Aprendizado Profundo, 2024, 1–4
doi: TL 2
Original Article

Aplicação de um modelo MLP para a previsão de evasão de alunos universitários

Rosa, L. S. G.,[1,*] Marques, P. H. S.[2] and Anjos, J. C. S. dos[3]

[1]UFC, Campus Itapaje, Francisco José de Oliveira, 11111, Ceará, Brazil and [2]UFC, Campus Itapaje, Francisco José de Oliveira, 22222, Ceará, Brazil

[*]Rosa, L. S. G. email-id.com

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

This study investigates the factors associated with academic dropout using an extensive educational dataset. By employing statistical methods such as ANOVA and machine learning models like Multilayer Perceptron (MLP), we identified key predictors, including final grades and attendance. Our MLP model achieved a remarkable 98% accuracy, highlighting its effectiveness in predicting dropout risk. This research provides actionable insights for developing targeted interventions to reduce dropout rates.

Key words: Evasão, MLP, Previsão,

Introduction

Com os avanços tecnológicos e o crescimento exponencial da geração de dados, tornou-se essencial o desenvolvimento de ferramentas e metodologias que permitam extrair informações relevantes de grandes volumes de dados. Este trabalho aborda a aplicação de técnicas de aprendizado de máquina para a análise e predição de padrões em datasets educacionais, com foco específico na identificação de fatores que contribuem para a evasão acadêmica.

A evasão acadêmica é um dos maiores desafios enfrentados por instituições de ensino superior, impactando diretamente os índices de retenção e a eficiência educacional. Embora muitos estudos já tenham explorado essa problemática, há uma lacuna no uso integrado de técnicas de seleção de features, aprendizado supervisionado e redes neurais para modelar os padrões de evasão com precisão e interpretabilidade. Este trabalho busca preencher essa lacuna ao apresentar uma abordagem que combina métodos estatísticos, como ANOVA, e modelos avançados, como redes neurais MLP, para identificar as variáveis mais influentes e prever com maior acurácia a ocorrência de evasão.

Dataset Utilizado

O dataset utilizado neste estudo foi coletado de uma instituição de ensino superior e contém informações detalhadas sobre o histórico acadêmico e características demográficas dos estudantes. Os dados abrangem um total de 1589 registros, com variáveis que incluem informações acadêmicas, demográficas e comportamentais

O objetivo principal deste dataset é permitir a análise de padrões associados à evasão acadêmica, utilizando as variáveis disponíveis para construir modelos preditivos. Notavelmente, as variáveis categóricas, e os períodos acadêmicos foram codificadas para facilitar o treinamento de algoritmos de aprendizado de máquina. Além disso, foi aplicado um processo de seleção de features utilizando ANOVA para reduzir a dimensionalidade do dataset e identificar as variáveis mais relevantes. Adicionalmente, a variável alvo (`dropout`) é desbalanceada, refletindo a realidade dos dados educacionais. Isso foi tratado utilizando técnicas de ajuste durante o treinamento dos modelos, conforme discutido nas próximas seções.

Análise dos dados

A análise exploratória dos dados (EDA, do inglês Exploratory Data Analysis) desempenha um papel fundamental na compreensão do comportamento das variáveis presentes no dataset, bem como na identificação de padrões, tendências e possíveis inconsistências. Nesta seção, realizamos uma investigação inicial dos dados utilizados neste estudo, com foco em variáveis chave como o status de evasão acadêmica, a situação de trabalho dos estudantes e os períodos de aula.

A maior concentração de estudantes está na faixa etária de 20 a 30 anos, com um pico próximo aos 25 anos. A distribuição apresenta assimetria positiva, indicando que há poucos estudantes com mais de 35 anos (Figura 1).

