Mineração de dados educacionais: integração de bancos de dados e análise de desempenho dos estudantes da UFSM

Luís Gustavo Werle Tozevich[1], Jaime Antonio Daniel Filho[1], Guilherme Meneghetti Einloft[1], Tobias Viero de Oliveira[1], Jefferson Menezes de Oliveira[2], Joaquim Vinicius Carvalho Assunção[3]

[1]Curso de Ciência da Computação, Universidade Federal de Santa Maria,
[2]Universidade Federal de Santa Maria,
[3]Departamento de Computação Aplicada, Universidade Federal de Santa Maria

{lgtozevich, jafilho, gmeinloft, tvoliveira, joaquim}@inf.ufsm.br

jefferson@ufsm.br

Abstract. *This paper investigates the impact of teacher and institutional variables on the pass rates in introductory mathematics courses at UFSM, using historical data (2021–2023). The integration of internal records and public data allowed the application of mining techniques, using k-means to identify clusters among classes. The evaluation of the clusters through ARI and NMI showed that variability in teaching assessments is the main factor associated with discrepancies in pass rates. Structuring data from different sources, we present key insights that pave the way for studies on educational and pedagogical policies.*

Resumo. *Este trabalho investiga o impacto de variáveis docentes e institucionais nas taxas de aprovação em disciplinas iniciais de matemática na UFSM, com dados históricos (2021–2023). A integração de registros internos e dados públicos permitiu aplicar técnicas de mineração, usando k-means para identificar agrupamentos entre turmas. A avaliação dos clusters via ARI e NMI evidenciou que a variabilidade nas avaliações docentes é o principal fator associado às discrepâncias nas taxas de aprovação. Estruturando dados de diferentes fontes, mostramos informações importantes que abrem margem para estudos de políticas educacionais e pedagógicas.*

1. Introdução

A evasão acadêmica em instituições públicas de ensino superior configura-se como um problema multidimensional, com impactos pedagógicos, operacionais e financeiros. Dados indicam que uma parcela expressiva dos ingressantes no ensino superior não conclui os cursos [INEP 2023], evidenciando a complexidade do fenômeno. Na Universidade Federal de Santa Maria (UFSM), observa-se uma variação substancial nas taxas de aprovação em disciplinas básicas de ciências exatas, com diferenças de até 97,73%. Essa discrepância sugere que o desempenho acadêmico nessas disciplinas pode estar correlacionado com a evasão universitária, possivelmente influenciado pela heterogeneidade dos perfis docentes ao longo da trajetória acadêmica dos discentes.

Estudos em *Educational Data Mining* (EDM) demonstram que fatores institucionais, pedagógicos e o perfil docente são determinantes na retenção estudantil

[Romero and Ventura 2020]. Pesquisas sugerem que a experiência docente pode reduzir as taxas de reprovação, embora ocorram variações acentuadas entre diferentes docentes e cursos [Koedinger et al. 2015, Souza et al. 2025]. Assim, o presente estudo propõe investigar o impacto das variáveis docentes, institucionais e do desempenho discente (avaliado por meio do ENEM) sobre as taxas de reprovação em disciplinas iniciais de matemática, contribuindo para a formulação de intervenções pedagógicas mais eficazes e, em um nível macro, para a reformulação de políticas publicas na educação superior.

2. Metodologia

A investigação adotou um delineamento quantitativo-exploratório, utilizando dados históricos anonimizados (2021-2023), fornecidos pela UFSM, totalizando 248 registros de 37 professores e 8504 alunos. O banco de dados original, estruturado em arquivos CSV, foi consolidado em ambiente *SQLite*, seguindo as etapas do *Knowledge Discovery in Databases* (KDD) [Fayyad et al. 1996]. Ademais, para ampliar o escopo analítico, integrou-se bases de dados públicos via *data linkage*, incorporando notas de corte do ENEM da ampla concorrência, a Classificação Internacional Normalizada da Educação (CINE) e indicadores do Exame Nacional de Desempenho dos Estudantes (Enade), como o Conceito Preliminar de Curso (CPC) e o Indicador de Diferença entre os Desempenhos Observado e Esperado (IDD)[1]. Consideraram-se apenas turmas com mais de 10 alunos e, na análise docente, professores que ministraram ao menos três turmas. O esquema do banco de dados é apresentado na Figura 1.

