Entendendo os Fatores de Previsão do TDI nas Escolas Públicas Brasileiras: Uma Abordagem Usando a Técnica SHAP

Abílio Nogueira Barros[1], **Gabriel Alves**[1], **Rafael Ferreira Mello**[1,2]

[1]Departamento de Computação - Universidade Federal Rural de Pernambuco (UFRPE)

[2]Centro de Estudos e Sistemas Avançados do Recife (CESAR)

abilionbarros@gmail.com, gabriel.alves,rafael.mello@ufrpe.br

Abstract. *The Age-Grade Distortion Rate (TDI) measures the number of students not in the expected grade for their age. This article employs machine learning techniques to predict TDI values from 2018 to 2023, covering pre-pandemic, pandemic, and post-pandemic periods. Using data from the Basic Education Census, we demonstrate the algorithm selection process and the application of SHAP for interpreting its metrics. Our objective is to identify the most important features highlighted by the predictive model and to stimulate discussion on qualitative and quantitative improvements in educational institutions, addressing structural, planning, and pedagogical aspects.*

Resumo. *A Taxa de Distorção Idade-Série (TDI) mede a quantidade de alunos fora do ano curricular esperado para sua idade. Este artigo utiliza técnicas de aprendizagem de máquina para prever os valores do TDI entre 2018 e 2023, abrangendo períodos pré, durante e pós-pandemia. Utilizando dados do Censo da Educação Básica, detalhamos a seleção do algoritmo e o uso do SHAP para interpretar suas métricas. Nosso objetivo é identificar as características mais importantes apontadas pelo modelo preditivo e fomentar a discussão sobre melhorias qualitativas e quantitativas nas instituições de ensino, abordando aspectos estruturais, de planejamento e pedagógicos.*

1. Introdução

Segundo o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) apenas no ano de 2023 o sistema básico educacional brasileiro registrou 47,3 milhões de matrículas divido em suas 178,5 mil escolas[1]. Tendo esses números, o sistema educacional brasileiro abarca a mesma quantidade de alunos quanto os habitantes da Espanha, e com esses números vários desafios já são vivenciados há décadas e que ainda permeiam o dia a dia escolar como citado em [Schwartzman and Brock 2005, Palomino et al. 2022].

Quando um sistema numeroso é atingido por um dos maiores desafios enfrentados pela humanidade nesse século, a pandemia de COVID-19 [Brito et al. 2020], novos desafios foram criados [Peres 2020], como também a necessidade de novas

[1]https://download.inep.gov.br/publicacoes

DOI: 10.5753/sbie.2024.242097

políticas educacionais [Galzerano 2021] e até metodologias foram desenvolvidas fosse possível manter o ensino durante tempos de educação remota e híbrida [Silva et al. 2022, Corrêa and Brandemberg 2021].

Após superar esse momento crítico é essencial avaliar fazer estudos para analisar os impactos da pandemia considerando indicadores registrados antes, durante e depois desse período. Isso nos permitirá acompanhar as mudanças e gerar aprendizados em um contexto histórico por meio de técnicas de Learning Analytics para extração e processamento desses dados, processo esse utilizado em trabalhos como o de [do Nascimento et al. 2018]. Diversos indicadores são disponibilizados pelo INEP, como taxas de rendimento escolar [Justino 2022], complexidade de gestão escolar [de Andrade et al. 2020], taxa de alfabetização [Bernardi and Luchese 2020], taxa de distorção idade-série (TDI) [NOGUEIRA and Silva 2022] e índice de desenvolvimento da educação básica (IDEB) [Rodrigues et al. 2016].

Entre os indicadores fornecidos pelo INEP, este trabalho focará na Taxa de Distorção Idade-Série (TDI). A TDI é calculada pelo Ministério da Educação (MEC) e mede a proporção de alunos que estão matriculados em séries com pelo menos dois anos de diferença em relação à idade adequada. O cálculo da TDI envolve a comparação entre a idade dos alunos e a série em que estão matriculados. O MEC utiliza dados do Censo Escolar, realizado anualmente, para obter essas informações. Com base nesses dados, a TDI é calculada usando a seguinte fórmula:

$$\text{TDI} = \frac{\text{Número de alunos com dois anos ou mais de atraso escolar}}{\text{Total de alunos matriculados na série}} \times 100$$

O TDI é um indicador bem consolidado na temática de mapeamento educacional por meio de dados abertos. Esse indicador também pode ser usado para avaliar a defasagem entre a educação urbana e rural. Em [NOGUEIRA and Silva 2022], o TDI foi utilizado como ferramenta de acompanhamento de escolas situadas em zonas rurais, expondo fatores intra e extraescolares para seus gestores.Foi utilizado também em outras áreas que não diretamente para resultados educacionais, em [Ferreira and Teixeira 2018] por exemplo, foi possível cruzar os indicadores de violência em localidades com alto TDI, buscando indicar uma relação de causa e efeito.Em [Barros et al. 2023] o TDI foi utilizado junto ao censo da educação básica e por meio de técnicas de Learning Analytics de modo a avaliar como esse indicador pode sugerir melhorias com base em valores médios de TDI.

