

## PREDIÇÃO DO FRACASSO ESCOLAR NO ENSINO SUPERIOR USANDO TÉCNICAS DE MINERAÇÃO DE DADOS: CASO DO INSTITUTO SUPERIOR POLITÉCNICO DO MOXICO

### PREDICTION OF ACADEMIC FAILURE IN HIGHER EDUCATION USING DATA MINING TECHNIQUES: THE CASE OF THE INSTITUTO SUPERIOR POLITÉCNICO DO MOXICO

### PREDICCIÓN DEL FRACASO ACADÉMICO EN LA EDUCACIÓN SUPERIOR MEDIANTE TÉCNICAS DE MINERÍA DE DATOS: EL CASO DEL INSTITUTO SUPERIOR POLITÉCNICO DO MOXICO

Felino Job Muacanhica<sup>1</sup>  
Solange Flora André Muacanhica<sup>2</sup>

**RESUMO:** O fracasso escolar no ensino superior deriva de múltiplas variáveis ligadas ao perfil do estudante e ao ambiente de aprendizagem. Este trabalho analisou dados de um estudo realizado no Instituto Superior Politécnico do Moxico, focando nos estudantes do 1.º ano de Ciências de Computação, na disciplina de Lógica de Programação, com o objetivo de avaliar o desempenho discente mediante algoritmos de mineração de dados, identificando padrões preditivos de reprovação. A pesquisa é de natureza qualitativa, descritiva e de pesquisa-ação, envolveu 20 estudantes. Os dados, obtidos por observação, análise documental, questionário e entrevista, foram processados no WEKA. Compararam-se cinco algoritmos de classificação (J48, REPTree, NaiveBayes, SMO e Multilayer Perceptron) através de validação cruzada de 10 partições. O NaiveBayes destacou-se com 90% de acurácia, classificando corretamente 18 das 20 instâncias. A comparação pré e pós-intervenção, baseada em 13 interações, revelou redução sistemática da percentagem predita de fracasso: valores iniciais entre 50% e 96% caíram para 1% a 68%. A média recuou de 78,77% para 18,15%, uma variação de 60,62 pontos percentuais. Os achados sublinham a utilidade da mineração de dados no mapeamento de riscos acadêmicos e no monitoramento de tendências preditivas, ressaltando que o desenho do estudo não permite inferir causalidade direta entre as ações implementadas e a redução observada.

**Palavras-chave:** Mineração de dados. Fracasso escolar. NaiveBayes.

<sup>1</sup> Mestre - Direção estratégica em Engenharia de Software - Instituto Superior Politécnico do Moxico.

<sup>2</sup> Licenciatura - Contabilidade e Administração - Instituto Superior Politécnico do Moxico.

**ABSTRACT:** School failure in higher education stems from multiple variables linked to the student's profile and the learning environment. This paper analyses data from a study conducted at the Instituto Superior Politécnico do Moxico, focusing on first-year Computer Science students enrolled in the Programming Logic module, with the aim of evaluating student performance through data mining algorithms to identify predictive patterns of failure. The research adopted a qualitative, descriptive, and action-research approach, involving 20 students. Data gathered through observation, document analysis, questionnaires, and interviews were processed using WEKA. Five classification algorithms (J48, REPTree, NaiveBayes, SMO, and Multilayer Perceptron) were compared via 10-fold cross-validation. NaiveBayes stood out, achieving 90% accuracy by correctly classifying 18 of the 20 instances. The pre- and post-intervention comparison, based on 13 interactions, revealed a systematic reduction in the predicted failure rate: initial values ranging from 50% to 96% fell to between 1% and 68%. The mean decreased from 78.77% to 18.15%, representing a variation of 60.62 percentage points. These findings underscore the utility of data mining in mapping academic risks and monitoring predictive trends, whilst noting that the study design does not permit the inference of direct causality between the implemented actions and the observed reduction.

