XXXIII Congresso de Iniciação Científica da Unicamp

UMA ABORDAGEM DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DA EVASÃO ESCOLAR

Palavras-Chave: EVASÃO ESCOLAR, APRENDIZADO DE MÁQUINA, PREDIÇÃO DE RISCO

Autores(as):

ANA CLARA MARTIN DA SILVEIRA, COTUCA - UNICAMP

DANIEL DORIGAN DE CARVALHO CAMPOS, COTUCA - UNICAMP

ION MATEUS NUNES OPREA, COTUCA - UNICAMP

JÚLIO PACHECO STEIN, COTUCA - UNICAMP

VINÍCIUS GUEDES PINHEIRO, COTUCA - UNICAMP

Prof. Me. GUILHERME DE OLIVEIRA MACEDO (orientador), COTUCA - UNICAMP

INTRODUÇÃO:

A evasão escolar é um problema de grande relevância para o cenário educacional e social, impactando diretamente o futuro dos estudantes e a construção de uma sociedade mais justa e desenvolvida. O nível de escolaridade alcançado está fortemente relacionado a oportunidades profissionais, qualidade de vida e inclusão social. Por isso, garantir que os alunos permaneçam na escola e concluam seus estudos é uma meta essencial das políticas públicas e das instituições de ensino.

Diversos fatores podem levar um aluno a abandonar a escola, como dificuldades econômicas, baixo rendimento acadêmico e desinteresse. A subjetividade desses motivos, aliada à dificuldade de mensurá-los com precisão, torna a tarefa de prever a evasão um grande desafio. Ainda assim, a identificação precoce de alunos com maior risco de abandono permite a adoção de medidas preventivas mais eficazes, especialmente quando associada à compreensão dos fatores que influenciam essa decisão.

Este trabalho desenvolveu uma abordagem utilizando técnicas de aprendizado de máquina para a previsão da evasão escolar, integrando dados acadêmicos, socioeconômicos e temporais para identificar padrões que indicam maior risco de abandono. Foi construída uma ferramenta capaz de apoiar instituições de ensino na detecção de alunos vulneráveis, possibilitando a implementação de intervenções direcionadas e personalizadas.

Os resultados obtidos demonstraram o potencial dessas tecnologias para apoiar a tomada de decisão educacional, contribuindo para a redução da evasão e para o fortalecimento da permanência estudantil. A conclusão do trabalho confirma a eficácia da abordagem proposta e abre caminho para sua aplicação em diferentes contextos escolares.

METODOLOGIA:

- **Extração de características:** Nesta etapa, foram coletadas e processadas variáveis relevantes para a predição da evasão escolar. Como não foi possível obter dados reais por questões administrativas, foi desenvolvido um gerador de dados sintéticos capaz de emular a relação entre atributos pessoais e acadêmicos dos alunos. Os dados gerados incluíram informações acadêmicas, como notas, frequência e disciplinas reprovadas; socioeconômicas, como renda familiar — ver Figura 1 —, uso de transporte público e distância até a escola; e outras características pessoais, como idade e composição familiar. O objetivo dessa etapa foi transformar os dados brutos em atributos significativos e úteis para os algoritmos de aprendizado de máquina. A Tabela 1 apresenta um exemplo dos atributos selecionados e de como seus valores estariam representados.

Fig. 1 - Gráfico mostrando frequência das faixas de renda familiar nos dados sintéticos.

id_aluno	12345	**nota_geral_media**	6.5
curso	19	**frequencia_geral_media**	85
serie_atual	3	**disciplinas_reprovadas_total**	2
distancia_ate_escola	10	**notas_abaixo_de_5_total**	3
usa_transporte_publico	Verdadeiro	**nota_matematica_media_total**	4.8
renda_familiar	5000	**nota_portugues_media_total**	7.6
tem_bolsa_auxilio	Verdadeiro	**evadiu**	Falso

Tabela 1 - apenas demonstrativa. Foram escolhidos mais de 30 atributos para representar cada aluno.

- **Divisão do conjunto de dados:** Após a geração e preparação dos dados, o conjunto sintético foi dividido em subconjuntos distintos: um para treinamento do modelo e outro para validação e teste. Essa separação foi essencial para avaliar o desempenho do modelo de forma imparcial, garantindo que ele generalizasse bem para novos dados.

