

Governança de Machine Learning e IA Preditiva na Educação Pública: operacionalizando equidade, transparência e responsabilização

David Melo da Luz - USP
Leila Rentroia Iannone - PUC-SP
Natalia de Mesquita Matheus - Instituto Par

Resumo: Este artigo discute como operacionalizar equidade, transparência e responsabilização na governança de sistemas preditivos de *machine learning* aplicados à educação pública. O ponto de partida é um resultado empírico anterior, no qual modelos treinados apenas com variáveis do questionário socioeconômico dos microdados do ENEM 2023 atingiram *F1-score* em torno de 0,76 na identificação do risco de desempenho abaixo da média. Esse sinal expressa desigualdades estruturais, e não relações causais. Ancorado na perspectiva da política como *enactment*, o trabalho analisa três casos críticos de falha no uso de sistemas preditivos no setor público (o *Dropout Early Warning System* de Wisconsin, o algoritmo de notas do Ofqual e o escândalo das creches nos Países Baixos) e um contraponto ilustrativo com salvaguardas, confrontando-os com instrumentos de governança algorítmica já estabelecidos. Como contribuição, propõe-se um conjunto de instrumentos aplicáveis por redes e gestores: Carta do Modelo, matriz de usos recomendados e proibidos, auditoria por subgrupos e implantação em ondas com revisão contínua, articulados ao marco regulatório brasileiro (LGPD e PL 2338/2023) e europeu (AI Act). Conclui-se que a legitimidade pública desses sistemas depende, antes, das condições de implementação que mantêm seu uso agregado, auditável e não punitivo.

Palavras-chave: governança algorítmica; *machine learning*; educação pública; equidade; responsabilização

Governance of Machine Learning and Predictive AI in Public Education: operationalizing equity, transparency, and accountability

Abstract: This article discusses how to operationalize equity, transparency, and accountability in the governance of predictive *machine learning* systems applied to public education. The starting point is a prior empirical result in which models trained only with socioeconomic questionnaire variables from the 2023 ENEM microdata reached an *F1-score* of around 0.76 in identifying the risk of below-average performance. This predictive signal expresses structural inequalities rather than causal relationships. Drawing on the perspective of policy as *enactment*, the study analyzes three critical failure cases of predictive systems in the public sector (Wisconsin's *Dropout Early Warning System*, the United Kingdom's Ofqual grading algorithm, and the Dutch childcare benefits scandal) and one illustrative counterpoint with safeguards, confronting them with established instruments of algorithmic governance. As a contribution, a set of instruments applicable by school networks and managers is proposed: a Model Charter, a matrix of recommended and prohibited uses, subgroup auditing, and wave-based deployment with continuous review, articulated with the Brazilian (LGPD and Bill 2338/2023) and European (AI Act) regulatory frameworks. The article concludes that the public legitimacy of these systems depends chiefly on the implementation conditions that keep their use aggregated, auditable, and non-punitive.

Keywords: algorithmic governance; machine learning; public education; equity; accountability

1. Introdução

A crescente disponibilidade de microdados públicos de larga escala e o avanço do poder computacional têm impulsionado o uso de *machine learning* e de inteligência artificial preditiva na formulação e no monitoramento de políticas públicas educacionais (RAHAL; VERHAGEN; KIRK, 2024). Quando essas ferramentas migram do ambiente de pesquisa para a gestão pública, porém, a pergunta central deixa de ser se o modelo funciona e passa a ser sob quais condições ele produz benefício público com legitimidade, equidade e transparência.

A pesquisa que origina este trabalho (LUZ, 2025) investigou em que medida é possível antecipar o desempenho de estudantes no ENEM a partir apenas de suas condições socioeconômicas. Utilizando os microdados do ENEM 2023 e somente as variáveis do questionário socioeconômico, o estudo definiu como alvo a identificação de estudantes com

desempenho abaixo da média e treinou modelos de aprendizado de máquina, com o uso de *embeddings* semânticos para representar variáveis categóricas. Os modelos alcançaram *F1-score* em torno de 0,76, métrica que equilibra precisão e revocação, o que indica sinal preditivo consistente antes mesmo de o estudante iniciar a prova.

Esse sinal resulta do modo como o modelo combina variáveis e reconhece regularidades, e não constitui evidência de causalidade. Quando atributos como renda, raça/cor, gênero, escolaridade dos responsáveis, tipo de escola e território aparecem entre os preditores relevantes, o que se observa é o reflexo de desigualdades estruturais registradas nos dados. Por isso, o resultado se presta a uma leitura populacional, voltada a mapear vulnerabilidades e orientar prioridades de política em escala, e não a decisões sobre estudantes tomados individualmente.

Em termos de desenho, o estudo definiu “desempenho abaixo da média” como nota global inferior a 522 pontos, a média do ENEM 2023, e tratou a tarefa como classificação binária, com classes aproximadamente equilibradas, o que levou a priorizar a acurácia balanceada, isto é, a média dos acertos em cada classe, e o *F1-score* da classe de risco. Foram comparados modelos de referência (classificador trivial, regressão logística e floresta aleatória) e modelos de *boosting*, família de algoritmos que combina em sequência muitos modelos simples (XGBoost, LightGBM e CatBoost), com e sem *embeddings* semânticos, representações numéricas geradas a partir dos próprios atributos socioeconômicos convertidos em texto (LUZ, 2025).