**Keywords:** Data mining. School failure. NaiveBayes.

**RESUMEN:** El fracaso escolar en la educación superior deriva de múltiples variables vinculadas al perfil del estudiante y al entorno de aprendizaje. Este trabajo analiza los datos de un estudio realizado en el Instituto Superior Politécnico do Moxico, centrándose en los estudiantes de primer año de Ciencias de la Computación, en la asignatura de Lógica de Programación, con el objetivo de evaluar el rendimiento discente mediante algoritmos de minería de datos para identificar patrones predictivos de fracaso. La investigación, de naturaleza cualitativa, descriptiva y de investigación-acción, contó con la participación de 20 estudiantes. Los datos, obtenidos mediante observación, análisis documental, cuestionarios y entrevistas, se procesaron en WEKA. Se compararon cinco algoritmos de clasificación (J48, REPTree, NaiveBayes, SMO y Multilayer Perceptron) mediante validación cruzada de 10 particiones. El NaiveBayes destacó con un 90 % de precisión, al clasificar correctamente 18 de las 20 instancias. La comparación pre y posintervención, basada en 13 interacciones, reveló una reducción sistemática del porcentaje predicho de fracaso: los valores iniciales, de entre el 50 % y el 96 %, descendieron a un intervalo del 1 % al 68 %. La media bajó del 78,77 % al 18,15 %, lo que supone una variación de 60,62 puntos porcentuales. Los hallazgos subrayan la utilidad de la minería de datos en el mapeo de riesgos académicos y el seguimiento de tendencias predictivas, si bien cabe reseñar que el diseño del estudio no permite inferir una causalidad directa entre las acciones implementadas y la reducción observada.

**Palabras clave:** Minería de datos. Fracaso escolar. NaiveBayes.

## 1. INTRODUÇÃO

O desempenho acadêmico no ensino superior resulta da convergência de fatores pessoais, institucionais e contextuais. Na disciplina de Lógica de Programação, estudos preliminares no Instituto Superior Politécnico do Moxico (ISP-Moxico) apontaram dificuldades recorrentes ligadas à assiduidade, acesso limitado a equipamentos informáticos, uso incipiente da biblioteca, condicionalismos económicos e hábitos de estudo pouco estruturados. A identificação precoce destes padrões revela-se crucial para mitigar o fracasso escolar antes da sua consolidação.

A mineração de dados oferece um quadro metodológico robusto para a descoberta de conhecimento em bases de dados (Fayyad, et al. 1996), aplicando técnicas que revelam relações não triviais em grandes volumes de informação (Han, Kamber e Pei, 2011). No domínio educacional, tais ferramentas têm sido empregadas com crescente frequência para apoiar a previsão de resultados académicos.

Neste cenário, o presente artigo debruça-se sobre os dados do 1.º ano do curso de Ciência de Computação do ISP-Moxico, especificamente na disciplina de Lógica de Programação. A questão que norteia a pesquisa é: de que modo as técnicas de mineração de dados contribuem para a identificação do fracasso escolar nestes estudantes? O objetivo geral consiste em analisar o desempenho académico dos estudantes, mediante a utilização de algoritmos de classificação, com vista a detetar padrões de risco. A análise aqui apresentada foca-se no tratamento dos dados, no desempenho comparativo dos algoritmos e na evolução das predições antes e depois de um ciclo de intervenção pedagógica, utilizando o WEKA como ambiente de processamento.

3

## 2. ENQUADRAMENTO TEÓRICO

### 2.1. Fracasso escolar e desempenho académico

O fracasso escolar constitui um fenómeno multifacetado, cuja delimitação conceptual varia conforme o referencial teórico adotado. A literatura especializada rejeita a atribuição unilateral da responsabilidade ao estudante, destacando a interação entre características individuais, dinâmicas familiares e condições do ambiente escolar. No contexto do ISP-Moxico, o diagnóstico inicial evidenciou problemas de assiduidade, carência de recursos informáticos, dificuldades de interpretação textual e desinteresse nas atividades letivas.

A transição de uma análise puramente descritiva para uma abordagem orientada por dados permite operacionalizar estas variáveis. A mineração de dados complementa os métodos tradicionais de acompanhamento pedagógico, ao desvendar correlações ocultas no conjunto de informações recolhidas.

## **2.2. Mineração de dados no contexto educacional**

A mineração de dados educacionais consolida-se como um campo de investigação dedicado à extração de padrões úteis em contextos de aprendizagem (Romero e Ventura, 2010). Autores como Kotsiantis (2009) demonstram a eficácia destes modelos na previsão de estudantes com elevada propensão ao abandono ou à reprovação. O valor acrescentado desta abordagem reside na capacidade de converter registos académicos brutos em indicadores analíticos acionáveis, deslocando o foco da mera identificação do problema para a antecipação de cenários de risco.

## **3. METODOLOGIA**

### **3.1. Delineamento e contexto**

A presente pesquisa, assumiu um delineamento qualitativo, descritivo e de pesquisa-  
ação. Para os fins deste artigo, os dados foram reorganizados para destacar a componente analítica, mantendo a coerência com o desenho original. A dimensão temporal ópera através da comparação entre duas fases: uma predição inicial (dados pré-intervenção) e uma predição final (dados pós-intervenção). O estudo decorreu no ISP-Moxico, com a participação de 20 estudantes do 1.º ano de Ciências de Computação, matriculados na disciplina de Lógica de Programação. A recolha de informação integrou a observação de aulas, a análise documental e uma entrevista semiestruturada aplicada ao docente da disciplina.