Em busca de utilizar as técnicas de Learning Analytics para gerar essa informações com base no contexto histórico do TDI entre os anos 2018, 2019 como anos antes da pandemia, 2020 e 2021 como anos pandêmicos e 2022 e 2023 os posteriores a pandemia. E buscando guiar esse estudo foram desenvolvidas duas perguntas de pesquisa para serem respondidas com esse trabalho:

> **Pergunta de Pesquisa 1 (PP1):** *As características mais importantes na previsão do TDI permanecem as mesmas ao longo dos anos?*
>
> **Pergunta de Pesquisa 2 (PP2):** *Quais as características que estão mais relacionadas a um menor TDI?*

Diante disso, este trabalho visa demonstrar todo o processo de tomada de decisão incluindo à escolha do algoritmo, sua execução e, principalmente, a elucidação de seus

resultados visando embasar a tomada de decisão com base nos registros dos contextos temporais apresentados.

2. Método

Para o desenvolvimento deste experimento, seguimos várias etapas essenciais. Primeiramente, selecionamos as bases de dados e as preparamos para a aplicação dos algoritmos de aprendizado de máquina. Em seguida, escolhemos o algoritmo de regressão mais adequado para o problema. Aplicamos esses algoritmos aos dados de cada ano, permitindo uma análise detalhada ao longo do tempo. Finalmente, avaliamos os resultados utilizando a técnica *SHAP*, que nos ajudou a entender a contribuição de cada variável no modelo de regressão.O objetivo dessa etapa é gerar os dados necessários para responder às perguntas levantadas na 1. Nesa seção apresentamos como os dados foram adquiridos e processados, o processo de escolha do algoritmo e a extração do *feature importance*.

2.1. Aquisição e processamento dos dados

Os dados utilizados para este projeto foram todos adquiridos do INEP, respeitando a faixa temporal definida por este estudo. As bases necessárias foram o TDI ao nível de entidade escolar[2] e o censo escolar da educação básica[3].

Em [Barros et al. 2022], é demonstrado como realizar o processamento do Censo da Educação Básica utilizando o dicionário de dados para avaliar a ausência de colunas ao longo dos anos. No entanto, em nosso experimento, a junção dos dados não foi necessária, uma vez que os resultados foram avaliados individualmente para cada ano. Isso permitiu simplificar o método de processamento dos dados do censo em comparação com o proposto no trabalho de base.

Assim, foram removidas as colunas do tipo *string* ou *char* que identificassem a escola, como nome da entidade e código definido pelo INEP, para evitar o enviesamento dos dados. Em seguida, também foram excluídas as colunas que apresentavam 80% de registros nulos. Colunas que apresentassem um valor não registrado como:8888 para variáveis quantitativas e 9 para qualitativas (que apresentava resultados entre 0, 1, 2) foram zeradas. Essa operação foi guiada pela informação disponível no dicionário de dados de modo a garantir que não fossem prejudicadas outras colunas. Por fim foram removidas as colunas que tratavam sobre a quantidade de matriculas nas instituições buscando não induzir uma separação via quantidade desses elementos.

Para a extração dos resultados do TDI, selecionamos os dados a nível escolar e, após esse filtro, aplicamos um filtro adicional para instituições públicas, pois nosso objetivo é avaliar a rede pública de forma geral, englobando os níveis municipal, estadual e federal. Por fim, selecionamos os valores do TDI do ensino fundamental anos finais (*TDI_FUN_AF*).

Os dados foram armazenados utilizando o banco de dados colunar DuckDB [Mühleisen and Raasveldt 2024], com o objetivo de realizar os testes de treinamento dos algoritmos de aprendizado de máquina de forma mais eficaz e confiável, simplificando o

[2]https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/taxas-de-distorcao-idade-serie

[3]https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar

processo em comparação com outros bancos de dados que requerem uma infraestrutura mais complexa de instalação ou a utilização de contêineres, como MySQL, ClickHouse ou MariaDB.

2.2. Algoritmo de aprendizagem de máquina

Nesta etapa, foram aplicados algoritmos de aprendizagem de máquina para prever o valor que uma instituição de ensino pode alcançar no TDI. É importante destacar que, ao contrário de outros indicadores como taxa de aprovação e taxa de alfabetização, onde valores mais altos indicam melhor desempenho, no caso do TDI, valores mais baixos são melhores, afinal indicam que naquela instituição de ensino existe uma menor quantidade de matrículas com idade em desacordo ao nível escolar esperado. Assim, quando uma instituição apresenta um TDI menor que a do ano anterior, é considerado uma melhoria na situação escolar daquela instituição. Por exemplo, se uma escola apresentou um *TDI_FUN_AF* de 50% em 2020 e, após três anos, esse indicador caiu para 35% em 2023, isso indica uma melhora, mostrando que existiu uma redução de 15% nas matrículas fora da faixa escolar esperada para sua idade.

