Análise Experimental de Modelos Preditivos para a Previsão da Demanda por Vagas na Rede Pública de Ensino.

Lucas Bruno F. Praxedes[1] Lenardo Chaves e Silva[1] Breno Jacinto Duarte da Costa[2] Nicholas Joseph Tavares da Cruz[2] Rafael de Amorim Silva[2] Bruno Almeida Pimentel[2]

[1]Programa de Pós Graduação em Ciência da Computação – Universidade Federal Rural do Semiárido e Universidade do Estado do Rio Grande do Norte (UFERSA-UERN)
Mossoró – RN – Brazil

[2]Universidade Federal de Alagoas (UFAL)
Maceió – AL – Brazil

lucas.praxedes@alunos.ufersa.edu.br, lenardo@ufersa.edu.br,
{breno.duarte,nicholas.cruz,rafael.amorim,bruno.pimentel}@nees.ufal.br

Abstract. *This study focuses on the validation of machine learning models for demand forecasting in basic education, through a multidimensional error analysis. Comparing linear and ensemble algorithms, the evaluation went beyond simple accuracy, focusing on the magnitude and consistency of errors through a suite of metrics (MAE, RMSE, MAPE). The results prove that the LightGBM model exhibits not only high precision but also consistently low errors, which attests to its reliability. It is concluded that this rigorous error evaluation methodology is indispensable for transforming predictive models into safe and practical tools for decision-making in educational management.*

Resumo. *Este estudo foca na validação de modelos de machine learning para a previsão de demanda na educação básica, através de uma análise de erro em várias dimensões diferentes. Comparando algoritmos lineares e de comitês, a avaliação não se limitou a simples acurácia, concentrando também na magnitude dos erros por meio de um conjunto de métricas (MAE, RMSE, MAPE). Os resultados comprovam que o LightGBM possui não apenas alta precisão, mas também erros consistentemente baixos, o que atesta sua confiabilidade. Conclui-se que esta metodologia de avaliação rigorosa do erro é importante para transformar modelos preditivos em ferramentas práticas para a tomada de decisão na gestão educacional.*

1. Introdução

A evasão escolar no Brasil vai além da questão pedagógica, representando um problema socioeconômico de primeira ordem. Com um custo anual para o país de R$ 214 bilhões (3% do PIB) e uma perda de renda vitalícia individual de R$ 372 mil por jovem evadido, o impacto é alarmante (Insper; Fundação Roberto Marinho, 2022). Essa interrupção educacional perpetua ciclos de desigualdade, associando-se a piores contextos sociais e de saúde. Para as instituições de ensino, a evasão compromete a saúde financeira e representa o uso ineficiente de fundos públicos, um ciclo vicioso agravado por cortes orçamentários que ameaçam os programas de assistência estudantil.

Diante deste cenário, a Mineração de Dados Educacionais (EDM) aparece como uma resposta científica, utilizando aprendizado de máquina para desenvolver modelos preditivos. Em vez de uma abordagem reativa, a modelagem preditiva permite uma postura proativa, identificando estudantes em alto risco de evasão antes que o abandono

ocorra, com base em seus dados históricos (Shao et al., 2023). A tese central desta revisão é que a aplicação rigorosa e ética de técnicas de ML é uma estratégia interessante para diminuir a evasão. A avaliação rigorosa dessas previsões, por meio de um conjunto abrangente de métricas de erro, é fundamental para garantir a confiabilidade e a utilidade prática desses modelos, transformando dados brutos em previsões úteis que podem quebrar o ciclo de perdas econômicas e sociais.

2. Conceitos Fundamentais

A construção de modelos preditivos eficazes depende de algoritmos fundamentais, com destaque para os métodos de comitês, que criam um "aprendiz forte" a partir da combinação de múltiplos "aprendizes fracos". O *Random Forest*, proposto por Breiman (2001), é um exemplo de comitê que opera em paralelo, construindo árvores de decisão descorrelacionadas. Sua eficiência deriva do uso de *bagging*, que é a amostragem com reposição, e da seleção aleatória de um subconjunto de atributos em cada nó, o que resulta em um modelo final com baixo erro de generalização e alta capacidade de capturar relações não-lineares.