### **3.2. Recolha e preparação dos dados**

A análise documental abrangeu processos individuais dos estudantes, actas de avaliação e registos de atividades. Em janeiro de 2025, aplicou-se um questionário de 30 itens aos 20 estudantes. Para a modelagem preditiva, seleccionaram-se as variáveis consideradas críticas para a pesquisa de base: (i) motivo da escolha do curso; (ii) participação nas aulas; (iii) situação de

frequência; (iv) uso da biblioteca; (v) rendimento económico mensal; e (vi) posse de computador. As tabelas de predição incluem ainda género e idade como atributos das interações analisadas.

### 3.3. Processamento e comparação dos algoritmos

Os dados foram estruturados para processamento no ambiente WEKA, comparando cinco algoritmos de classificação: J48, REPTree, NaiveBayes, SMO e Multilayer Perceptron (MLP). O conjunto de treino compreendeu 20 instâncias, submetidas a validação cruzada de 10 partições (10-fold cross-validation).

A seleção do modelo para as predições subsequentes baseou-se no desempenho durante o treino. O algoritmo NaiveBayes atingiu a maior acurácia (90%), classificando corretamente 18 instâncias e errando apenas 2. Os demais algoritmos apresentaram desempenhos inferiores. Nota de correção metodológica: Para garantir a coerência matemática entre a acurácia e o número de instâncias (total de 20), os valores originais foram ajustados. O J48 registou 80% de acurácia (16 classificações corretas e 4 incorretas). Os algoritmos REPTree, SMO e MLP obtiveram 75% de acurácia (15 classificações corretas e 5 incorretas), corrigindo a inconsistência aritmética presente no documento original, que indicava 79% para 15/20 instâncias.

**Tabela 1.** Desempenho dos algoritmos de classificação no treinamento

| Algoritmo  | Acurácia | Classificações corretas | Classificações incorretas |
|------------|----------|-------------------------|---------------------------|
| J48        | 80%      | 16                      | 4                         |
| REPTree    | 75%      | 15                      | 5                         |
| NaiveBayes | 90%      | 18                      | 2                         |
| SMO        | 75%      | 15                      | 5                         |
| MLP        | 75%      | 15                      | 5                         |

**Fonte:** Elaboração própria, com correção aritmética a partir dos resultados da pesquisa, 2025.

## 4. PROCESSO DE INTERVENÇÃO E ESTRATÉGIA DE ANÁLISE

A análise temporal estruturou-se em dois momentos. O primeiro utilizou os dados da fase diagnóstica para gerar a predição inicial de fracasso. Após esta etapa, seguiu-se um ciclo de intervenção (de Outubro de 2025 a Junho de 2026) direcionado às lacunas identificadas. As ações

incluíram: abertura das salas de informática em horário extracurricular, orientação para o uso regular da biblioteca e ao acesso às bases de dados digitais, atribuição de responsabilidades, aulas de reforço (duas vezes na semana), fomento de grupos de estudo (com acompanhamento permanente), seminários sobre assiduidade e orientação vocacional (duas vezes no ano acadêmico).

Após este período, geraram-se novas predições com base nas combinações das características atualizadas. A comparação entre a Percentagem da Predição Inicial (PPI) e a Percentagem da Predição Final (PPF) permite observar a variação do risco predito. É imperativo salientar que esta comparação é de natureza descritiva. A ausência de grupo de controlo, testes inferenciais ou intervalos de confiança impede a atribuição de causalidade estatística entre a intervenção e a redução das percentagens.

## 5. RESULTADOS

### 5.1. Desempenho dos algoritmos

Conforme demonstrado na Tabela 1, o algoritmo NaiveBayes superou os demais modelos, justificando a sua seleção para a etapa preditiva. Convém reiterar que uma acurácia de 90% num conjunto de apenas 20 instâncias reflete o comportamento do modelo neste contexto específico, não constituindo um parâmetro generalizável para outras turmas ou instituições sem validação externa.

6

### 5.2. Predição antes e depois da intervenção

Os Dados apresentados na tabela 2, detalham as 13 interações com as percentagens de fracasso previstas antes e depois do processo de intervenção da investigação. A inspeção direta realizada aos dados recolhidos revela um intervalo real de 50% a 96%. Para assegurar o rigor analítico, este artigo adota os valores desagregados das tabelas como base numérica.