No conjunto de teste, com 628.793 estudantes, o melhor modelo, uma variante do CatBoost (C-CatBoost), alcançou ROC-AUC, medida da capacidade de ordenar corretamente os casos de maior e de menor risco, de 0,816, acurácia balanceada de 0,736 e *F1-score* de 0,758 na classe de risco, superando a regressão logística (ROC-AUC de 0,805) e a floresta aleatória (0,801) por margens pequenas; a inclusão dos *embeddings* trouxe ganhos apenas marginais. A convergência dos modelos de *boosting* em torno dos mesmos valores sugere retornos decrescentes e um possível teto empírico para a predição baseada apenas em variáveis socioeconômicas. A qualidade probabilística do melhor modelo (*log loss* de 0,522; PR-AUC de 0,822), com limiar de decisão de 0,387, o ponto de corte que converte a probabilidade estimada em classificação, indica o cuidado necessário ao transformar probabilidades em decisões individuais (LUZ, 2025).

O problema de pesquisa nasce na passagem da evidência técnica para a política implementada. Mesmo com bom desempenho preditivo e interpretação normativa cuidadosa, a implementação pode inverter o sentido do instrumento. Pressões por listas nominais, usos fora da finalidade declarada, ausência de auditoria por subgrupos, baixa transparência sobre dados e critérios e arranjos de contratação que bloqueiam a revisão contínua são fatores capazes de transformar um instrumento de diagnóstico em mecanismo de classificação e punição.

A pergunta que orienta este artigo pode ser formulada assim: quais mecanismos de governança e implementação, expressos em regras de uso, rotinas de auditoria, controles de acesso, práticas de comunicação e arranjos de responsabilização, permitem que sistemas preditivos baseados em *machine learning*, treinados com variáveis socioeconômicas, sejam utilizados de forma agregada na educação pública, com equidade, transparência e prestação de contas, sem efeitos de individualização e estigmatização? O objetivo é duplo: analisar casos documentados de uso de sistemas preditivos no setor público, identificando padrões recorrentes de falha de governança, e propor, a partir dessa análise e de instrumentos já estabelecidos na literatura, um conjunto de salvaguardas aplicáveis por redes e gestores públicos brasileiros.

2. Referencial teórico

O uso de sinais preditivos na gestão educacional situa-se no encontro entre três corpos de evidência e de teoria: a documentação consistente da associação entre condições socioeconômicas e desempenho, o amadurecimento de métodos capazes de explorar relações múltiplas e não lineares em grandes bases, e um campo normativo que condiciona o uso legítimo desses sinais à finalidade explícita, à transparência e à responsabilização. Os dois eixos

a seguir organizam esses elementos: o analítico, que trata a implementação como prática situada, e o técnico-normativo, que aproxima capacidade preditiva e legitimidade pública.

2.1 Política como enactment e a centralidade das métricas

No plano analítico, este trabalho ancora-se na sociologia das políticas educacionais e entende a implementação como tradução e recontextualização, e não como execução linear de um desenho técnico. As políticas operam como textos e discursos que ganham sentido na prática, interpretados e reconstruídos pelos atores em cada contexto (BALL, 1993; BALL; MAGUIRE; BRAUN, 2012). Mainardes e Gandin (2013) destacam a pertinência dessa abordagem no Brasil para examinar efeitos, inclusive distributivos, das políticas educacionais.

A perspectiva é produtiva para a governança de sistemas preditivos porque desloca a atenção do artefato técnico para as condições institucionais de seu uso. Um modelo de bom desempenho pode produzir efeitos opostos aos pretendidos conforme é traduzido em rotinas, conforme os incentivos que cercam os gestores e conforme o lugar que as métricas ocupam na prática. A centralidade dos indicadores e a performatividade que dela decorre podem converter instrumentos de diagnóstico em mecanismos de pressão e classificação (BALL, 2003). Compreender a governança de IA na educação pública como prática situada, atravessada por disputas, é condição para desenhar salvaguardas que resistam às pressões do dia a dia da implementação.

2.2 Predição, viés e o requisito de legitimidade pública

No plano técnico, *machine learning* designa um conjunto de métodos em que algoritmos aprendem padrões a partir de dados para executar uma tarefa sob critérios de desempenho (MITCHELL, 1997). Em bases educacionais extensas, marcadas com frequência por relações múltiplas e não lineares, esses métodos podem oferecer ganhos preditivos em relação a abordagens tradicionais (FERNÁNDEZ-DELGADO et al., 2014; ALMALAWI et al., 2024). No Brasil, a literatura documenta de forma consistente que desigualdades sociais e territoriais se refletem em resultados acadêmicos distintos, inclusive em estudos com microdados do ENEM (NOVAES; PITON-GONÇALVES, 2021; RUBIM; MANTOVANI; ALAVARSE, 2025).

Esse potencial vem acompanhado de um risco bem caracterizado. Ao aprender com dados que carregam desigualdades históricas, os modelos podem reproduzir e até amplificar vieses, com efeitos distributivos assimétricos sobre grupos definidos por raça, gênero e renda (ALMALAWI et al., 2024; SIMÕES-GOMES; ROBERTO; MENDONÇA, 2020).

A consequência normativa é direta: predição não equivale a causalidade, e o uso de sinais preditivos no setor público só se sustenta quando acompanhado de finalidade explícita, transparência, auditoria e responsabilização. O campo da analítica da aprendizagem antecipou parte dessas exigências. Slade e Prinsloo (2013) sistematizam dilemas éticos do uso de dados de estudantes, e Drachler e Greller (2016) propõem a lista de verificação DELICATE, que organiza em oito pontos princípios de transparência, consentimento, finalidade e confiança para uma analítica responsável. Esses referenciais fornecem o vocabulário para traduzir princípios abstratos em rotinas verificáveis.

