UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

DANILO CESAR VITORINO DE ARRUDA

PREVISÃO DE TAXAS DE APROVAÇÃO DE ESCOLAS PÚBLICAS COM APRENDIZADO DE MÁQUINA: explorando os Valores de Shapley para aumentar a explicabilidade do modelo

Recife
2024

DANILO CESAR VITORINO DE ARRUDA

PREVISÃO DE TAXAS DE APROVAÇÃO DE ESCOLAS PÚBLICAS COM APRENDIZADO DE MÁQUINA: explorando os Valores de Shapley para aumentar a explicabilidade do modelo

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Engenharia de Produção.

Área de concentração: Engenharias III.

Orientador (a): Prof. Dr. Francisco Sousa Ramos.

Recife

2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Vitorino de Arruda, Danilo Cesar.
Previsão de taxas de aprovação de escolas públicas com aprendizado de máquina: explorando os valores de shapley para aumentar a explicabilidade do modelo / Danilo Cesar Vitorino de Arruda. - Recife, 2024.
103f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Tecnologias e Geociências, Programa de Pós Graduação em Engenharia de Produção, 2024.
Orientação: Francisco de Sousa Ramos.
Inclui referências.

DANILO CESAR VITORINO DE ARRUDA

PREVISÃO DE TAXAS DE APROVAÇÃO DE ESCOLAS PÚBLICAS COM APRENDIZADO DE MÁQUINA: explorando os Valores de Shapley para aumentar a explicabilidade do modelo

Dissertação ou Tese apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, como requisito parcial para a obtenção do título de Mestre em Engenharia de Produção. Área de concentração: Engenharias III.

Aprovada em: 29/08/2024.

BANCA EXAMINADORA

_____

Prof. Dr. Francisco de Sousa Ramos (Orientador)
Universidade Federal de Pernambuco

_____

Profa. Dra. Maisa Mendonça Silva (Examinadora Interna)
Universidade Federal de Pernambuco

_____

Prof. Dr. Jevuks Matheus de Araujo (Examinador Externo)
Universidade Federal da Paraíba

Dedico esse trabalho ao meu pai, Francisco José de Arruda Filho que navegava entre águas carinhosas e exigentes, comemorando as minhas vitórias, sempre seguro em uma corda de juízo. Me disse, em meio às celebrações da minha aprovação no vestibular, “mas não pare por aí, siga além da graduação” antes de algumas doses de uísque. Com muita saudade, ofereço, inundado em saudade, mais esta realização.

AGRADECIMENTOS

Agradeço, em especial, à minha mãe, Vanderleide Vitorino da Silva, por tudo que me proporcionou ao longo da minha vida e a todo apoio e torcida durante este caminho.

Extendo meus agradecimentos à minha avó materna, uma mulher destemida que, analfabeta, negou as limitações do seu mundo e rumou em busca de oferecer uma educação de qualidade aos seus filhos. Todos nós que viemos depois somos frutos da sua perseverança e teimosia, resistentes a todas as adjetivações e previsões negativas.

Coletivamente, agradeço a todas as pessoas que, de uma forma ou de outra, conviveram comigo durante este período, muitas vezes estressante, e me foram árvore e estrela, me ajudando a fincar os pés no chão para me acalmar e continuar caminhando rumo aos objetios.

> O treinamento de usuários consiste em parte do processo de educação, em base repetitiva, compreende ações e/ou estratégias para desenvolver determinadas habilidades ou habilidades específicas do usuário por desconhecer situações específicas de uso da biblioteca e seus recursos informacionais, que envolvem o conjunto de meios necessários para tal (DIAS; PIRES, 2004, p. 36).

RESUMO

A educação tem potencial de trazer benefícios para toda uma sociedade. Entretanto, a falha nesse campo pode representar efeitos bastante negativos. Pessoas mais escolarizadas tendem a desfrutar de maiores rendas e uma melhor saúde, enquanto o coletivo mais escolarizado está relacionado ao usufruto de comunidades mais seguras e com economia aquecida. Por outro lado, a desistência da escolarização pode trazer malefícios tanto para indivíduos quanto para grupos. Para evitar a concretização desse risco, este estudo explora os dados disponíveis, desenvolve e aplica um modelo de previsão para avaliar as taxas de aprovação nas escolas públicas brasileiras com o intuito de compreender a contribuição dos diversos fatores relacionados a estas instituições para os seus resultados educacionais. Para este último propósito, utiliza os Valores de Shapley para ponderar essas contribuições a partir da análise das coalizões possíveis dos aspectos característicos de sistemas educacionais. A biblioteca PyCaret foi explorada no processo de escolha do melhor modelo de Aprendizado de Máquina para o problema. Usando dados de 2015 a 2022, o modelo escolhido foi o CatBoost Regressor, que apresentou um coeficiente de determinação de 0,492 para os dados de treino e 0,467 para os dados de teste e foi utilizado para prever as taxas de aprovação escolar do ano de 2023. O método SHAP foi utilizado para interpretar o modelo e aumentar a sua explicabilidade, identificando os fatores-chave para os resultados de previsão obtidos, o que possibilita uma alocação mais eficiente dos recursos públicos. Os resultados mostram que escolas com altos índices de distorção idade-série apresentam grande influência negativa às taxas de aprovação. Diferentemente, a média de alunos por turma não tem uma análise trivial, pois menores números contribuem, majoritariamente, de forma positiva, mas há muitos pontos com influência negativa, indicando que deve haver uma diminuição nesse número de maneira cautelosa. A taxa de docentes com ensino superior apresentou um resultado nulo, quando são altas, mas negativo quando menores, assim como o fornecimento de internet. Foi observado, também, que a região geográfica exerceu uma influência considerável, com a região Sudeste se destacando positivamente e Norte negativamente. A localização da escola apresentou resultados contraintuitivos, com escolas urbanas contribuindo negativamente para o modelo, enquanto as escolas rurais contribuíram positivamente. Por fim, os resultados do trabalho foram discutidos em termos das suas implicações em sugestões para políticas públicas visando melhorar o indicador educacional em questão.

Palavras-chave: educação pública; educação básica; aprendizado de máquina; Valores de Shapley; explicabilidade; modelo de previsão.

ABSTRACT

Education has the potential to bring benefits to society as a whole. However, failure in this field can result in significantly negative effects. More educated individuals tend to enjoy higher incomes and better health, while a collectively educated society is associated with safer communities and a robust economy. On the other hand, dropping out of school can harm both individuals and groups. To mitigate this risk, this study explores available data, develops, and applies a predictive model to evaluate approval rates in Brazilian public schools, aiming to understand the contribution of various institutional factors to educational outcomes. Shapley Values were used to weigh these contributions through the analysis of possible educational system coalitions. The PyCaret library was utilized to select the best Machine Learning model for the problem. Using data from 2015 to 2022, the chosen model was the CatBoost Regressor, which demonstrated a determination coefficient of 0.492 for training data and 0.467 for test data, and was used to predict school approval rates for 2023. The SHAP method was employed to interpret the model and enhance its explainability, identifying key factors for the obtained prediction results, allowing for more efficient public resource allocation. The results indicate that schools with high age-grade distortion indices significantly negatively influence approval rates. In another way, the average number of students per class did not yield straightforward analysis, as lower numbers mostly contributed positively, but there were many points with negative influence, suggesting a cautious approach to reducing class sizes. The percentage of teachers with higher education showed a null effect when high, but negative when lower, similar to internet availability. Geographic region also had a considerable influence, with the Southeast region standing out positively and the North negatively. The school's location yielded counterintuitive results, with urban schools contributing negatively and rural schools positively. Finally, the study's results were discussed in terms of their implications for public policy suggestions aimed at improving the educational indicator in question.

Keywords: public education; basic education; machine learning; Shapley Values; explainability; predictive model.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O Brasil, apesar de ser uma nação rica em diversidade cultural e recursos naturais, enfrenta profundas desigualdades sociais e econômicas que persistem como obstáculos à realização de seu pleno potencial. O Global Wealth Report (2023) aponta que o país figura na primeira posição do ranking referente à maior concentração de riqueza pelo 1% mais rico, com 48,3% da renda e patrimônio nacional. Tal feito o coloca à frente de nações como Índia (41%), Estados Unidos (34,3%) e China (31,1%).

Além da desigualdade, a pobreza também é um problema presente nas famílias brasileiras. De acordo com o IBGE (2023), a sua renda média domiciliar per capita é de apenas R$ 1625. A condição, que já é alarmante, se mostra ainda mais espantosa quando se trata da metade mais pobre da distribuição. Em 2022, essa parcela da população recebeu um rendimento médio de R$ 537 por pessoa (BRASIL, 2023).

A disparidade entre diferentes camadas da população, bem como a carência que é imposta a muitas comunidades, são questões urgentes que exigem soluções eficientes e inovadoras. A constituição do Sistema Único de Saúde (SUS) foi um marco nessa luta pela redução das desigualdades. Apesar de críticas acerca da viabilidade orçamentária do programa, os mais de 30 anos de vigência demonstraram a sua capacidade de atingir, praticamente, acesso universal aos serviços de saúde de maneira gratuita e sua expansão possibilitou a adaptação de maneira ágil às mudanças nas necessidades da população (CASTRO *et al.*, 2019).

Segundo Engbom e Moser (2022), uma política que acompanhou a redução significativa na desigualdade de renda foi o aumento real de 128% do salário-mínimo entre 1996 e 2018. Mais recentemente, o programa Bolsa Família teve um impacto positivo não só na redução da miséria extrema, mas também na frequência escolar, pois esta é uma das prerrogativas para receber o benefício (SANTOS *et al.*, 2019). São várias as experiências que se mostram eficazes no combate às mazelas sociais.

Uma das áreas de políticas públicas que tem potencial de superar essa realidade é aquela destinada a tratar de questões relacionadas à educação. Esta, que faz parte da vida cotidiana de todas as pessoas e, em especial, dos brasileiros, se manifesta como um poderoso instrumento para moldar o desenvolvimento de uma sociedade. Paulo Freire (1971), um dos pensadores da educação mais reconhecidos e citados no mundo, elaborou o que chamou de educação libertadora, defendendo que, a partir dela, é possível mudar as estruturas de uma sociedade de forma a criar as condições para construir uma outra mais justa, inclusiva e próspera.

Se, por um lado, a educação possui caráter transformador para o desenvolvimento intelectual e o combate às desigualdades, resultados problemáticos nessa área têm o potencial de gerar consequências devastadoras a uma sociedade. Esse perigo iminente – que será descrito mais à frente – está em alerta por uma grave ameaça que tem se mostrando presente no Brasil: o aumento da evasão escolar.

Em documento feito pelo INEP (2023b), houve uma queda na quantidade de matrículas no sistema público de educação. De 2019 a 2022, a queda foi de mais de 350 mil matrículas. Ao comparar as matrículas de 2019 com 2023, a redução foi de mais de 850 mil matrículas. Há, também, a observação de um preocupante número de estudantes que enfrentaram o impulso por desistir da escolarização, como observado pela Unicef (2022). A pesquisa verificou que uma fração de 21% deles disseram ter pensado em desistir da escola nos últimos três meses antes da indagação – parcela que piora progressivamente com a faixa etária, englobando 27% dos 15 aos 19 anos e 16% dos 11 aos 14. Essas evidências sugerem que a pandemia da COVID-19 pode ter agravado a magnitude desse risco.

Entre os fatores potencialmente relacionados à evasão escolar, estão aspectos acadêmicos e exteriores ao ambiente escolar. Mudanças graves e persistentes nas realidades pessoais e familiares dos estudantes, como problemas de saúde e outras vulnerabilidades, se mostraram como evidências relevantes para a desistência da vida escolar (DUPÉRÉ *et al.*, 2015). No cenário do brasileiro, o panorama econômico que levou a choques de desemprego nas famílias, já provou exercer influência na vida acadêmica de crianças e adolescentes, que foram impelidas a deixar a vida escolar para trabalhar (DURYEA; LAM; LEVISON, 2007).

Por outro lado, questões como o baixo envolvimento dos pais na escola, reprovação escolar e baixo desempenho acadêmico estão relacionadas ao absenteísmo durante o ano letivo e, em última instância, ao abandono da escolarização (GUBBELS; PUT; ASSINK, 2019; NO; TANIGUCHI; HIRAKAWA, 2016). A relação entre professores e estudantes é outro aspecto relevante, observado tanto por docentes na função de professor quanto aqueles que exerciam cargos de gestão (ALFONSO J. GIL; PÉREZ-NAVÍO, 2019). Diante da complexidade de agentes e elementos que podem levar ao abandono – e evasão – da educação básica, o papel da escola é essencial para evitar esse problema. Este trabalho destacará, para isto, as questões relativas ao rendimento acadêmico como aspectos centrais que demandam atenção aos sistemas educacionais.

A evasão, entre outras consequências que serão discutidas mais à frente, impede a realização plena do direito à educação. Esse contexto desafiador e inquietante provoca uma questão para investigação: como melhorar o rendimento escolar da educação pública para,

consequentemente, diminuir o risco de evasão escolar? Em um cenário de recursos escassos, a priorização da peças-chave para a melhoria dos indicadores educacionais na ocasião da distribuição dos recursos é central para a tomada de decisão.

Os fatores que compõem todo o sistema educacional são abundantes e atraem a atenção dos gestores públicos, frequentemente sobrecarregados por dados e informações complexas.. Tal conflito pela prioridade merece uma avaliação detalhada devido ao caráter transformador da educação como investimento no potencial humano e na transformação do futuro coletivo. Na busca por uma distribuição mais eficiente dos recursos, a decomposição em Valores de Shapley (SHAPLEY, 1953) emerge como uma técnica capaz de suprir essa necessidade.

Esse método avalia a contribuição de um fator para um resultado em cada uma das possíveis coalizões formadas, ponderando a média dessas contribuições. A partir dessa análise voltada aos sistemas educacionais brasileiros, é possível avaliar quais fatores devem ser priorizados. Ademais, as possíveis sequelas que a pandemia pode ter deixado nos sistemas educacionais fortalece a demanda urgente do esforço de todos os setores da sociedade – entre eles, a universidade pública – para melhorar os sistemas educacionais, visando garantir o direito à educação de qualidade para todos.

1.1 JUSTIFICATIVA E RELEVÂNCIA

No Brasil, educação é tida como um direito de todos os residentes do país, tendo o Estado o dever de prover esse serviço com qualidade (BRASIL, 1988). Se, pela perspectiva da universalização, o sistema educacional brasileiro alcançou resultados satisfatórios, do ponto de vista do nível do serviço prestado, o cenário ainda é desalentador. O relatório realizado pelo Todos Pela Educação (2022) revela que, entre crianças de quatro e cinco anos, mais de 90% estão matriculadas na escola, dado que se comporta de maneira constantemente crescente. Entre crianças e adolescentes de 6 a 16 anos, mais de 98% estão matriculadas no Ensino Fundamental (EF) desde 2012.

Já o percentual de estudantes com aprendizagem adequada em português e matemática figura abaixo de 40% e 20% (respectivamente) ao final do EF e pouco acima de 30% e 5% ao final do Ensino Médio (EM). Tais dificuldades relativas à qualidade da educação brasileira têm consequências negativas em outros campos, bem como a melhoria pode alavancar múltiplos indicadores sociais.