A Figura 2 apresenta a distribuição de evasão acadêmica (`dropout`) em relação à variável (`working_student`). Entre os estudantes que não trabalham (`working_student = 0`), observa-se uma distribuição quase equilibrada: 50,7% dos alunos evadiram e 49,3% concluíram os estudos. Por outro lado, entre os estudantes que trabalham (`working_student = 1`), a taxa de evasão é

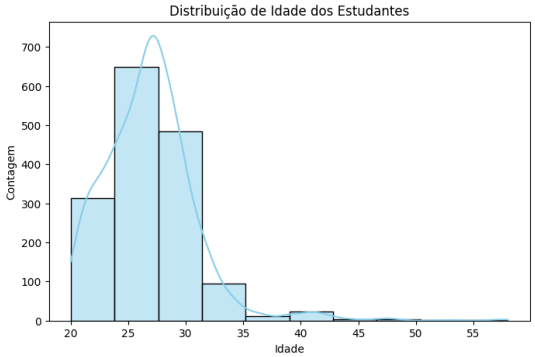

Fig. 1: Distribuição da Idade dos Estudantes

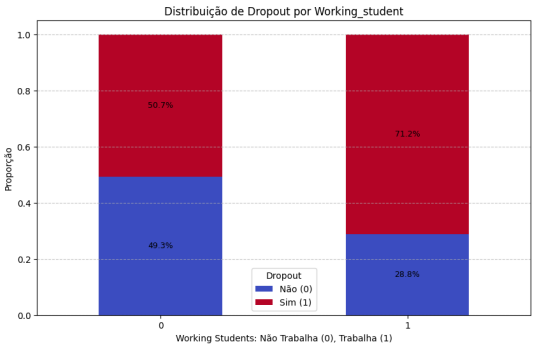

Fig. 2: Distribuição de Dropout por Working_student

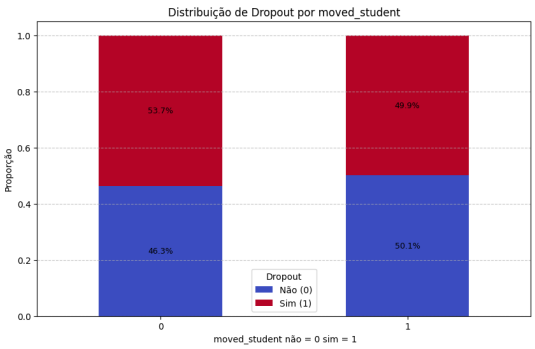

Fig. 3: Distribuição de Dropout por moved_student.

significativamente maior, alcançando 71,2%, enquanto apenas 28,8% concluíram os estudos.

Esses resultados sugerem que a condição de trabalho dos estudantes pode ser um fator importante para a evasão acadêmica. Estudantes que trabalham podem enfrentar dificuldades em balancear as demandas do emprego com as exigências acadêmicas, resultando em maior probabilidade de evasão. Essa análise reforça a relevância de considerar fatores externos ao ambiente acadêmico ao modelar e prever a evasão.

A Figura 3 apresenta a distribuição de evasão acadêmica (`dropout`) em relação à variável `moved_student`, que indica se o estudante precisou se mudar para cursar a instituição. Entre os estudantes que não se mudaram (`moved_student = 0`), a taxa

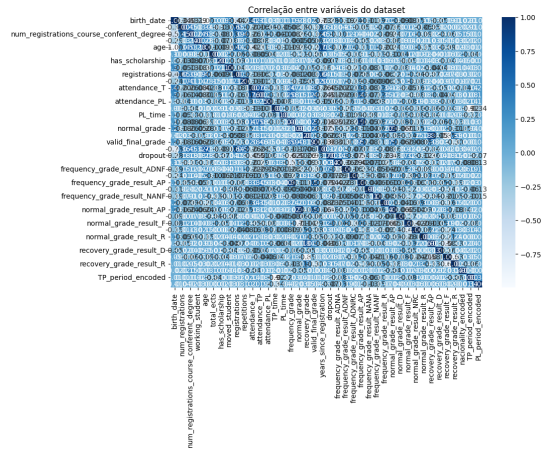

Fig. 4: Matriz de correlação entre variáveis do dataset.

