AVALIAÇÃO DE MODELOS DE LINGUAGEM DE PEQUENO PORTE PARA A PREDIÇÃO DE EVASÃO EM DADOS TABULARES ACADÊMICOS

FABRÍCIO GONZALES SCAGLIONI[1]; JÚLIO CARLOS BALZANO DE MATTOS[2]; MARILTON SANCHOTENE DE AGUIAR[2]

[1] *Universidade Federal de Pelotas – fabricio.scaglioni@inf.ufpel.edu.br*
[2] *Universidade Federal de Pelotas – {julius, marilton}@inf.ufpel.edu.br*

1. INTRODUÇÃO

A evasão no ensino superior gera desperdícios sociais e econômicos e compromete a formação de capital humano (SILVA FILHO et al., 2007). A identificação precoce de estudantes em risco de abandono fundamenta boa parte das políticas de retenção adotadas pelas instituições, e a predição a partir de dados administrativos e de desempenho já é uma aplicação consolidada da Mineração de Dados Educacionais (BAKHSHINATEGH et al., 2018), apoiada, em sua maioria, em modelos clássicos de aprendizado de máquina como Random Forest, XGBoost e SVM.

Com a consolidação dos modelos de linguagem, a tarefa passou a poder ser tratada também por modelos de linguagem de pequeno porte (SLMs, do inglês *Small Language Models*, entre 7 e 14 bilhões de parâmetros), adaptados por meio de técnicas eficientes de ajuste como a *Low-Rank Adaptation* (LoRA) (HU et al., 2022; DETTMERS et al., 2023). A literatura sobre LLMs aplicados a tarefas preditivas tabulares, contudo, concentra-se em modelos de dezenas ou centenas de bilhões de parâmetros e sem restrição de quantização; falta mensuração empírica sobre a eficácia de SLMs adaptados por LoRA em cenários institucionais com recursos limitados, que exigem quantização de 4 bits e GPUs de entrada.

Este estudo investiga essa lacuna com dados acadêmicos reais da Universidade Federal de Pelotas (UFPel), totalizando 6.011 registros: foram avaliados dois SLMs (Phi-4 e Qwen2.5-7B) ajustados com LoRA, três estratégias de representação tabular, um modelo fundacional especializado (Mitra) e classificadores clássicos, em uma comparação sistemática sob restrição de hardware. As questões de pesquisa são: (RQ1) SLMs ajustados com LoRA são eficazes na predição de evasão? (RQ2) qual representação, entre *scaled*, *labeled* e *storytelling*, maximiza o desempenho preditivo? (RQ3) como os SLMs se comparam a modelos tabulares especializados sob restrições computacionais?

2. METODOLOGIA

2.1. Dataset

O *dataset* reúne registros acadêmicos de graduação da UFPel com predição binária de evasão: “Formado” compõe a classe negativa (0), e abandono, cancelamento e transferência compõem a classe positiva (1). O arquivo original tinha 7.191 registros, dos quais restaram 6.011 após a unificação e a limpeza. A estrutura é transversal: cada linha agrega os totais acumulados de um estudante no primeiro semestre letivo, com uma única linha por aluno e sem séries longitudinais, o que elimina o risco de vazamentos temporais. Concentrar-se no período inicial permite antecipar o risco de evasão o quanto antes na trajetória acadêmica. As variáveis dividem-se entre dados demográficos e de ingresso (idade, estado civil, sexo, cor, cotas, turno, tempo entre o ensino médio e a graduação e área fundamental do curso) e variáveis de desempenho do semestre (disciplinas, créditos, média global, aprovações, dispensas, reprovações, trancamentos, infrequências e exames). A evasão corresponde a cerca de 60% da base limpa, o que torna a classe de abandono majoritária e orienta tanto a escolha do F1-*Score* ponderado (Seção 2.3) quanto o descarte do SMOTE, discutido a seguir.