Por um lado, diversos são os estudos que fortalecem a ideia de correlação positiva entre os resultados da educação formal e a melhoria em outras áreas da sociedade. Hanushek e

Woessmann (2020), por exemplo, concluíram que o aumento de um desvio padrão no resultado do Programa Internacional de Avaliação de Estudantes (PISA, na sigla em inglês) de um país – equivalente a 42 pontos no teste de matemática do PISA 2018 – está associado a um aumento no crescimento do PIB per capita de dois pontos percentuais. Barros (2017), avaliando, pelo contrário, os efeitos do mau desempenho educacional, encontrou um custo social estimado em R$ 100 bilhões por ano gerado pelo alto grau de evasão escolar, entre menores rendas, alta criminalidade e gastos com saúde. Esse valor é pouco menos da metade do orçamento para o ano de 2024 do Ministério da Educação (BRASIL, 2024).

A longevidade da escolarização e suas consequências também foram objeto de estudo para compreensão da sua relação com a realidade de indivíduos e comunidades. O relatório da UNESCO (2011) Monitoramento Global de Educação para Todos aponta que uma criança cuja mãe saiba ler tem uma chance de sobreviver até depois dos cinco anos de idade 50% maior em detrimento de outra de mãe analfabeta. Ainda, uma criança filha de quem não possui nenhuma educação formal tem uma taxa de mortalidade antes de completar cinco anos pelo menos três vezes maior em relação àquelas filhas de mães com alguma educação de nível médio.

Dada a relevância intrínseca e efeitos colaterais da área, urge a necessidade de que ações sejam tomadas para que os sistemas educacionais cumpram o seu papel social. Day, Gu e Sammons (2016), buscando entender os impactos da liderança escolar – representada pela gestão – nos resultados estudantis, concluíram que a capacidade das escolas melhorarem e manterem uma performance satisfatória no longo prazo não é, essencialmente, um produto do estilo de liderança do(a) gestor(a). Na verdade, eles perceberam que isso depende da sua compreensão das necessidades da escola. Apesar de parecer uma tarefa simples para profissionais que vivenciam o contexto da instituição, as condições podem torná-la um grande desafio.

O Banco Mundial (2018), por exemplo, afirmou que sistemas educacionais, sejam escolas ou redes, possuem pouca informação acerca de quem está aprendendo e quem não está. O documento produzido pela entidade aponta para a grande quantidade de dados que tem sido produzida nos últimos anos em sistemas educacionais e como eles têm sido mal aproveitados pelos tomadores de decisão. Essa oferta crescente de dados pode fornecer informações valiosas aos gestores, de maneira que tenham a capacidade tanto de melhorar os seus processos gerenciais, visando entregar melhores efeitos ao longo dos anos, quanto corrigir o curso dentro do período letivo antes que o risco da reprovação ou abandono escolar se concretize.

Todos esses dados e informações têm potencial de transformar sistemas de ensino que sofrem com problemas crônicos relacionados à evasão escolar. Por outro lado, eles podem

produzir um ambiente caótico em que representam uma dificuldade a ser enfrentada, ao invés de auxiliar a tomada de decisão. O presente trabalho visa, portanto, realizar uma análise dos dados abertos das escolas públicas brasileiras, fornecendo, ao debate público, informações qualificadas para formulação de políticas direcionadas à redução da evasão escolar, potencializando as decisões a partir de dados e evidências.

1.2 DESCRIÇÃO DO PROBLEMA

A educação pública vem acompanhando com as inovações institucionais do Estado brasileiro. Na história mais recente do país, junto à redemocratização, promulgou-se a Constituição Federal, em 1988. Nela estabeleceu-se que o EF seja "obrigatório e gratuito, inclusive para os que a ele não tiveram acesso na idade própria", sendo expandido, em 2009, também para o EM (BRASIL, 1988, p. 124). Apesar de ter sido imposto tardiamente, o EM gratuito vem sendo gradativamente ampliado desde meados da década de 90, com a Lei de Diretrizes e Bases da Educação (BRASIL, 1996).

Entretanto, o aumento das matrículas incorreu no problema do subfinanciamento da educação básica, visto que o orçamento para educação não cresceu paralelamente. Para enfrentar essa dificuldade, em 2007, foi criado o Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (FUNDEB), em substituição ao Fundo de Manutenção e Desenvolvimento do Ensino Fundamental e de Valorização do Magistério (FUNDEF), criado em 1997, pela Emenda Constitucional número 14 (FILHO, 2008).

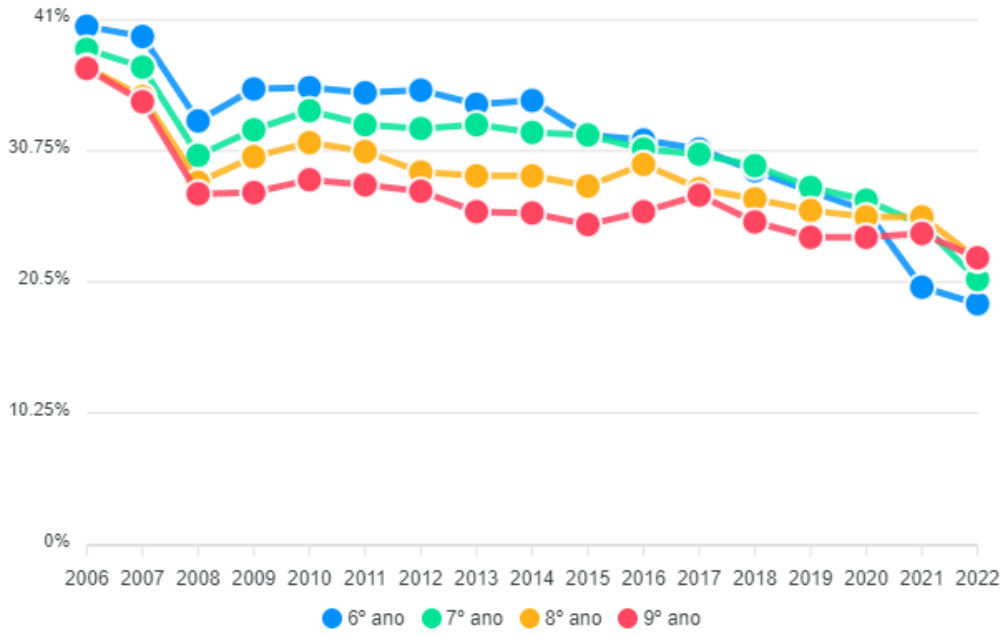
Com o problema do financiamento relativamente controlado, em 2014, foi aprovada a lei que estabeleceu o Plano Nacional de Educação (PNE). Este traçou as diretrizes e metas para a educação básica brasileira para um ciclo de 10 anos (BRASIL, 2014). O plano trata de metas dispostas a alcançar o pressuposto de uma educação pública, universal e de qualidade para a população. Neste trabalho, terá maior importância a Meta 2: até 2024, 95% dos adolescentes de 16 anos devem ter concluído o EF. Os dados mais recentes apontam que, em 2020, este número chegou a 82% (TODOS PELA EDUCAÇÃO, 2022). Contudo, após a pandemia da COVID-19, há o receio de que não haja a evolução necessária para bater a meta.

Esse número está diretamente relacionado a outro indicador: a distorção idade-série. Este representa os estudantes com dois ou mais anos à frente da idade recomendada para a série na qual estão matriculados. Uma publicação do Inep (2021) – o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – expõe que a situação de distorção idade-série é algo

dificilmente reversível. Em geral, um aluno que atrasa os estudos no início da educação básica, por conta da reprovação ou abandono de um ano escolar, permanece nessa situação até a conclusão da etapa de ensino ou, eventualmente, até uma evasão.

Em 2022, de acordo com o QEdu (2023), 21% dos estudantes matriculados nos anos finais do EF de escolas públicas estiveram em situação de distorção idade-série. Este número vem caindo, como indica a Figura 1. Porém, a queda drástica observada entre 2006 e 2008 foi sucedida por um declínio brando, o que também pode comprometer o alcance da meta.

Figura 1 - Taxas de distorção idade-série ao longo dos anos



Fonte: QEdu (2023)

Como indicado anteriormente, a distorção idade-série está diretamente ligada à não aprovação de um estudante, seja por reprovação ou abandono do ano letivo, ambas as situações componentes das taxas de rendimento escolar. Como as próprias denominações sugerem, é classificado como: aprovado aquele estudante que obtiver desempenho, ao final do ano, suficiente para cursar a série seguinte; reprovado aquele cujo desempenho, ao final do ano letivo, for insuficiente para cursar a série seguinte; e em situação de abandono aquele que, por qualquer motivo, deixar de frequentar a escola naquele ano letivo, não tendo sido formalmente desvinculado. Nestes dois últimos casos, quando o(a) estudante tem que repetir o mesmo ano devido a reprovações ou abandonos em dois ou mais anos, passa a ser contabilizado(a) em situação de distorção idade-série (UNICEF, 2018).

Além do não cumprimento do objetivo delineado pelo PNE, o déficit de aprendizado gerado por reprovações e abandonos ao longo da trajetória escolar afeta a confiança do(a) estudante e resulta em um aumento na evasão escolar (BARROS, 2017). Destarte, a situação descrita neste tópico apresenta o seguinte problema de pesquisa: como utilizar dados e evidências de maneira a oferecer aos decisores de sistemas de educação básica informações que contribuam ao aprimoramento das taxas de rendimento escolar?

1.3 OBJETIVOS

1.3.1 Geral

O trabalho visa apresentar uma ferramenta de previsão da taxa de aprovação das escolas públicas do país que auxilie na proposição de políticas públicas a partir da priorização do investimento de intervenções e recursos para melhoria desse indicador de maneira mais eficiente.

1.3.2 Específicos

Objetivos específicos:

- Explorar a literatura em busca de metodologias inovadoras e adequadas ao tema de pesquisa;
- Construir uma base de dados de fonte pública, unindo características e indicadores das escolas públicas brasileiras com o intuito de relacioná-las às suas respectivas taxas de rendimento;
- Analisar a base construída à procura de *insights* que podem nortear a investigação e modelagem posterior;
- Construir um modelo de previsão, a partir dos dados selecionados, que sirva como base para inferir sobre o valor que os atributos educacionais possuem, seja individualmente ou coletivamente;
- Prever as taxas de aprovação das escolas públicas brasileiras para o ano de 2023 com base naquelas que possuem dados divulgados pelo INEP;
- Aplicar um modelo matemático que consolide a análise anterior e gere informação qualificada acerca dos aspectos dos sistemas educacionais que

contribuem em maior magnitude, seja positiva ou negativamente, para as suas taxas de aprovação escolar;

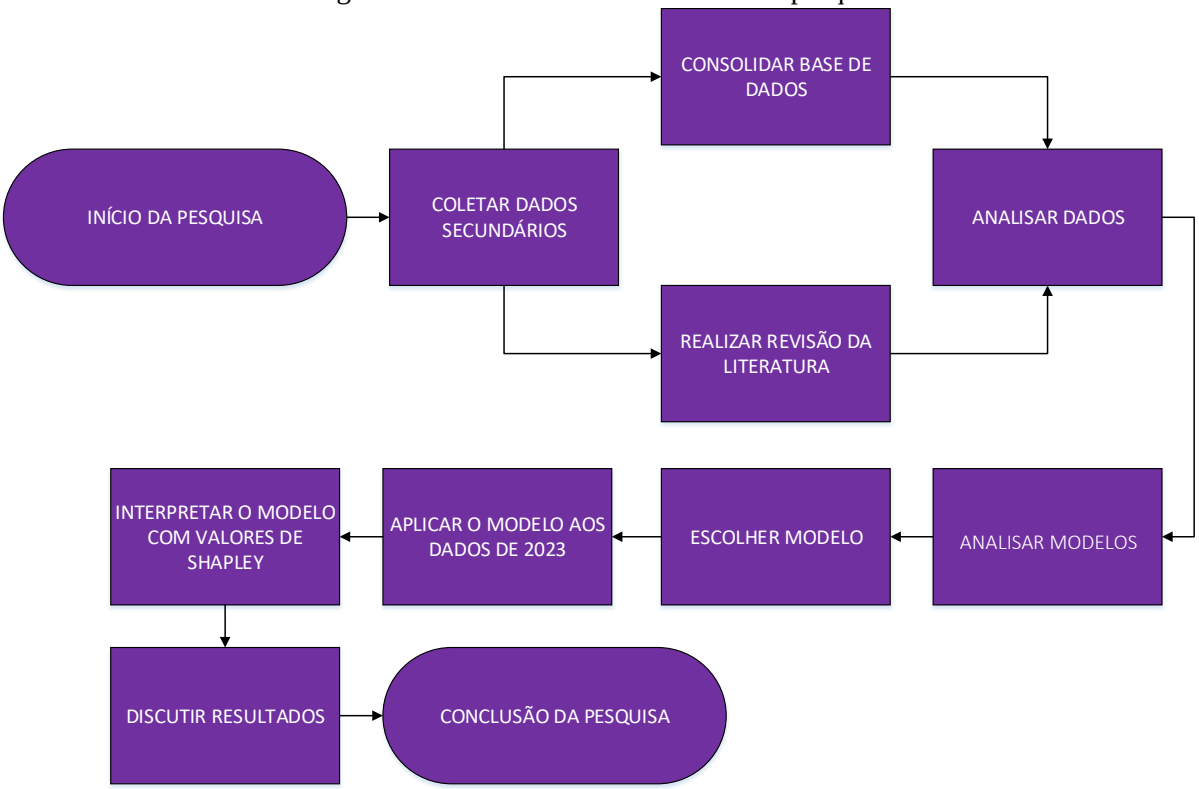
- Oferecer os resultados da pesquisa à comunidade e ao debate público dedicado ao aprimoramento da educação brasileira

1.4 METODOLOGIA

Primeiramente, o trabalho fará um levantamento dos dados secundários referentes ao contexto educacional brasileiro. A base que será construída consistirá na união dos chamados indicadores educacionais com as informações estruturais e de matrícula das escolas brasileiras. Em paralelo, por meio de uma revisão da literatura acadêmica preliminar, serão buscadas evidências e *insights* que sirvam de parâmetro à posterior aplicação do modelo matemático. Em seguida, uma análise será executada a partir dos dados consolidados com o intuito de alcançar os objetivos da pesquisa.

Então, será construído um modelo de previsão a partir da base de dados criada e das informações provenientes do diagnóstico produzido. Nesta fase, a revisão de literatura servirá para entender qual modelo melhor se encaixa para os dados disponíveis. A biblioteca Pycaret, da linguagem de programação em código aberto Python será usada para selecionar o modelo adequado. Este será aplicado aos dados referentes ao ano de 2023. Finalmente, os Valores de Shapley serão calculados para fornecer informações qualificadas aos gestores públicos e tomadores de decisão. A partir dos resultados encontrados pela pesquisa, serão delineadas as discussões e conclusões do trabalho. A Figura 2 ilustra o percurso percorrido pelo trabalho.

Figura 2 - Fluxo do desenvolvimento da pesquisa



Fonte: O Autor (2024)

Portanto, o trabalho realizará uma pesquisa aplicada (quanto à natureza) – visando auxiliar aos gestores públicos na tomada de decisão baseada em dados e evidências – exploratória (quanto ao objetivo) – buscando conhecer melhor as dificuldades, limitações e oportunidades de aplicar a Teoria dos Jogos em um contexto educacional. É uma pesquisa de abordagem quantitativa, explorando a coleta, limpeza e tratamento de dados secundários para construção do modelo preditivo.

Este trabalho está estruturado em seis capítulos, além desta introdução. No segundo capítulo, será apresentada a fundamentação teórica, explorando a Teoria dos Jogos, com foco em jogos cooperativos e a decomposição de Shapley, assim como os fundamentos do *Machine Learning* (Aprendizado de Máquina), abrangendo modelos de previsão e o conceito de *explainability* (explicabilidade), além de uma revisão da literatura, buscando casos de aplicação dos métodos componentes da base teórica do trabalho e que o norteiam.

