Predição de evasão de alunos de cursinhos populares da USP a partir de dados de simulados online

Title: Prediction of student dropout from popular USP preparatory courses based on online simulated test data

Título: Predicción de la deserción de estudiantes de cursos preuniversitarios populares de la USP a partir de datos de exámenes simulados en línea

Ana Clara Segal Vidal Pessanha
Universidade de São Paulo
anaclarasvp@usp.br

Giovanna Almeida Albuquerque
Universidade de São Paulo
gi.albuqq@usp.br

Gustavo Pompermayer Fulanetti Silva
Universidade de São Paulo
gustavpompermayer@usp.br

Julia Du Bois Araujo Silva
Universidade de São Paulo
juliaduboisas@usp.br

Resumo

Este plano de trabalho descreve o problema da evasão escolar e no contexto de cursinhos populares, especialmente afetados por conta da combinação de fatores econômicos, pedagógicos e sociais que levam a esse fenômeno. Sugere-se uma linha de pesquisa baseada em aprendizado de máquina, com o uso dos classificatórios binários árvores de decisão e regressão logística, para criar um modelo que preveja antecipadamente os alunos com maior risco de evasão, permitindo que o provedor do cursinho popular possa tomar medidas preventivas. Os dados examinados provêm do Cursinho Popular EACH USP e consistem de resultados de simulados ENEM aplicados e tabelas de frequência dos alunos.

Palavras-chave: *Evasão escolar; Cursinho Popular; Aprendizado de máquina; Árvores de decisão; Regressão logística*

Abstract

This work plan describes the problem of student dropout in the context of popular preparatory courses, which are especially affected by a combination of economic, pedagogical, and social factors that lead to this phenomenon. A research approach based on machine learning is suggested, using binary classifiers such as decision trees and logistic regression, to create a model that can predict, in advance, the students at higher risk of dropping out. This would allow the popular preparatory course provider to take preventive measures. The data examined comes from the Cursinho Popular EACH USP and consists of results from applied ENEM practice exams and student attendance records.

Keywords: *Student Dropout; Popular Preparatory Course; Machine Learning; Decision Trees; Logistic Regression*

Resumen

Este plan de trabajo describe el problema del abandono escolar en el contexto de los cursos preuniversitarios populares, los cuales se ven especialmente afectados por la combinación de factores económicos, pedagógicos y sociales que conducen a este fenómeno. Se propone una línea de investigación basada en el aprendizaje automático, con el uso de clasificadores binarios como los árboles de decisión y la regresión logística, para crear un modelo que pueda predecir con antelación a los estudiantes con mayor riesgo de abandono. Esto permitiría al proveedor del curso preuniversitario popular tomar medidas preventivas. Los datos examinados provienen del Cursinho Popular EACH USP y consisten en los resultados de simulacros del ENEM aplicados y en las tablas de asistencia de los estudiantes.

Palabras clave: *Abandono Escolar; Curso Preuniversitario Popular; Aprendizaje Automático; Árboles de Decisión; Regresión Logística*

Cite as: PESSANHA, A. C. S. V., ALBUQUERQUE, G. A., SILVA, G. P. F. & SILVA, J. D. B. A. (2025). Predição de evasão de alunos de cursinhos populares da USP a partir de dados de simulados online. Revista Brasileira de Informática na Educação, vol, pp-pp. https://doi.org/10.5753/rbie.yyyy.id.

1 Introdução

O Cursinho Popular EACH USP, fundado em 2015, tem como missão "democratizar o acesso ao ensino superior, sobretudo, às pessoas de baixa renda, de escola pública e/ou da Zona Leste da cidade de São Paulo, através da vivência e preparação pré-universitária gratuita e de qualidade"(CURSINHO POPULAR EACH USP, 2025), por meio do oferecimento de aulas e materiais gratuitos para estudantes em situação de vulnerabilidade social. Apesar disso, o Cursinho Popular EACH USP enfrenta, como diversos outros provedores educacionais, grandes desafios para evitar a evasão escolar.

Para auxiliar a prevenção de evasão de alunos no Cursinho, esse plano propõe o uso de técnicas de aprendizado de máquina classificatórios binários, especialmente as árvores de decisão e a regressão logística, para apontar os alunos sob maior risco de evasão. Para realizar o estudo, serão utilizados os desempenhos dos alunos do Cursinho Popular EACH USP em simulados com questões do ENEM por meio da plataforma Eduqo, um sistema de gestão de aprendizado, combinados com os registros internos de frequência dos mesmos alunos.