Em contraste, as *Gradient Boosting Machines* (GBM) adotam uma abordagem sequencial. Implementações eficientes como o *LightGBM* treinam cada nova árvore para corrigir os erros residuais do comitê anterior, focando nas instâncias com maior erro. Esse processo aditivo, onde o modelo é refinado a cada passo, pode ser representado pela seguinte equação:

$$f_m(x) = f_{m-1}(x) + \nu \cdot h_m(x)$$

Nesta equação, $f_m(x)$ é o modelo na etapa m, $f_{m-1}(x)$ é o modelo anterior, $h_m(x)$ é a nova árvore de decisão e ν é a taxa de aprendizado que controla a contribuição de cada árvore. Essa técnica frequentemente alcança maior precisão, mas exige um ajuste cuidadoso de hiperparâmetros para evitar *overfitting* (Friedman, 2001).

2.1. Preparação de dados e de atributos

A preparação de dados é um pré-requisito para uma modelagem confiável, transformando dados brutos em um formato limpo através do pré-processamento. O processo inclui a seleção de atributos com Informação Mútua (MI), que captura dependências não-lineares e quantifica a redução na incerteza sobre a variável alvo, sendo definida como demonstrado abaixo. (Cover & Thomas, 2006).

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Outras etapas importantes são a remoção de atributos sem variância, a transformação logarítmica para estabilizar dados assimétricos, a imputação de valores ausentes pela mediana e a aplicação de Winsorização para diminuir a influência de *outliers*.

2.2. Validação de Modelos e Métricas de Erro para Regressão

Para evitar *overfitting*, a validação cruzada K-Fold estratificada é o padrão-ouro para a validação interna (Kohavi, 1995), garantindo que o modelo generalize para dados não vistos dentro da mesma distribuição. Embora a validação externa (com dados de outros

períodos ou regiões) seja o teste definitivo para a generalização, a validação interna bem executada é um pré-requisito para confirmar a estabilidade do modelo.

No contexto de regressão, a avaliação não se baseia em acurácia, mas em um conjunto de métricas de erro que fornecem uma visão completa da performance. Neste estudo, foram utilizadas as métricas padrão para tarefas de regressão: Erro Absoluto Médio (MAE), que mede o erro médio absoluto; Raiz do Erro Quadrático Médio (RMSE), que penaliza erros maiores; Erro Percentual Absoluto Médio (MAPE), que expressa o erro em termos percentuais; e o Coeficiente de Determinação (R^2), que indica a proporção da variância explicada pelo modelo. A análise conjunta dessas métricas permite uma avaliação robusta e multidimensional do erro.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

2.3. Aplicação do contexto

A pesquisa em Mineração de Dados Educacionais (EDM) no Brasil utiliza microdados do INEP para criar modelos preditivos. O foco atual da área está no aprimoramento da avaliação e na interpretabilidade dos resultados, garantindo sempre a conformidade com a LGPD através da anonimização dos dados.

2.4. Interpretabilidade de Modelos com SHAP

Além de avaliar a performance preditiva, é crucial entender como os modelos chegam a suas conclusões. A interpretabilidade abre a "caixa-preta" de algoritmos complexos como o *LightGBM*. Para isso, utiliza-se a abordagem SHAP (*SHapley Additive exPlanations*), uma técnica baseada na teoria dos jogos que atribui a cada atributo um valor de importância para uma previsão específica (Lundberg & Lee, 2017). O SHAP garante que as explicações sejam consistentes e precisas, permitindo analisar tanto o impacto global de cada variável no modelo quanto sua influência em previsões individuais.

3. Metodologia

O experimento foi executado em um computador com 70GB de memória RAM, Processador *Intel Xeon* 2.20Ghz de 20 núcleos e Sistema Operacional *Canonical Ubuntu* 22.04, hospedado por meio de uma *Virtual Machine* do Laboratório de Redes e Sistemas Distribuídos da Universidade do Estado do Rio Grande do Norte. A metodologia deste estudo foi estruturada como um algoritmo computacional em *Python* (usando *Pandas, NumPy* e *Scikit-learn*), projetado para garantir a reprodutibilidade. O fluxo de trabalho foi dividido em três fases: Preparação, Modelagem e Avaliação, conforme ilustrado na Figura 1.