O terceiro capítulo abordará o desenvolvimento do trabalho, detalhando a metodologia aplicada na análise dos dados e na implementação das estratégias teóricas propostas. No capítulo seguinte, os resultados obtidos serão minuciosamente discutidos, oferecendo insights e interpretações relevantes. O quinto capítulo se dedicará a explorar possíveis impactos resultantes da pesquisa, bem como suas limitações e as lacunas e oportunidades para trabalhos

futuros, visando avanços na pesquisa. Por fim, o sexto capítulo consolidará as conclusões deste estudo, apresentando uma síntese das descobertas, suas implicações e contribuições para a compreensão dos fenômenos em questão.

Essa estrutura proporcionará uma abordagem abrangente e aprofundada para a investigação proposta, contribuindo para o avanço do conhecimento nas áreas de Teoria dos Jogos, Aprendizado de Máquina (AM) e suas aplicações na área de educação básica do setor público.

2 FUNDAMENTAÇÃO TEÓRICA E REVISÃO DA LITERATURA

2.1 FUNDAMENTAÇÃO TEÓRICA

Na exploração dos conceitos com potencial benéfico ao trabalho em questão, algumas áreas do conhecimento emergiram. Uma delas é a Teoria dos Jogos, com a sua aplicabilidade abrangente e sua estrutura conceitual flexível, permitindo sua adaptação ao contexto estudado. Em especial, os jogos cooperativos se demonstraram um campo de grande valia ao estudo, como será detalhado posteriormente. Além disso, a pesquisa se utiliza do poder de modelos de previsão como instrumentos para extrair padrões a partir de dados complexos e investiga esses modelos de maneira a compreender melhor como suas previsões são geradas.

Nesta seção, exploraremos conceitos-chave da Teoria dos Jogos, aprofundando para uma análise dos jogos cooperativos. Em particular, examinaremos como jogadores que buscam alcançar objetivos comuns podem formar coalizões e negociar de maneira colaborativa para maximizar seus ganhos conjuntos. Além disso, abordaremos os Valores de Shapley como uma técnica central para uma distribuição das recompensas de um jogo aos jogadores de maneira mais eficiente com base em suas contribuições marginais à coalizão total. Ao compreender esses conceitos fundamentais, estaremos preparados para investigar mais profundamente as decisões e dinâmicas presentes em situações de priorização em alocação de recursos.

Em seguida, o trabalho abordará a ideia de AM, explorando modelos de previsão como instrumentos poderosos para extrair padrões a partir de dados complexos. Será discutido como a transparência dos modelos de previsão é uma questão na sua análise e como os Valores de Shapley são uma forma de lidar com essa problemática. A integração dessas teorias proporcionará uma base sólida para o alcance dos objetivos da pesquisa.

2.1.1 Teoria dos Jogos

Diversas são as situações da vida cotidiana em que pessoas, grupos ou organizações têm a necessidade de interagir partindo de interesses conflituosos. Sejam casos corriqueiros como uma negociação de preço de aluguel de imóvel, ou trabalhadores exigindo melhores salários dos seus empregadores, as soluções para esses desacordos podem não ser obtidas simplesmente. Muitas vezes, as partes interessadas se beneficiam pela influência de uma racionalidade externa

na decisão tomada, de maneira a encontrar uma solução possível que satisfaça – mesmo que parcialmente – a todas elas. Nesse contexto, esta seção aborda como a disciplina da Teoria dos Jogos analisa e modela situações dessa natureza.

A ciência em questão oferece um arcabouço teórico-matemático para apoiar a resolução do conflito entre as partes interessadas. Apesar do seu marco originário ter ocorrido na década de 1940, alguns estudos anteriores colaboraram para a sua gênese. Desde os trabalhos de Cournot (1838), com seu modelo de duopólio que conferiu base para elaborações posteriores, autores contribuíram, crescentemente, para a construção de uma teoria abrangente das interações estratégicas entre agentes sociais. Borel (1921), inclusive, foi pioneiro ao fornecer essa noção de estratégia, tão presente na Teoria dos Jogos moderna, quando falou em "método de jogo" – correspondente a todas as ações e circunstâncias possíveis de um jogador.

Mais tarde, em meados da década de 1940, surge o trabalho que é considerado a pedra fundamental da Teoria dos Jogos, em coautoria por Neumann e Morgenstern (1945). Nele, além de tratarem de jogos de soma zero e jogos na forma extensiva, os autores discutem uma forma de jogo que é de suma importância para este trabalho: cooperação e formação de coalizões entre jogadores, tema aprofundado no tópico a seguir. Ainda tratando de aspectos gerais da Teoria dos Jogos, os autores propõem uma estrutura simples, capaz de modelar qualquer jogo competitivo: um conjunto de jogadores, cada um possuindo um conjunto de estratégias possíveis, e uma função que representa a recompensa resultante da combinação das suas estratégias com as dos demais. Como pode ser conferido a seguir, trabalhos recentes seguem esse mesmo arcabouço.

A importância da definição de um modelo para análise pela perspectiva de um jogo é alertada por Myerson (1991). Seu destaque mostra como o ato de esmiuçar o cenário no qual o jogo está inserido – e que deve servir como analogia à conjuntura organizacional de interesse – precede qualquer avaliação a ser feita. Logo, são as premissas do modelo que determinam as possíveis inferências sobre ele aplicáveis, assim como elas permitem aos seus analistas a compreensão da lógica que o estrutura e como esta é análoga ao contexto examinado.

Sintetizando o que seria a Teoria dos jogos, Dufwenberg (2011) a entende como uma coleção de ferramentas utilizada para examinar situações em que decisores interagem de modo que suas decisões influenciam uns aos outros. Esse tipo de interação também é explorado no conceito cunhado por Gibbons (1992). Para ele, essa é a matéria que estuda as interações estratégicas entre diferentes jogadores racionais em conflito. Mais uma vez, palavras como "estratégia", "jogadores" e "racionais" surgem no cerne da Teoria dos Jogos. Mesmo que não

despertem um olhar mais atento, seus conceitos são de extrema importância para estruturar um modelo.

O cuidado necessário para definir todos os elementos constitutivos de jogos e que são imprescindíveis para a sua análise foi uma preocupação de Fiani (2015). No seu livro, conceitos básicos são explicitados para que o leitor compreenda do que se trata cada expressão no mundo de jogos. Para ele, um jogo é um modelo formal que descreve uma situação envolvendo interações entre agentes racionais que se comportam estrategicamente – em especial, jogos de estratégia. Agentes (ou jogadores), nessa conjunção, são indivíduos ou um grupo de indivíduos capazes de tomar decisões que afetam outros. Tais decisões são realizadas a partir de um conjunto de ações possíveis, interferindo nos outros agentes, portanto representando interações. O comportamento dos jogadores nessas interações é considerado estratégico, o que significa que eles consideram as consequências dos seus atos nos outros jogadores, bem como aquelas dos outros sobre si.

Partindo desse arcabouço conceitual, o autor enxerga a Teoria dos Jogos como um campo da Pesquisa Operacional destinado à análise do comportamento de agentes envolvidos em jogos. O seu potencial pode ser verificado pela abrangência das suas técnicas de modelagem, o que permite uma variedade imensurável de aplicações nas ciências sociais. Sua utilização em campos como ciência política, economia, sociologia e áreas correlatas não são uma novidade. Entretanto, a versatilidade de aplicação também já foi vista em áreas nas quais a utilização não é trivial, como a filosofia. Bruin (2005) observou como a Teoria dos Jogos, por exemplo, pode, a partir do equilíbrio de Nash, tratar de eficiência e liberalismo; dos jogos de barganha, lidar com moralidade e racionalidade; da teoria dos jogos evolucionária, abordar justiça distributiva e igualitarismo. Esta pesquisa, na sua revisão de literatura, abordará casos de aplicação da Teoria dos jogos no ramo educacional.

Como já exposto, jogos podem ser classificados como não cooperativos, quando cada um dos jogadores busca os melhores resultados para si em antagonismo aos resultados dos demais jogadores; ou cooperativos, quando os jogadores formam coalizões buscando recompensas compartilhadas em benefício do grupo. Esta última circunstância conduz o trabalho a estudar a base teórica que a originou.

2.1.2 Jogos Cooperativos e Valores de Shapley

Como argumentado ao longo do texto até então, a amplitude de aplicação da Teoria dos Jogos permite seu emprego em situações e contextos bastante distintos entre si. Os problemas oriundos de cada tipo de jogo exigem uma modelagem e técnica próprias a eles. Considerando-se um cenário no qual agentes escolhem as suas estratégias de maneira independente, sem que haja comunicação ou colaboração entre eles, para que os seus ganhos sejam os maiores possíveis, o conflito reside na possibilidade de haver um choque entre os interesses das partes envolvidas. Dessa forma, a racionalidade conduz a análises individualizadas para cada jogador.

Por outro lado, no caso de jogos nos quais os participantes unem esforços em busca do benefício comum a todos, não há interesses contraditórios em relação ao resultado do jogo. Esse panorama descreve os chamados jogos cooperativos, nos quais as análises giram em torno das coalizões formadas pelos jogadores (NASH, 1951). Neste cenário, eclode a necessidade de formular a divisão dos frutos do empenho coletivo de tal sorte que esta seja mais justa em concordância com a contribuição de cada jogador – e suas coalizões – ao resultado obtido. Curiel (2013), antes de apresentar uma definição formal acerca de jogos cooperativos, traz um exemplo anedótico para demonstrar a situação-problema na qual repousa o conflito. Nele, cinco pessoas se unem em prol da criação de um negócio, investindo seu capital e competências. A divisão dos retornos óbvia seria 20% para cada uma, entretanto os participantes notam que, caso formassem grupos menores do que a totalidade, as coalizões representariam uma parcela dos proventos maior que a soma das parcelas dos indivíduos. Tendo essa informação em mãos, os participantes utilizam-na como argumento para obter uma maior quantia dos retornos para a sua coalizão – e, consequentemente, para si. Tal impasse envolvendo a divisão mais justa de acordo com a contribuição dos agentes e as possíveis coalizões merece uma atenção meticulosa de maneira a superar o litígio.

Finalmente, um jogo cooperativo na sua forma de função característica é um par ordenado $< N, v >$ tal que N representa o conjunto de todos os jogadores, v designa o valor da coalizão do subconjunto de jogadores S $(v(S))$ ou da grande coalizão N $(v(N))$ e G^N o conjunto de todas as coalizões possíveis (CURIEL, 2013). Considerando o exemplo anterior, imagina-se que cada jogador i avalie a sua contribuição em todas as coalizões possíveis. A equação 2.1 retrata essa situação, calculando a diferença entre o valor da coalizão S, da qual o jogador i não faz parte, $(v(S))$ e o valor da mesma coalizão adicionando i $(v(S \cup \{i\}))$, como mostra a equação

$$v(S \cup \{i\}) - v(S) \tag{2.1}$$

para todas as G^N coalizões possíveis. Shapley (1953) desenvolveu um método para o cálculo dessa contribuição marginal para cada jogador, sintetizado na equação 2.2 a seguir

$$\varphi_i(v) = \frac{1}{n!} \sum_{S \subset N \setminus \{i\}} \left[(s!\,(n\ - s - 1)!) \big(v(S \cup \{i\}) - \ v(S) \big) \right] \tag{2.2}$$

na qual:

- $\varphi_i(v)$ é o valor de Shapley para o jogador i;
- S é a coalizão considerada previamente à participação do jogador i;
- n é o número total de jogadores em N;
- s é o número de jogadores na coalizão S
- $v(S \cup \{i\})$ é o valor da coalizão quando o elemento i é adicionado a S;
- $v(S)$ é o valor da coalizão S.

O valor de Shapley, portanto, é um cálculo da contribuição marginal de um jogador i a partir da análise da diferença entre coalizão S com e sem i ($v(S \cup \{i\}) - v(\{i\})$) somada para todas as coalizões que i não faz parte ($S \subset N \setminus \{i\}$). Cada um desses produtos é ponderado por todas as maneiras possíveis de sua obtenção ($s!\,(n\ - s - 1)!$) divididas por todas as coalizões possíveis ($n!$). No total, a soma de todas as contribuições marginais $\Phi_{i \ldots N}$ será igual ao valor da coalizão total $v(N)$.

A partir da utilização dessa técnica que pretende distribuir as recompensas de um jogo de maneira mais eficiente, é necessário discutir a forma como serão calculadas as recompensas para cada coalizão, bem como quem serão seus jogadores e em que contexto o jogo se insere. Tais reflexões, além de serem de importância central a este trabalho, representam o princípio do que, posteriormente, culminará na modelagem do jogo cooperativo basilar para as análises do estudo.

Como pontuado na introdução deste texto, os objetivos da pesquisa versam, em linhas gerais, em fornecer informações para o debate sobre a educação pública, seja a partir de um instrumento de previsão do resultado gerado pelas escolas, ou pela informação de quais fatores contribuem mais (ou menos) para esses resultados. Tais atributos “competem” entre si pela atenção dos decisores, por meio da qual as decisões para investimentos públicos na área são tomadas.

É possível, então, representar os atributos das escolas como jogadores em um jogo cooperativo no qual a taxa de aprovação é o resultado que todos buscam maximizar e a recompensa a ser dividida entre eles é a atenção que cada um receberá dos gestores públicos. Por exemplo, diversas escolas possuem atributos como a sua região geográfica, estado, rede administrativa, se possui biblioteca, quadra de esportes, serviço de internet etc., com taxas de aprovação variadas. Considerando como objetivo comum de uma escola elevar a sua taxa de aprovação à mais próxima possível de 100%, o resultado obtido é imputado aos atributos na medida das suas contribuições.

Quanto maior a magnitude da contribuição para índices mais ou menos próximos de 100%, os atributos recebem maior atenção e, possivelmente, intervenções e investimentos públicos. Se for verificado que o fator regional é relevante, os gestores devem direcionar esforços para a melhoria das regiões com piores rendimentos e aprender com a experiência das regiões com melhores resultados; se for observado que a localidade das escolas exerce grande influência no seu resultado, deve-se investigar o que faz escolas urbanas renderem mais do que as rurais – ou o contrário. É possível, também, a partir dos resultados do estudo, direcionar as decisões a serem tomadas, como a execução de projetos de construção de bibliotecas em detrimento de quadras de esportes, ou ainda a construção de mais escolas em uma rede ao invés da expansão de vagas nas escolas a ela pertencentes.

Assim, o jogo visa compreender melhor quais atributos são responsáveis pela maior ou menor aprovação escolar, em maior ou menor escala para alocar-se recursos proporcionalmente. Este estudo, portanto, utiliza o arcabouço teórico de jogos cooperativos para sugerir quais fatores devem ser priorizados quando da necessidade de decidir como será realizada a distribuição eficiente dos investimentos públicos.

Quanto às recompensas, os objetivos do estudo demandam que seja construída uma ferramenta que aponte para políticas futuras, em especial para as discussões do novo PNE, com vigência de 2025 a 2034. Assim sendo, não basta apenas descrever o cenário revelado pelos dados, mas também prever o que estes indicam para o futuro. Isto é, fornecer aos gestores públicos um cenário possível de como a aprovação escolar responderá se alguns dos atributos forem alterados como uma simulação de políticas propostas. Para tal fim, os retornos do jogo serão calculados utilizando o conceito de AM, interpretadas como a previsão do modelo. Seus atributos (ou *features*) serão considerados os jogadores. Esse tema, junto à *explainability* (explicabilidade, em tradução livre) será abordado de maneira mais profunda no tópico a seguir.

