Análise Preditiva da Evasão em Cursos de Licenciatura em Letras no Brasil: Uma Abordagem de Aprendizado de Máquina Focada em Características Institucionais

Fernando Toba Lopez – 14804486
Universidade de São Paulo

Henrique Naoki Teruya – 14578324
Universidade de São Paulo

Karen Miyuki Massuda – 14585868
Universidade de São Paulo

Vivian Ayumi Miamura – 14835140
Universidade de São Paulo

Resumo
A alta taxa de evasão em cursos de ensino superior, especialmente em licenciaturas, representa um desafio significativo para o sistema educacional brasileiro. Este projeto de pesquisa propõe a criação de um modelo preditivo, utilizando técnicas de Machine Learning, para identificar os fatores institucionais que mais influenciam a evasão nos cursos de licenciatura em Letras. Utilizando os microdados do Censo da Educação Superior do INEP (2021-2023), o estudo focará em variáveis das instituições e dos cursos, como categoria administrativa, qualificação docente e modalidade de ensino. O resultado esperado é um modelo de regressão capaz de prever a taxa de evasão, oferecendo insights valiosos para gestores educacionais e para a formulação de políticas de permanência estudantil. O desenvolvimento será realizado na linguagem de programação Python.

Palavras-chave: Evasão; Ensino Superior; Aprendizado de Máquina; Licenciatura em Letras.

Abstract
The high dropout rate in higher education courses, especially in teaching degrees, represents a significant challenge for the Brazilian educational system. This research project proposes the creation of a predictive model, using Machine Learning techniques, to identify the institutional factors that most influence dropout in undergraduate Language and Literature programs. Using Higher Education Census microdata from INEP (2021–2023), the study will focus on variables related to institutions and courses, such as administrative category, faculty qualifications, and teaching modality. The expected result is a regression model capable of predicting dropout rates, providing valuable insights for educational managers and policy formulation on student retention. The development will be carried out in Python.

Keywords: Dropout; Higher Education; Machine Learning; Language and Literature Degrees.

1 Introdução

A evasão no ensino superior é um fenômeno complexo e multifatorial que acarreta perdas sociais, acadêmicas e financeiras. Em cursos de licenciatura, como Letras, que são fundamentais

para a formação de educadores no país, as taxas de abandono são frequentemente preocupantes. Compreender as causas da evasão é essencial para o desenvolvimento de estratégias eficazes de retenção de alunos. Enquanto muitos estudos se concentram no perfil socioeconômico do estudante, esta pesquisa parte da hipótese de que as características da própria instituição de ensino e a estrutura do curso desempenham um papel igualmente crucial.

Diante disso, a questão central que norteia este trabalho é: Quais características institucionais e de curso são os melhores preditores da taxa de evasão em licenciaturas de Letras no Brasil?

O objetivo geral é desenvolver um modelo de Aprendizado de Máquina para prever a taxa de evasão nesses cursos, com base em dados públicos. Para tal, será necessário a realização das seguintes etapas: Coletar e pré-processar os dados do Censo da Educação Superior (INEP) de 2021 a 2023. Identificar as variáveis institucionais com maior impacto na evasão. Treinar e avaliar diferentes algoritmos de regressão para encontrar o modelo com melhor performance preditiva. Gerar conhecimento aplicável para a gestão de políticas de permanência estudantil.

2 Fundamentos Teóricos

A construção de modelos preditivos em educação demanda a compreensão de conceitos fundamentais de Aprendizado de Máquina (Machine Learning), bem como dos algoritmos utilizados no processo de modelagem. Nesta seção, são apresentados os princípios conceituais que sustentam a metodologia adotada neste estudo, com ênfase nos tipos de aprendizado de máquina e nos algoritmos selecionados para a análise da evasão em cursos de Licenciatura em Letras.

Em termos gerais, o aprendizado de máquina pode ser dividido em três paradigmas principais: Aprendizado Supervisionado: ocorre quando o modelo é treinado a partir de dados rotulados, ou seja, pares de entrada e saída conhecidos. O objetivo é aprender uma função que relacione as variáveis explicativas (inputs) com a variável-alvo (output). Este paradigma é o mais adequado ao presente estudo, uma vez que a evasão discente pode ser caracterizada como uma variável de saída observável (permanência ou desligamento).

Aprendizado Não Supervisionado: utilizado quando os dados não apresentam rótulos pré-definidos. Nesse caso, os algoritmos buscam agrupar observações semelhantes ou identificar estruturas latentes nos dados. Exemplos comuns são métodos de clusterização (como K-means e K-modes). Embora não seja o foco deste trabalho, tais técnicas têm sido aplicadas em estudos educacionais para segmentar perfis de estudantes ou cursos.