2.3 Trabalhos relacionados

Três frentes de literatura dialogam com esta pesquisa. A primeira reúne estudos de predição educacional: Bowers, Sprott e Taff (2013) mostraram a grande variação de precisão entre sistemas de alerta precoce de evasão, Almalawi et al. (2024) confirmam a predominância de trabalhos centrados no desempenho dos modelos e, no Brasil, estudos com microdados do ENEM (NOVAES; PITON-GONÇALVES, 2021; RUBIM; MANTOVANI; ALAVARSE, 2025) documentam o peso das condições socioeconômicas sobre os resultados. A segunda trata de viés e equidade algorítmica na educação, com o mapeamento de Baker e Hawn (2022) sobre as origens do viés nesses sistemas e o balanço brasileiro de Simões-Gomes, Roberto e Mendonça (2020). A terceira propõe instrumentos de governança, como os *Model Cards* (MITCHELL et al., 2019), os *datasheets for datasets* (GEBRU et al., 2021), a lista DELICATE (DRACHSLER; GRELLER, 2016) e as rotinas de auditoria interna de Raji et al. (2020).

Nos trabalhos examinados, essas frentes aparecem tratadas de forma separada. A contribuição deste artigo está em articulá-las em um mesmo desenho: parte-se de um sinal preditivo obtido em base pública brasileira de larga escala, confronta-se esse sinal com casos documentados de implementação e traduz-se o resultado em instrumentos operacionais ajustados à LGPD e ao PL 2338/2023.

3. Material e Métodos

Adota-se um percurso teórico-propositivo, organizado em quatro movimentos articulados, cujo produto é um conjunto de instrumentos de governança e uma agenda de investigação empírica. O desenho prioriza a aplicação em diferentes redes e contextos, com mecanismos concretos para que o uso de sistemas preditivos permaneça agregado, equitativo e auditável.

O primeiro movimento toma o estudo que origina este artigo (LUZ, 2025) como caso de referência. Dele se extraem, de forma sistematizada, a finalidade adequada do sistema, os limites de interpretação do sinal preditivo, os riscos de uso indevido e os requisitos mínimos de governança que dali decorrem. Esse caso fornece o problema concreto a partir do qual os instrumentos são desenhados.

O segundo movimento é uma revisão narrativa de casos, de natureza qualitativa e ilustrativa, e não uma revisão sistemática. A busca foi conduzida entre março e maio de 2026 em fontes acadêmicas (Google Scholar, Scopus e SciELO) e em fontes institucionais e jornalísticas de reconhecida credibilidade (relatórios de organizações de direitos digitais, órgãos reguladores e veículos de jornalismo investigativo), combinando os descritores, em português e inglês, "predictive analytics", "early warning system", "algorithmic governance", "risk scoring", "education" e "public sector" com os termos próprios de cada caso.

Foram incluídos casos que atendiam a quatro critérios: (i) uso de modelos preditivos ou de pontuação de risco por órgão público ou rede educacional; (ii) escala real de operação, e não protótipo; (iii) documentação pública robusta e verificável sobre desenho, uso e efeitos; e (iv) variação de finalidade e de arranjo institucional entre os casos. Foram excluídos relatos sem fonte primária localizável, experimentos restritos a ambiente de pesquisa e duplicações. Casos amplamente citados ficaram de fora quando a documentação disponível não permitia avaliar as seis dimensões de análise, como em sistemas cujos detalhes técnicos não são públicos.

A seleção é intencional e orientada a casos críticos. Três casos foram escolhidos por expor, de forma documentada, falhas de governança, e um caso adicional foi incluído como contraponto ilustrativo, de modo a permitir o contraste entre arranjos com e sem salvaguardas. O número reduzido de casos é uma decisão metodológica: na lógica dos casos críticos, poucos casos bem documentados, escolhidos por sua capacidade de revelar mecanismos, valem mais para a construção teórica do que muitos casos superficiais (FLYVBJERG, 2006). Os três casos de falha variam em país, finalidade e arranjo institucional, e a recorrência dos mesmos padrões de inversão em contextos tão distintos funciona como replicação analítica, sustentando a generalização dos mecanismos, ainda que não de frequências.

Cada caso foi codificado de forma estruturada nas seis dimensões descritas adiante, a partir de um protocolo comum que registra, para cada dimensão, a evidência, a fonte e o trecho de origem. Para reduzir o viés interpretativo, adotaram-se três procedimentos: priorização, em cada caso, da fonte documental mais autorizada, complementada por relatos independentes quando disponíveis; distinção explícita entre fato documentado e inferência do autor; e uso das mesmas categorias de análise para todos os casos, inclusive o positivo. Reconhece-se, ainda assim, o limite de uma leitura baseada em fontes secundárias, sem acesso aos sistemas e aos dados originais. O Quadro 1 sintetiza o procedimento de seleção e as fontes.