2.1.3 Aprendizado de Máquina e Explicabilidade

A partir da extensa quantidade de dados disponibilizados, a possibilidade de direcionar esforços para as oportunidades e desafios da educação aumenta o poder interpretativo de pesquisas em educação (DANIEL, 2019). Entretanto, apesar de haver evidência de que esses dados podem servir à análise e posterior melhoria dos sistemas educacionais (PARK, 2020), a sua vasta disponibilidade pode gerar, também, uma dificuldade para analisá-los eficientemente. Nas ciências sociais, esse fator representa uma dificuldade ao uso de métodos iterativos há décadas (EDDY, 2001). Deste modo, os decisores de organizações educacionais podem ceder ao impulso de enfrentar o desafio de definir as prioridades como um mero exercício de preferências pessoais ao invés de uma decisão baseada em evidências.

A Inteligência Artificial (IA) surge como alternativa para lidar com tais tarefas complexas (DIETTERICH, 1996), seja com independência ou mínima interferência humana (RUSSELL e NORVIG, 2016). Jordan e Mitchell (2015) apontam que o AM, dando um passo à frente, lida com a tarefa de construir sistemas que não só realizam tarefas, como melhoram a si mesmos automaticamente. Os autores ainda mostram que esses métodos têm transformado o processo de tomada de decisão em algo cada vez mais baseado em evidências em diversas áreas do cotidiano, incluindo educação.

Hastie *et al.* (2009) apresentam o AM como um conjunto de técnicas matemáticas que dão a algoritmos computacionais a habilidade de aprender. A capacidade de aprender dos sistemas é provocada a partir da experiência que dados de treino relacionados a um problema específico geram (JANIESCH; ZSCHECH e HEINRICH, 2021). Portanto, pode-se resumir os métodos de AM como algoritmos capazes de manipular um conjunto de dados de treinamento para que possam classificar corretamente dados que não foram observados anteriormente (COPPIN, 2015).

Coppin (2015) descreve como diferentes tipos de modelos de AM, realmente “aprendem” na prática. Para este trabalho, os conceitos mais relevantes e que serão abordados se limitam ao aprendizado supervisionado e aprendizado por treinamento. Neste a tarefa gira em torno de classificar as entradas a partir de um conjunto de dados já classificados. O sistema, portanto, buscará padrões nas classificações a ele fornecidas e, a partir delas, tentará aplicar estes padrões para classificá-los, bem como – e principalmente – novas entradas. Treinamento é um caso particular do aprendizado supervisionado, no qual o algoritmo é apresentado às classificações previamente definidas para que ele aprenda como classificar. No caso de extensas quantidades

de dados, mais uma vez esses modelos se destacam, pois a diversidade de entradas deixa a base de treino mais robusta, gerando previsões mais assertivas.

Apesar dos benefícios de lidar com grandes conjuntos de dados e prever resultados prováveis (CAMACHO *et al.*, 2018), os modelos mais complexos construídos possuem baixa ou nenhuma transparência acerca de como obtêm suas saídas. Os modelos de previsão, por mais que apresentem resultados altamente precisos, são uma caixa-preta quando busca-se entender como eles os entregam. Rudin (2018) defende que, em casos cujas decisões podem resultar em consequências graves para a vida humana, como justiça criminal e saúde, não sejam utilizados nenhum modelo considerado uma caixa-preta, tamanho o risco. Em áreas que não oferecem risco direto à vida – como educação – emerge a necessidade de aprimorar a explicabilidade da IA. A ideia é elucidar as ações de um modelo de maneira que possam ser compreendidas qualquer que seja a pessoa consumindo suas previsões (MUNN; PITMAN, 2022).

É neste cenário de desenvolvimento do conhecimento científico que os Valores de Shapley se apresentam não apenas como uma ótima técnica para explicar e interpretar os modelos de previsão de AM, como também são diversas as formas de utilizá-la, como mostram Sundararajan e Najmi (2020). A aplicação dessa técnica aumenta a confiança e transparência das previsões (AAS; JULLUM e LØLAND, 2019), suprindo a demanda que consumidores da explicabilidade possuem de entenderem os modelos construídos, sejam eles seus praticantes, observadores ou usuários finais (MUNN e PITMAN, 2022).

Tendo em vista esses conceitos, a seção seguinte explorará a literatura acadêmica mais recente buscando entender como esses conhecimentos têm sido aplicados e como este trabalho pode contribuir ao avanço dos campos da Teoria dos Jogos e AM. A revisão da literatura também tem o intuito de conduzir a pesquisa ao alcance dos seus objetivos, iluminando-a em direção à melhor maneira de aplicar as técnicas para o cenário proposto.

2.2 REVISÃO DA LITERATURA

Em busca de desenhar a metodologia do estudo e compreender como outros autores usufruíram dos conceitos apresentados anteriormente, uma revisão da literatura mais recente foi realizada. Esta seção, portanto, mergulha no universo do AM empregado à educação, explorando as maneiras pelas quais essa vertente da IA vem sendo utilizada no meio acadêmico. Em especial, quais foram os fatores utilizados em pesquisas anteriores para construção dos modelos. A partir dessa investigação, uma lacuna na produção científica será observada para ser confrontada por esta pesquisa. Em seguida, será apresentada a biblioteca PyCaret – uma

ferramenta da linguagem de programação Python – como forma de facilitar a escolha do método de AM a ser implementado.

Para entendimento de como os Valores de Shapley podem ser usados para aumentar a transparência do modelo, serão explorados casos de aplicação dessa técnica, em especial associada a modelos de previsão. Para otimizar a busca por artigos relevantes, será utilizado o software ASReview, um aliado valioso na organização e análise da literatura. Através dele, será possível identificar e selecionar, entre estudos de alta qualidade, quais se encaixam no escopo da pesquisa.

Com esta revisão de literatura, espera-se ter uma clareza dos benefícios, desafios e do potencial de uso da AM no campo da educação, bem como entender como essa técnica pode ser combinada aos Valores de Shapley para aumentar a explicabilidade do modelo de previsão. Em seguida, esses aprendizados irão nortear o detalhamento da metodologia da pesquisa e o seu desenvolvimento, fase que será abordada no próximo capítulo.

2.2.1 Aprendizado de máquina e educação

O uso de técnicas de IA e AM tem sido, crescentemente, popularizado nos diversos setores da vida cotidiana. Desde ferramentas de geração de texto, imagens e vídeos até softwares sofisticados de classificação de risco para crédito ou previsão de falha de máquina, o tema tem sido explorado no mundo acadêmico. Este tópico investiga os trabalhos dessa área sobrepostos ao contexto educacional para intuir os tipos de usos desses conceitos e seus modos de aplicação.

Com este fim, foram pesquisados trabalhos publicados nos últimos cinco anos, em jornais indexados à base *Web of Science*, cujo título ou resumo contenham uma combinação das palavras *machine*, *learning* e *education* e que as expressões *machine learning* e *education* estejam nas palavras-chave. A pesquisa se limitou a essa base porque nela estão os periódicos mais rigorosos e de maior prestígio na área. A busca resultou em 234 saídas, que foram, então, utilizadas para alimentar o software ASReview (VAN DE SCHOOT *et al.*, 2021), empregado com o fim de tornar mais eficiente o processo de triagem dos registros. A partir da classificação de artigos em relevante (e irrelevante), o software apresentou ao autor, dentre as mais de 200 saídas, aquelas com maior probabilidade de serem relevantes. Ressalta-se que apenas o pesquisador possui a autonomia de classificar um texto como relevante ou irrelevante. A seguir, são apresentados alguns deles.

Pektas (2023) produziu uma revisão de artigos publicados desde 1979 até 2023 em jornais de rigor e prestígio para extrair quais são os principais temas de pesquisa de AM relacionados

a educação. Suas conclusões deixam claro como o campo vem evoluindo, visto que a maior parcela dos trabalhos produzidos no período de 45 anos se concentra nos últimos seis, com a explosão de publicações depois de 2017. Ainda, comparando-se os anos de 2017 e 2022, o número de estudos publicados foi de 15 a 145, o que corrobora o aumento da exploração do tema. Os temas basilares encontrados na pesquisa foram: modelos de previsão de taxa de evasão, performance acadêmica e admissões a partir de perfis dos estudantes; técnicas de avaliação automatizadas e adaptáveis, seja de estudantes ou professores; sistemas inteligentes de tutoria; cursos abertos *online* em massa (do inglês, MOOCs); NLP (sigla, em inglês, para processamento de linguagem natural); e previsões em educação à distância.

Outros estudos demonstram o alto potencial que pesquisadores têm enxergado para o setor educacional do seu usufruto do AM. Fahd *et al*. (2022) também observaram um crescimento exponencial a partir de 2017 do número de estudos nesse campo de aplicação. Além disso, trazem a informação de que apenas dois estudos foram encontrados no Brasil, de acordo com a limitação que o trabalho se impôs. Rahul e Katarya (2023) notaram um – menos íngreme – crescimento consolidado a partir de 2015 e com 70% dos trabalhos analisados sendo concentrados na previsão da performance acadêmica dos estudantes, ponto em comum entre este e os anteriores.

Luan e Tsai (2021) analisaram 40 estudos relativos à personalização da educação e concluíram que a maioria deles se dedicam à previsão do nível de aprendizado ou de abandonos. Além disso, em geral estão focados em ambientes de educação remota. AL Mutawa e Sruthi (2023) utilizaram AM para prever a satisfação e as emoções dos estudantes na sua interação com o computador para aprimorar a experiência do usuário nesse formato. Munir, Vogel e Jacobsson (2022) fizeram uma revisão sistemática apenas de aplicações executadas em ambiente virtual, delineando os principais temas utilizados nesses estudos: tutores inteligentes, previsão de abandono e de performance, adaptação de estilos de aprendizagem, aprendizado baseado em grupo e automação.

Há evidência de que o nível de adaptação de estudantes ao ambiente virtual de aprendizagem pode ser influenciado pela condição socioeconômica, o que está relacionado à sua moradia, bem como a qualidade de fornecimento de internet, o dispositivo de acesso e a estrutura física para acessar a estrutura digital (IPARRAGUIRRE-VILLANUEVA *et al.*, 2023). Fatores como saúde mental e diversidade no acesso ao ensino superior também foram observados como relevantes no contexto da restrição instituída pela pandemia (VILLEGAS-CH; GOVEA; REVELO-TAPIA, 2023). Neste, a sua aplicação foi capaz de proporcionar, à gestão da instituição, a capacidade de intervir a tempo, aumentando a retenção dos estudantes.

Saúde mental foi utilizada junto a atributos relativos a moradia, conexão com internet e com o ambiente virtual para prever a performance acadêmica (HOSSAIN *et al.*, 2023).

Essa conjuntura na qual o aproveitamento do AM não representou nenhuma ação sobre a infraestrutura digital das instituições – além daquelas causadas pela própria pandemia – foi observada em outros trabalhos publicados. Zaman *et al*. (2023) utilizaram, como objeto de estudo, um Distrito do nordeste indiano, região que tem sofrido historicamente de mazelas sociais decorrentes de conflitos armados, entre outras questões. Essa circunstância privou, em especial, a juventude do acesso a instituições educacionais de nível primário e secundário.

O estudo tirou proveito, de maneira pioneira, de técnicas de IA Geoespacial e AM para extrair informações acerca da educação no Distrito e construir um sistema de apoio a decisão que visa auxiliar as partes interessadas a melhorarem a qualidade da educação escolar na região. A análise foi baseada em fatores como o histórico de aprovações e reprovações, localização e demografia vizinha, informações dos discentes, informações dos professores, a qualificação destes, se são permanentes ou temporários, instalações e infraestrutura, acessibilidade, orçamento alocado à escola. A variável-alvo a ser prevista foi a taxa de aprovação.

O modelo de IA Geoespacial focou no estudo dos problemas de acessibilidade. Já o algoritmo de AM foi usado para encontrar os fatores responsáveis pela maior ou menor taxa de aprovação escolar. Este obteve uma exatidão de 93%. Fatores importantes para melhores resultados foram a existência de muros no perímetro escolar, número de docentes qualificados e a sua quantidade total, categoria da escola, razão pupilo/professor, matrículas, presença de *playground*, número de salas-de-aula, rede de distribuição elétrica, água potável e razão alfabetizado/analfabeto.

A contribuição mais crítica no estudo foram as sugestões de intervenção para melhorar a qualidade da educação escolar. Foram propostas as seguintes ações: priorizar as áreas para intervenção a partir das lacunas produzidas pelo modelo geoespacial; proporcionar aos estudantes opções de transporte seguras e confiáveis nas áreas mais remotas para garantir a sua frequência regular; desencorajar a matrícula direta em escolas secundárias, dado a menor taxa de abandono dos estudantes que seguiam um percurso contínuo entre as etapas de ensino na mesma escola; agir sobre as escolas com infraestrutura precária; garantir equidade ao alocar recursos baseado nos resultados do modelo geoespacial; desenvolver políticas focadas no treinamento e desenvolvimento docente, em especial nas regiões com menor taxa de aprovação; e mobilizar recursos com organizações e comunidade para apoiar escolas com baixas taxas de aprovação.

Estes casos levantam a curiosidade sobre outras aplicações nas quais as variáveis que explicam o modelo de previsão não são apenas aquelas que possuem caráter educacional. Um outro exemplo disso é o trabalho de Poudyal, Mohammadi-Aragh e Ball (2022), que considera, além de aspectos educacionais, os demográficos, pessoais, psicológicos e outras influências do meio no qual vive o(a) estudante para prever sua performance acadêmica. Destacam-se, no escopo desta pesquisa, os atributos demográficos: gênero, idade, raça, local de residência etc.

O trabalho de Abraha (2019), em contrapartida, fortalece a ideia de que fatores estruturantes do próprio ambiente escolar possuem grande influência na qualidade da educação. No seu estudo, características como quantidade limitada de docentes, baixa motivação e qualificação docente, liderança escolar ineficiente, capacidade do prédio, baixa participação familiar, instalações inadequadas, baixa qualidade de salas-se-aula, alta razão discente/docente e ausência de materiais de instrução foram consideradas como responsáveis pela baixa qualidade da educação. Tais informações serão cruciais para a construção da base de dados que servirá de entrada no desenvolvimento deste trabalho.

Tais estudos indicam grupos ou tipos de características que podem ser relevantes para as previsões do modelo que será produzido. Essa triagem é de suma importância, dado o problema da dimensionalidade dos dados que exige um maior esforço computacional tanto quanto maior for o tamanho da base de dados. Cardona *et al*. (2023) inferem, a partir de revisão sistemática, que a eficiência de algoritmos para previsão do sucesso estudantil está circunscrita a um limite de até 68 variáveis, indicando que essa escolha está relacionada e induz quais análises podem ser realizadas.

Modelos de previsão também foram estudados no contexto da gestão educacional (XU; DENG, 2023). Com a necessidade de lidar com a pressão causada pela quantidade de dados para gerenciar a educação superior e afim de melhorar a eficiência e o trabalho docente, eles criaram um sistema de gestão educacional. Este tem o papel de propor percursos personalizados de aprendizado a partir de decisões baseadas em dados e evidências. Utilizando o sistema de gestão de bases de dados relacional SQL como servidor de retaguarda, o modelo proposto obteve uma exatidão de 94,7% na previsão da qualidade do trabalho docente.

Excetuando a pesquisa realizada na Índia, todas têm o ensino superior como objeto de estudo. Esse ramo de exploração reforça o fato de que questões de natureza socioeconômica estão entre as mais relevantes para o sucesso – ou fracasso – acadêmico (PALACIOS *et al.*, 2021). Trabalhos utilizados na educação básica são escassos, mas também apresentam resultados destacados (YOUSAFZAI; HAYAT; AFZAL, 2020).