No pré-processamento, foram removidas instâncias nulas, linhas duplicadas e registros anômalos (idade de ingresso menor ou igual a zero ou maior ou igual a 100, e tempo negativo entre o ensino médio e a graduação), além das colunas identificadoras (*id*, código do curso e situação original). Variáveis contínuas receberam compressão logarítmica (*log1p*); categóricas, *Label Encoding* seguido de padronização. Uma divisão estratificada 80:20, executada uma única vez antes de qualquer experimento, gerou o *holdout* fixo com N=1.203 instâncias, usado por todos os modelos; a partição única, em vez de validação cruzada, evita variância entre arquiteturas heterogêneas e garante comparação sobre o mesmo conjunto de teste. A sobreamostragem (SMOTE) foi descartada: vizinhos sintéticos distorceriam a inferência baseada na prevalência real da evasão.

2.2. Estratégias de representação dos dados

Os três formatos testam hipóteses distintas sobre RQ2 (representação numérica pura, codificação semântica categórica ou linguagem natural densa) e foram derivados estritamente das matrizes do particionamento único, mantidos imutáveis ao longo dos experimentos. No formato *scaled*, os dados seguem o pipeline clássico e alimentam os modelos como pares chave-valor padronizados. No *labeled*, as variáveis contínuas são discretizadas em quintis fixos, sem busca exaustiva por pontos de corte ótimos, e transpostas para chaves semânticas no padrão *between_X_and_Y*. Já no *storytelling*, cada registro é convertido em um parágrafo narrativo em inglês, idioma fixo adotado para reduzir ruído multilingue, seguindo um modelo rígido de texto: *"This student is {age} years old, identifies as {gender} and {color}. The student started the course in {year}. Over this reference semester, they enrolled in {nr_tot} courses with an overall average grade of {media}. Are they at risk of dropping out? Answer strictly yes or no."* Antes do veredito final (*yes* ou *no*), é injetado um encadeamento analítico de raciocínio. As mensagens de sistema foram parametrizadas de forma estática sobre inferência assíncrona em lote, e o veredito foi extraído da resposta textual por expressão regular.

2.3. Modelos e configuração experimental

Foram avaliados o Phi-4 (Microsoft), com cerca de 14 bilhões de parâmetros, e o Qwen2.5-7B (Alibaba Cloud), com cerca de 7 bilhões, ambos com quantização de 4 bits (*bitsandbytes*) para viabilizar o treinamento em GPU NVIDIA Tesla T4 (16 GB de VRAM) no Google Colaboratory, com cerca de uma hora por experimento. O ajuste fino utilizou LoRA (posto 32, *alpha* 16, *dropout* nulo e módulos alvo automáticos) pela biblioteca Unsloth (HAN et al., 2023), com *sequence packing* e *gradient checkpointing* para agrupar *tokens* e economizar VRAM, além de 100 passos, taxa de aprendizado de 2×10^{-4} e decaimento cosseno, *warmup* de 5 passos, *batch* efetivo 8, sequência máxima de 1.024 *tokens* e *seed* 3407, sem parada antecipada. No formato *scaled*, sem cadeia de raciocínio antes do veredito, a taxa padrão levou o Qwen2.5-7B a um colapso por sobreajuste imediato à classe dominante, o que motivou a redução para 2×10^{-5}. Como baseline especializado, o Mitra (ZHANG; ROBINSON, 2025), modelo fundacional tabular que abstrai *features* de domínios desconhecidos por codificação contínua, com arquitetura baseada em árvores de decisão, operou em *zero-shot* sobre os dados brutos, com automação (AutoML) e ramos GBDT desabilitados. *Random Forest*, *XGBoost* e SVM foram avaliados no formato *scaled*, com hiperparâmetros padrão das bibliotecas, mantendo comparabilidade com os SLMs. Ao todo, foram testadas 10 configurações: sete de Mitra e SLMs e três *baselines* clássicos. A métrica primária foi o F1-*Score* ponderado, já que a formulação macro distorce comparações sob amostragem desbalanceada,

enquanto o F1 ponderado consolida o F1 de cada classe conforme o suporte real de teste. Acurácia, precisão, recall e AUC-ROC complementam a análise. Para os SLMs, a AUC foi calculada sobre a predição binária fixa extraída da resposta (*yes*/*no*), sem logits calibrados; valores achatados de AUC não devem ser interpretados como calibração de probabilidades.