Aprendizado por Reforço: baseia-se na interação de um agente com um ambiente, em que as decisões tomadas geram recompensas ou penalidades. Esse tipo de aprendizado é aplicado em contextos de tomada de decisão sequencial, não sendo o enfoque desta pesquisa. Dessa forma, o presente trabalho se insere no campo do aprendizado supervisionado, utilizando algoritmos de regressão e classificação para prever a probabilidade de evasão discente a partir de variáveis institucionais e de curso.

Neste estudo, três algoritmos supervisionados foram selecionados para análise comparativa: Regressão Logística, Árvore de Decisão e Random Forest.

A Regressão Logística é um modelo estatístico amplamente utilizado para problemas de classificação binária, nos quais a variável dependente assume apenas dois valores possíveis (por exemplo: evasão ou permanência). O método estima a probabilidade de ocorrência do evento de interesse a partir de uma combinação linear das variáveis independentes, transformada por meio da função logística (sigmóide). Uma das principais vantagens deste algoritmo é a interpretabilidade dos coeficientes, que permitem estimar o efeito marginal de cada variável explicativa sobre a probabilidade da evasão (HOSMER; LEMESHOW, 2000).

As Árvores de Decisão constituem uma abordagem não paramétrica baseada na divisão recur-

siva do espaço de atributos em subconjuntos cada vez mais homogêneos em relação à variável-alvo. O modelo é construído por meio de nós de decisão, que testam condições em atributos, e folhas, que representam a classe prevista. A principal vantagem dessa técnica é a facilidade de interpretação, pois as regras geradas podem ser visualizadas e comunicadas de forma intuitiva, o que favorece sua aplicação em contextos educacionais (BREIMAN et al., 1984).

O Random Forest é um algoritmo de ensemble que combina múltiplas Árvores de Decisão construídas a partir de subconjuntos aleatórios de dados e variáveis. O resultado final é obtido pela agregação das previsões individuais de cada árvore (votação ou média). Essa técnica apresenta, em geral, maior capacidade preditiva e robustez contra sobreajuste em comparação com modelos individuais. Além disso, possibilita a análise da importância das variáveis, permitindo identificar quais atributos institucionais e de curso exercem maior influência sobre a evasão (BREIMAN, 2001).

3 Trabalhos Relacionados

A literatura acadêmica sobre evasão no ensino superior brasileiro é vasta. Estudos como os de Silva Filho et al. (2007) já apontavam para a magnitude do problema e suas múltiplas causas. Mais recentemente, com o avanço da tecnologia, pesquisadores têm aplicado técnicas de mineração de dados para investigar o tema. Por exemplo, Alencar & Netto (2020) utilizaram algoritmos de classificação para prever a evasão com base em dados de desempenho acadêmico e perfil de estudantes em uma universidade específica. Já Passos et al. (2024) apresenta um levantamento empírico sobre evasão e retenção nos cursos de Licenciatura em Letras na Bahia entre 2010 e 2022, discutindo causas locais e tendências temporais. Esse trabalho é diretamente relevante por tratar da mesma área (Letras) e por fornecer insumos empíricos e variáveis contextuais (por exemplo: turnos, locais e efeitos regionais) que podem ser incorporados ou contrastados na análise preditiva proposta neste estudo. A principal contribuição desse tipo de estudo é contextualizar os determinantes regionais e qualitativos da evasão, complementando a abordagem quantitativa preditiva.

No entanto, a maioria desses estudos tende a focar em uma única instituição ou no perfil individual do aluno. São menos comuns as pesquisas que realizam uma análise em larga escala, com dados nacionais, e que se concentram primordialmente nos fatores institucionais como preditores. Este trabalho busca preencher essa lacuna, oferecendo uma perspectiva macro sobre como as características das universidades e cursos em todo o Brasil se correlacionam com as taxas de evasão em Letras.

4 Método

O delineamento metodológico desta pesquisa está estruturado para atingir dois objetivos centrais: o desenvolvimento de um modelo preditivo para a taxa de evasão em cursos de Letras e a identificação dos fatores de maior influência neste fenômeno. Para isso, o trabalho seguirá um processo de Mineração de Dados Educacionais (MDE), fundamentado nos procedimentos descritos por Aggarwal (2015) e adaptado a partir das aplicações práticas de Lima et al. (2021). O processo está organizado em quatro fases: 1) Coleta e Definição dos Dados; 2) Pré-processamento e Transformação; 3) Modelagem Preditiva; e 4) Análise e Avaliação dos Resultados.