A partir desta investigação, foi enxergada uma lacuna na literatura acadêmica. Há poucos trabalhos que tratam da educação básica, em especial no Brasil. Em especial, não foi encontrado nenhum trabalho de abrangência nacional que possa nortear uma política de financiamento e gestão educacional integrada entre os entes federativos. Este trabalho, portanto, irá contribuir para preencher a lacuna observada, conjuntamente ao apoio no desenho de políticas públicas educacionais no nível federal. A seguir, artigos que exploram os Valores de Shapley no contexto de políticas públicas serão revisados para dar embasamento à inovação proposta.

2.2.2 Valores de Shapley

Chantreuil *et al*. (2021) buscaram verificar a parcela de responsabilidade de diversos fatores para a desigualdade de renda nos Estados Unidos (EUA). A partir da decomposição dessas características em Valores de Shapley, se destacaram as desigualdades de renda associadas à raça/cor e, em maior escala, gênero – enquanto esta diminuiu, aquela aumentou entre 2005 e 2017. Os resultados embasam a sugestão de caminhos para reduzir as disparidades econômicas por meio de políticas públicas. Recomendações ao setor incluíram políticas e ações afirmativas, em especial no ramo de educação – uma das origens das desigualdades sociais – e ações do poder público que objetivem a igualdade de gênero.

Gambau et al. (2022) ponderaram as influências de alguns indicadores sociais em termos de vulnerabilidade econômica, especialmente associada às consequências geradas pandemia do COVID-19 nos Estados Unidos (EUA). Compararam-se as participações de características da demografia entre grupos sociais, tais quais raça/cor, gênero, grau de escolarização e estado de moradia. Suas conclusões apontam para indicadores educacionais contribuindo por uma parcela maior para o aumento da desigualdade ocorrido no período pandêmico. Um exemplo foi o maior grau de escolarização – como nível superior – representou uma menor exposição ao risco de pobreza.

A constatação do estudo fortalece a ideia de que uma ciência para a mudança social exige um maior investimento em pesquisas educacionais. O esforço de entender um sistema tão complexo como a educação pública, no entanto, demanda uma investigação minuciosa do âmago desse aparelho. Apenas através da coleta de dados e produção de evidências a mediação dos problemas poderá ser conduzida por um processo decisório que realmente compreenda todas as nuances do que se busca melhorar, ao invés de propor soluções impulsivamente (BRYK *et al.*, 2015).

Aleu *et al*. (2021), seguindo esse encadeamento de ideias, buscaram os fatores que exercem influência na satisfação dos estudantes em uma universidade mexicana e observaram que cinco são os pilares dessa percepção: limpeza do local; habilidades de ensino do corpo docente; serviços noturnos; funções no gerenciamento de estudantes de mestrado; e idade dos estudantes de mestrado. A decomposição dessas vertentes da vida universitária pode conferir mais robustez ao diagnóstico feito pelos autores. É nesse contexto que os Valores de Shapley exercem um papel cuja valorização é mais estimada.

Esse método foi aplicado seguindo o eixo de alguns dos trabalhos apresentados nesta dissertação. Kar *et al.* (2023) avaliaram as questões críticas para a adaptação ao ambiente virtual, impelidos pela realidade pós-pandemia. Construíram um modelo do tipo *Random Forest* para classificar os estudantes em baixa, média ou alta adaptação e, em seguida, utilizaram os Valores de Shapley para explicar os resultados. Esta técnica não só expôs os parâmetros mais determinantes para a classificação total, como também foi capaz de indicar a contribuição de cada atributo dentro de cada classe. Decorre, então, do produto desse trabalho, o portencial de direcionar esforços para as questões prioritárias específicas de cada classe de estudantes ao invés de tratar do conjunto total de discentes.

Um outro benefício da utilização dos Valores de Shapley é a redução de subjetividade do processo decisório em sistemas multicritério de apoio à decisão (MCDSS) (MUHAMEDYEV *et al.*, 2020). Os autores uniram AM ao método Shapley Additive Explanation (SHAP) em um MCDSS, abreviando o trabalho de priorização de fatores mais relevantes para os indicadores de qualidade de educação. Assim, foi possível aumentar a explicabilidade do modelo e fornecer recomendações personalizadas ao contexto estudado acerca de quais características devem ser alteradas para melhorar a qualidade da educação.

Com a emergência da educação à distância, surge a discussão sobre as condições nas quais ela é possível. Países de baixa e média renda, em detrimento dos ditos países ricos, possuem várias desvantagens estruturais que, na prática, impedem o acesso universal a esse modelo de ensino (MCINTYRE, 2023). Mendonça *et al.* (2023) observaram, também, a importância das tecnologias digitais no desenvolvimento estudantil e na qualidade da educação. Este estudo tem maior valor para esta dissertação por ter sido realizado no Brasil e ter aplicabilidade nacional.

Buscando entender quais fatores exercem maior influência positiva para o IDEB, principal indicador de qualidade da educação brasileira, utilizaram a abordagem SHAP. Concluíram, então, que as variáveis mais impactantes no modelo final foram: número de computadores por estudante, tempo de serviço dos professores, acesso a internet do tipo banda-larga, investimento no treinamento em tecnologias para os professores e laboratórios de

informática na escola. Isso indica, portanto, que esses fatores são relevantes para o cenário nacional.

Enquanto este estudo se absteve de análises socioeconômicas, também há evidências acerca das desigualdades sociais e demográficas na literatura explorada. Desigualdades entre zona urbana e rural (GARG; CHOWDHURY; KANCHAN, 2022), grupos raciais e étnicos, regiões geográficas (VITORINO DE ARRUDA; RAMOS, 2023) e gêneros (MCINTYRE, 2023). Esses dados nortearão a construção da base que será usada como fonte de dados para o modelo desta dissertação.

A seguir, serão abordados trabalhos recentes acerca da escolha da técnica de AM a ser utilizada e como aumentar a capacidade interpretativa desse modelo a partir da explicabilidade. O uso da biblioteca PyCaret na produção científica será explorado para entender como ela pode facilitar o processo da pesquisa e, em especial, como ela tem sido utilizada em conjunto com a abordagem SHAP.

2.2.3 PyCaret e Explicabilidade

A biblioteca PyCaret é uma alternativa em código aberto para possibilitar aplicações de AM com poucas linhas de código, visando a democratização do conhecimento (ALI, 2020). Dado que a sua criação tem menos de cinco anos, as aplicações na literatura ainda são poucas. Realizando uma busca bastante abrangente na base *Web of Science* utilizando como termo de pesquisa apenas a expressão “PyCaret”, 44 trabalhos foram encontrados, compreendidos entre os anos de 2021 e 2024. O número de aplicações mais do que duplicou de um ano para o outro e, até a metade deste último, a quantidade de publicações envolvendo o tema na base citada já é a mesma do ano anterior.

A técnica de aplicação da documentação descrita foi utilizada para categorizar, detectar ou prever questões relacionadas a diabetes (JOSE *et al.*, 2024; TABACARU *et al.*, 2024; WHIG *et al.*, 2023). Também foi explorada no campo dos impactos ambientais (GUPTA *et al.*, 2023; XIN *et al.*, 2024), assim como a análise de dados relativos a distúrbios neurológicos (RADHAKRISHNAN; BORUAH; RAMAMURTHY, 2022). Entretanto, não foi encontrado, nessa base, nenhum trabalho que trate do emprego da técnica no setor educacional, o que representa uma lacuna a ser preenchida e expõe o caráter inovador desta pesquisa. Entretanto, os trabalhos revisados foram importantes para esta dissertação pois eles mostram como a PyCaret foi aplicada ao contexto da produção acadêmica.

A biblioteca foi usada, basicamente, comparando uma quantidade de modelos de AM em termos de um conjunto de métricas de avaliação. Algoritmos como *Decision Tree*, *Random Forest*, *K-Nearest Neighbors*, entre vários outros, são avaliados em termos de conceitos de mensuração de desempenho tais quais *Accuracy*, *Precision*, *Recall*, *F1-score* etc. A pessoa responsável pela pesquisa, então, pode discernir entre as técnicas analisadas de acordo com a sua performance.

A biblioteca também pode ser utilizada para avaliar um modelo pré-determinado e, a fim de evitar o caráter de "caixa-preta" do algoritmo, pode-se adicionar o método SHAP (KILIC *et al.*, 2023). A técnica, seguindo o conceito de Valores de Shapley, calcula a magnitude da contribuição positiva ou negativa de cada atributo.

Esse foi o caminho seguido por Tsai *et al.* (2022), que buscaram prever o tempo de resposta de laboratórios entre o momento que o técnico solicita o resultado até a entrega deste. A documentação PyCaret avaliou e comparou múltiplos modelos de regressão para determinar qual seria o melhor entre eles e os valores SHAP explicaram o modelo no que diz respeito à contribuição de suas variáveis. A regressão *ExtraTrees* destacou-se pela melhor performance entre os algoritmos preditores e a análise de valores SHAP indicou um tempo de resposta cuja maior influência é dada pela carga de trabalho na fase de pré-análise e no número de visitas aos módulos de análise das amostras.

No setor alimentício, a hipótese sobre a capacidade de classificar 10 variedade de queijo por meio de algoritmos supervisionados de AM foi investigada. Utilizou-se PyCaret para avaliar que os modelos *ExtraTrees* e *Random Forest* alcançaram 90% de acerto nas classificações (FROHLICH-WYDER; BACHMANN; SCHMIDT, 2023). Além disso, o módulo SHAP aferiu a importância dos atributos para a caracterização das variedades de queijo suíço.

Para ponderar sobre as perturbações em uma floresta na Turquia, Eker, Alkis e Aydin (2024) utilizaram o PyCaret para comparar 14 algoritmos de AM e ordená-los de acordo com o parâmetro *AUC* – área sob a curva, uma métrica de avaliação da exatidão de um modelo preditivo. Posteriormente, a análise SHAP pôde verificar os fatores-chave relativos às previsões do modelo tanto para danos causados por incêndios, quanto por insetos e vento. Os fatores importantes são a densidade e distância de estradas, tipo da floresta, precipitação anual, temperatura média no trimestre mais quente e velocidade máxima do vento. Partindo, então, do conjunto de trabalhos relacionados a esta pesquisa, o resultado desta revisão de literatura serão sintetizados no tópico a seguir.

2.2.4 Síntese e próximos passos

Essa revisão de literatura foi capaz de desvelar lacunas na literatura, mostrar ferramentas úteis para a realização da pesquisa e como elas são utilizadas no meio acadêmico. Além disso, foram encontrados trabalhos que unem as técnicas estudadas para potencializar os resultados e proporcionar uma análise de dados mais complexa.

No campo do AM, o ritmo de publicações revela, de maneira incontestável, que essa área está sendo cada vez mais explorada pela produção científica. Virtualmente em todas as áreas do saber é possível encontrar exemplos de aplicação desse tipo de método computacional e IA. Não é diferente no campo da educação. Contudo, as pesquisas inscritas a esse limite do conhecimento têm explorado o crescimento do ensino remoto – o que fornece mais dados que servem de insumo para investigação. Porém, dada a realidade de países subdesenvolvidos ou em desenvolvimento – como o Brasil –, que não permite o usufruto integral das tecnologias digitais, este trabalho escolhe não se limitar ao contexto de ambientes virtuais de aprendizagem.

Complementando-se a isso está o fato de que poucas pesquisas se dedicam a fornecer informações qualificadas para contribuir ao debate público de uma região administrativa como um município, estado ou país. Pelo contrário, estão comumente direcionadas a instituições e seus contextos peculiares. As escassas aplicações mais abrangentes se concentram em outros países, todavia trazem benefícios pelos seus objetos de pesquisa estarem em países com as mesmas características sociais e econômicas do Brasil, superficialmente falando.

A biblioteca PyCaret se mostrou uma boa alternativa, especialmente pela sua facilidade de uso e baixa exigência computacional, para a comparação e escolha do modelo de previsão para esta dissertação. Sua combinação à abordagem SHAP possibilita uma análise sofisticada acerca dos dados e do modelo por eles gerado. Além disso, os casos de aplicação dessas ferramentas indicam a este trabalho quais são os tipos de atributos mais comuns e que têm maior potencial de produzir informações qualificadas a serem extraídas neste estudo, de acordo com a literatura mais recente. Se destacaram fatores geográficos, demográficos, estruturais, assim como aqueles relacionados ao trabalho docente. No capítulo a seguir, será abordado o desenvolvimento detalhado do trabalho, utilizando as informações descritas até aqui.

3 APLICAÇÃO DA METODOLOGIA PROPOSTA

Neste capítulo, serão detalhadas as etapas cruciais para a realização deste trabalho, divididas em três principais grupos: a limpeza e o tratamento dos dados, as análises estatísticas e a construção do modelo preditivo. A primeira constituirá o conjunto de escolas que fará parte das análises subsequentes. A segunda será fundamental para determinar quais variáveis irão compor a base de dados que será usada para a construção do modelo. Por fim, serão descritas as tarefas realizadas para, usufruindo do auxílio da biblioteca PyCaret, construir o modelo de previsão mais eficiente possível.

3.1 LIMPEZA E TRATAMENTO DOS DADOS

Este trabalho se limitou ao nível de detalhamento "escola" devido à impossibilidade de acesso aos dados dos estudantes brasileiros, protegidos pela Lei Geral de Acesso à informação. Por conseguinte, alguns dados demográficos – como auto declaração racial e de gênero – foram adaptados para não serem desconsiderados na pesquisa, inspirado na literatura (VITORINO DE ARRUDA; RAMOS, 2023). Além dessa limitação, o trabalho considerou dados dos oito primeiros anos de vigência do PNE: 2015 a 2022, visto que não estão disponíveis os indicadores de taxas de rendimento – essenciais para a construção da variável resposta do trabalho – para o ano de 2023.

Os dados foram encontrados no site basedosdados.org, uma organização dedicada a tratar e limpar bases de dados com o intuito de democratizar e universalizar o acesso a dados de alta qualidade (DAHIS *et al.*, 2022). A partir de uma consulta feita em Linguagem de Consulta Estruturada (SQL, do inglês *Structured Query Language*), foram unidas três tabelas de bases diferentes: da base do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Foram extraídos os indicadores educacionais e o censo escolar, ambos no nível de escola; e, da base do IBGE, as informações acerca das Unidades Federativas (UF). O Apêndice A fornece informações sobre a base de dados resultante da extração. A quantidade de registro, de variáveis, suas descrições e as classes dos dados.

Foram extraídas, dessa consulta, 1.572.806 linhas, representando os registros de escolas em cada um dos oito anos componentes do período estudado, junto às 77 colunas que representam os fatores que descrevem as características das escolas. Começa, a partir de então, o processo de adequar a base às condições necessárias para aplicar a metodologia do estudo.

Primeiramente, foram modificados os tipos das variáveis, iniciando pelas que, apesar de serem representadas por números, são variáveis categóricas, sejam elas inteiras ou decimais. Em seguida, foram alterados os tipos de variáveis numéricas, pelo rigor conceitual, deixando de serem decimais para se tornarem inteiras.

Como a pesquisa foca seu interesse em escolas públicas, foram excluídos todos os registros cuja variável "rede" indica que a dependência administrativa da escola é privada, permanecendo 1.231.837 registros. Em seguida, iniciou o tratamento das variáveis que possuem um ou mais valores ausentes para que a construção do modelo não enfrente maiores problemas. As primeiras duas variáveis a serem preenchidas foram aquelas que versam sobre a quantidade de matrícula nas duas etapas de ensino consideradas: EF e EM. Os valores ausentes nessas duas colunas indicam que a escola não oferece essas etapas de ensino e, portanto, foram preenchidos com zero matrícula. Começar por essas duas foi estratégico, pois elas dão possibilidade de avaliar o significado dos valores ausentes nas outras variáveis.