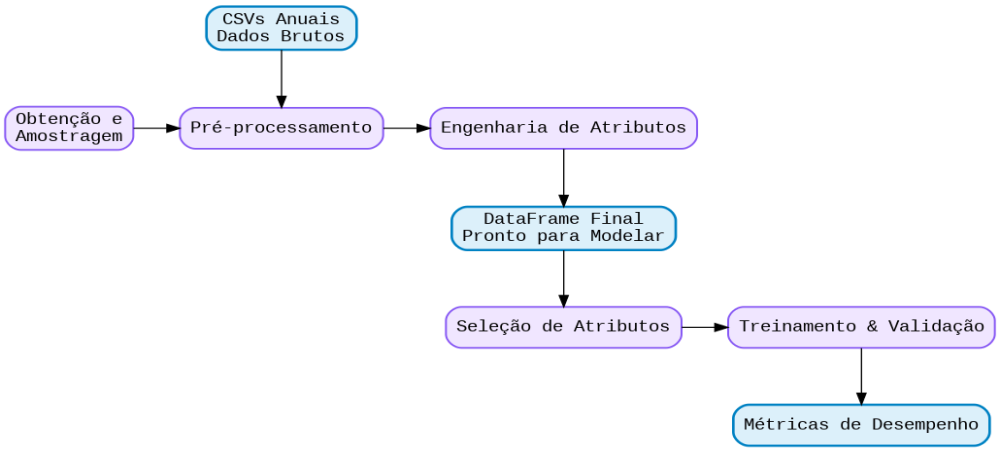

Figura 1 – Algoritmo Metodológico Completo, da obtenção dos dados à avaliação final.

3.1. Obtenção e Preparação dos Dados

A base de dados foi construída a partir dos Microdados do Censo da Educação Básica do INEP (2011-2021). Foi extraída uma amostra aleatória de 30% dos códigos de escolas únicos. A seguir, executou-se o pré-processamento:

1. Limpeza Inicial: Remoção de colunas com mais de 90% de valores nulos e de atributos com variância nula após os tratamentos, pois não possuem poder preditivo.
2. Tratamento de Indicadores: Imputação de valores ausentes (*NaN*) como 0 em colunas binárias (prefixo *IN_*), partindo do pressuposto de que a ausência de informação implica a não existência do atributo.
3. Codificação Categórica: Atributos textuais foram convertidos para números via codificação de rótulos. Colunas com alta cardinalidade (>200 valores únicos), geralmente identificadores, foram removidas.
4. Tratamento de *Outliers*: Para mitigar o efeito de valores extremos em variáveis quantitativas (*QT_*), foi aplicada a técnica de Winsorização, substituindo valores acima do percentil 99 pelo próprio valor do percentil.
5. Imputação Final: Valores ausentes restantes em colunas numéricas foram preenchidos com a mediana da respectiva coluna, escolhida por sua eficácia a *outliers* em distribuições assimétricas.

3.2. Manipulação de Atributos e Modelagem

Foram criados atributos defasados (*_lag1*) e de taxa de crescimento (*_growth_rate_lag1*). A modelagem foi estruturada para prever a quantidade de matrículas (*QT_MAT*) para cada etapa de ensino. O ciclo de experimentação, detalhado na Figura 2, envolveu os seguintes passos:

1. Seleção de Atributos: Uso de Informação Mútua para selecionar os 25 atributos mais relevantes.
2. Transformação da Variável Alvo: Aplicação da função logarítmica *log1p* para estabilizar a variância.
3. Treinamento Comparativo: Foram treinados e comparados três algoritmos: Regressão Linear, *LightGBM* e *Random Forest Regressor*. O desempenho de

cada modelo foi quantificado utilizando as métricas de erro detalhadas na seção 2.2, permitindo uma análise comparativa robusta de suas capacidades preditivas.

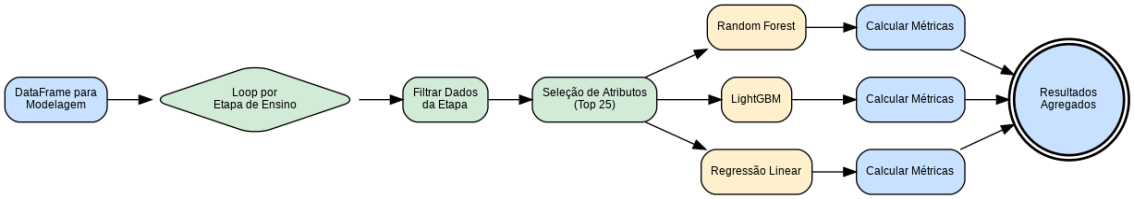

Figura 2 – Ciclo de experimentação para cada etapa de ensino, incluindo seleção de atributos, treinamento com múltiplos algoritmos e validação cruzada.