Por esse trabalho buscarmos a explicabilidade dos resultados dos algoritmos e não apenas o melhor desempenho, optamos por selecionar algoritmos de árvore de decisão, pois em sua natureza são algoritmos de melhor interpretação e explicabilidade [Mahbooba et al. 2021]. Em nosso caso foi utilizado um conjunto de algoritmos da família das árvores de regressão:

- **CatBoost**: É um algoritmo de gradient boosting desenvolvido pela Yandex que se destaca no manuseio de variáveis categóricas sem a necessidade de pré-processamento extenso [Prokhorenkova et al. 2018].
- **LightGBM**: É uma variante do gradient boosting, desenvolvido pela Microsoft, que foca em alta eficiência e escalabilidade [Ke et al. 2017].
- **XGBoost**: É uma implementação otimizada de gradient boosting que tem sido amplamente utilizada em competições de machine learning devido ao seu desempenho superior [Chen and Guestrin 2016].
- **Random Forest**: É um *ensemble* (uma técnica de aprendizado de máquina que combina múltiplos modelos-base para melhorar a precisão e a robustez das previsões em comparação com um único modelo individual) de árvores de decisão que se destaca pela sua simplicidade e robustez. Ele constrói múltiplas árvores de decisão durante o treinamento e produz a média das previsões de todas as árvores para regressão [Breiman 2001].
- **Extra Trees**: Extra Trees (Extremely Randomized Trees) é uma variante do Random Forest que randomiza ainda mais a construção das árvores. Em vez de procurar o ponto de corte ótimo, ele seleciona pontos de corte aleatoriamente, o que pode reduzir a variância do modelo e aumentar a diversificação entre as árvores[Geurts et al. 2006].
- **Decision Tree**: O Decision Tree Regressor é um modelo de árvore de decisão simples que divide o espaço dos dados em regiões homogêneas, com base em um conjunto de regras de decisão [Quinlan 1986].

Após a definição dos algoritmos foram definidas quais as métricas seriam utilizadas pelo *framework pycaret*[4], responsável por realizar os testes de cada algoritmo

[4]https://pycaret.gitbook.io/docs

ponderando sobre as seguintes métricas:

- **MAE (Mean Absolute Error)**: O MAE calcula a média das diferenças absolutas entre valores previstos e observados.
- **MSE (Mean Squared Error)**: O MSE calcula a média dos quadrados dos erros entre valores previstos e observados.
- **RMSE (Root Mean Squared Error)**: O RMSE é a raiz quadrada do MSE e representa o desvio padrão dos resíduos.
- **R² (Coefficient of Determination)**: O R² indica a proporção da variância dos dados dependentes explicada pelo modelo.
- **RMSLE (Root Mean Squared Logarithmic Error)**: O RMSLE calcula o erro entre os logaritmos dos valores previstos e observados.
- **MAPE (Mean Absolute Percentage Error)**: O MAPE calcula a média das diferenças percentuais absolutas entre valores previstos e observados.

A busca pelo melhor algoritmo foi feito com base nos dados de 2022, que foi a base anual que apresentou uma menor quantidade de mudança de colunas entre toda a faixa temporal avaliada, vide inclusão e exclusão, com base em nossa faixa temporal escolhida.

2.3. Análise de características

Como última etapa do processo, foi utilizado o *SHAP values*, mais especificamente sua versão para algoritmos de regressão. Os SHAP values, ou valores SHAP (SHapley Additive exPlanations), são uma técnica para a interpretação de modelos de aprendizado de máquina, baseada no conceito de valores de Shapley da teoria dos jogos cooperativos [Leite 2022]. Técnica essa já consolidada para a avaliação de modelos de aprendizagem de máquina como vemos em [Wang et al. 2022], [Lubo-Robles et al. 2020] e [Hamilton and Papadopoulos 2023].

A utilização dessa técnica foi escolhida para esse trabalho justamente trazer uma interpretação dos valores SHAP é direta e intuitiva. Cada valor SHAP atribuído a uma variável representa a mudança esperada na predição quando essa variável é incluída no modelo. Onde:

- **Valores Positivos e Negativo**s: Um valor SHAP positivo indica que a variável aumenta a predição do modelo, enquanto um valor negativo sugere que a variável reduz a predição.
- **Magnitude**: A magnitude do valor SHAP reflete a importância da variável para a predição específica. Valores absolutos maiores indicam maior influência.

Fornecendo para a nossa interpretação, diferentemente da técnica de *feature_importance* dos modelos, é sua explicabilidade em detrimento de uma quantidade maior ou menor de cada variável, para variáveis quantitativas como também para valores 0 e 1 na definição de se a entidade escolar possui ou não possui aquela característica informada, fazendo com que seja possível essa visualização já direta nos gráficos.

3. Resultados

Uma vez que todos os dados tenham sido calculados e os gráficos gerados, podemos responder às perguntas elencadas na seção 1.O dicionário de dados das *características* apontadas podem ser encontrados em https://docs.google.com/spreadsheets/d/1zzvwGIStfOshGum7zbkyr0vayrsZ0ENrgdFSNdaA5w0/edit?gid= de características.

