



Estudo Comparativo de Algoritmos de Classificação para a Previsão de Rendimento Escolar

Henrique Oliveira Rodrigues

Universidade Federal do Vale do São Francisco
(henrique.rodrigues@discente.univasf.edu.br)

Thiago Magalhães Amaral

Universidade Federal do Vale do São Francisco
(thiago.magalhaes@univasf.edu.br)

CONTEXTUALIZAÇÃO: Um dos desafios mais críticos enfrentados pelas instituições de ensino é a evasão escolar e a reprovação, que geram impactos sociais e econômicos significativos. Estudos apontam que este fenômeno não apenas representa um custo social elevado para o país, mas também afeta negativamente a trajetória individual do aluno, associando-se a menores salários futuros e a um ciclo vicioso de desmotivação que frequentemente culmina no abandono escolar (PEREIRA, 2022). Diante disso, a Mineração de Dados surge como uma abordagem promissora para a identificação de fatores de risco associados a esse fenômeno (SILVA, 2019). Esta área de pesquisa, conhecida como Mineração de Dados Educacionais (MDE), foca na aplicação de métodos computacionais para analisar dados de ambientes de ensino e compreender melhor os estudantes (ROMERO; VENTURA, 2010). Diversos estudos demonstram a eficácia de algoritmos de aprendizado de máquina, como a Árvore de Decisão e o k-NN, que são frequentemente utilizados para construir modelos preditivos de alta acurácia (DE MARI; LIMA, 2020). A capacidade de identificar, de forma precoce, estudantes com alto risco de reprovação permite a implementação de estratégias de intervenção personalizadas, como tutorias, planos de estudo adaptativos e apoio psicopedagógico, antes que o fracasso escolar se concretize.



Anais do Simpósio Acadêmico de Engenharia de Produção (SAEPRO) da EEL-USP

IX SAEPRO – 25 e 26 de novembro de 2025

OBJETIVO: Desenvolver e avaliar modelos preditivos para a classificação do desempenho acadêmico de estudantes a partir de variáveis socioeconômicas e acadêmicas, utilizando os algoritmos k-NN e Árvore de Decisão (*Tree*).

MÉTODO: Trata-se de um estudo quantitativo desenvolvido a partir da análise do *dataset* público "*Student Performance Data Set*" (CORTEZ; SILVA, 2008), composto por 395 registros de estudantes e 33 atributos que abrangem informações demográficas, sociais e de desempenho acadêmico. Toda a análise foi conduzida no software de mineração de dados Orange. A etapa de pré-processamento foi fundamental para adequar os dados ao problema de classificação. Inicialmente, a variável-alvo foi criada a partir da nota final contínua 'G3', sendo transformada em uma variável categórica binária, "Resultado", que classifica os alunos em 'Aprovado' (nota ≥ 10) ou 'Reprovado' (nota < 10). Na sequência, a variável 'G3' original foi removida do conjunto de atributos preditores para evitar o vazamento de dados (*data leakage*), garantindo que o modelo aprendesse a prever o resultado com base em informações prévias, e não na própria resposta. Para a modelagem, foram treinados e comparados os algoritmos k-NN (testado com $k=5$ e $k=10$) e o algoritmo Árvore de Decisão que foi configurado para gerar a árvore com um número mínimo de 2 instâncias por folha. Como critério de parada, a divisão dos nós não era realizada em subconjuntos menores que 5 instâncias e era interrompida quando a pureza de um nó atingia 95%. A avaliação da performance dos modelos foi realizada através da técnica de validação cruzada (*10-fold cross-validation*), que divide o *dataset* em 10 partes e realiza 10 ciclos de treino e teste. Para uma análise comparativa completa, foram utilizadas métricas como a Acurácia (CA), *Area Under the Curve* (AUC), F1-Score (métrica utilizada para medir o equilíbrio do modelo) e a Matriz de Confusão, a fim de determinar não apenas a precisão geral, mas também a capacidade de cada modelo em lidar com as diferentes classes de resultado.

RESULTADOS E DISCUSSÕES: A avaliação comparativa dos modelos, realizada com validação cruzada de 10 partições, indicou um desempenho geral satisfatório para todas as abordagens testadas. O modelo de Árvore de Decisão obteve a maior Acurácia de Classificação (CA) e o maior F1-Score, atingindo um pouco mais de 90% em ambas



Anais do Simpósio Acadêmico de Engenharia de Produção (SAEPRO) da EEL-USP

IX SAEPRO – 25 e 26 de novembro de 2025

as métricas. Em contrapartida, o modelo k-NN com $k=10$ demonstrou uma capacidade superior de discriminação entre as classes, alcançando o maior valor de AUC (0.951). No entanto, a análise dos tipos de erro via Matriz de Confusão foi decisiva para a discussão. O modelo de Árvore de Decisão destacou-se por minimizar o erro mais crítico – prever que um aluno seria aprovado quando na verdade foi reprovado (Falso Negativo) –, cometendo apenas 21 erros deste tipo, contra 35 do kNN $k=10$ e 25 do kNN $k=5$. A eficácia preditiva dos modelos, por sua vez, está diretamente relacionada à relevância dos atributos, onde uma análise de importância (*Rank*) revelou que as notas parciais (G1 e G2) e o histórico de reprovações (*failures*) foram os fatores mais determinantes.

CONSIDERAÇÕES FINAIS: O objetivo de desenvolver e avaliar modelos preditivos para a classificação de desempenho de estudantes teve resultados satisfatórios. A análise comparativa demonstrou que, embora todos os algoritmos tenham apresentado alta performance, o modelo de Árvore de Decisão se mostrou a solução mais robusta e confiável para o problema proposto. Essa conclusão se baseia não apenas em sua acurácia superior (90,9%), mas principalmente em sua maior capacidade de identificar corretamente os alunos em risco de reprovação, minimizando os erros de maior impacto para uma possível intervenção pedagógica. O estudo reforça, portanto, o potencial da mineração de dados como ferramenta de apoio à gestão educacional, permitindo a criação de sistemas proativos que visam reduzir as taxas de insucesso escolar.

PALAVRAS-CHAVE: Mineração de Dados Educacionais, Desempenho Acadêmico, Algoritmos de Classificação, Árvore de Decisão, k-NN.

REFERÊNCIAS

- CORTEZ, P.; SILVA, A. Using Data Mining to Predict Secondary School Student Performance. In: FUTURE BUSINESS TECHNOLOGY CONFERENCE (FUBUTEC), 5., 2008, Porto. Anais [...]. Porto: FUBUTEC, 2008. p. 5-12.
- DE MARI, T. L.; LIMA, L. H. B. Predição de Desempenho Acadêmico de Alunos Utilizando Mineração de Dados. Revista Latino-Americana de Tecnologia Educacional, [S. l.], v. 19, n. 2, p. 177-190, 2020.



Anais do Simpósio Acadêmico de Engenharia de Produção (SAEPRO) da EEL-USP

IX SAEPRO – 25 e 26 de novembro de 2025

PEREIRA, V. Causas e consequências do abandono e da evasão escolar. Rio de Janeiro: Instituto Mobilidade e Desenvolvimento Social, 2022. (Texto para Discussão, n. A002-2022).

ROMERO, C.; VENTURA, S. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), [S. l.], v. 40, n. 6, p. 601-618, nov. 2010.

SILVA, F. G. F. da. Mineração de dados para identificação de fatores de reprovação no ensino superior. 2019. Dissertação (Mestrado em Ciência da Computação) – Faculdade de Computação, Universidade Federal de Uberlândia, Uberlândia, 2019.