4. Resultados

A estratégia de modelagem foi aplicada com sucesso a 10 etapas de ensino distintas. A análise comparativa concentra-se na capacidade dos modelos em minimizar os erros de previsão.

4.1. Análise Comparativa Geral dos Modelos

A Figura 3 apresenta um panorama do desempenho agregado dos modelos. O gráfico utiliza um eixo Y duplo para comparar simultaneamente as métricas de erro (RMSE, MAE, MAPE) no eixo esquerdo e a métrica de acurácia (R^2) no eixo direito.

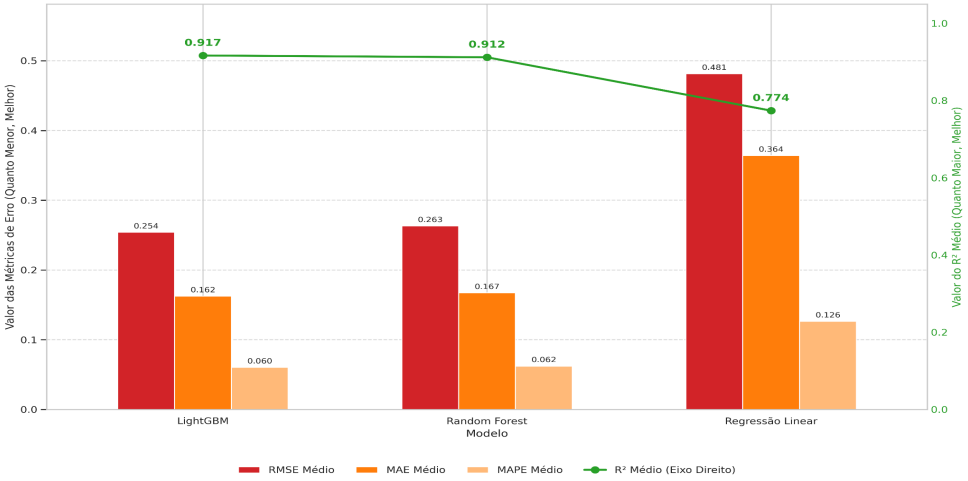

Figura 3 – *Comparação Agregada de Métricas de Desempenho e Erro. As barras de erro (vermelho, laranja) estão no eixo esquerdo, enquanto a barra de R^2 (verde) está no eixo direito. Valores menores são melhores para as métricas de erro, e maiores são melhores para o R^2.*

A visualização evidencia uma clara distinção de performance. Os modelos de comitê, *LightGBM* e *Random Forest*, superam a Regressão Linear de forma expressiva em todas as métricas. Eles não apenas alcançam um R^2 médio superior a 0.91, indicando alta acurácia, como também apresentam erros (RMSE, MAE, MAPE) drasticamente menores. Entre os dois, o *LightGBM* supera com uma ligeira vantagem, registrando os menores valores médios em todas as métricas de erro.

4.2. Análise Detalhada do Erro por Etapa de Ensino

Para aprofundar a análise, a Tabela 1 detalha o desempenho do melhor modelo, *LightGBM*, em cada uma das 10 etapas de ensino.

Etapa	R^2 Médio	RMSE Médio	MAE Médio	MAPE Médio
IN_FUND	0.981	0.186	0.12	0.036
IN_MED	0.972	0.167	0.107	0.026
IN_FUND_AI	0.971	0.209	0.136	0.044
IN_FUND_AF	0.968	0.198	0.12	0.037
IN_INF_PRE	0.932	0.3	0.21	0.099
IN_INF_CRE	0.918	0.336	0.229	0.113
IN_PROF_TEC	0.902	0.29	0.171	0.051
IN_EJA_FUND	0.887	0.32	0.228	0.072
IN_EJA_MED	0.866	0.303	0.191	0.056
IN_ESP_CE	0.77	0.23	0.112	0.069

Tabela 1 – *Desempenho Detalhado do Modelo LightGBM por Etapa de Ensino.*

Os resultados confirmam a eficiência do *LightGBM*, que mantém um R^2 acima de 0.86 e baixos erros na maioria dos contextos. O desempenho é notável nas etapas centrais do sistema educacional, como Ensino Fundamental (*IN_FUND*) e Médio (*IN_MED*), onde os erros são mínimos.

Para visualizar a performance comparativa, a Figura 4 apresenta um *heatmap* do RMSE para todos os modelos e etapas.