3. RESULTADOS E DISCUSSÃO

A Tabela 1 consolida o desempenho de todos os modelos no *holdout* de teste. O Mitra, sobre dados brutos e sem qualquer ajuste, alcançou F1-*Score* de 0,8203 e AUC de 0,8943, resultado equivalente ao do XGBoost, do SVM e do *Random Forest* (F1-*Score* em torno de 0,80 e AUC acima de 0,88). Sua matriz de confusão registrou 541 evasões corretamente identificadas e 55 falsos positivos.

Tabela 1: Desempenho dos modelos avaliados no *holdout* de teste.

Modelo	Formato	Acurácia	Precisão	Recall	F1-*Score*	AUC-ROC
Mitra	Bruto	**0.8030**	**0.8050**	**0.8200**	**0.8203**	**0.8943**
XGBoost	*Scaled*	0.8071	0.8000	0.8100	0.8000	0.8915
SVM	*Scaled*	0.8055	0.8000	0.8200	0.8000	0.8836
Random Forest	*Scaled*	0.8047	0.8000	0.8100	0.8000	0.8973
Phi-4	*Scaled*	0.7207	0.7201	0.7295	0.7176	0.7295
Phi-4	*Labeled*	0.7190	0.7124	0.7200	0.7134	0.7200
Phi-4	*Storytelling*	0.7099	0.7183	0.7264	0.7087	0.7264
Qwen2.5-7B	*Scaled*	0.6293	0.6412	0.8700	0.7383	0.5683
Qwen2.5-7B	*Labeled*	0.7016	0.9062	0.5615	0.6934	0.7370
Qwen2.5-7B	*Storytelling*	0.7232	0.8707	0.6335	0.7334	0.7459

Nota: precisão, recall e F1-*Score* seguem a convenção de origem de cada linha: média macro entre as classes (baselines clássicos e Phi-4) ou valores da classe de evasão (Qwen2.5-7B); no caso do Mitra, o F1-*Score* corresponde à classe de evasão. A AUC-ROC equivale à acurácia balanceada do limiar binário fixo (Seção 2). As colunas, portanto, não são diretamente comparáveis entre linhas.

Frente ao piso de 0,8203 estabelecido pelo Mitra, os SLMs ficaram cerca de 10 pontos percentuais abaixo em F1-*Score*. O Phi-4 obteve o melhor resultado no formato *scaled* (F1-*Score* de 0,7176 e acurácia de 72,07%), identificando 496 das evasões reais do teste, com 109 falsos positivos. Os formatos *labeled* (0,7134) e *storytelling* (0,7087) ficaram praticamente empatados, refutando a hipótese de ganho preditivo da semântica densa (RQ2). O Qwen2.5-7B foi o modelo mais instável. No formato *scaled*, sem cadeia de raciocínio, a taxa padrão levou o modelo a classificar 100% dos casos como evasão; mesmo após a redução para 2×10^{-5}, o F1-*Score* de 0,7383 ainda esconde um forte viés pela classe majoritária: acurácia de 62,93%, AUC de 0,5683, recall de 0,8700 e precisão de 0,6412. No formato *labeled*, a precisão elevada (0,9062) veio acompanhada de recall baixo (0,5615). O maior equilíbrio ocorreu no *storytelling* (acurácia de 72,32%, F1-*Score* de 0,7334 e AUC de 0,7459), superando o melhor resultado do Phi-4 (0,7334 contra 0,7176), o que responde a RQ1 de forma mista: a eficácia dos SLMs depende tanto da representação quanto do modelo específico.

