tinto (1975), fornecem uma estrutura para categorizar os múltiplos fatores que influenciam essa trajetória, dividindo-os em dimensões individuais (demográficas), socioeconômicas, de integração acadêmica e institucionais. Serão exploradas as particularidades do contexto brasileiro, contrastando os desafios específicos enfrentados por estudantes de IES públicas e privadas. Em segundo lugar, define-se a MDE como a aplicação de métodos computacionais para descobrir padrões e extrair conhecimento de grandes volumes de dados educacionais (Han et al., 2011). O foco será em tarefas de classificação, cujo objetivo é prever uma categoria discreta (e.g., "evadiu"vs. "permaneceu"), e na análise de importância de atributos (*feature importance*), que permite hierarquizar a influência de cada variável no modelo. Técnicas como Árvores de Decisão e Random Forest são particularmente adequadas para este problema, devido à sua capacidade de lidar com dados complexos e fornecer resultados interpretáveis.

3 Trabalhos Relacionados

A literatura sobre evasão no ensino superior brasileiro é vasta, com muitos estudos utilizando abordagens estatísticas tradicionais para identificar fatores de risco em contextos específicos. Pesquisas frequentemente apontam para a relevância de variáveis como renda familiar, escolaridade

dos pais e desempenho no ensino médio. Paralelamente, as bases de dados do INEP, como o ENEM e o Censo da Educação Superior, têm sido amplamente validadas como fontes ricas para análises educacionais em larga escala, embora nem sempre de forma integrada. No campo da MDE, estudos nacionais e internacionais têm demonstrado a eficácia de algoritmos de Aprendizagem de Máquina para prever o desempenho acadêmico e o risco de evasão, como evidenciado por Rodrigues e Gouveia (2021). Contudo, identifica-se uma lacuna na literatura: a carência de estudos que realizem uma análise preditiva e *comparativa* dos fatores de permanência entre IES públicas e privadas, utilizando um conjunto de dados nacional e integrado. Este trabalho visa preencher essa lacuna, oferecendo uma visão sistêmica e contrastiva que pode informar políticas públicas e institucionais de forma mais abrangente.

4 Método

A metodologia desta pesquisa será quantitativa e seguirá um processo estruturado de Mineração de Dados Educacionais, inspirado no modelo KDD (Knowledge Discovery in Databases), similar ao adotado por Souza e Santos (2021). O processo é dividido em quatro etapas principais: (1) Coleta e Seleção dos Dados, (2) Pré-processamento e Transformação, (3) Modelagem e Mineração, e (4) Avaliação e Interpretação dos Resultados.

4.1 Coleta e Seleção dos Dados

A base de dados para este estudo será construída a partir da integração de duas fontes de dados públicos de grande escala, disponibilizadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP):

- **Censo da Educação Superior (CES):** Esta será a fonte primária para a variável alvo. O CES coleta anualmente informações detalhadas sobre instituições, cursos, docentes e, crucialmente, sobre os alunos. Serão utilizados os microdados do aluno para rastrear a trajetória acadêmica, identificando o ano de ingresso e a situação de vínculo em anos subsequentes (cursando, formado, trancado, falecido ou desvinculado). A partir disso, será possível construir a variável dependente (alvo) da nossa predição.
- **Exame Nacional do Ensino Médio (ENEM):** Esta base de dados fornecerá o conjunto de variáveis preditoras (atributos). Os microdados do ENEM contêm não apenas as notas dos estudantes nas diferentes áreas do conhecimento, mas também um rico questionário socioeconômico, que abrange desde a renda familiar e escolaridade dos pais até hábitos culturais e acesso a bens.

A integração entre as bases será realizada utilizando um identificador único do estudante (ID do aluno), que permite o cruzamento seguro e anônimo dos dados. Será selecionada uma coorte de estudantes ingressantes em um ano específico (e.g., 2017) e sua situação será acompanhada nos dados do CES dos anos seguintes (e.g., até 2022) para determinar o status de permanência. Adicionalmente, como fonte de dados complementar para contextualização socioeconômica regional, a **Pesquisa Nacional por Amostra de Domicílios (PNAD Contínua)** do IBGE poderá ser consultada, embora não permita um cruzamento em nível individual.

4.2 Pré-processamento e Transformação

Esta etapa é fundamental para garantir a qualidade dos dados e prepará-los para a modelagem. As seguintes subtarefas serão executadas:

- **Limpeza de Dados:** Tratamento de valores ausentes (missing values) através de técnicas de imputação (média/mediana para variáveis numéricas, moda para categóricas) ou remoção de registros, caso a quantidade de dados faltantes em uma variável seja excessiva.
- **Seleção e Engenharia de Atributos:** Será realizada uma seleção criteriosa de atributos das bases integradas, agrupados em categorias: (a) *Demográficos* (idade, gênero, cor/raça); (b) *Socioeconômicos* (renda familiar, escolaridade dos pais, tipo de escola no ensino médio); (c) *Desempenho Acadêmico Prévio* (notas do ENEM por competência e nota da redação); e (d) *Institucionais* (tipo de IES – pública/privada, modalidade do curso, área de conhecimento). A variável alvo será definida categoricamente, como, por exemplo: 'PERMANECEU' (aluno ainda vinculado ou formado no tempo esperado) e 'EVADIU' (aluno com status "desvinculado").
- **Transformação de Dados:** As variáveis categóricas nominais (ex: região do país) serão convertidas em formato numérico através da técnica de *One-Hot Encoding*. As variáveis numéricas contínuas (ex: notas do ENEM) serão normalizadas (escalonadas para um intervalo entre 0 e 1) para garantir que algoritmos sensíveis à escala, como SVM ou redes neurais, funcionem corretamente.