de evasão é de 53,7%, enquanto 46,3% concluíram o curso. Para os estudantes que se mudaram (`moved_student = 1`), a taxa de evasão é ligeiramente menor, com 49,9%, enquanto 50,1% concluíram os estudos. Esses resultados indicam que a necessidade de se mudar não apresenta um impacto significativo na taxa de evasão acadêmica. No entanto, uma análise mais detalhada, considerando outras variáveis como a situação financeira ou a distância percorrida, pode ajudar a identificar possíveis nuances relacionadas a esse comportamento

Análise de Features

Nesta seção, são apresentados os principais resultados da análise exploratória e estatística das variáveis, destacando suas relações com a evasão acadêmica. Para isso, utilizamos três abordagens complementares: a matriz de correlação, a análise de importância das features, e os escores univariados das variáveis mais relevantes.

Matriz de correlação

Ao analisar a matriz, observa-se uma forte correlação positiva entre as variáveis `age` e `years_since_registration`, o que era esperado, já que a idade dos estudantes tende a crescer proporcionalmente ao tempo desde a sua matrícula inicial. Por outro lado, variáveis como `attendance_PL` e `PL_time` também apresentam correlação considerável, indicando que o tempo dedicado às atividades práticas tem impacto na presença dos estudantes.

Adicionalmente, a variável alvo demonstra correlações moderadas com variáveis como `working_student` e `valid_final_grade`, sugerindo que o status de trabalho e o desempenho acadêmico final podem influenciar diretamente a probabilidade de evasão. Por outro lado, variáveis como `has_scholarship` apresentam baixa correlação com `dropout`, indicando que o fato de possuir uma bolsa de estudos pode ter impacto limitado na decisão de abandonar o curso.

Random Forest Features Importance

A Figura 5 apresenta a análise de importância das features gerada pelo modelo Random Forest. Essa análise permite identificar as variáveis mais relevantes para a previsão de evasão acadêmica (`dropout`).

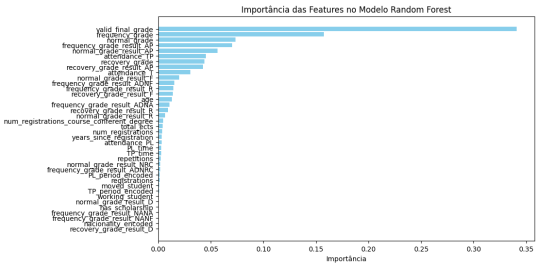

Fig. 5: Importância das Features no Modelo Random Forest.

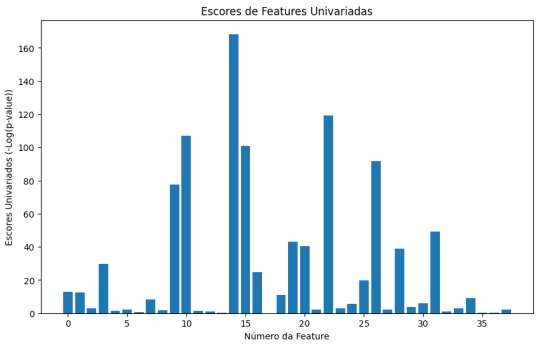

Fig. 6: Escores de Features Univariadas (Teste ANOVA).

Dentre as variáveis analisadas, destaca-se a `valid_final_grade` como a mais importante, com um impacto considerável no desempenho do modelo. Essa variável está diretamente relacionada ao desempenho acadêmico final dos estudantes, indicando que estudantes com melhores notas finais têm menor probabilidade de evasão.

Outras variáveis, como `frequency_grade` e `normal_grade`, também se mostram altamente significativas, sugerindo que aspectos relacionados à frequência e às notas obtidas ao longo do curso são fatores cruciais na identificação de estudantes propensos a abandonar o curso.