2 Arcabouço teórico

Em termos gerais, a evasão escolar corresponde à interrupção do processo formativo anterior à conclusão da etapa educacional prevista (FILHO; ARAÚJO, 2017). Trata-se de um fenômeno multifatorial e complexo, resultante da interação de aspectos econômicos, pedagógicos e sociais, cuja gravidade não reside em um único elemento isolado, mas na combinação de diversos fatores (FILHO; ARAÚJO, 2017). No âmbito de cursinhos populares, como os cursinhos populares da Universidade de São Paulo (USP), esses desafios se tornam ainda mais evidentes, uma vez que o público atendido é majoritariamente composto por estudantes oriundos da rede pública de ensino e de contextos socioeconômicos vulneráveis (JORNAL DA USP, 2017). Nesse sentido, uma abordagem promissora para identificar padrões comuns entre estudantes que evadiram e, consequentemente, detectar potenciais candidatos à evasão, é a aplicação de modelos de classificação no âmbito do aprendizado de máquina supervisionado. Tais modelos têm se mostrado eficazes na antecipação de riscos educacionais, permitindo a adoção de estratégias preventivas e de intervenções focalizadas pelos seus instrutores (Ramos et al., 2018).

A classificação, no contexto do aprendizado supervisionado, é um processo de modelagem preditiva que organiza objetos em categorias ou classes previamente definidas e conhecidas. Nesse processo, o modelo utiliza as variáveis independentes (também chamadas de atributos ou *features*) disponíveis no conjunto de dados para predizer o rótulo ou classe (variável dependente) de novas observações. Diferentemente da regressão, em que a variável resposta é numérica e contínua, a classificação lida com variáveis categóricas. Um exemplo clássico de aplicação ocorre na área financeira, onde, a partir de características de clientes, pode-se prever se um futuro cliente será classificado como adimplente ou inadimplente (SICSÚ; SAMARTINI; BARTH, 2023). Como evidenciado no exemplo anterior, os classificadores são ótimos preditores para resolverem problemas binários, como o de evasão.

Dentre os algoritmos de classificação, destaca-se a Árvore de Decisão, que organiza os

objetos em uma estrutura hierárquica contendo raiz, nós internos, arestas e folhas. O primeiro atributo testado constitui o nó raiz, que se ramifica em nós internos de acordo com condições estabelecidas nas arestas, até alcançar os nós do tipo folha, que representam os rótulos finais (FERNANDES, 2017).

Outro algoritmo consolidado para resolução de problemas de classificação é a regressão logística, um modelo estatístico de predição em que a variável dependente geralmente apresenta apenas duas categorias (dicotômica) (GONZALEZ, 2018), embora existam extensões que possibilitem a classificação multiclasse. Este modelo tem como objetivo estimar a probabilidade de ocorrência de um resultado favorável com base nas características observadas nas variáveis independentes (MIRANDA, 2023).

A eficácia desses modelos pode ser avaliada por métricas como acurácia, precisão, revocação (ou sensibilidade) e F1-score, fundamentais para mensurar tanto a taxa geral de acertos do modelo quanto a sua capacidade de classificar corretamente objetos em suas respectivas categorias (COSTA, 2025). Especificamente no contexto da evasão, a escolha das métricas de avaliação assume papel central, sobretudo no que se refere à correta classificação dos objetos. Nesse sentido, métricas como a precisão ganham relevância, pois permitem analisar a proporção de estudantes corretamente identificados como evadidos entre todos aqueles previstos nessa classe.

Considerando esse panorama, a aplicação de algoritmos de classificação em dados provenientes de simulados online dos cursinhos populares da USP representa uma oportunidade promissora para investigar padrões associados à evasão, bem como propor intervenções baseadas em evidências empíricas.

3 Trabalhos correlatos

Com o objetivo de identificar pesquisas relevantes sobre a aplicação de *Machine Learning* - ML (Aprendizado de Máquina) na classificação de dados em estudos sobre evasão escolar, foram realizadas revisões e mapeamentos da literatura. Os trabalhos encontrados serviram de referência para orientar este estudo, permitindo identificar, por exemplo, os algoritmos mais utilizados para esse tipo de problema. Segundo os artigos selecionados, destacam-se a Regressão Linear Múltipla e os algoritmos baseados em Árvores de Decisão.

O estudo de Jesus e Gusmão (2024) apresenta o estado da arte sobre a aplicação de Mineração de Dados (MD) e ML em pesquisas de evasão escolar, evidenciando que os algoritmos de árvore de decisão são os mais recorrentes na construção de modelos de classificação. Já o trabalho de Ramos et al. (2018) compara cinco classificadores na predição de risco de evasão em cursos de graduação a distância, concluindo que a Regressão Logística obteve ligeira vantagem.