3.1. Seleção do regressor

Modelo	MAE	MSE	RMSE	R2	RMSLE	MAPE	Tempo de Execução (Sec)
CatBoost	**5.83**	**76.98**	**8.77**	**0.832**	**0.441**	**0.342**	13.05
LGBM	6.001	81.33	9.01	0.822	0.456	0.355	258.20
XGBoost	6.260	89.25	9.44	0.805	0.470	0.384	1.740
Random Forest	6.600	93.01	9.63	0.797	0.475	0.416	3.650
Extra Trees	6.989	111.07	10.53	0.757	0.485	0.418	3.630
Decision Tree	8.695	168.50	12.97	0.632	0.609	0.496	**2.780**

Tabela 1. Resultados das Métricas de Desempenho dos Modelos Testados

Como pode ser visto na Tabela 1, o algoritmo *CatBoost* apresentou o melhor resultado. A escolha desse algoritmo é corroborada por artigos que destacam sua eficácia em comparação a outros regressores para dados categóricos, como mostrado em [Hancock and Khoshgoftaar 2020] e [Prokhorenkova et al. 2018].Tendo um MAE de 5,83 sugere que, em média, as previsões do modelo estão a 5,83 unidades de distância dos valores reais. O MSE de 76,98 é um bom indicativo de que o modelo está conseguindo minimizar erros grandes. Um RMSLE de 0,441 é bastante competitivo e indica que o modelo lida bem com grandes variações. Um valor de R^2 de 0,832 significa que o modelo explica 83,2% da variação nos dados, indicando um bom ajuste sobre os dados.

Uma vez definido o *CatBoost* como algoritmo a ser utilizado, foi aplicado o protocolo de separação dos dados em 80%, da base de dados criada, para treino e o restante para teste. Com isso para cada ano um modelo foi gerado e treinado com os dados do ano correspondente, gerando assim seis modelos já treinados.

Feature	Importância por Ano					
	2018	2019	2020	2021	2022	2023
IN_DESKTOP_ALUNO		8				
IN_ENERGIA_REDE_PUBLICA	3	6	6	4	3	2
IN_EQUIP_MULTIMIDIA						9
IN_FUNDAMENTAL_CICLOS	1	1	1	3		AUSENTE
IN_INTERNET	7					
IN_INTERNET_APRENDIZAGEM		3	3	2	2	5
IN_LABORATORIO_INFORMATICA	8					
IN_NOTURNO	6					
IN_ORGAO_ASS_PAIS_MESTRES					7	6
IN_ORGAO_GREMIO_ESTUDANTIL		2	2	1	1	1
IN_QUADRA_ESPORTES_COBERTA	2	4	4	9	9	7
QT_DESKTOP_ALUNO		7	5	5	6	4
QT_DOC_FUND_AF			10		10	
QT_DOC_FUND_AI	5					
QT_EQUIP_MULTIMIDIA	4	10	9	6	5	3
QT_EQUIP_TV	9					
QT_TRANSP_PUBLICO	AUSENTE	AUSENTE	AUSENTE	AUSENTE	AUSENTE	8
QT_TUR_FUND_AF	10			10		
QT_TUR_FUND_INT	AUSENTE	AUSENTE	AUSENTE	AUSENTE	AUSENTE	10
TP_ATIVIDADE_COMPLEMENTAR		5	7	7	8	
TP_PROPOSTA_PEDAGOGICA		9	8	8	4	

Tabela 2. Importância das Características por Ano

3.2. Pergunta de pesquisa 1

Para respondermos essa primeira pergunta, foram levantadas as dez principais características apontadas pelo método SHAP. E que podem ser conferidas na Tabela 2.

Ao analisarmos o contexto dos anos pré-pandemia e pós-pandemia, observamos que a característica IN_FUNDAMENTAL_CICLOS, que define o funcionamento da escola em ciclos do ensino fundamental, foi a mais importante para a previsão do indicador TDI antes da pandemia. No entanto, essa variável não aparece mais nos anos pós-pandêmicos. Em 2021, por exemplo, a característica mais relevante passou a ser IN_ORGAO_GREMIO_ESTUDANTIL, que indica a existência ou não de um grêmio estudantil na instituição educativa. Essa variável já mostrava um alto grau de impacto desde 2019.

Ao analisar as três primeiras características, vemos a presença constante de variáveis estruturais, como IN_QUADRA_ESPORTES_COBERTA e IN_ENERGIA_REDE_PUBLICA, além de uma principal característica pedagógica: IN_INTERNET_APRENDIZAGEM, que indica se a escola utiliza a internet como ferramenta educacional.

Expandindo a análise para as demais características, observamos a importância da quantidade de equipamentos eletrônicos, número de turmas, apresentação da proposta pedagógica, e a existência de organizações de alunos e de pais e mestres.

O ano de 2023 trouxe mais mudanças, principalmente devido à introdução de novas variáveis, como transporte público e a quantidade de equipamentos por aluno, incluindo desktops, portáteis e multimídia em geral.Em contrapartida, onde em 2018 ainda existiam características que eram relacionados ao funcionamento da instituição escolar, IN_NOTURNO, onde esse tipo de característica não foi vista mais ao decorrer dos anos.

Por fim, para formulamos uma resposta a nossa pergunta, a evolução das características entre os anos sofreu poucas alterações quando falamos do objetivo geral, onde a organização escolar e participação dos alunos (IN_ORGAO_GREMIO_ESTUDANTIL, IN_ORGAO_ASS_PAIS_MESTRES), técnicas pedagógicas (IN_INTERNET_APRENDIZAGEM, TP_PROPOSTA_PEDAGOGICA, TP_ATIVIDADE_COMPLEMENTAR), estrutural (IN_QUADRA_ESPORTES_COBERTA, IN_ENERGIA_REDE_PUBLICA, IN_FUNDAMENTAL_CICLOS) e por fim, na quantidade de equipamentos da escola (QT_DESKTOP_ALUNO, QT_EQUIP_MULTIMIDIA, QT_EQUIP_TV).