Na maioria dos casos, estas foram preenchidas, também, com o valor "0", visto que são consequência da ausência de matrículas em uma etapa de ensino ou na outra. O único atributo que não seguiu essa regra foi o índice de regularidade docente, pois é um valor médio e a substituição dos respectivos valores ausentes pelo valor zero iria alterar a média de maneira mais ou menos intensa de acordo com o grupo analisado. Portanto, tais ausentes foram alterados segundo a sua dependência administrativa: se a escola faz parte da rede municipal, pela média daquela rede municipal no ano correspondente; se estadual, pela média da rede estadual naquele ano; e, se federal, pela média das escolas federais no respectivo ano. Essa abordagem foi feita em um esforço de representar uma realidade mais próxima à conjuntura local.

Como último trabalho de limpeza dos dados, a variável correspondente à complexidade da gestão, recebendo valores de texto, precisou de uma atenção mais detalhada. Seus dados vão do Nível 1 ao Nível 5, porém não estavam padronizados, seja pelo uso de letras minúsculas e maiúsculas, acentuação ou pelo espaçamento. Assim, todas foram adequadas ao modelo "Nível_#", ou seja, com a primeira letra maiúscula, acento agudo na letra "i", seguido de um espaço sublinhado e a numeração correspondente ao nível.

Por fim, com o intuito de direcionar o trabalho para a o seu objetivo, foi calculada a taxa de aprovação geral de cada escola através da quantidade de matrículas em cada etapa de ensino e das suas respectivas taxas de aprovação. Como a variável resposta do modelo construído será esta taxa, os registros nos quais não foi possível calcular esse número foram excluídos. Essa situação representa escolas que não possuem, em um ano, informações sobre as matrículas ou taxas de aprovação em nenhuma das etapas de ensino (EF e EM) avaliadas. Assim, foi possível

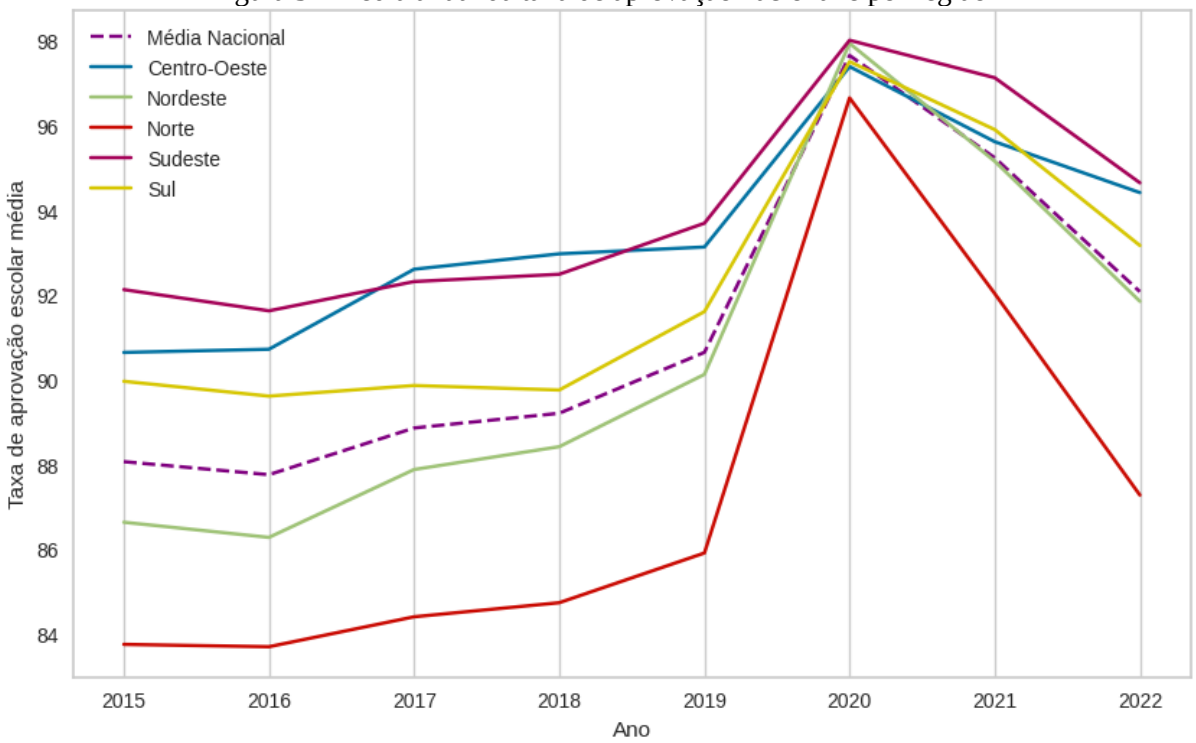
enxugar ainda mais o conjunto de dados, permanecendo apenas aqueles que são interessantes ao estudo pretendido.

Por conseguinte, a quantidade de linhas foi reduzida a 882.065, anterior à última redução, realizada ao perceber que apenas uma escola considerava estudantes de Educação Profissional (EP) como pertencentes ao EM. Ao final, restaram 882.064 registros que realmente estavam alinhados aos objetivos do trabalho. A eliminação dos registros, apesar de ter representado uma importante etapa para o alcance do objetivo do trabalho, pode representar algumas problemáticas que não devem ser desconsideradas. Algumas assimetrias causadas, como um déficit na coleta de dados em escolas de uma determinada localidade, podem representar um viés no estudo. Não obstante, a pesquisa assume esse risco em nome de uma maior fidelidade aos dados oficiais. No próximo tópico serão discutidas as modificações nas variáveis explicativas do modelo, quais foram incluídas e excluídas.

3.2 ANÁLISE ESTATÍSTICA

A análise descritiva dos dados possibilitou verificar a evolução dos indicadores ao longo dos anos que compreendem o estudo. A taxa de aprovação média das escolas que fazem parte do escopo do trabalho apresenta, visualmente, uma tendência de crescimento. Porém, quando essa avaliação é decomposta por regiões, é possível notar que há uma grande disparidade entre elas, como mostra a Figura 3. Regiões como Norte e Nordeste estiveram, praticamente em todos os anos, abaixo da média nacional – esta, particularmente, se aproximando da média ano após ano. Os anos de 2020 e 2021 apresentam níveis altíssimos de aprovação, o que é explicado pela política de avaliação em ciclos adotada durante esse período, em razão da pandemia (INEP, 2023a). Essas diferenças instigam esta pesquisa à indagação acerca da relevância do fator regional para as taxas de aprovação escolar.

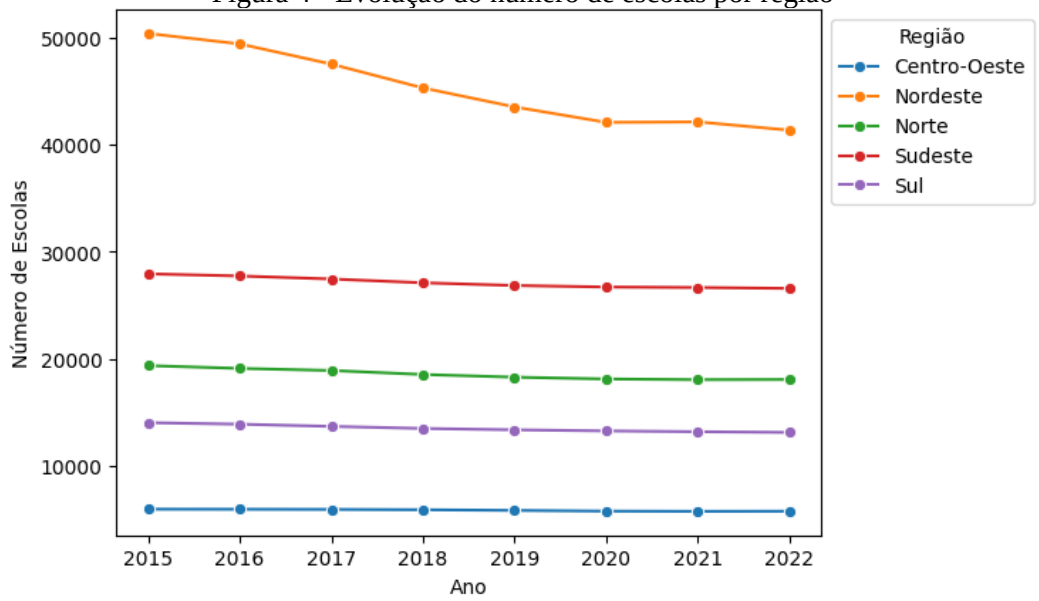
Figura 3 - Média anual da taxa de aprovação nacional e por região



Fonte: O Autor (2024)

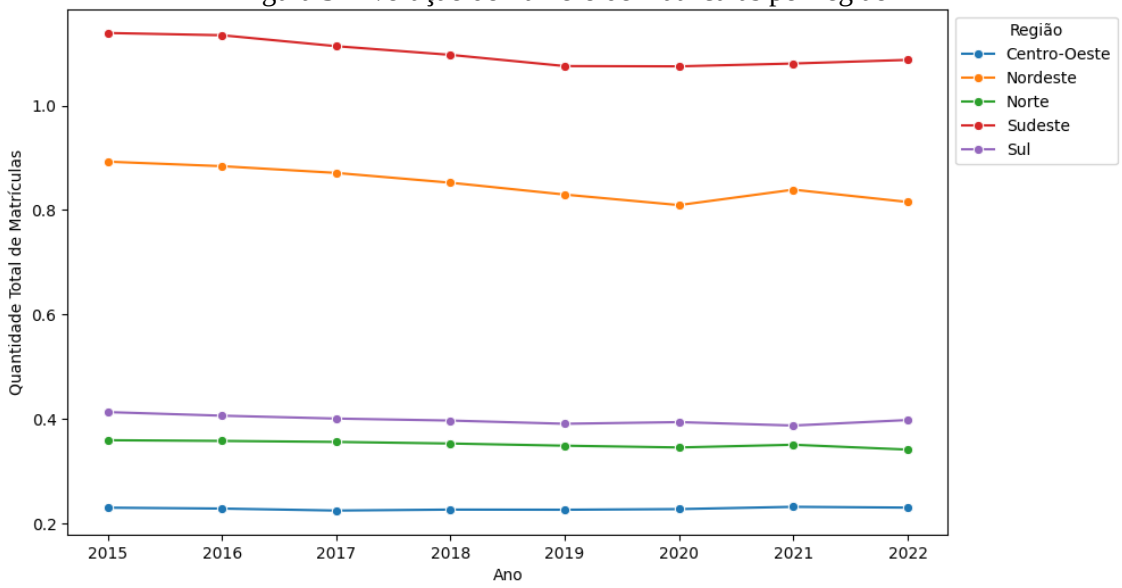
A quantidade de escolas também foi uma informação que trouxe curiosidade. Como mostra a Figura 4, a quantidade de escolas por ano, dentro do escopo do trabalho, diminuiu em quase 13 mil unidades no período avaliado. Vale observar que, apesar de ter havido uma queda no número de escolas públicas ao longo dos anos (INEP, 2023b), esse valor pode ter sido causado pela limpeza dos dados realizada na fase anterior da metodologia. Destaca-se a região Nordeste, que concentra 9 mil escolas a menos de 2015 a 2022. Houve, paralelamente, queda na quantidade de matrículas, exposta na Figura 5. Apesar disso, essa diminuição não ocorreu de maneira proporcional. O número de escolas em 2022 é 89,15% da quantidade observada em 2015, uma queda de 10,85%, enquanto a redução das matrículas foi de apenas 8,65%. Isso leva a inferir que o número de matrículas por escola tenha sido dilatado.

Figura 4 - Evolução do número de escolas por região



Fonte: O Autor (2024)

Figura 5 -Evolução do número de matrículas por região

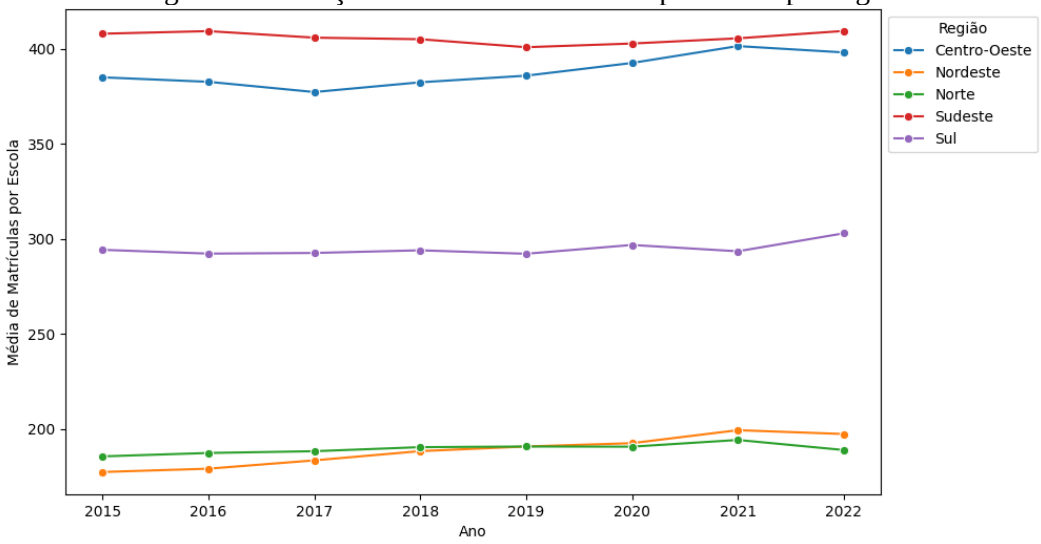


Fonte: O Autor (2024)

A Figura 6 exibe esse aumento na média de matrículas por escola, que foi mais acentuada na região Nordeste – que teve um aumento médio de 20 estudantes – e centro-oeste – com 13 estudantes a mais. Essa realidade indica a possível influência do número de estudantes dentro de uma mesma escola para a sua aprovação. Em especial, o número de estudantes dentro de cada sala-de-aula pode ter contribuição importante. O indicador que reflete a média de alunos por turma nas escolas surge, então, como um candidato a ser avaliado com atenção. A Figura 7 mostra a evolução deste índice. Apesar de não ter havido uma mudança muito significativa na

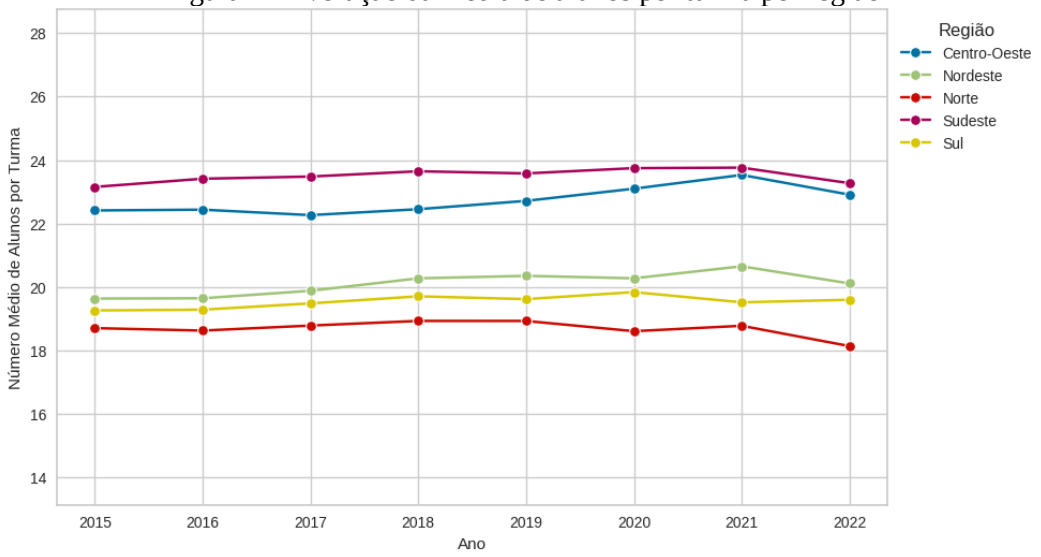
média, é possível que haja uma relação quando a contribuição individual do índice for verificada.