Esses resultados evidenciam o risco de avaliar os modelos por uma métrica isolada em bases desbalanceadas; apenas a leitura conjunta das métricas evita atribuir qualidade preditiva a modelos cujo comportamento é dominado pelo desbalanceamento das classes. Na literatura, o GPT-3.5 ajustado por MA et al., 2024 alcançou F1-*Score* de 0,92 com cerca de 175 bilhões de parâmetros e sem quantização restritiva, contra 0,8203 do Mitra no teto de 14 bilhões e 4 bits deste estudo;

a distância sugere que o gargalo observado em RQ3 é mais de hardware do que de arquitetura.

4. CONCLUSÃO

Este estudo avaliou empiricamente SLMs de 7 a 14 bilhões de parâmetros, ajustados com LoRA, para a predição de evasão a partir de dados tabulares sob restrições severas de hardware. Nas condições avaliadas, a eficácia dos SLMs mostrou-se dependente do modelo e da representação (RQ1): o Phi-4 manteve-se estável nos três formatos, enquanto o Qwen2.5-7B oscilou entre colapso e equilíbrio, chegando a superar o melhor resultado do Phi-4 no formato *storytelling*. Ainda assim, o desempenho ficou cerca de 10 pontos percentuais abaixo dos modelos especializados e clássicos (RQ3), com destaque para o Mitra (F1-*Score* de 0,8203 e AUC aproximada de 0,89). Quanto à representação (RQ2), o efeito não foi uniforme: no Phi-4, os três formatos produziram desempenhos equivalentes, refutando a hipótese de vantagem narrativa; no Qwen2.5-7B, a escolha do formato alterou substancialmente o equilíbrio entre precisão e recall, sugerindo que a representação importa menos para modelos maiores e estáveis e mais para modelos menores e sensíveis à configuração. Diante desses resultados, instituições com orçamento computacional limitado devem priorizar modelos tabulares para tarefas preditivas estruturadas e reservar modelos de linguagem para análises qualitativas complementares. O custo-benefício do ajuste fino precisa ser reavaliado à medida que os SLMs se tornarem mais capazes e as otimizações de inferência superarem os gargalos da quantização; por ora, os modelos especializados permanecem a escolha tecnicamente justificada. Trabalhos futuros incluem escalas intermediárias (30 a 70 bilhões de parâmetros) sem quantização agressiva, arquiteturas híbridas e técnicas de calibração.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BAKHSHINATEGH, B.; ZAIANE, O. R.; ELATIA, S.; IPPERCIEL, D. (2018). **Educational data mining applications and tasks: A survey of the last 10 years**. Em: **Education and Information Technologies** 23, pp. 537–553.

DETTMERS, T.; PAGNONI, A.; HOLTZMAN, A.; ZETTLEMOYER, L. (2023). **QLoRA: Efficient Finetuning of Quantized LLMs**. arXiv: 2305.14314 [cs.LG].

HAN, D.; HAN, M.; TEAM, U. (2023). **Unsloth**.

HU, E. J.; SHEN, Y.; WALLIS, P.; ALLEN-ZHU, Z.; LI, Y.; WANG, S.; WANG, L.; CHEN, W. (2022). **LoRA: Low-rank Adaptation of Large Language Models**. Em: **International Conference on Learning Representations (ICLR)**.

MA, X.; LIU, W.; ZHAO, C.; TUKHVATULINA, L. R. (2024). **Can Large Language Model Predict Employee Attrition?** Em: **Proceeding of the 2024 5th International Conference on Computer Science and Management Technology**. Xiamen Guangdong China: ACM, pp. 1164–1172. ISBN: 979-8-4007-0999-9.

SILVA FILHO, R. L. L. e.; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. d. C. M. (2007). **A evasão no ensino superior brasileiro**. Em: **Cadernos de Pesquisa** 37.132, pp. 641–659. ISSN: 0100-1574.

ZHANG, X.; ROBINSON, D. M. (2025). **Mitra: Mixed Synthetic Priors for Enhancing Tabular Foundation Models**.