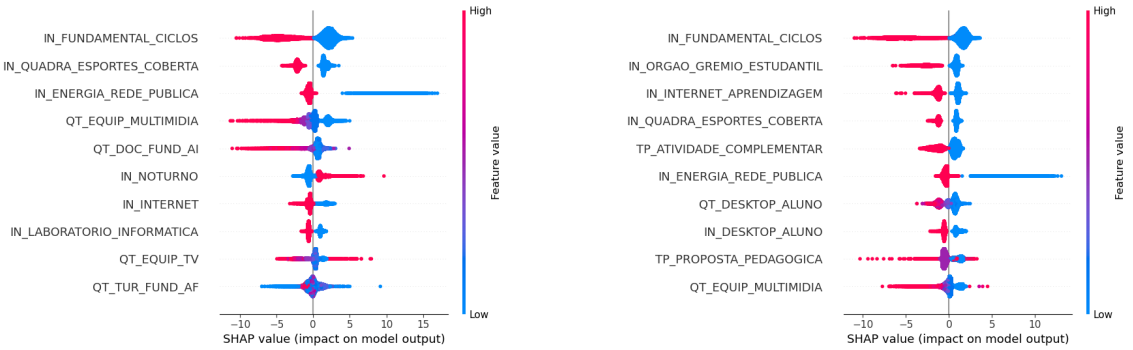
3.3. Pergunta de Pesquisa 2

Para responder a essa pergunta, analisaremos um dos tipos dos gráficos SHAP, o *Beeswarm*. É uma importante ferramenta do conjunto SHAP que auxilia na interpretação de modelos menos transparentes como Redes Neurais, ou mais transparentes como árvore de decisão, que foi o escolhido para esse trabalho. O gráfico permite identificar quais características têm o maior impacto nas previsões do modelo e se essas características estão aumentando ou diminuindo a previsão. Esse tipo de rastreamento é o que precisamos para avaliar a interação das características mais importantes com a previsão do TDI de cada unidade escolar entre cada dupla de anos, caracterizados como: pré-pandêmico, pandêmico e pós-pandêmico.

A estrutura do gráfico SHAP *Beeswarm* é composta por dois eixos principais e um conjunto de pontos que fornecem uma visualização detalhada das contribuições das características para as previsões do modelo. O eixo X representa os valores SHAP, as quais são as contribuições individuais das características para a previsão do modelo; valores

positivos indicam um aumento na previsão, que para nosso caso é um pior desempenho, enquanto valores negativos indicam uma diminuição, que em nosso caso indicam um melhor desempenho da instituição. O eixo Y lista as dez principais características do modelo, permitindo uma análise comparativa de sua influência. Cada ponto no gráfico representa uma instância de dados, que em nosso caso é uma instituição de ensino, e a cor dos pontos indica o valor da característica correspondente, com tons vermelhos representando valores altos e tons azuis representando valores baixos.

Para respondermos a essa pergunta analisaremos a presença das características de duas formas diferentes, pois em nosso conjunto de análise, para variáveis com prefixo **IN**, que indicam a existência utilizando o valor 1, e a ausência com o valor 0. Já para variáveis com prefixo **QT** demostraram variáveis quantitativas. Como a observação dessas variáveis são divergentes, a seguir iremos abordá-las como qualitativas (IN) e quantitativas (QT).



(a) SHAP 2018 **(b)** SHAP 2019

Figura 1. Análise SHAP 2018 e 2019: Pré pandêmico

Iniciando pelas Figuras 1 a e b, podemos analisar as variáveis qualitativas, destacando a manutenção do IN_FUNDAMENTAL_CICLOS como 1, indicando que instituições que mantêm o ensino fundamental em ciclos preveem um menor TDI. A implementação de um sistema escolar de ensino fundamental em ciclos representa uma abordagem progressista que visa promover um aprendizado contínuo e inclusivo, ao evitar a repetência direta, o sistema de ciclos tem em vista reduzir a evasão escolar e os impactos negativos associados à reprovação. Esse tipo de estratégia tem um alto teor de impacto junto ao TDI visto que reduzirá a quantidade de alunos, que são reprovados e caso já sejam o segundo ano de atraso escolar.

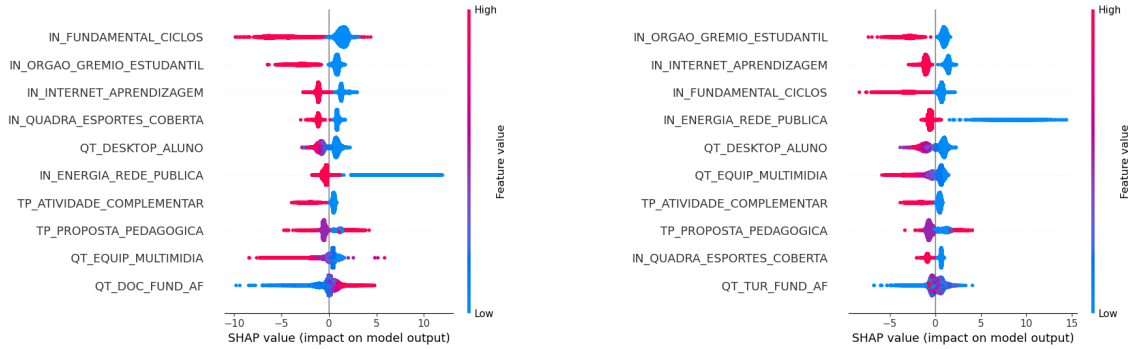
Outras variáveis qualitativas demonstram que: possuir quadras cobertas, laboratórios de informática, acesso à internet, uso da internet como ferramenta pedagógica, atividades complementares e proposta pedagógica atualizada anualmente são fatores que indicam uma previsão de menor TDI, impactando positivamente para o resultado do indicador na instituição.