- **Treinamento do modelo:** Com os dados preparados e divididos, foram treinados diferentes algoritmos de aprendizado de máquina para identificar padrões associados à evasão escolar. Foram escolhidos os modelos *Random Forest* e *XGBoost*, devido à sua capacidade de lidar com variáveis complexas e distintos tipos de dados. Além disso, foi utilizado o método de *stacking* para combinar os dois modelos e obter uma previsão mais precisa e robusta.

- **Avaliação do modelo:** O modelo foi avaliado utilizando métricas de desempenho como acurácia, precisão, recall e *AUC-ROC* para verificar sua eficácia na predição da evasão escolar. Também foi realizada uma análise interpretativa dos resultados por meio da técnica *SHAP*, que permitiu entender os fatores mais influentes na previsão — ver Figura 2 — e apoiar ações preventivas a partir das características identificadas.

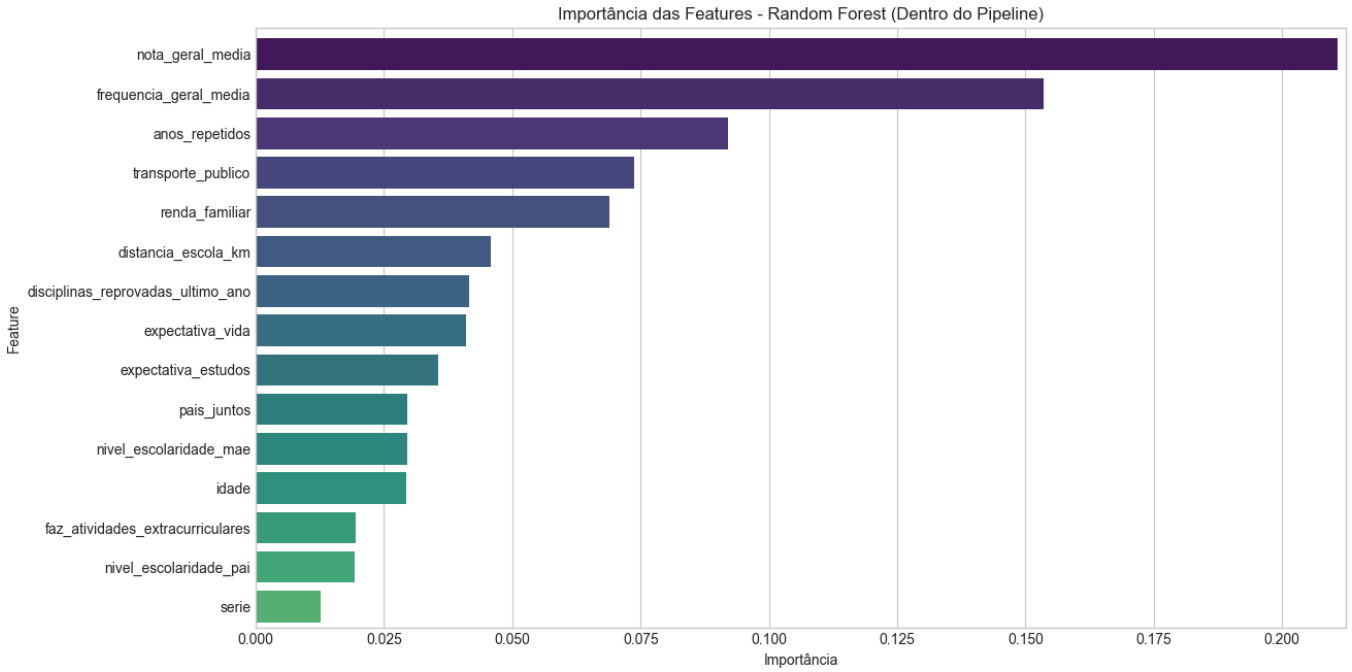

Fig. 2 - Gráfico de importância das *features* para a IA do *Random Forest.*

- **Implantação do modelo:** A interface gráfica da aplicação foi desenvolvida em Python, utilizando o *framework Streamlit*, com o objetivo de facilitar a visualização geral e individual dos dados dos alunos. Foram criadas telas de login, visão geral e previsão, integrando o modelo preditivo para consumo e visualização dos dados. Apesar das limitações estéticas da biblioteca, a plataforma resultante é funcional e intuitiva, permitindo a utilização prática da ferramenta para acompanhamento e prevenção da evasão escolar. A figura 3 mostra a parte da interface que exibe as informações de um aluno específico.