4.1 Coleta e Definição dos Dados

A etapa inicial consistirá na obtenção e delimitação do conjunto de dados que servirá como base para todas as análises subsequentes. Os microdados serão obtidos a partir do Censo da Educação Superior, disponibilizados pelo INEP, em arquivos CSV delimitados por ponto e vírgula, referentes aos anos de 2018, 2019, 2022 e 2023. Após o download, será realizada a filtragem para isolar exclusivamente os cursos de graduação da área de Letras, utilizando as variáveis NO_CURSO (Nome do Curso) e CO_CURSO (Código do Curso). Como os microdados não trazem a taxa de evasão de forma direta, esta será inferida a partir da variável de situação de vínculo do estudante, permitindo identificar casos de permanência e de desligamento.

4.2 Pré-processamento dos Dados

Nesta fase, o conjunto de dados será preparado para aplicação dos modelos. Serão realizadas tarefas de limpeza, tratamento de valores ausentes e correção de inconsistências. As variáveis categóricas serão convertidas para formato numérico por meio de técnicas de codificação, enquanto variáveis numéricas poderão ser normalizadas. Além disso, será feita uma seleção criteriosa de atributos, considerando fatores institucionais, características do curso e perfil discente.

4.3 Modelagem Preditiva

Com o conjunto de dados tratado, ele será dividido em duas subamostras: 80

4.4 Análise e Avaliação dos Resultados

A etapa final compreenderá a avaliação de desempenho dos modelos e a interpretação dos resultados. A performance será medida na amostra de teste, utilizando métricas de classificação como Acurácia, Precisão, Revocação, F1-score e AUC-ROC. Os resultados serão comparados para determinar qual algoritmo apresenta maior capacidade preditiva. Adicionalmente, será conduzida uma análise de importância das variáveis, identificando os fatores que mais influenciam a evasão. Por fim, os achados serão discutidos e contextualizados à luz da literatura, de forma a responder aos objetivos propostos.

5 Cronograma

Agosto (21/08 - 31/08)

- Estruturação preliminar do projeto (introdução, justificativa, metodologia).
- Download e exploração inicial dos microdados do INEP.
- Limpeza e organização dos microdados.

Setembro (01/09 a 30/09)

Semanas 1 e 2 (01/09 – 14/09)

- Continuação da limpeza e organização dos microdados. Seleção das variáveis relevantes para evasão. Criação da base consolidada de análise.

Semanas 3 e 4 (15/09 – 30/09)

- Análise exploratória dos dados (estatísticas descritivas, correlações, padrões).
- Escrita mais profunda do capítulo de Fundamentos Teóricos (evasão, regressão logística, machine learning).
- Início da implementação do modelo preditivo.

Outubro (01/10 a 31/10)

Semanas 1 e 2 (01/10 – 14/10)

- Implementação dos modelos preditivos (Regressão Logística, Árvore de Decisão, Random Forest. . .).
- Avaliação inicial dos modelos (AUC, F1, precisão, recall, . . .).

Semanas 3 e 4 (15/10 – 31/10)

- Interpretação dos resultados (importância das variáveis, perfis de risco).
- Escrita da seção de Metodologia e início da seção de Resultados.

Novembro (01/11 a 17/11)

Semana 1 (01/11 – 09/11)

- Finalização da seção de Resultados e Discussão.
- Redação das Considerações Finais.

Semana 2 (10/11 – 16/11)

- Revisão geral do texto (clareza, coesão e ortografia).
- Normalização de referências.
- Preparação de gráficos e tabelas finais.
- Preparação e organização da apresentação.

Dia 17/11 - Entrega final do trabalho.

Referências

[1] Censo da Educação Superior. (n.d.). Instituto Nacional de Estudos E Pesquisas Educacionais Anísio Teixeira | Inep. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior.

[2] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. doi: 10.1023/A:1010933404324.

[3] Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O., & Lobo, M. B. C. M. (2007). A evasão no ensino superior brasileiro. *Cadernos de Pesquisa*, 37(132), 641–659. Disponível em: https://publicacoes.fcc.org.br/cp/article/view/346.

[4] Alencar, T. C., & Netto, J. F. M. (2020). Análise de técnicas de mineração de dados para predição da evasão de alunos. *Revista Brasileira de Informática na Educação*, 28, 58–84.

[5] Passos, M., Letti, A. G., Castro, A. L., & Bueno, C. J. G. (2024). Evasão e retenção nos cursos de licenciatura em Letras – Língua Portuguesa e Literaturas da Bahia no período de 2010 a 2022. *Encontro de Discentes Pesquisadores e Extensionistas*, 2(1). Disponível em: https://revistas.uneb.br/edpe/article/view/19146.

[6] Aggarwal, C. C. (2015). Data mining : the textbook. Springer.