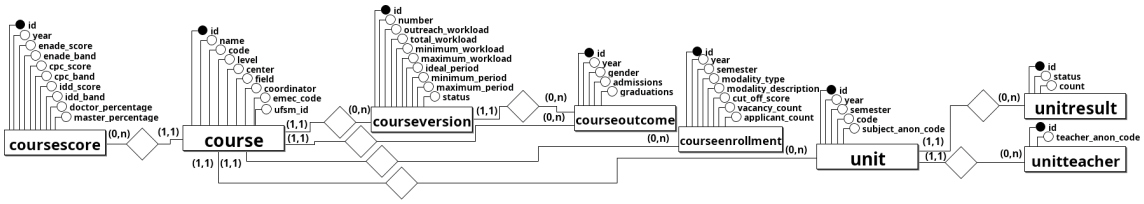

Figura 1. Esquema relacional do banco de dados

Na análise dos dados, adotamos uma abordagem que mescla métodos estatísticos e técnicas de ML (*Machine Learning*), implementados em *Python* 3.13.1, com o apoio das bibliotecas *Pandas*, *Scikit-learn* e *Matplotlib*. Inicialmente, foi calculada a correlação entre turmas da mesma disciplina, com base na taxa média de aprovação e nas notas de corte. Em seguida, aplicou-se o *k-means*, definindo o número de clusters conforme o total de categorias distintas de cada atributo categórico (professor, curso, área CINE e nota ENADE), com o objetivo de comparar os agrupamentos gerados com as classificações originais. Os *clusters* foram formados por meio da distância euclidiana e as variáveis normalizadas com o *StandardScaler*. Para avaliar a consistência dos agrupamentos, empregamos duas métricas de similaridade: o Índice Rand Ajustado (ARI) [Hubert and Arabie 1985], que ajusta a contagem de concordâncias entre duas partições ao considerar a probabilidade de coincidências aleatórias, apresentando valores próximos de 1 quando os agrupamentos são robustos e a Informação Mútua Normalizada (NMI), que mede a quantidade de informação compartilhada entre as partições, normalizando os resultados em uma escala de 0 a 1, onde valores mais elevados indicam uma forte similaridade [Strehl and Ghosh 2002].

[1]https://www.ufsm.br/app/uploads/sites/735/2022/06/Indicadores-de-Qualidade.pdf

3. Resultados e discussões

A investigação teve como objetivo analisar a relação entre a nota obtida no ENEM – principal porta de entrada para as universidades federais brasileiras – e o desempenho acadêmico dos alunos em disciplinas iniciais de matemática, como Cálculo A, Cálculo B e Álgebra Linear. Para isso, representaram-se graficamente as turmas, utilizando um espaço bidimensional no qual a nota de corte do curso foi disposta no eixo horizontal e a porcentagem de alunos aprovados na disciplina no eixo vertical. A análise dos gráficos, como exemplificado na Figura 2, não revelou a existência de correlação linear, uma vez que o coeficiente de Pearson resultou em 0,239 para Cálculo A, 0,121 para Cálculo B e 0,230 para Álgebra Linear. Além disso, observou-se uma grande variação nas taxas de aprovação, independentemente da nota de corte do curso, sugerindo que outros fatores podem influenciar mais significativamente o desempenho dos alunos nessas disciplinas.