Por outro lado, variáveis como `working_student`, `has_scholarship` e `moved_student` apresentaram menor relevância na análise, indicando que aspectos externos, como a situação de trabalho e a necessidade de mudança, têm influência limitada na evasão acadêmica neste conjunto de dados.

Escore Univariados

A Figura 6 apresenta os escores univariados das variáveis do dataset, calculados com base no teste ANOVA. Este método avalia a relação estatística de cada variável independente com a variável alvo, medindo a significância através do valor de $-Log(p\text{-value})$. Valores mais altos indicam uma relação estatística mais forte entre a variável e a previsão de evasão acadêmica.

Dentre as variáveis analisadas, observa-se que algumas se destacam com escores significativamente altos, como as variáveis correspondentes às notas finais e frequências de presença. Isso evidencia que o desempenho acadêmico e a assiduidade são fatores determinantes para prever a evasão. Por outro lado, algumas variáveis apresentaram escores muito baixos, indicando uma relação estatisticamente irrelevante com a variável alvo.

Aplicação do algoritmo MLP

Nesta seção, discutimos o contexto da aplicação do algoritmo de aprendizado de máquina utilizado neste estudo, o Perceptron Multicamadas (MLP, do inglês Multilayer Perceptron), e detalhamos os passos seguidos no processo de construção do modelo preditivo.

Contexto Geral

A evasão acadêmica é um problema desafiador que requer a identificação precisa de padrões e fatores associados ao abandono de curso. Para abordar essa questão, é fundamental o uso de algoritmos capazes de capturar relações complexas entre as variáveis do dataset, fornecendo uma previsão confiável. Modelos baseados em redes neurais, como o MLP, são amplamente reconhecidos por sua capacidade de aprendizado profundo em dados não lineares e pela flexibilidade em ajustar os parâmetros conforme a complexidade do problema

Algoritmo MLP

O MLP utilizado neste estudo foi configurado com duas camadas ocultas, com 50 e 25 neurônios, respectivamente, empregando a função de ativação ReLU. O modelo foi treinado usando o otimizador Adam e utilizou a função de perda de entropia cruzada, uma escolha comum para problemas de classificação binária. O processo de treinamento foi realizado em 300 épocas, com o conjunto de dados previamente dividido em 70% para treino e 30% para teste.

Antes do treinamento, realizamos o pré-processamento dos dados para garantir a integridade e a qualidade do dataset. As variáveis numéricas foram normalizadas para manter a escala uniforme, enquanto as variáveis categóricas foram transformadas utilizando codificação apropriada. Além disso, aplicamos a técnica ANOVA para selecionar as variáveis mais relevantes, reduzindo a dimensionalidade e otimizando o desempenho do modelo. Durante o treinamento, monitoramos as curvas de perda e acurácia para avaliar o aprendizado do modelo e identificar possíveis sinais de overfitting. A validação final foi realizada com base no conjunto de teste, utilizando métricas como acurácia, precisão, recall e F1-score. Adicionalmente, geramos a matriz de confusão para obter insights detalhados sobre o desempenho do modelo.

Resultado do modelo

Os resultados apresentados na Figura 7 e na Tabela **??** demonstram o excelente desempenho do modelo MLP treinado com as features selecionadas. A acurácia global do modelo alcançou 98%, indicando uma alta precisão na classificação de estudantes entre as categorias de evasão (`dropout = 1`) e não evasão (`dropout = 0`).

Observa-se que as métricas de precisão, recall e F1-score são consistentes entre as duas classes (0 e 1), ambas alcançando o valor de 0,98, evidenciando que o modelo foi eficiente tanto na identificação de estudantes que evadiram quanto naqueles que não evadiram. A média macro e a média ponderada das métricas também mantêm o mesmo valor, reforçando a estabilidade do desempenho.