Na mesma linha, Silva et al. (2025) analisam os determinantes da evasão no ensino a distância em instituições de ensino superior brasileiras, utilizando Regressão Linear Múltipla. Os autores observaram que fatores como idade, gênero e origem escolar exercem influência significativa sobre as taxas de evasão, as quais também variam entre instituições. Complementarmente, Paula e Picanço (2024) investigam a relação entre origem social dos estudantes e evasão no ensino superior por meio de Regressão Logística Multinível.

Outros trabalhos utilizam metodologias distintas. Araújo, Mariano e Oliveira (2021) aplicam um modelo de escolha multinomial ordenado para identificar determinantes da retenção em instituições federais de cursos presenciais. Já Menolli et al. (2025) exploram a redução de variáveis preditoras para previsão da evasão, destacando que a correlação entre a duração da matrícula e a porcentagem de conclusão do curso se apresenta como principal fator explicativo.

Com foco no desempenho de técnicas, Souza e Santos (2021) comparam abordagens de ML e Deep Learning (Aprendizado Profundo), concluindo que este último apresenta leve superioridade. De forma semelhante, Baldasso (2019) aplica random forests (Florestas Aleatórias) para identificar alunos em risco de evasão em cursos pré-vestibular online.

Alguns estudos assumem perspectivas mais qualitativas. Campos e Cruz (2020), por exemplo, elaboram um instrumento de autoavaliação para estudantes de cursinhos populares vinculados à USP, buscando compreender fatores relacionados a dificuldades, desmotivação e evasão. Por fim, Prenkaj et al. (2020) oferecem uma análise aprofundada da literatura internacional sobre *Student Dropout Prediction* - SDP (Previsão de Evasão Estudantil), ampliando a compreensão sobre tendências, limitações e desafios do campo.

Diferentemente dos estudos revisados, que em sua maioria analisam a evasão em cursos de graduação presenciais ou a distância, ou ainda em contextos amplos de instituições de ensino superior, nosso trabalho concentra-se em um cenário específico e pouco explorado: os cursinhos populares vinculados à USP. O foco na predição da evasão a partir de dados de simulados online representa uma inovação metodológica, pois utiliza informações de desempenho acadêmico coletadas de forma contínua e sistemática por meio da plataforma Eduqo (antigo QMágico). Dessa forma, além de contribuir para o avanço da literatura sobre predição de evasão estudantil, este estudo busca oferecer uma ferramenta prática de apoio à gestão pedagógica do Cursinho Popular da EACH USP (e outros que tiverem interesse), possibilitando intervenções mais precisas e tempestivas no acompanhamento dos estudantes em situação de risco.

4 Método

O presente estudo utiliza como base metodológica a Mineração de Dados Educacionais (MDE), aplicada à predição da evasão de estudantes do Cursinho Popular da EACH USP, por meio dos registros de desempenho em simulados online da plataforma Eduqo (antiga QMágico). Embora haja a intenção de, futuramente, coletar dados de outros cursinhos populares vinculados à USP, nesta primeira etapa consideramos apenas o conjunto de dados do cursinho da EACH, que já foi objeto de análise em pesquisas anteriores sobre clusterização de estudantes.

4.1 Coleta de dados

Os dados serão obtidos junto ao Cursinho Popular da EACH USP por meio do banco de registros da plataforma Eduqo, que armazena os resultados dos simulados aplicados ao longo do curso preparatório. Esses registros incluem variáveis de desempenho (notas por disciplina por simulado), além de dados de interação com a plataforma (tempo de resolução médio por disciplina). Somados a esses registros, serão considerados também os dados internos de frequência às aulas, coletados fora da plataforma, que funcionam como importantes indicadores de evasão. Para

garantir a manutenção da Lei Geral de Proteção de Dados e a anonimidade dos alunos cujos dados serão utilizados, serão excluídas informações de identificação pessoal, mantendo-se apenas atributos acadêmicos e de desempenho.

4.2 Pré-processamento e organização

Os dados coletados passarão por procedimentos básicos de preparação, incluindo limpeza, padronização e organização em um formato adequado para análise. Esse processo garante que os registros estejam consistentes e prontos para serem utilizados nos modelos de predição. As etapas de pré-processamento e modelagem serão implementadas em linguagem de programação Python, utilizando bibliotecas consolidadas da área de ciência de dados e aprendizado de máquina.

4.3 Algoritmos aplicados

Para a análise preditiva, serão empregados algoritmos de aprendizado supervisionado, com foco em métodos amplamente utilizados na literatura sobre evasão estudantil, como regressão logística e árvores de decisão. A escolha se deve à capacidade desses modelos de relacionar variáveis e apontar padrões que indicam risco de evasão.

Os modelos serão avaliados utilizando divisão de dados em 70% para treinamento e 30% para teste, sendo o conjunto de treinamento submetido ao procedimento de validação cruzada (k-fold cross validation). Essa abordagem garante que todos os subconjuntos de dados sejam utilizados tanto para treinamento quanto para validação, aumentando a robustez e a confiabilidade dos resultados.