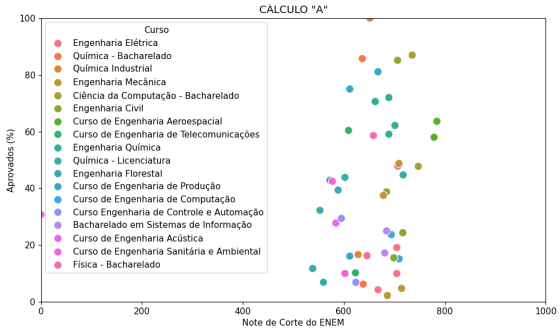

Figura 2. Aprovação por nota de corte no ENEM em Cálculo A

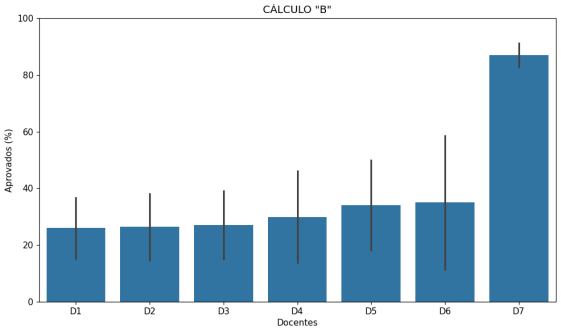

Figura 3. Aprovação por docente em Cálculo B

Na sequência, aplicou-se o algoritmo *k-means* para mensurar os indices NMI e ARI e relação aos atributos categóricos de cada turma, conforme detalhado na seção 2. O índice ARI, que se aproximou de zero em todas as análises, evidenciou a ausência de uma distribuição uniforme nas amostras, enquanto os valores do NMI apontaram para uma possível correlação entre o desempenho dos alunos e o professor, com índices que variaram consideravelmente entre as disciplinas, chegando a 0,64 em Álgebra Linear, 0,42 em Cálculo A e 0,16 em Cálculo B. Em contrapartida, a análise da área CINE e da nota do Enade não trouxe indícios de correlações significativas, pois os respectivos índices aproximaram-se de zero ou apresentaram valores modestos.

A dificuldade em estabelecer padrões claros através do *clustering* e demais técnicas de mineração de dados se deve, primordialmente, à alta variabilidade dos índices de aprovação dos docentes entre diferentes turmas e cursos, o que impede a identificação de agrupamentos consistentes, além de refletir a limitação da quantidade de dados disponíveis para uma análise estatística mais robusta. A fim de evidenciar essa variabilidade, a Figura 3 apresenta a porcentagem de alunos aprovados por professor na disciplina de Cálculo B, representada no eixo vertical. Além disso, as barras são acompanhadas por velas que indicam o desvio padrão, ilustrando a flutuação na taxa de aprovação de cada professor entre diferentes turmas. Essa variabilidade exposta nas velas revelou, contudo, que alguns professores mantêm uma taxa de aprovação constante, enquanto outros apresentam oscilações significativas de acordo com a turma e o curso, fato que sugere a relevância da taxa de aprovação como um potencial indicador da qualidade do ensino.

No entanto, a subjetividade inerente às avaliações e a insuficiência de dados impedem um julgamento conclusivo acerca do desempenho docente ou da eficácia das turmas.

Diante deste cenário, propõe-se que a investigação avance por meio da análise diferenciada dos professores cujas taxas de aprovação se desviam significativamente da média, comparando o desempenho entre suas diferentes turmas e disponibilizando tais informações à coordenação dos cursos. Esse procedimento permitiria compreender melhor as práticas de ensino adotadas e sugerir ajustes que possam contribuir para a melhoria do desempenho acadêmico. De forma complementar, recomenda-se uma investigação direcionada aos docentes que apresentam taxas de aprovação constantes, a fim de avaliar se essa consistência é reflexo de uma abordagem pedagógica eficaz, podendo ser enriquecida com a coleta de *feedbacks* dos alunos para compreender se os métodos empregados atendem de maneira satisfatória às necessidades de aprendizagem.

4. Conclusão

Nesta pesquisa, a integração de dados externos com os registros históricos da UFSM, juntamente à aplicação de técnicas de mineração de dados, evidenciou que variações na conduta e nas avaliações dos docentes podem estar associadas às taxas de aprovação, superando os efeitos dos indicadores institucionais e dos exames padronizados. Além disso, há técnicas e resultados que, devido a limitação de espaço, serão investigados em trabalhos futuros, explorando diferenças de desempenho entre os cursos de graduação e demais relações que corroborem para estipular padrões capazes de prever o desempenho dos discentes. A integração de *feedbacks* dos alunos poderá auxiliar a avaliar se bons desempenhos são reflexo de uma abordagem pedagógica eficaz. Além disso, pode-se aplicar modelos estatísticos mais avançados e técnicas de ML para refinar as predições e fornecer uma base mais sólida para o desenvolvimento de políticas educacionais direcionadas.

Referências

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

INEP (2023). Resumo Técnico do Censo da Educação Superior.

Koedinger, K. R., Kim, J., Jia, J., McLaughlin, E., and Bier, N. (2015). Learning is not a spectator sport: Doing is better than watching for learning from a mooc. *Proceedings of the Second ACM Conference on Learning @ Scale*, pages 111–120.

Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1355.

Souza, M. A. d., Machado, P. A. d. S., Santos, L. C. G. d. S., Lima, T. H. d., Rodrigues, J. M., and Diniz, H. A. G. (2025). A influência da qualificação docente no desempenho acadêmico em cursos de engenharia de produção: análise comparativa regional no brasil. *Revista de Gestão e Secretariado*, 16(1):e4466.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.