Quadro 1 - Procedimento de seleção e fontes dos casos

Parâmetro	Descrição
-----------	-----------

Bases e fontes consultadas	Google Scholar, Scopus e SciELO (literatura acadêmica); portais legislativos e regulatórios (Senado Federal, Câmara dos Deputados e EUR-Lex); organização de direitos digitais (Amnesty International); e veículos de jornalismo investigativo (The Markup e Chalkbeat).
Termos de busca	Combinações, em português e inglês, de "predictive analytics", "early warning system", "algorithmic governance", "risk scoring", "education" e "public sector", associadas aos nomes dos casos (DEWS, Ofqual, toeslagenaffaire e GPS Advising).
Período da busca	Março a maio de 2026, com verificação atualizada do status regulatório.
Triagem e leitura documental	Cerca de 40 a 60 registros triados por título e resumo; 12 a 15 lidos integralmente, com uma fonte documental principal por caso e relatos de apoio quando disponíveis.
Seleção final dos casos	Quatro casos no total: três de falha (DEWS, Ofqual e toeslagenaffaire) e um contraponto ilustrativo com salvaguardas (GPS Advising), escolhidos por contraste de escala, finalidade e arranjo institucional e por documentação pública robusta e verificável.

O terceiro movimento é a análise comparativa por mecanismos. As evidências de cada caso foram organizadas em seis dimensões observáveis, derivadas do problema de pesquisa: finalidade declarada; usos permitidos e proibidos; auditoria por subgrupos; controles de acesso e proteção de dados; comunicação e transparência; e responsabilização. A comparação procura padrões que se repetem em contextos distintos.

O quarto movimento é a proposição. Os padrões identificados são confrontados com instrumentos de governança algorítmica já estabelecidos, em particular os *Model Cards* (MITCHELL et al., 2019) e as fichas de dados nos moldes dos *datasheets for datasets* (GEBRU et al., 2021), além da lista DELICATE (DRACHSLER; GRELLER, 2016) e dos marcos regulatórios pertinentes (LGPD; PL 2338/2023; AI Act europeu). Desse confronto deriva o conjunto de instrumentos aplicáveis apresentado na seção seguinte.

Convém demarcar o alcance do trabalho. A validação empírica dos instrumentos constitui a etapa seguinte da pesquisa, detalhada como agenda na conclusão. As proposições deste artigo têm, portanto, o estatuto de hipóteses de desenho, conceito retomado e explicado na Seção 4.3, e não de recomendações já testadas em campo.

4. Resultados e Discussão

4.1 Casos críticos de falha e um contraponto ilustrativo com salvaguardas

O primeiro caso é o *Dropout Early Warning System* (DEWS), utilizado pela rede pública do estado de Wisconsin, nos Estados Unidos. Em operação desde 2012, o sistema combina algoritmos de *machine learning* que empregam dados históricos, entre eles notas, registros disciplinares, situação de gratuidade da merenda e raça, para estimar a probabilidade de cada estudante do ensino fundamental concluir o ensino médio no tempo esperado. Duas vezes por ano, as escolas recebem listas nominais com a previsão de cada aluno em código de cores, do baixo ao alto risco.

Investigação jornalística mostrou que, ao prever que um estudante não se formaria no prazo, o sistema errava em cerca de três quartos das vezes, com taxa de falsos alarmes bem maior para estudantes negros e hispânicos; educadores relataram receber pouca informação sobre como as previsões eram geradas, e estudantes descreveram os rótulos de alto risco como estigmatizantes (FEATHERS, 2023).

O segundo caso é o algoritmo de notas adotado pelo Ofqual, regulador de exames da Inglaterra, em 2020. Com a suspensão das provas presenciais durante a pandemia, o regulador substituiu as notas estimadas pelos professores por notas calculadas a partir de um algoritmo que combinava o desempenho histórico de cada escola com os ordenamentos fornecidos pelos docentes, um procedimento de moderação estatística, e não de aprendizado de máquina, mas sujeito aos mesmos dilemas de governança algorítmica. Cerca de 40% das notas foram rebaixadas em relação às avaliações dos professores, e estudantes de escolas com desempenho histórico mais baixo, em geral de contextos menos favorecidos, tenderam a ser os mais prejudicados, pois o procedimento reproduzia o desempenho pretérito das escolas; a validação era frágil em centros pequenos e os recursos só podiam ser interpostos pela escola, não pelo estudante (KELLY, 2021). A reação pública levou o governo a recuar e a adotar as notas atribuídas pelos professores.

O terceiro caso, ainda que não educacional, é exemplar dos riscos de inversão do instrumento: o escândalo das creches nos Países Baixos, conhecido como *toeslagenaffaire*. A administração tributária utilizou um modelo de classificação de risco para sinalizar suspeitas de fraude no auxílio-creche, no qual indicadores como a nacionalidade funcionavam como fatores de risco; havia ainda uma lista de sinalização em que constar elevava de forma automática a pontuação de risco da pessoa.

No encadeamento entre a sinalização algorítmica, a investigação e a aplicação administrativa rígida, entre 26 mil e 35 mil famílias foram acusadas indevidamente e obrigadas a devolver valores integrais, muitas vezes com encargos, o que provocou endividamento severo e perdas que chegaram à guarda de crianças; o episódio levou à renúncia do governo em 2021 e ao reconhecimento de racismo institucional (AMNESTY INTERNATIONAL, 2021). É a inversão precisa que se procura evitar: um sinal de risco, próprio de uma leitura agregada, foi convertido em decisão individual, punitiva e auto-reforçada, sem mecanismos efetivos de contestação.