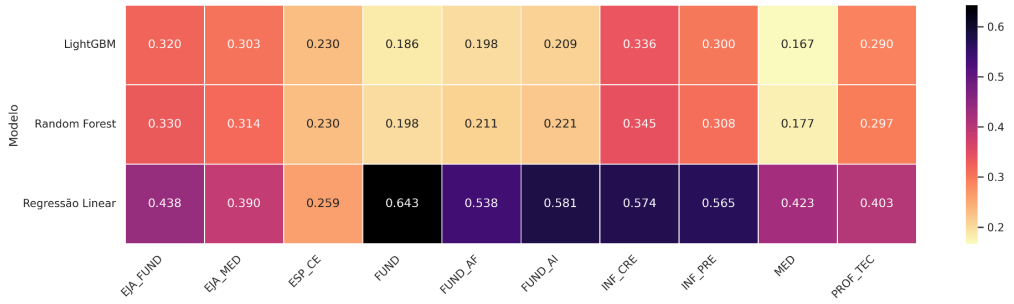

Figura 4 – *Análise de Erro (RMSE Médio) por Modelo e Etapa de Ensino. Cores mais escuras indicam menor erro (melhor desempenho).*

O *heatmap* ilustra de forma conclusiva a superioridade dos modelos de comitês. As células da Regressão Linear apresentam cores mais claras, indicando erros significativamente mais altos em todos os cenários. Em contraste, *LightGBM* e *Random Forest* exibem cores escuras, representando baixo erro. A figura também permite identificar outros aspectos, como o desempenho ligeiramente inferior dos modelos na etapa de Educação Especial (*IN_ESP_CE*), que, embora ainda bom, apresenta um erro relativo maior, possivelmente devido à menor quantidade de amostras e maior variação dos dados.

4.3. Análise de Interpretabilidade (SHAP)

Para abrir a "caixa-preta" do modelo, a análise com SHAP foi aplicada ao *LightGBM*. A Figura 5 apresenta a importância geral dos atributos, revelando padrões consistentes, sendo o eixo horizontal a representação do valor médio absoluto do SHAP, e o eixo vertical a representação dos atributos que o modelo usou. Em todas as etapas, o preditor de maior impacto foi o número de matrículas do ano anterior (ex:

QT_MAT_FUND_lag1), confirmando uma forte inércia histórica. Em segundo lugar, surgem variáveis de estrutura escolar, como o número de turmas (ex: *QT_TUR_FUND*). Isso indica que o modelo aprendeu que a demanda futura é fortemente influenciada pelo histórico e pela capacidade instalada da escola. Notavelmente, para a EJA, atributos geográficos como *CO_CEP* ganharam relevância, sugerindo uma demanda mais sensível a fatores locais. Essa clareza sobre os fatores que direcionam a previsão é fundamental para transformar o resultado do modelo em ações de gestão concretas, como será detalhado na discussão a seguir.