Em contraste, o IN_NOTURNO, que indica funcionamento noturno, mostrou um impacto inverso. A existência desse turno se correlaciona com previsões de maior TDI, levantando discussões sobre os desafios do ensino noturno, conforme abordado em [dos Santos and Pouchain 2011]. Isso sugere a necessidade de revisar estratégias pe-

dagógicas diferenciadas para os turnos diurnos e noturnos.

Na análise das variáveis quantitativas, especialmente de equipamentos físicos, observamos que uma maior quantidade de equipamentos multimídia (QT_EQUIP_MULTIMIDIA) e computadores (QT_DESKTOP_ALUNO) tende a reduzir o TDI. Em 2018, a quantidade de TVs (QT_EQUIP_TV) não teve grande importância, pois esses equipamentos se tornaram obsoletos com a adoção de projetores, os quais são mais versáteis e adequados a diferentes espaços escolares, como auditórios e laboratórios.

Por fim, ao analisar o fator humano, em 2018, a quantidade de docentes no ensino fundamental anos iniciais (QT_DOC_FUND_AI) indicou que uma maior quantidade de professores correlaciona-se com um menor TDI. No mesmo ano, a quantidade de turmas do fundamental anos finais (QT_TUR_FUND_AF) também foi importante, embora com um impacto variável. Menos turmas podem afetar o TDI tanto positiva quanto negativamente, mas, em geral, uma menor oferta de vagas não é necessariamente positiva, pois o TDI é analisado com base na totalidade das matrículas e sua conformidade com o fluxo regular esperado.



(a) SHAP 2020 **(b)** SHAP 2021

Figura 2. Análise SHAP 2020 e 2021: Durante a pandemia

Analisando agora o que foi disposto na figura 2 a e b, temos a avaliação das características mais importantes para anos mais críticos no que tangem a educação brasileira [Ludovico et al. 2020].

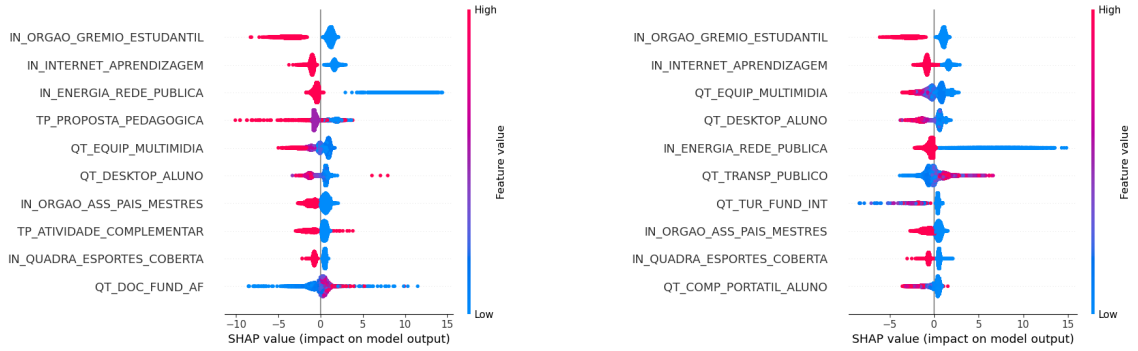
Semelhante ao modo apresentado na seção anterior, ao analisarmos primeiramente as características qualitativas apresentadas. Temos a manutenção das variáveis visto anteriormente, onde a presença dos fatores levantados favorecem uma previsão de menores valores do TDI. Destacando o aparecimento da variável que mensura a existência de atividades complementares oferecidas pela instituição, sejam de forma exclusiva ou parcial, onde avaliando ambos os anos mostra que esse fator impacta em sua totalidade para a previsão de menores valores de TDI.

Já quando avaliamos as quantitativas temos a manutenção da quantidade de desktops por aluno, onde prevalece que uma maior quantidade favorece uma previsão menor do indicador. O mesmo é observado para equipamento de multimídia, característica essa que em 2021 se torna mais importante, podendo assim estar relacionada com a vota parcial dos alunos para a sala de aula.

Indo para a análise do fator humano dentre os quantitativos, temos o aparecimento

da quantidade de docentes do fundamental anos finais(QT_DOC_FUND_AF) em 2020, onde ao interpretarmos vemos que uma quantidade maior de professores impactaria num maior TDI, trazendo uma análise contrária a lógica, de que mais professores trariam um melhor acompanhamento dos alunos e consequentemente um melhor resultado, mas nesse ano em específico esse não foi o resultado apresentado. Onde essa, possa ser uma peculiaridade vista nesse ano, pois ao visitarmos novamente a figura 2 b, vemos que essa variável em questão já não mais figura nosso gráfico. Com isso podemos interpretar também que a maior quantidade de professores já se deem em virtude de uma maior quantidade de alunos e um esforço para a melhoria situacional da escola e buscando assim reduzir o índice, visto que não é mantido para os anos sucessores.