Como contraponto condicionado, o programa *GPS Advising (Graduation Progression Success)*, da Georgia State University, uma universidade pública dos Estados Unidos, ilustra um arranjo com salvaguardas no uso de sinais preditivos. Desde 2012, a instituição acompanha diariamente centenas de fatores de risco e gera alertas individuais; a diferença decisiva está no que se faz com o alerta. Em vez de rótulo ou sanção, ele dispara o contato proativo de um orientador acadêmico, e um alerta não coloca o estudante em situação de advertência nem aciona qualquer consequência automática, funcionando como gatilho para uma conversa de apoio.

É preciso, porém, não confundir o sistema preditivo com o ecossistema institucional que o cerca. Os ganhos de equidade relatados pela universidade, como a aproximação das taxas de graduação entre estudantes de minorias e de baixa renda e os demais e a elevação da graduação de homens negros de cerca de 18% para 55% (GEORGIA STATE UNIVERSITY, 2024), decorrem de um pacote amplo de políticas, que inclui a contratação de dezenas de orientadores, a reorganização do serviço de aconselhamento, bolsas de conclusão e o reforço do apoio acadêmico, e não do modelo isoladamente. O caso não deve, portanto, ser lido como prova de que o modelo produziu os resultados, e sim como indício de que, quando o sinal preditivo é subordinado a uma finalidade de cuidado, a apoio humano qualificado e à ausência de punição, ele pode integrar um arranjo associado a ganhos de equidade, em contraste com os usos que falharam em Wisconsin.

Examinados em conjunto, os casos indicam que o dano, quando ocorre, não decorreu em primeiro lugar de baixa acurácia preditiva, mas do modo como as saídas foram governadas; o contraponto ilustra que salvaguardas adequadas podem reduzir esse risco. A Tabela 1 sintetiza o que se observou em cada dimensão de análise.

Tabela 1 - Casos analisados: padrões de falha e contraponto ilustrativo com salvaguardas

Caso (país)	Finalidade e forma de uso	Falha de governança predominante	Consequência observada
DEWS (EUA, Wisconsin)	Predição de evasão por aluno; listas nominais com rótulo de risco enviadas às escolas	Uso individualizado; variáveis sensíveis sem auditoria; baixa transparência; sem avaliação de efeito	Falsos alarmes desiguais por raça; estigmatização
Ofqual (Reino Unido)	Cálculo de notas individuais a partir de histórico da escola e ordenamento docente	Desvio de finalidade; opacidade; supervisão humana e contestação insuficientes	40% das notas rebaixadas; desvantagem para alunos pobres; recuo do governo
Toeslagenaffaire (Países Baixos)	Score de risco de fraude no auxílio-creche; lista de sinalização individual	Inversão de agregado para individual e punitivo; proxies sensíveis; sem contestação efetiva	Dezenas de milhares de famílias acusadas; danos graves; queda do governo
GSU (EUA): GPS Advising	Alertas individuais de risco que disparam orientação acadêmica proativa	Salvaguardas presentes: apoio humano, finalidade de cuidado, ausência de punição e investimento em pessoal	Redução das diferenças de graduação por raça e renda, atribuída ao conjunto do programa

Três regularidades distinguem os casos de falha do caso ilustrativo com salvaguardas. A primeira é o destino do sinal: nos casos de falha, saídas agregadas viraram rótulos individuais, e por vezes nominais; no contraponto, o alerta individual existia apenas para acionar apoio humano, nunca punição. A segunda é o tratamento das variáveis sensíveis, com ausência de auditoria por subgrupos de um lado e atenção explícita às diferenças de raça e renda de outro. A terceira é a presença de transparência e de canais de contestação, frágeis nos casos de falha. Nos casos examinados, o que separa os arranjos com salvaguardas dos que falharam não é a acurácia do modelo, e sim o desenho institucional que cerca o seu uso: finalidade de cuidado, supervisão humana qualificada e responsabilização.

4.2 Instrumentos de governança aplicáveis

Os padrões de falha orientam a seleção dos instrumentos segundo uma lógica de correspondência, em que cada instrumento responde a um modo específico de inversão observado nos casos. Quatro deles compõem a proposta.

O primeiro é a Carta do Modelo, adaptação dos *Model Cards* (MITCHELL et al., 2019) complementada por uma ficha de dados nos moldes dos *datasheets for datasets* (GEBRU et al., 2021). Documento público e acessível, a Carta reúne a finalidade científica e social do sistema, os dados mínimos utilizados, os limites de decisão, o desempenho por subgrupos, a frequência de revisão, as responsabilidades institucionais e, de forma expressa, os usos proibidos, entre eles a produção de listas nominais para punição ou exclusão. Ao tornar legíveis a finalidade e os limites, a Carta responde à opacidade e ao desvio de finalidade vistos nos casos DEWS e Ofqual.

O segundo instrumento é a matriz de usos, que distingue, de modo operacional, usos recomendados, usos de risco (admissíveis apenas sob revisão humana qualificada) e usos proibidos. A matriz fixa a fronteira entre o regime agregado e o regime individualizado e atua

como resposta direta à inversão exemplificada pelo caso holandês. O Quadro 2 apresenta uma formulação inicial.