O conjunto de dados final será dividido em dois subconjuntos: um para treinamento dos modelos (tipicamente 80% dos dados) e outro para teste (20% restantes), garantindo que a avaliação seja feita em dados não vistos durante o treinamento.

4.3 Modelagem e Mineração

Nesta fase, serão aplicados diferentes algoritmos de Aprendizagem de Máquina supervisionada para a tarefa de classificação. A escolha dos algoritmos visa comparar modelos de diferentes naturezas (lineares, baseados em árvores, ensembles) para identificar o de melhor desempenho e interpretabilidade. Os algoritmos a serem explorados incluem:

- **Regressão Logística:** Como um modelo de base (baseline) para comparação.
- **Árvores de Decisão (CART):** Pela sua alta interpretabilidade, permitindo visualizar as regras de decisão.
- **Random Forest:** Um método de ensemble que combina múltiplas árvores de decisão para melhorar a acurácia e controlar o sobreajuste (overfitting).
- **Gradient Boosting Machines (XGBoost/LightGBM):** Outro método de ensemble, conhecido por seu alto desempenho preditivo em competições de ciência de dados.

Os modelos serão treinados separadamente para os dados de IES públicas e privadas, permitindo uma análise comparativa direta dos resultados e da importância dos atributos em cada contexto.

4.4 Avaliação e Interpretação dos Resultados

A avaliação do desempenho dos modelos será realizada no conjunto de teste. Serão utilizadas as seguintes métricas de classificação:

- **Acurácia:** Percentual geral de predições corretas.
- **Matriz de Confusão:** Para analisar os tipos de erros (falsos positivos e falsos negativos).
- **Precisão, Recall e F1-Score:** Métricas importantes, especialmente se as classes (evadiu/-permaneceu) forem desbalanceadas.
- **Curva ROC e AUC:** Para avaliar a capacidade do modelo de discriminar entre as classes.

Além da performance preditiva, será realizada uma análise da **importância dos atributos** (feature importance), extraída principalmente dos modelos baseados em árvores (Random Forest, Gradient Boosting). Esta análise revelará quais fatores têm maior poder preditivo para a permanência estudantil, permitindo responder à questão central da pesquisa e comparar os resultados entre o setor público e privado.

5 Resultados Esperados e Discussão

Espera-se que os modelos de classificação alcancem uma acurácia satisfatória, permitindo a predição da permanência estudantil com um grau de confiança útil para intervenções pedagógicas. O principal resultado será a identificação e o ranqueamento dos atributos mais preditivos para a evasão em cada setor (público e privado). A hipótese inicial é que, nas IES públicas, fatores relacionados ao capital cultural e ao custo de oportunidade (como a necessidade de trabalhar) terão maior peso, enquanto nas IES privadas, variáveis diretamente ligadas à condição financeira (renda familiar, financiamento estudantil) serão mais proeminentes. A discussão dos resultados irá contrastar esses achados com a literatura existente, explorando as implicações das diferenças encontradas. Por exemplo, se o desempenho no ENEM for um forte preditor em ambos dos setores, isso reforça a importância de políticas de nivelamento acadêmico. Se a escolaridade dos pais for mais relevante no setor público, isso pode indicar a necessidade de programas de mentoria e acolhimento específicos para estudantes de primeira geração. A análise permitirá, assim, a formulação de recomendações de políticas públicas e institucionais segmentadas e baseadas em evidências.

6 Conclusão

Este trabalho se propõe a realizar uma análise comparativa e preditiva dos fatores que influenciam a permanência de estudantes no ensino superior brasileiro, distinguindo entre IES públicas e privadas. A principal contribuição da pesquisa será a geração de um modelo empírico, baseado em dados nacionais, que não apenas identifica os perfis de estudantes com maior risco de evasão, mas também quantifica a importância relativa dos fatores de risco em cada contexto. Como limitação,

o estudo reconhece que os dados administrativos não capturam variáveis psicossociais (motivação, saúde mental) que também influenciam a permanência. No entanto, os resultados fornecerão um panorama robusto que pode guiar futuras investigações qualitativas. Como trabalhos futuros, sugere-se a aplicação de modelos mais complexos, como redes neurais profundas, e a incorporação de dados longitudinais mais granulares, caso se tornem disponíveis, para modelar a trajetória do estudante ao longo do tempo.

Agradecimentos

Esta seção será preenchida na versão final do artigo para agradecer às agências de fomento, instituições e indivíduos que contribuíram para a realização da pesquisa.

7 Cronograma

Tabela 1: Cronograma de Execução da Pesquisa..

Atividade	**Período**
Revisão Bibliográfica e Coleta de Dados	20/08 a 16/09
Pré-processamento e Limpeza dos Dados	17/09 a 07/10
Modelagem e Avaliação dos Modelos	08/10 a 28/10
Análise dos Resultados e Escrita do Artigo	29/10 a 10/11
Revisão Final e Submissão	11/11 a 17/11

Referências

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd). Morgan Kaufmann.

Rodrigues, E. M., & Gouveia, R. M. (2021). Técnicas de machine learning para predição do tempo de permanência na graduação no âmbito do ensino superior público brasileiro. *Congresso sobre Tecnologias na Educação (Ctrl+e)*.

Souza, V. F. d., & Santos, T. C. B. d. (2021). Processo de Mineração de Dados Educacionais aplicado na Previsão do Desempenho de Alunos: Uma comparação entre as Técnicas de Aprendizagem de Máquina e Aprendizagem Profunda. *Revista Brasileira de Informática na Educação - RBIE*, *29*, 519–546. https://doi.org/10.5753/RBIE.2021.29.0.519

Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, *45*(1), 89–125. https://doi.org/10.3102/00346543045001089