Por fim, vemos em 2021, que a quantidade de turmas do fundamental anos finais novamente contrasta entre a quantidade menor de turmas para um TDI maior ou menor, onde no gráfico define uma pequena diferença ao favor de uma quantidade menor de turma trazer um resultado ligeiramente menor.



(a) SHAP 2022 **(b)** SHAP 2023

Figura 3. Análise SHAP 2022 e 2023: Pós pandemia

Seguindo a forma da análise anteriores, analisemos a última faixa temporal definida, e a mais recente durante o desenvolvimento desse trabalho. Os gráficos estão presentes na figura 3 a e b.

Ao começarmos pelas características qualitativas, a existência do grêmio estudantil se perpetuou nesses anos, sempre para impactar em um TDI mais baixo. Como também a utilização da internet como ferramenta educacional. O impactado dessas variáveis quantitativas seguem positivo, com a adição de uma que trate sobre organização de pais e mestres (IN_ORGAO_ASS_PAIS_MESTRES) onde também atua positivamente na previsão de um menor valor para o indicador alvo.

O desaparecimento da variável IN_FUNDAMENTAL_CICLOS da lista de variáveis deve-se que para o ano de 2022 ela já não teve 80% dos dados preenchidos, e em 2023 a variável foi definitivamente removida dos campos disponibilizados no censo da educação básica.

Passando para as variáveis quantitativas relacionado a objetos pertencentes a escola, temos a manutenção do comportamento, que onde tivermos uma maior quantidade de equipamentos multimídia, desktops, e para 2023, computadores portáteis trazem uma previsão menor do modelo sobre o TDI.

E ao analisarmos agora quanto a fatores quantitativos e não de patrimonio da instituição, temos a quantidade de docentes do ensino fundamental anos finais ainda trazendo um impacto, mesmo que pequeno, para valores maiores do TDI, comportamento visto anteriormente, favorecendo a interpretação de que a simples inserção de profissionais de ensino seja a resposta para a regressão do indicador. Para o ano de 2023 tivemos a ausência dessa feature entre as dez mais importantes, dando lugar a quantidade de turmas fundamentais de ensino integral (QT_TUR_FUND_INT), onde temos uma certa sobreposição dos valores, mas indicando que valores mais alto desse aspecto tendem a ter um valor de diminuição durante a previsão do TDI. Nesse mesmo ano temos também a adição da variável que trata da quantidade de transporte público utilizado pelos alunos da instituição (QT_TRANSP_PUBLICO), ele oferece uma relação oposta ao que pode ser indicado em [Evangelista et al. 2017] que ressalta os desafios enfrentados pelos alunos entre a ida e volta do âmbito escolar. Essa feature, criada nesse ano de 2023, carece de um melhor acompanhamento em seus próximos resultados, pois hoje traz seu resultado que quanto matriculas nesse tipo de serviço indicaria um resultado maior na previsão de seu TDI.

4. Considerações Finais e Trabalhos Futuros

Este trabalho se propôs a análise do indicador TDI das escolas públicas brasileiras durante os dois anos que precederam a pandemia de COVID-19, os dois anos de ápice da pandemia e os dois anos subsequentes. Após testar algoritmos e adequar a base de dados, utilizamos a técnica de interpretação de modelos SHAP para identificar as dez principais características de cada ano, evidenciando suas mudanças e impactos, positivos ou negativos, na previsão do indicador.

A discussão dessas características ao longo dos anos visa fornecer um ponto de vista baseado em dados para gestores escolares, gestores de redes educacionais e a comunidade em geral. O objetivo é promover melhorias na educação, considerando características como a presença de quadras cobertas. Destacamos a importância da participação da comunidade escolar, por meio de associações de pais e mestres e grêmios estudantis. Vimos também a estrutura disponível na escola, e também seus equipamentos se mostraram de fundamental importância da redução do TDI, por fim a utilização da internet como ferramenta educacional e manter sua proposta pedagógica atualizada complementa um perfil pedagógico mais proveniente a menores resultados do indicador.

Aspiramos continuar este trabalho, analisando e mapeando como essas características se comportam em outros níveis educacionais, comparando escolas em zonas rurais e urbanas. Também aspiramos aplicar o método para categorizar variáveis do censo da educação básica em pilares administrativos, estruturais e pedagógicos. Por fim, visamos validar os resultados com gestores para captar suas opiniões e aprimorar a aplicação prática das descobertas.

Referências

Barros, A. N., Alencar, A., Nascimento, A., de Albuquerque, A. F., and Mello, R. F. (2022). Elaboração do conjunto de dados agregados do censo da educação básica. In *Anais do IV Dataset Showcase Workshop*, pages 35–45. SBC.

Barros, A. N., Xavier, E. L. S., Alves, G., and Mello, R. F. (2023). Aplicação de learning analytics para identificação de tomada de decisão sobre a distorção idade-série

no brasil. In *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 21–31. SBC.