Quadro 2 - Matriz de usos recomendados, de risco e proibidos

Usos (agregados)	recomendados	Usos de risco (exigem revisão humana)	Usos proibidos
Priorização territorial de recursos e programas		Qualquer pontuação ou recomendação referida a indivíduo	Listas nominais para punição, exclusão ou triagem
Monitoramento de desigualdades por rede e região		Pilotos de focalização que toquem o nível da pessoa	Decisões automatizadas sobre acesso ou permanência
Dimensionamento de políticas universais proporcionais		Cruzamentos com novas bases, que exigem reavaliar a finalidade	Uso de raça/cor e gênero para rotular pessoas

O terceiro instrumento é a auditoria por subgrupos, organizada em duas camadas, na qual se concentra a resposta ao risco de viés de atributos. Na primeira camada, o modelo que orienta políticas é treinado sem variáveis sensíveis: raça/cor e gênero são excluídos do conjunto de preditores. Como a exclusão não elimina o problema, pois atributos como renda, escolaridade dos responsáveis e território podem funcionar como *proxies*, variáveis que carregam de forma indireta a informação do atributo sensível, a segunda camada mantém um modelo paralelo de auditoria que incorpora as variáveis sensíveis com finalidade única e declarada: medir diferenças injustificadas entre grupos nas saídas do modelo de decisão.

A auditoria compara, por subgrupo de raça/cor, renda, gênero, tipo de escola e território, as taxas de falso positivo e de falso negativo, o *FI-score* e a calibração, entendida como a correspondência entre a probabilidade prevista e a frequência observada; diferentemente do estudo de origem, no qual atributos sensíveis integraram os preditores (LUZ, 2025), aqui sua exclusão do modelo operacional é uma salvaguarda proposta para o uso em políticas públicas.

Disparidades entre subgrupos devem acionar investigação das causas, revisão dos dados, atributos ou limiares, nova auditoria e, se persistirem, suspensão do uso. Os resultados são reportados em painéis agregados, sem rastreamento individual, e registrados na Carta do Modelo. O arranjo responde aos riscos observados no DEWS e converge com os princípios da lista DELICATE (DRACHSLER; GRELLER, 2016) e com as rotinas de auditoria interna propostas por Raji et al. (2020).

O quarto instrumento é a implantação em ondas, com pilotos regionais, monitoramento, gatilhos explícitos de pausa e revisão (por exemplo, limiares de disparidade entre subgrupos) e prestação de contas pública. Esse desenho cobre duas lacunas recorrentes: a ausência de avaliação de efeito, evidente no DEWS, e o caráter de difícil reversão dos danos, evidente no caso holandês. Tratar o risco como grau, calibrar e revisar de forma periódica transforma a previsão em capacidade de cuidar, ao permitir priorizar melhor, apoiar melhor e reduzir desigualdades de aprendizagem com responsabilidade.

Para que esses instrumentos passem de desenho conceitual a ferramenta adotável, é preciso fixar parâmetros mínimos de operacionalização: quem aprova, quem audita, com que periodicidade, sob quais métricas e limiares, por qual canal se contesta e o que é público ou restrito. O Quadro 3 propõe um conjunto inicial desses parâmetros, compatível com as exigências da LGPD e do PL 2338/2023, a ser calibrado por cada rede conforme sua capacidade institucional.

Quadro 3 - Parâmetros mínimos de operacionalização dos instrumentos

Parâmetro	Definição proposta
-----------	--------------------

Aprovação de uso	Comitê de governança de dados da rede, intersetorial (gestão pedagógica, jurídico e encarregado de dados), mediante Carta do Modelo publicada e Avaliação de Impacto Algorítmico aprovada.
Responsável auditoria	pela Equipe técnica independente da que desenvolve o modelo, com segregação de funções e relatório público.
Periodicidade	Revisão a cada ciclo letivo, com reauditoria obrigatória a qualquer mudança de dados, de finalidade ou de modelo.
Métricas mínimas	F1-score, calibração e taxas de falso positivo e falso negativo, reportadas no agregado e por subgrupo (raça/cor, renda, gênero, tipo de escola e território).
Limiar de pausa	Disparidade entre subgrupos acima de patamar fixado previamente (valor ilustrativo, como diferença relativa superior a 20% na taxa de erro, a ser fixado ex ante por cada rede) ou queda de calibração além do tolerado suspende o uso até correção.
Contestação	Canal formal sob o encarregado de dados da rede, com prazo de resposta definido, direito de revisão da decisão nos termos do art. 20 da LGPD e, como salvaguarda adicional aqui proposta, revisão por pessoa natural.
Dados públicos	Carta do Modelo, finalidade, métricas agregadas por subgrupo, relatórios de auditoria e registro dos gatilhos acionados.
Dados restritos	Microdados pessoais, escores individuais e qualquer saída que permita reidentificação, acessíveis apenas a perfis autorizados e registrados.
Registro da revisão humana	Registro datado de toda decisão que se afaste da recomendação do modelo, com justificativa e responsável identificável, arquivado para auditoria.

Em uma rede estadual ou municipal, a adoção pode seguir um roteiro em quatro ondas. Na onda preparatória, a secretaria constitui o comitê de governança, publica a Carta do Modelo e a matriz de usos e realiza a Avaliação de Impacto Algorítmico (recomendação, à espera da vigência do PL 2338/2023). Na primeira onda, o sistema opera como piloto em um recorte territorial e produz apenas painéis agregados para as equipes de planejamento, sem devolutiva nominal às escolas. Na segunda, o uso se expande para a rede, com a auditoria por subgrupos ativa, o canal de contestação em funcionamento e os limiares de pausa monitorados a cada ciclo; na terceira, o instrumento entra em rotina, com reauditoria anual e prestação de contas pública, e qualquer mudança de dados ou de finalidade reinicia o ciclo de aprovação.