4.4 Análise e interpretação dos resultados

Após o treinamento e teste, os modelos serão comparados quanto à sua capacidade preditiva e interpretabilidade. Também será realizada uma análise dos atributos mais relevantes, buscando responder quais fatores — como baixa frequência, desempenho reduzido em disciplinas-chave ou tempo excessivo de resolução — estão mais associados à evasão. Essa análise permitirá, em etapas futuras, propor intervenções pedagógicas baseadas em evidências, tais como reforços direcionados.

5 Cronograma

Tabela 1: Tabela 1: Cronograma.

Data	Atividade
20/08/2025	Entrega do Plano de Trabalho
27/08/2025	Finalização da coleta dos dados junto ao Cursinho Popular da EACH USP
28/08/2025 - 25/09/2025	Aplicação dos algoritmos e criação de modelos
26/09/2025 - 07/09/2025	Análise dos modelos
08/10/2025	Entrega do relatório parcial e vídeo
09/10/2025 - 30/10/2025	Análise dos resultados obtidos
31/10/2025 - 16/11/2025	Conclusões finais
17/11/2025	Entrega do relatório final
19/11/2025 - 01/12/2025	Apresentação do trabalho (Seminário interno de socialização)

6 Referências

Referências

Araújo, A. C. P. L. D., Mariano, F. Z., & Oliveira, C. S. D. (2021). Determinantes acadêmicos da retenção no Ensino Superior. *Ensaio: Avaliação e Políticas Públicas em Educação*, *29*(113), 1045–1066.

Baldasso, R. O. (2019). Aplicação de algoritmo de machine learning na identificação de alunos em risco de evasão [Trabalho Acadêmico].

Campos, L. M. L., & da Cruz, N. H. (2020). Instrumento de autoavaliação para estudantes de cursinhos populares: a evasão como problemática. *Cadernos CIMEAC*, *10*(2), 31–58.

da Silva, A. R., Pereira, F. C., de Souza Mendonça, R., de Almeida, R. G. M., dos Santos Barbosa, E. H., & Delgado, K. V. (2025). Uma Análise Quantitativa dos Determinantes da Evasão no Ensino Superior EaD. *EaD em Foco*, *15*(1), e2322–e2322.

de Jesus, J. A., & de Gusmão, R. P. (2024). Investigação da evasão estudantil por meio da mineração de dados e aprendizagem de máquina: Um mapeamento sistemático. *Revista Brasileira de Informática na Educação*, *32*, 807–841.

Fernandes, F. (2017). *Emprego de diferentes algoritmos de árvores de decisão na classificação da atividade celular in vitro para tratamentos de superfícies em titânio* [Dissertação (Mestrado em Engenharia)]. Universidade Federal do Rio Grande do Sul [Disponível em: https://lume.ufrgs.br/handle/10183/165456. Acesso em: 18 ago. 2025].

Gonzales, L. (2018). *URegressão Logística e suas aplicações*.

Menolli, A., Dionísio, G. M., da Paz Floriano, A. S., & Coleti, T. A. (2025). Exploring Feature Reduction for Dropout Predicting in Higher Education in Brazil. *Revista Brasileira de Informática na Educação*, *33*, 106–129.

Miranda, F. (2023). *Utilização de Regressão Logística na Análise da Percepção da Comunidade da UFOP acerca da Pandemia de Covid-19*.

Paula, G. B. D., & Picanço, F. (2024). Desigualdades após o acesso: origem social e evasão do sistema de ensino superior. *Educação & Sociedade*, *45*, e281915.

Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A survey of machine learning approaches for student dropout prediction in online courses. *ACM Computing Surveys (CSUR)*, *53*(3), 1–34.

Ramos, J. L. C., Silva, J., Prado, L., Gomes, A., & Rodrigues, R. (2018). Um estudo comparativo de classificadores na previsão da evasão de alunos em EAD. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, *29*(1), 1463.

Sicsú, A., Samartini, A., & Barth, N. (2023). *Técnicas de Machine Learning*. Blucher.

Silva Filho, R., & Araújo, R. (2017). Evasão e abandono escolar na educação básica no Brasil: fatores, causas e possíveis consequências [Disponível em: https://revistaseletronicas.pucrs.br/porescrito/article/view/24527. Acesso em: 18 ago. 2025]. *Revista Educação por Escrito*, *8*(1), 35–48.

Souza, V. F., & dos Santos, T. C. B. (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*, *29*, 519–546.

USP, C. P. E. (2025). *Quem somos – Cursinho EACH USP*. Cursinho Each. Recuperado agosto 21, 2025, de https://cursinhoeach.com.br/quem-somos/