**Tabela 2.** Comparação das predições inicial e final nas 13 interações

| Interação | PPI (%) | PPF (%) | Redução (p.p.) |
|-----------|---------|---------|----------------|
| 1         | 82      | 12      | 70             |
| 2         | 78      | 1       | 77             |

|    |    |    |    |
|----|----|----|----|
| 3  | 67 | 16 | 51 |
| 4  | 50 | 4  | 46 |
| 5  | 85 | 68 | 17 |
| 6  | 67 | 2  | 65 |
| 7  | 85 | 68 | 17 |
| 8  | 92 | 1  | 91 |
| 9  | 90 | 1  | 89 |
| 10 | 82 | 5  | 77 |
| 11 | 96 | 24 | 72 |
| 12 | 81 | 28 | 53 |
| 13 | 69 | 6  | 63 |

**Fonte:** Elaboração própria a partir dos Outputs dos dados treinados pelo Weka, 2026.

#### Legenda:

PPI = Percentagem da Predição Inicial;

PPF = Percentagem da Predição Final.

p.p. = Pontos Percentuais

Por tanto, é importante salientar que, em todas as 13 interações, a percentagem predita de fracasso foi inferior no segundo momento. A média aritmética da PPI foi de 78,77%, caindo para 18,15% na PPF, o que corresponde a uma redução descritiva de 60,62 pontos percentuais. A mediana recuou de 82% para 6%. A maior redução pontual foi de 91 pontos percentuais (de 92% para 1%), e a menor foi de 17 pontos percentuais (de 85% para 68%, observada em duas interações).

## 6. DISCUSSÃO

Os achados confirmam que a mineração de dados auxilia na identificação do fracasso escolar ao converter variáveis discentes em padrões que são classificáveis e indicadores de risco. A seleção do algoritmo NaiveBayes, fundamentada por ter alcançado uma acurácia superior que os demais durante o treino realizado ao conjunto de dados e, validou a etapa de modelagem.

A análise comparativa entre PPI e PPF evidencia uma redução sistemática do risco predito após o ciclo de intervenções pedagógicas realizados com os estudantes. Este resultado

ganha relevância ao demonstrar não apenas um diagnóstico estático, mas a evolução do perfil de risco.

Todavia, impõe-se uma leitura cautelosa. A redução das predições não equivale a uma prova de eficácia causal da intervenção. O estudo carece de grupo de controlo, testes de significância estatística e trabalha com uma amostra reduzida ( $n=20$ ).

## 7. LIMITAÇÕES DO ESTUDO

A principal limitação reside na dimensão reduzida da amostra (20 instâncias) e na ausência de validação externa. A falta de testes inferenciais impede a estimação do nível de significância estatística das diferenças observadas entre os dois momentos. A comparação pré-pós é estritamente descritiva e não controla variáveis de confusão externas. As inconsistências documentais da fonte original, incluindo a alteração de género registada em duas interações entre as tabelas, levaram a que a análise se centrasse exclusivamente nas variáveis PPI e PPF, evitando a reconstrução arbitrária dos atributos individuais.

## 8. CONCLUSÃO

A aplicação de técnicas de mineração de dados revelou-se viável para apoiar a identificação de padrões associados ao fracasso escolar no contexto do ISP-Moxico. Entre os cinco algoritmos testados, o NaiveBayes destacou-se com 90% de acurácia no treino, fundamentando a sua seleção para as predições.

A análise temporal das 13 interações documentadas registou uma diminuição generalizada da percentagem predita de fracasso, com a média a recuar de 78,77% para 18,15%. Estes indicadores sugerem que a mineração de dados pode funcionar como um instrumento válido de diagnóstico e monitoramento de risco académico.

Reitera-se, contudo, que os dados não sustentam afirmações de causalidade entre a intervenção pedagógica e a melhoria das predições. Pesquisas futuras devem privilegiar amostras de maior dimensão, validação cruzada externa, definição clara de variáveis de resultado observadas e métodos estatísticos inferenciais robustos para testar a generalização dos modelos.

## 9. REFERÊNCIAS

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, Nova Iorque, v. 39, n. 11, p. 27-34, 1996.

GOULART, A. Aplicando mineração de dados para predição da evasão escolar no ensino superior em uma instituição federal de ensino. *Revista de Sistemas de Informação da FSMA*, São Paulo, v. 29, p. 12-24, 2022.

HAN, J.; KAMBER, M.; PEI, J. *Data mining: concepts and techniques*. 3. ed. Waltham: Morgan Kaufmann, 2011.

HEREDIA, D.; AMAYA, Y.; BARRIENTOS, E. Student dropout predictive model using data mining techniques. *IEEE Latin America Transactions*, Nova Iorque, v. 13, n. 9, p. 3127-3134, 2015.

KOTSIANTIS, S. Educational data mining: a case study for predicting dropout-prone students. *International Journal of Knowledge Engineering and Soft Data Paradigms*, Inderscience, v. 1, n. 2, p. 101-111, 2009.

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)*, Nova Iorque, v. 40, n. 6, p. 601-618, 2010.

YAĞCI, M. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, Springer, v. 9, n. 1, p. 11, 2022.