Bernardi, M. C. and Luchese, T. Â. (2020). A taxa de alfabetização de antônio prado, rio grande do sul (1895-1920). *Revista Educação em Questão*, 58(56).

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brito, S. B. P., Braga, I. O., Cunha, C. C., Palácio, M. A. V., and Takenami, I. (2020). Pandemia da covid-19: o maior desafio do século xxi. *Vigilância Sanitária em Debate: Sociedade, Ciência & Tecnologia*, 8(2):54–63.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Corrêa, J. N. P. and Brandemberg, J. C. (2021). Tecnologias digitais da informação e comunicação no ensino de matemática em tempos de pandemia: desafios e possibilidades. *Boletim Cearense de Educação e História da Matemática*, 8(22):34–54.

de Andrade, M. C. B., Silva, L. F., Fecury, A. A., de Oliveira, E., Dendasck, C. V., de Araújo, M. H. M., da Souza, K. O., da Silva, I. R., de Medeiros Moreira, E. C., Pascoal, R. M., et al. (2020). Indicadores de complexidade de gestão em escolas públicas e privadas de duas cidades do estado do amapá entre 2014 e 2018. *Research, Society and Development*, 9(9):e856998112–e856998112.

do Nascimento, R. L. S., da Cruz Junior, G. G., and de Araújo Fagundes, R. A. (2018). Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. *RENOTE*, 16(1).

dos Santos, M. J. C. and Pouchain, J. F. (2011). Evasão escolar no ensino médio noturno: Um estudo de caso na escola de ensino fundamental e médio prof. jáder moreira de carvalho. *Conhecer: debate entre o público e o privado*, 1(01):295–329.

Evangelista, J. C. S., Santos, C. R., Silva, L. R., and Santos, A. R. d. (2017). A política do transporte escolar na educação do campo: impactos e desafios na realidade escolar. *Seminário Nacional e Seminário Internacional Políticas Públicas, Gestão e Práxis Educacional*, 6(6).

Ferreira, V. B. and Teixeira, E. C. (2018). O impacto da distorção idade-série sobre a criminalidade nos municípios de minas gerais. *Revista Brasileira de Segurança Pública*, 12(2):269–291.

Galzerano, L. S. (2021). Políticas educacionais em tempos de pandemia. *Argumentum*, 13(1):123–138.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Hamilton, R. I. and Papadopoulos, P. N. (2023). Using shap values and machine learning to understand trends in the transient stability limit. *IEEE Transactions on Power Systems*.

Hancock, J. T. and Khoshgoftaar, T. M. (2020). Catboost for big data: an interdisciplinary review. *Journal of big data*, 7(1):94.

Justino, M. R. (2022). A relação do esforço docente e da infraestrutura escolar nas taxas de rendimento escolar: uma análise para a cidade do natal no ano de 2019. B.S. thesis, Universidade Federal do Rio Grande do Norte.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154.

Leite, G. B. (2022). Jogos cooperativos: uma introdução ao valor de shapley. *Informe Econômico (UFPI)*, 44(1).

Lubo-Robles, D., Devegowda, D., Jayaram, V., Bedle, H., Marfurt, K. J., and Pranter, M. J. (2020). Machine learning model interpretability using shap values: Application to a seismic facies classification task. In *SEG international exposition and annual meeting*, page D021S008R006. SEG.

Ludovico, F. M., Molon, J., Barcellos, P. D. S. C. C., Franco, S. R. K., et al. (2020). Covid-19: desafios dos docentes na linha de frente da educação. *Interfaces Científicas-Educação*, 10(1):58–74.

Mahbooba, B., Timilsina, M., Sahal, R., and Serrano, M. (2021). Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, 2021:1–11.

Mühleisen, H. and Raasveldt, M. (2024). *duckdb: DBI Package for the DuckDB Database Management System*. R package version 1.0.0.9000, https://github.com/duckdb/duckdb-r.

NOGUEIRA, M. D. O. E. and Silva, L. C. (2022). Escolarização em áreas rurais: a distorção idade-série na ótica dos gestores. *Estudos em Avaliação Educacional*, 33.

Palomino, P., Falcao, T. P., Medeiros, R., Uehara, M., Bittencourt, I., and Mello, R. F. (2022). Plataformas de dados educacionais: Análise com foco no plano nacional de educaçao. In *Anais do I Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 60–68. SBC.

Peres, M. R. (2020). Novos desafios da gestão escolar e de sala de aula em tempos de pandemia. *Revista de Administração Educacional*, 11(1):20–31.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Rodrigues, E. C. et al. (2016). Indicadores educacionais e contexto escolar: uma análise das metas do ideb. *Estudos em Avaliação Educacional*, 27(66):662–688.

Schwartzman, S. and Brock, C. (2005). Os desafios da educação no brasil. *Rio de Janeiro: Nova Fronteira*, 1320.

Silva, D. S. M. d., Sé, E. V. G., Lima, V. V., Borim, F. S. A., Oliveira, M. S. d., and Padilha, R. d. Q. (2022). Metodologias ativas e tecnologias digitais na educação médica: novos desafios em tempos de pandemia. *Revista Brasileira de Educação Médica*, 46:e058.

Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., and Tysklind, M. (2022). Towards better process management in wastewater treatment plants: Process analytics based on shap values for tree-based machine learning methods. *Journal of Environmental Management*, 301:113941.