Figura 6 - Evolução da média de matrículas por escola por região



Fonte: O Autor (2024)

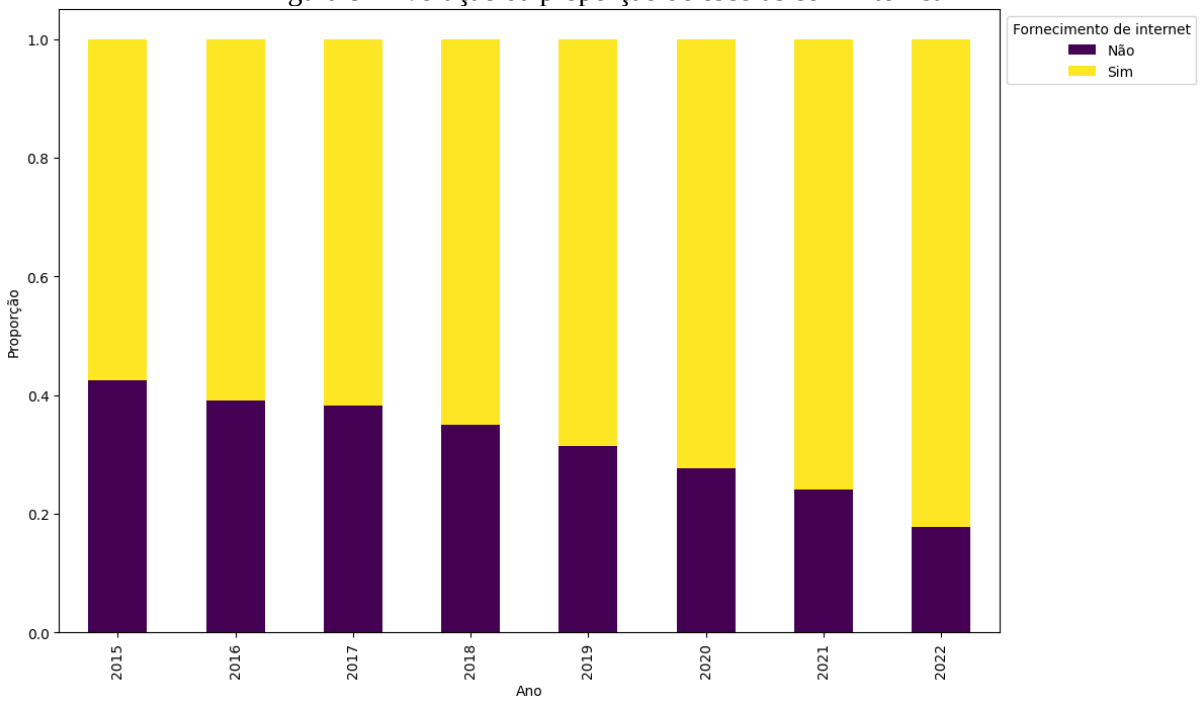
Figura 7 - Evolução da média de alunos por turma por região



Fonte: O Autor (2024)

Um aspecto excepcionalmente importante no debate sobre educação pública diz respeito ao fornecimento de internet nas escolas. A crescente participação do mundo digital na vida dos brasileiros é uma realidade também para estudantes da educação básica (AGÊNCIA BRASIL, 2021). Na base de dados utilizada, a proporção de escolas com fornecimento de internet tem crescido ano após ano, alcançando mais de 80% das escolas em 2022. A Figura 8 apresenta essa dinâmica.

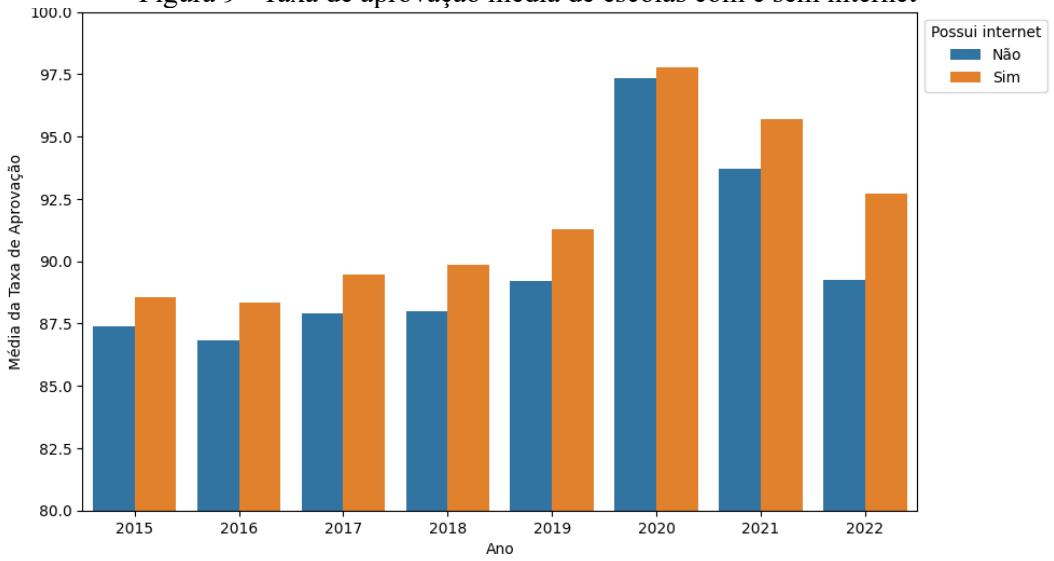
Figura 8 - Evolução da proporção de escolas com internet



Fonte: O Autor (2024)

Ao compararmos a taxa de aprovação média das escolas com e sem acesso à internet, verifica-se uma leve vantagem daquelas sobre estas, ano após ano, como mostra a Figura 9. No entanto, o prejuízo resultante da falta de conexão de banda larga se acentua no último ano de análise. Curiosamente, o primeiro após a política de avaliação em ciclos. Levanta-se, então, o questionamento sobre o tamanho da contribuição para as taxas aprovação escolar, a presença ou ausência de acesso à internet numa escola.

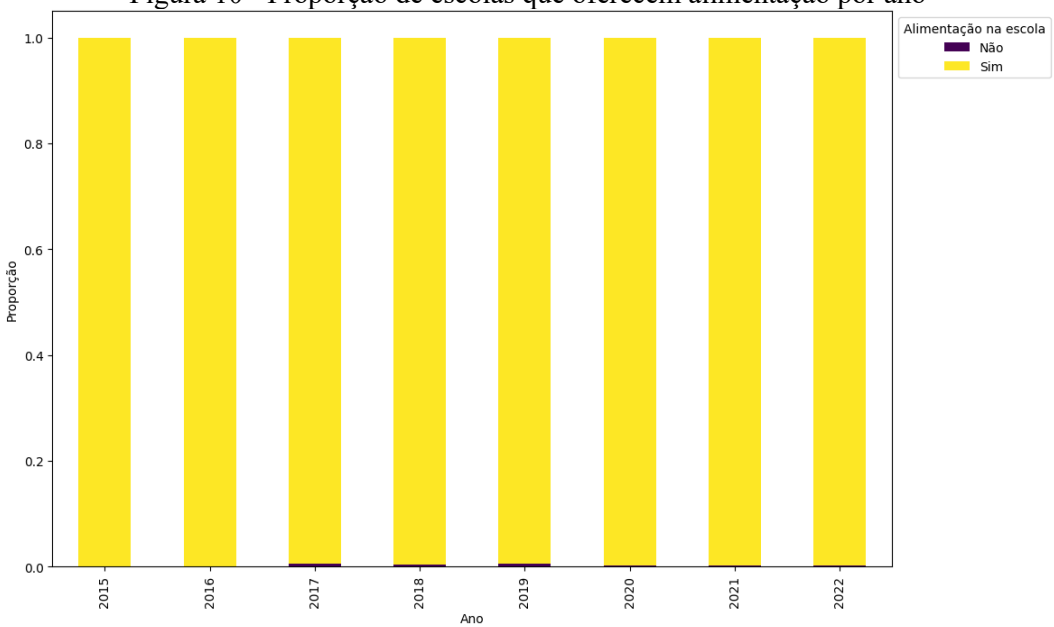
Figura 9 - Taxa de aprovação média de escolas com e sem internet



Fonte: O Autor (2024)

As Figuras 10, 11, 12 e 13 tratam, respectivamente, da proporção das escolas que possuem abastecimento de água, alimentação no prédio escolar, banheiro e sala multiuso. No primeiro caso, percebe-se que a rede de distribuição de água é uma política virtualmente universal e já está consolidada, portanto, esse fator não será considerado para o modelo que será construído. O mesmo acontece na segunda informação. Já o terceiro caso denota uma mudança brusca de 2015 a 2018, partindo de zero escolas com banheiro e passando a patamares acima de 95%. Por fim, esse cenário se repete quando avaliamos as salas multiuso. Assim sendo, o trabalho considerou que, até 2018, esse dado não era coletado e também não estará presente no modelo.

Figura 10 - Proporção de escolas que oferecem alimentação por ano

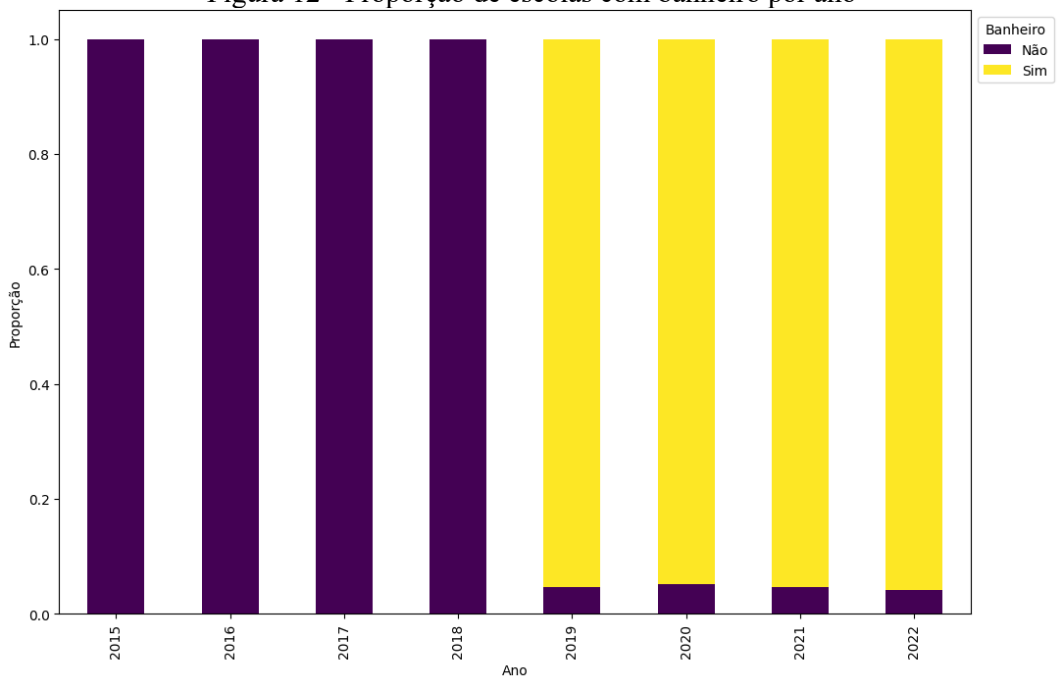


Fonte: O Autor (2024)

Figura 11 - Proporção de escolas com abastecimento de água

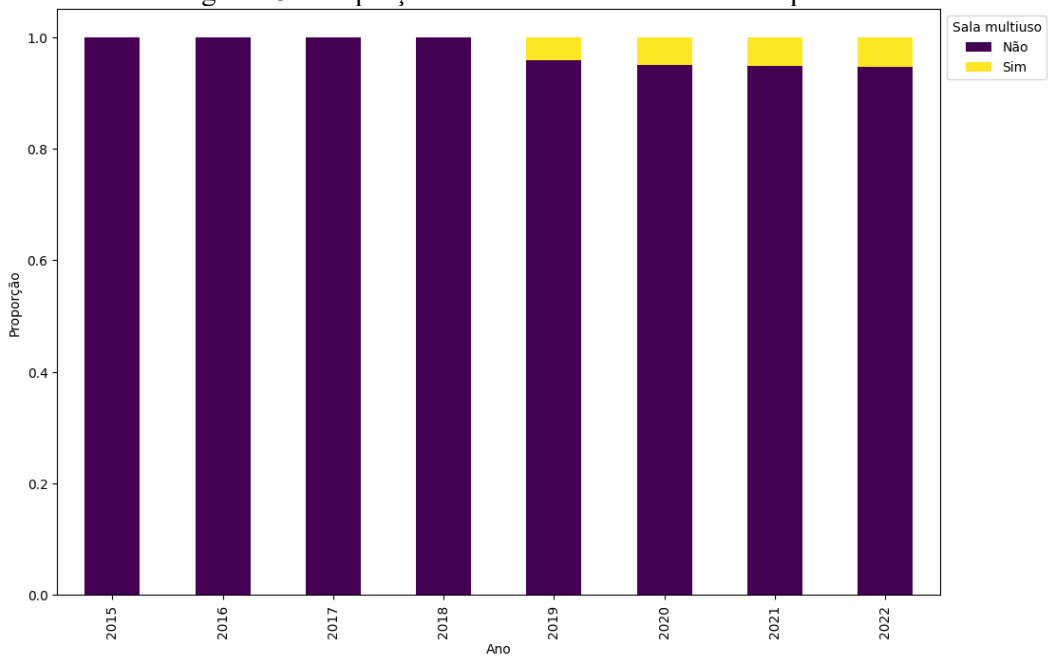
Fonte: O Autor (2024)

Figura 12 - Proporção de escolas com banheiro por ano



Fonte: O Autor (2024)