Os instrumentos conversam com o marco regulatório em consolidação. No plano internacional, o Regulamento (UE) 2024/1689, conhecido como AI Act, classifica como de alto risco os sistemas de IA destinados a avaliar resultados de aprendizagem, a definir o acesso à educação e a monitorar a conduta em avaliações, e impõe gestão de risco, governança de dados, documentação técnica, transparência, supervisão humana e robustez (UNIÃO EUROPEIA, 2024). No Brasil, o PL 2338/2023, aprovado pelo Senado em dezembro de 2024 e em tramitação na Câmara dos Deputados, ainda não tem vigência; se aprovado, adotará abordagem baseada em risco e exigirá, para sistemas de alto risco, a Avaliação de Impacto Algorítmico,

além de direitos de explicação, contestação e revisão humana de decisões automatizadas (BRASIL, 2023). Por ora, no que toca a decisões automatizadas que tratem dados pessoais, o direito de solicitar a revisão dessas decisões já encontra base no artigo 20 da LGPD (BRASIL, 2018). A Carta do Modelo e a matriz de usos podem ser lidas como uma operacionalização antecipada das exigências de documentação, finalidade e supervisão humana que esses marcos tendem a impor.

Resta reconhecer os limites do próprio conjunto de instrumentos. Documentar não garante uso responsável, e os dispositivos podem reduzir-se a cumprimento formal, sobretudo onde a maturidade institucional em governança de dados é baixa; a eficácia depende de autoridade real para a revisão, de incentivos alinhados e da economia política da implementação, já que pressões por eficiência e por corte de gastos podem reinverter o uso, como mostra o caso holandês. São os parâmetros do Quadro 3 que dão auditabilidade aos instrumentos e reduzem esse risco, e é a etapa empírica, detalhada na conclusão, que poderá verificá-lo.

4.3 Limitações

O alcance deste trabalho é delimitado por alguns fatores. A análise apoia-se em fontes secundárias, sem acesso aos sistemas, aos dados e às decisões internas de cada caso. A revisão é narrativa e a seleção intencional de casos críticos favorece a identificação de padrões de dano, mas limita a generalização. O contraponto ilustrativo com salvaguardas é único e condicionado, sem avaliação causal independente que isole o efeito do sistema preditivo dos demais componentes do programa. Por fim, os instrumentos propostos têm o estatuto de hipóteses de desenho: proposições prescritivas do tipo “se a rede adotar o instrumento X sob as condições Y, espera-se reduzir o risco Z”, formuladas a partir de evidência secundária, ainda sem teste empírico direto. Resta avaliar em campo sua inteligibilidade, sua viabilidade operacional e sua efetividade em manter o uso agregado, auditável e não punitivo. Essas limitações reforçam a necessidade da etapa empírica detalhada na conclusão.

5. Conclusões

Este artigo deslocou a discussão sobre sistemas preditivos na educação pública do desempenho do modelo para as condições de sua implementação. A análise de três casos críticos de falha e de um contraponto ilustrativo com salvaguardas sugere que a legitimidade pública desses sistemas se decide no desenho institucional que cerca o seu uso, e não na acurácia que eles alcançam.

O padrão de falha que se repete é a inversão do instrumento, isto é, a passagem de um sinal agregado, que mapeia desigualdades estruturais, a um rótulo individual, punitivo e de difícil contestação. Contra essa inversão, o artigo propôs um conjunto de instrumentos de governança que operacionaliza equidade, transparência e responsabilização e que antecipa exigências da LGPD, do PL 2338/2023 e do AI Act.

Do ponto de vista ético, a orientação é clara: as previsões devem servir à priorização de políticas com foco territorial e sociodemográfico, sem personalização punitiva e sem que pessoa alguma seja excluída por um algoritmo.

Como trabalhos futuros, sugerem-se quatro frentes. A primeira é o estudo de implementação com múltiplos casos em redes públicas, com análise documental, entrevistas com gestores, encarregados de dados e educadores e oficinas de validação, sob critérios de inteligibilidade, viabilidade e aderência regulatória. A segunda é o protótipo de painel agregado com a Carta do Modelo, testado com uma rede parceira segundo o roteiro em ondas da Seção 4.2. A terceira é a calibração empírica dos limiares de disparidade do Quadro 3, aqui fixados por convenção. A quarta é a avaliação de efeito do arranjo completo sobre as decisões de alocação e as desigualdades de aprendizagem, para que as hipóteses de desenho se convertam em recomendações testadas.