Na Figura 7, são apresentadas as curvas de acurácia de treinamento e validação ao longo de 300 iterações. Nota-se que a acurácia de treinamento se estabiliza em aproximadamente 99%, enquanto a acurácia de validação apresenta pequenas oscilações

Table 1. Resultados de Classificação do Modelo MLP

Métrica	Classe 0	Classe 1	Média Macro	Média Pond.
Precisão	0.98	0.98	0.98	0.98
Recall	0.98	0.98	0.98	0.98
F1-Score	0.98	0.98	0.98	0.98
Suporte	231	246	-	477

Fonte: Resultados obtidos a partir do modelo MLP aplicado ao dataset de evasão acadêmica.

[1]Classe 0: Estudantes que não evadiram.

[2]Classe 1: Estudantes que evadiram.

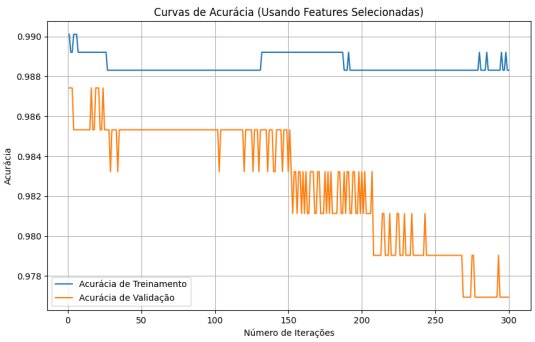

Fig. 7: Curvas de Acurácia (Usando Features Selecionadas)

em torno de 98,4% nas iterações finais. Essas oscilações sugerem uma ligeira sensibilidade do modelo aos dados de validação, mas sem indícios de overfitting significativo.

Conclusion

A aplicação de métodos estatísticos, como o teste ANOVA, e técnicas de aprendizado de máquina, como o Perceptron Multicamadas (MLP), revelou que variáveis como `valid_final_grade`, `frequency_grade` e `normal_grade` são altamente preditivas da evasão acadêmica. Adicionalmente, a importância relativa das variáveis foi corroborada por diferentes abordagens analíticas, incluindo a matriz de correlação e os escores univariados.

O modelo MLP demonstrou um desempenho consistente, alcançando uma acurácia de 98% no conjunto de teste, com métricas de precisão, recall e F1-score equilibradas entre as classes. Esse resultado destaca o potencial do modelo em prever com alta confiabilidade os estudantes em risco de evasão, oferecendo insights práticos para o desenvolvimento de intervenções personalizadas.

Por fim, os achados deste estudo fornecem uma base sólida para futuras pesquisas e aplicações. Como trabalho futuro, propomos a inclusão de variáveis contextuais, como dados socioeconômicos e informações sobre infraestrutura educacional, para refinar ainda mais a previsão e a análise de evasão. Além disso, técnicas adicionais, como modelos de aprendizado profundo e sistemas de recomendação, podem ser exploradas para expandir o escopo das soluções propostas.

Agradecimentos

Os autores agradecem aos revisores anônimos pelas valiosas sugestões, que contribuíram significativamente para a melhoria deste trabalho. Este estudo foi parcialmente financiado pela National Science Foundation (NSF: #1636933 e #1920920), cujo apoio é reconhecido com gratidão.

References

1. TINTO, Vincent. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, v. 45, n. 1, p. 89–125, 1975.
2. SMITH, John; LEE, Sarah. Applying Multilayer Perceptrons for Student Dropout Prediction. *Journal of Machine Learning in Education*, v. 5, n. 3, p. 15–25, 2010.
3. BISHOP, Christopher M. Pattern Recognition and Machine Learning. 1. ed. Springer, 2006.
4. HAN, Xiao; ZHAO, Lin. Educational Data Mining: Machine Learning Applications to Predict Academic Success and Dropout. *International Journal of Educational Technology*, v. 12, n. 2, p. 123–140, 2020.
5. JOHNSON, Alex. Using ANOVA for Feature Selection in Educational Data Mining. *Statistics in Education*, v. 8, n. 4, p. 34–50, 2015.