RESULTADOS:

Após a avaliação do modelo de aprendizado de máquina, verificou-se que o modelo preditivo alcançou 76% de acurácia, indicando que foi capaz de prever corretamente a evasão ou permanência do aluno na maioria dos casos. A precisão foi de 55%, o que revela que aproximadamente metade das previsões de evasão foram falsos positivos — resultado intencional, obtido pela escolha de um limiar de classificação mais baixo, com o objetivo de incluir sob supervisão também alunos com probabilidade média de evasão. O recall atingiu 75%, indicando que o modelo deixou de identificar 25% dos alunos em risco, o que foi considerado aceitável devido ao alto nível de ruído presente nos dados. A *AUC-ROC* foi de 82%, mostrando que o modelo conseguiu discriminar entre alunos evasores e não evasores na maioria das vezes, evidenciando uma boa capacidade avaliativa.

Considerando o conjunto dos resultados, conclui-se que o modelo preditivo atendeu aos critérios de acurácia e cumpriu corretamente sua função principal: priorizar o alerta para alunos com chance média de evasão, permitindo sua inclusão em ações preventivas.

Fig. 3 - O risco de evasão de um aluno selecionado com os principais fatores da avaliação e gráficos que relacionam os atributos do aluno à distribuição da turma.

Com base nas análises das avaliações *SHAP*, observou-se que o modelo final atribuiu maior importância a variáveis como nota, frequência e condição familiar dos alunos ao tomar decisões, o que está em conformidade com a estrutura dos dados utilizados. Esse alinhamento sugere que o modelo de aprendizado de máquina foi eficaz em identificar os fatores mais relevantes para a evasão, além de apresentar baixo risco de *overfitting*. Esse aspecto é relevante, pois indica que o modelo não apenas obteve bons resultados com os dados de teste, mas também possui capacidade de generalização, o que o torna adaptável a novos dados fora do conjunto de treino. Além disso, o fato de o modelo ser capaz de explicar por que classificou um aluno como evasor é especialmente valioso para que as instituições de ensino compreendam melhor os riscos enfrentados pelos estudantes e adotem medidas mais direcionadas de apoio.

Apesar do desempenho positivo, é importante ressaltar que o treinamento e a validação do modelo foram realizados exclusivamente com dados sintéticos. Assim, embora os resultados sejam promissores, é possível que o uso de dados reais traga desafios adicionais, relacionados à imprevisibilidade do comportamento humano — um aspecto que pode não ter sido plenamente representado pelo ruído adicionado aos dados sintéticos.

CONCLUSÕES E PERSPECTIVAS FUTURAS:

Analisando os resultados obtidos, concluímos que este trabalho foi capaz de cumprir seus principais objetivos. O sistema final permite a análise geral dos dados, como a quantidade de alunos em alto risco de evasão, a demografia dos estudantes e a porcentagem de evasão por curso e por atributos socioeconômicos. Além disso, possibilita a análise individual dos dados dos alunos para estimar suas chances de evasão, apresentando uma lista dos fatores que mais contribuíram para a avaliação e permitindo a realização de simulações ao recalcular a probabilidade de evasão com atributos modificados. Isso facilita a visualização de possíveis estratégias de intervenção e recuperação por parte das instituições de ensino.

Destaca-se o grande potencial da solução desenvolvida para aplicação real em ambientes educacionais. O bom desempenho e a capacidade de explicação do modelo preditivo — construído com *Ensemble Learning* via *Stacking* e treinado com dados sintéticos — sugerem que é plenamente viável o desenvolvimento de um modelo semelhante utilizando dados reais de estudantes. No entanto, esse modelo deve ser compreendido como uma base para a criação de uma plataforma de apoio à gestão pedagógica, na qual a interface gráfica e a inteligência artificial sejam desenvolvidas com o objetivo de facilitar a identificação de alunos com risco de evasão e a adoção de medidas adequadas pelas equipes educacionais.

BIBLIOGRAFIA

FELIPE, Anderson da Silva; VIEIRA, Marcelo Campos; OLIVEIRA, Daniel Ferreira de; PEREIRA, Lucas da Silva; RODRIGUES, Edson Vilela; MOREIRA, Cláudio Antônio. **A machine learning-based approach to predict student dropout in higher education**. Expert Systems with Applications, v. 219, p. 119659, 2023.

ZHANG, Cha; MA, Yunqian (eds.). **Ensemble Machine Learning: Methods and Applications.** New York: Springer, 2012.