Referências Bibliográficas

- ALMALAWI, A.; SOH, B.; LI, A.; SAMRA, H. Predictive Models for Educational Purposes: A Systematic Review. *Big Data and Cognitive Computing*, v. 8, n. 12, artigo 187, 2024. DOI: 10.3390/bdcc8120187.
- AMNESTY INTERNATIONAL. *Xenophobic machines: discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal*. Londres: Amnesty International, 2021. Disponível em: <https://www.amnesty.org/en/documents/eur35/4686/2021/en/>. Acesso em: 5 jun. 2026.
- BAKER, R. S.; HAWN, A. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, v. 32, p. 1052-1092, 2022. DOI: 10.1007/s40593-021-00285-9.
- BALL, S. J. What is policy? Texts, trajectories and toolboxes. *Discourse: Studies in the Cultural Politics of Education*, v. 13, n. 2, p. 10-17, 1993.
- BALL, S. J. The teacher's soul and the terrors of performativity. *Journal of Education Policy*, v. 18, n. 2, p. 215-228, 2003.
- BALL, S. J.; MAGUIRE, M.; BRAUN, A. *How schools do policy: policy enactments in secondary schools*. Abingdon: Routledge, 2012.
- BOWERS, A. J.; SPROTT, R.; TAFF, S. A. Do we know who will drop out? A review of the predictors of dropping out of high school. *The High School Journal*, v. 96, n. 2, p. 77-100, 2013.
- BRASIL. *Lei n. 13.709, de 14 de agosto de 2018*. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/113709.htm. Acesso em: 5 jun. 2026.
- BRASIL. *Projeto de Lei n. 2338, de 2023*. Dispõe sobre o uso da inteligência artificial. Brasília: Senado Federal, 2023. Disponível em: <https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>. Acesso em: 5 jun. 2026.
- DRACHSLER, H.; GRELLER, W. Privacy and analytics: it's a DELICATE issue. A checklist for trusted learning analytics. In: *Proceedings of the 6th International Conference on Learning Analytics and Knowledge (LAK '16)*. Edimburgo: ACM, 2016. p. 89-98. DOI: 10.1145/2883851.2883893.
- FEATHERS, T. False alarm: how Wisconsin uses race and income to label students high risk. *The Markup*, 27 abr. 2023. Disponível em: <https://themarkup.org/machine-learning/2023/04/27/false-alarm-how-wisconsin-uses-race-and-income-to-label-students-high-risk>. Acesso em: 4 jun. 2026.
- FERNÁNDEZ-DELGADO, M.; CERNADAS, E.; BARRO, S.; AMORIM, D. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, v. 15, p. 3133-3181, 2014.
- FLYVBJERG, B. Five misunderstandings about case-study research. *Qualitative Inquiry*, v. 12, n. 2, p. 219-245, 2006. DOI: 10.1177/1077800405284363.
- GEBRU, T.; MORGENSTERN, J.; VECCHIONE, B.; VAUGHAN, J. W.; WALLACH, H.; DAUMÉ III, H.; CRAWFORD, K. Datasheets for datasets. *Communications of the ACM*, v. 64, n. 12, p. 86-92, 2021. DOI: 10.1145/3458723.
- GEORGIA STATE UNIVERSITY. *Student success: GPS Advising*. Atlanta: Georgia State University, 2024. Disponível em: <https://success.students.gsu.edu/gps-advising/>. Acesso em: 5 jun. 2026.
- KELLY, A. A tale of two algorithms: the appeal and repeal of calculated grades systems in England and Ireland in 2020. *British Educational Research Journal*, v. 47, n. 3, p. 725-741, 2021. DOI: 10.1002/berj.3705.
- LUZ, D. M. da. *Previsão do desempenho dos estudantes no ENEM por meio de modelos de aprendizado de máquina e embeddings semânticos: uma abordagem baseada em variáveis socioeconômicas*. 2025. Dissertação (Mestrado) - Instituto de Ciências Matemáticas e de

Computação, Universidade de São Paulo, São Carlos, 2025. DOI: 10.11606/D.55.2025.tde-17062026-164821.

MAINARDES, J.; GANDIN, L. A. Contributions of Stephen J. Ball to the research on educational and curriculum policies in Brazil. *London Review of Education*, v. 11, n. 3, p. 256-264, 2013. DOI: 10.1080/14748460.2013.840985.

MITCHELL, M.; WU, S.; ZALDIVAR, A.; BARNES, P.; VASSERMAN, L.; HUTCHINSON, B.; SPITZER, E.; RAJI, I. D.; GEBRU, T. Model cards for model reporting. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Atlanta: ACM, 2019. p. 220-229. DOI: 10.1145/3287560.3287596.

MITCHELL, T. M. *Machine Learning*. Nova York: McGraw-Hill, 1997.

NOVAES, A. A.; PITON-GONÇALVES, J. Predição do desempenho de candidatos do ENEM por regressão múltipla. *Sigmae*, v. 10, n. 1, p. 82-98, 2021.

RAHAL, C.; VERHAGEN, M.; KIRK, D. The rise of machine learning in the academic social sciences. *AI & Society*, v. 39, n. 2, p. 799-801, 2024. DOI: 10.1007/s00146-022-01540-w.

RAJI, I. D.; SMART, A.; WHITE, R. N.; MITCHELL, M.; GEBRU, T.; HUTCHINSON, B.; SMITH-LOUD, J.; THERON, D.; BARNES, P. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Barcelona: ACM, 2020. p. 33-44. DOI: 10.1145/3351095.3372873.

RUBIM, J. A. D. O.; MANTOVANI, D.; ALAVARSE, O. M. Exclusão digital e seus impactos sobre a proficiência no ENEM: um estudo com concluintes do ensino médio de baixa renda entre 2015 e 2023. *Educação em Revista*, v. 41, e59221, 2025.

SIMÕES-GOMES, L.; ROBERTO, E.; MENDONÇA, J. Viés algorítmico: um balanço provisório. *Estudos de Sociologia*, Araraquara, v. 25, n. 48, 2020. DOI: 10.52780/res.13402.

SLADE, S.; PRINSLOO, P. Learning analytics: ethical issues and dilemmas. *American Behavioral Scientist*, v. 57, n. 10, p. 1510-1529, 2013. DOI: 10.1177/0002764213479366.

UNIÃO EUROPEIA. *Regulamento (UE) 2024/1689* do Parlamento Europeu e do Conselho (Lei da Inteligência Artificial). Jornal Oficial da União Europeia, 2024. Disponível em: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>. Acesso em: 5 jun. 2026.