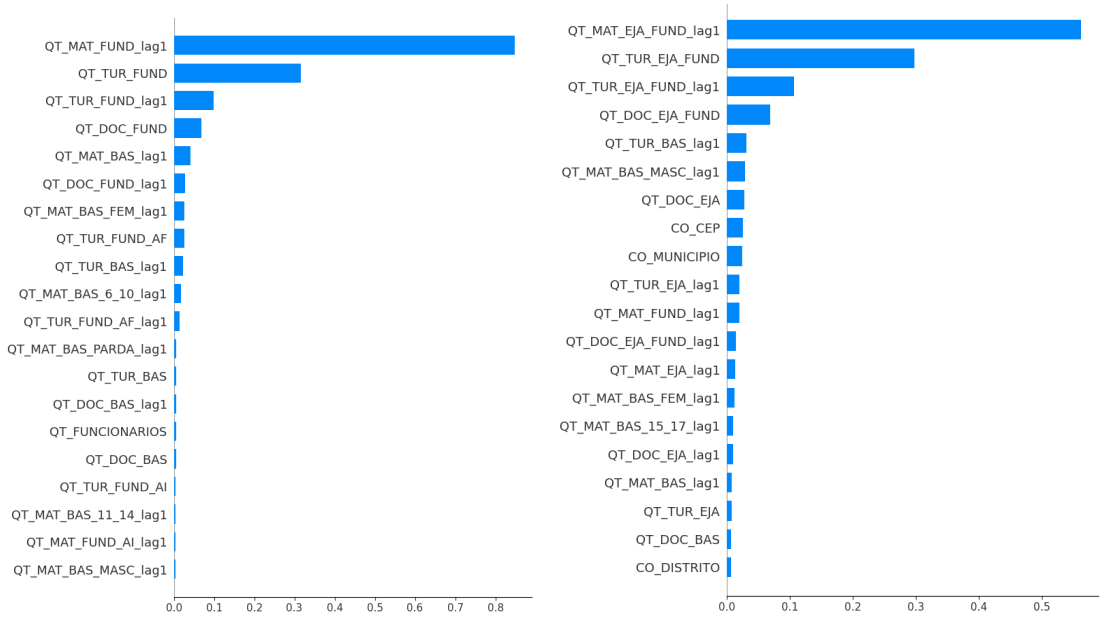

Figura 5 – *Importância geral dos atributos segundo o SHAP para o Ensino Fundamental (gráfico à esquerda) e Ensino EJA de nível fundamental (à direita).*

5. Considerações finais

Este estudo demonstrou a eficácia superior de modelos de comitês, com destaque para o *LightGBM*, na previsão da demanda por matrículas na educação básica. A validação do modelo, contudo, foi além da alta acurácia (R^2), focando na magnitude dos erros de previsão (RMSE, MAE e MAPE). A confiabilidade do *LightGBM* foi confirmada pela sua capacidade de manter erros consistentemente baixos, como um MAPE médio de apenas 3.6% no Ensino Fundamental, fornecendo aos gestores educacionais uma ferramenta de alta confiabilidade para o planejamento de recursos.

A análise de vários ângulos do erro permitiu uma compreensão mais profunda da performance do modelo, respondendo à pergunta principal de qual a margem de erro esperada na prática, algo que o R^2, de forma isolada, não consegue fazer. A análise de interpretabilidade com SHAP serviu para conectar os resultados à prática gerencial. Ao confirmar que as previsões são guiadas por fatores como a inércia de matrículas e o número de turmas, o modelo transforma a previsão numérica em um diagnóstico claro. Isso oferece aos gestores uma base de evidências para, por exemplo, planejar a alocação de recursos ou a expansão da infraestrutura com maior segurança. Concluímos que a combinação de algoritmos avançados e interpretabilidade gera modelos de alto desempenho e, mais importante, de aplicabilidade real. O estudo reforça que uma avaliação focada em múltiplas métricas de erro é um requisito para validar a utilidade de

modelos de Mineração de Dados Educacionais no mundo real, podendo garantir que as decisões sejam baseadas em previsões confiáveis.

Apesar da alta performance, a principal limitação do estudo é a ausência de uma validação externa, etapa importante para a implementação prática. A validação cruzada *K-Fold* assegurou o desempenho interno do modelo, mas sua generalização para novos períodos ou contextos geográficos precisa ser testada. Portanto, os trabalhos futuros devem priorizar essa validação externa, aplicando o modelo a dados de anos subsequentes e de outras regiões. A integração de dados socioeconômicos, o aprofundamento com técnicas de XAI e a condução de um estudo piloto em um cenário real são bons passos para transformar o protótipo em uma ferramenta de gestão precisa, idealmente com recursos computacionais que permitam a análise em maior escala de dados.

Referências

BREIMAN, L. Random Forests. Machine Learning, v. 45, n. 1, p. 5–32, 2001.

COVER, T. M.; THOMAS, J. A. Elements of Information Theory. 2. ed. Hoboken, NJ: Wiley-Interscience, 2006.

HOERL, A. E.; KENNARD, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, v. 12, n. 1, p. 55–67, 1970.

FUNDAÇÃO ROBERTO MARINHO; INSPER. Consequências da Violação do Direito à Educação. Rio de Janeiro: Fundação Roberto Marinho, 2021.

KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2. Montreal, Canada: Morgan Kaufmann Publishers Inc., 1995. p. 1137–1143.

KOTSIANTIS, S. B. Supervised Machine Learning: A Review of Classification Techniques. Informatica, v. 31, n. 3, p. 249–268, 2007.

SHAO, L. et al. Machine Learning Methods for Course Enrollment Prediction. Strategic Enrollment Management Quarterly

SOUZA, A. M. et al. Beyond scores: A machine learning approach to comparing educational system effectiveness. PLOS ONE, v. 18, n. 10, p. e0289260, 2023.

LUNDBERG, S. M.; LEE, S.-I. A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017.