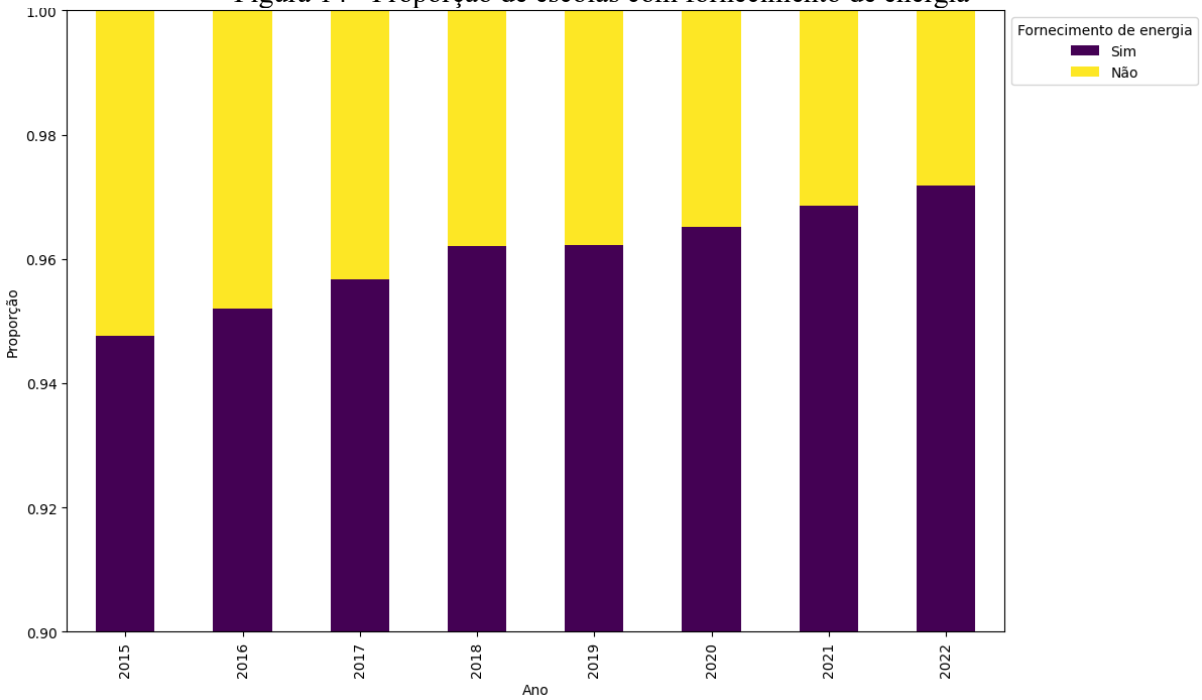
Figura 13 - Proporção de escolas com sala multiuso por ano



Fonte: O Autor (2024)

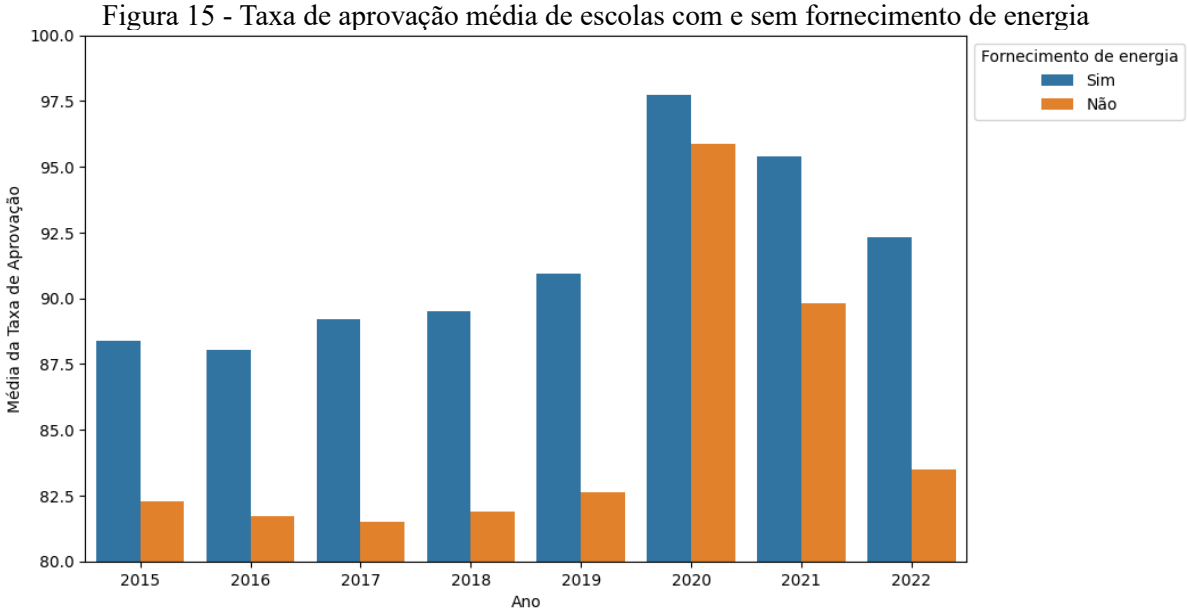
O fornecimento de energia também pode ser considerado como uma política praticamente universal. As escolas que não desfrutam desse fator representam menos de 5% do total considerado por esta dissertação e esse número tem caído ao longo dos anos, alcançando mais de 98% de escolas contempladas com energia (Figura 14). Entretanto, as diferenças nas taxas de aprovação entre esses dois grupos são as maiores apresentadas até aqui (Figura 15), o que leva a pesquisa a considerar esse como um fator possivelmente relevante na análise da sua contribuição.

Figura 14 - Proporção de escolas com fornecimento de energia



Fonte: O Autor (2024)

Figura 15 - Taxa de aprovação média de escolas com e sem fornecimento de energia

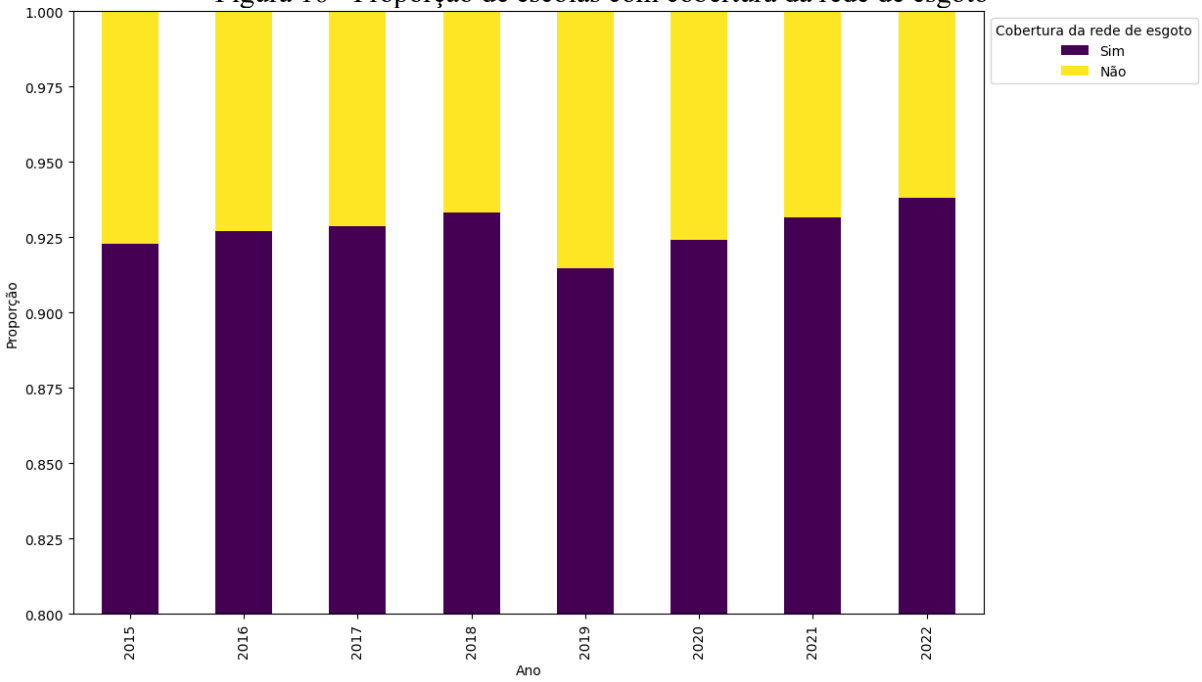


Fonte: O Autor (2024)

O caso do atendimento pela rede de esgoto também merece uma atenção especial. Apesar de haver imensa maioria de escolas atendidas, houve um aumento considerável do número de escolas fora da abrangência da rede de esgoto de 2017 para 2018 (Figura 16). Por causa disso, este trabalho não considerarará uma política universalizada ou em processo de universalização. Ademais, as taxas de aprovação de escolas que são contempladas pela rede de esgoto são maiores em todos os anos, aumentando essa diferença no período em que houve uma

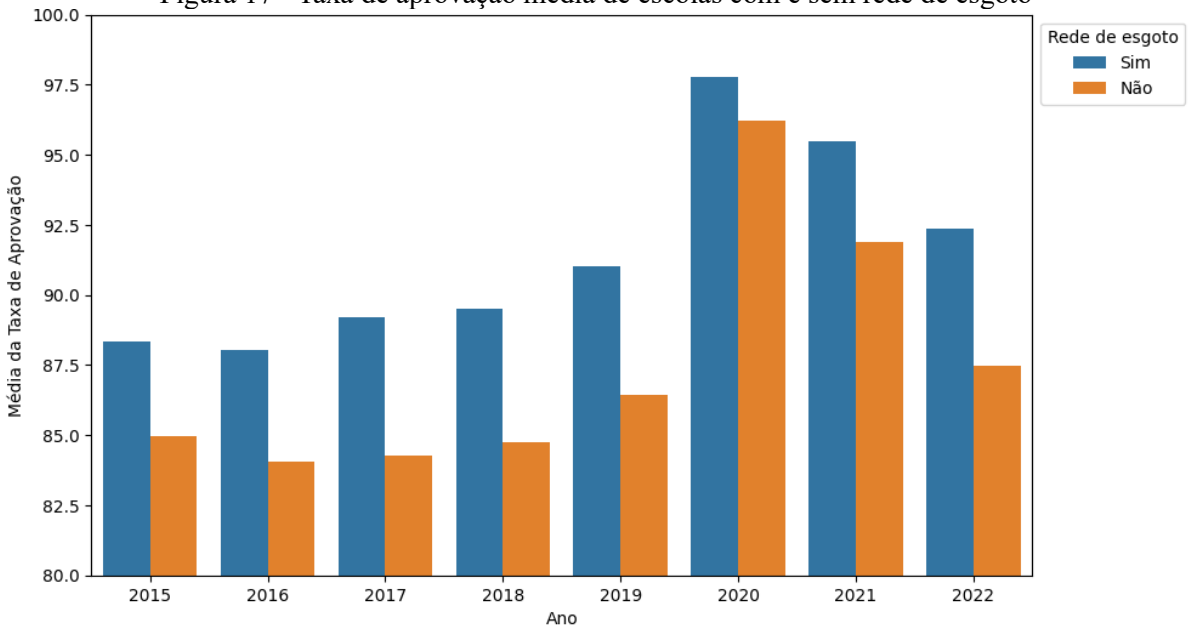
diminuição da proporção dessas escolas abrangidas pela rede de saneamento básico (Figura 17). Isso indica que esse fator pode ter alguma relevância.

Figura 16 - Proporção de escolas com cobertura da rede de esgoto



Fonte: O Autor (2024)

Figura 17 - Taxa de aprovação média de escolas com e sem rede de esgoto

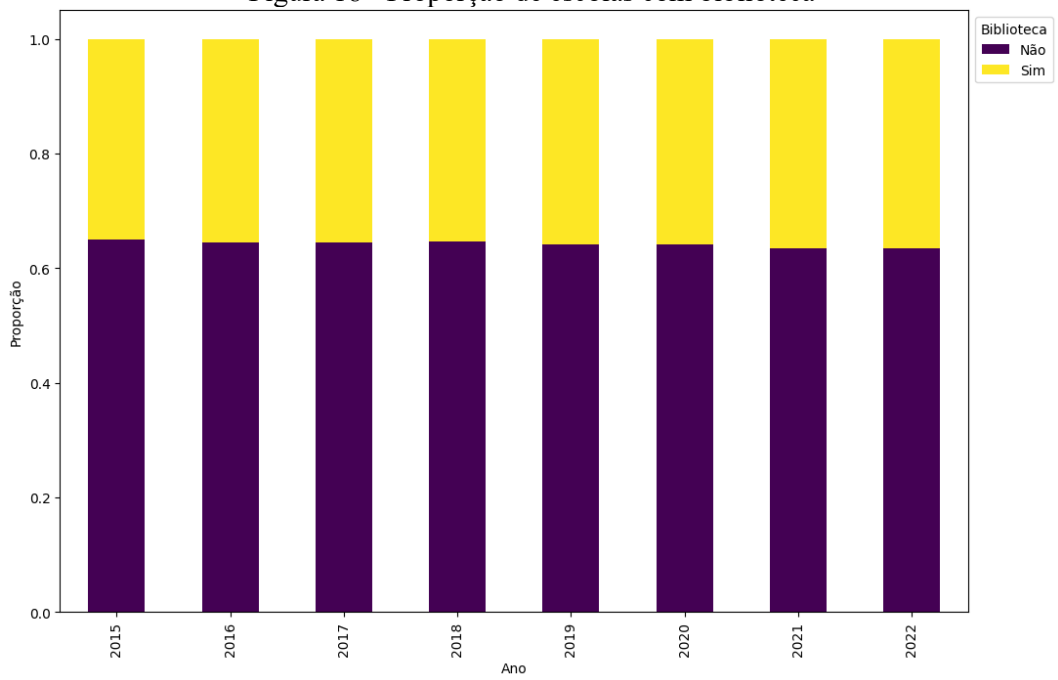


Fonte: O Autor (2024)

Uma análise que resultou em números contraintuitivos foi a presença de uma biblioteca na escola e a média da sua taxa de aprovação. A Figura 18 exibe como a proporção de escolas com bibliotecas é relativamente constante, a uma taxa pouco acima de 1/3 das escolas. Além de não parecer haver uma política de investimento em bibliotecas nas escolas públicas brasileiras,

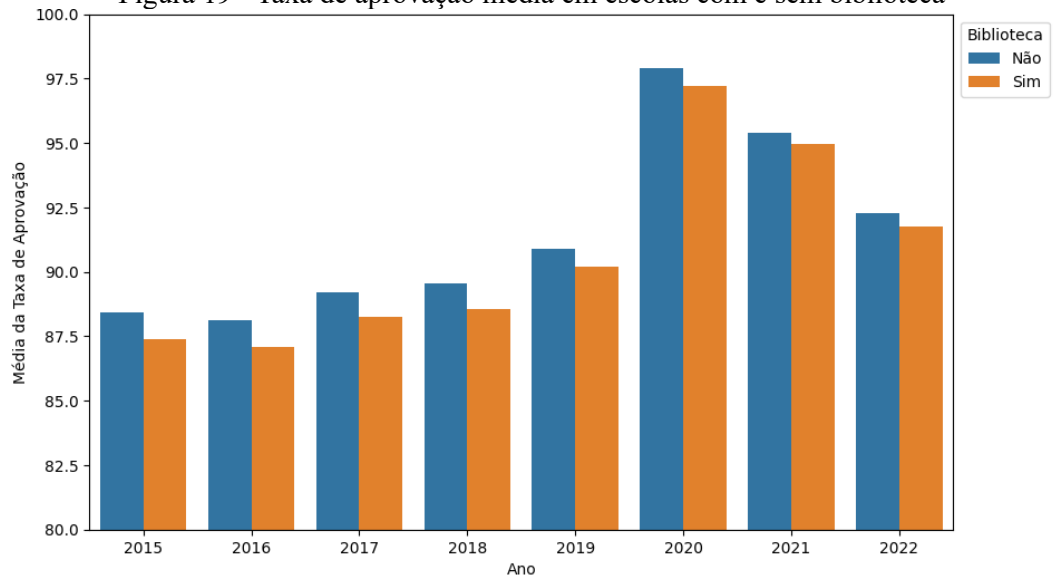
a média da taxa de aprovação nas escolas que não possuem esse espaço é ligeiramente maior em todos os anos considerados, como prova a Figura 19. Isso faz este trabalho indagar se a sua existência é realmente relevante e positiva.

Figura 18 - Proporção de escolas com biblioteca



Fonte: O Autor (2024)

Figura 19 - Taxa de aprovação média em escolas com e sem biblioteca

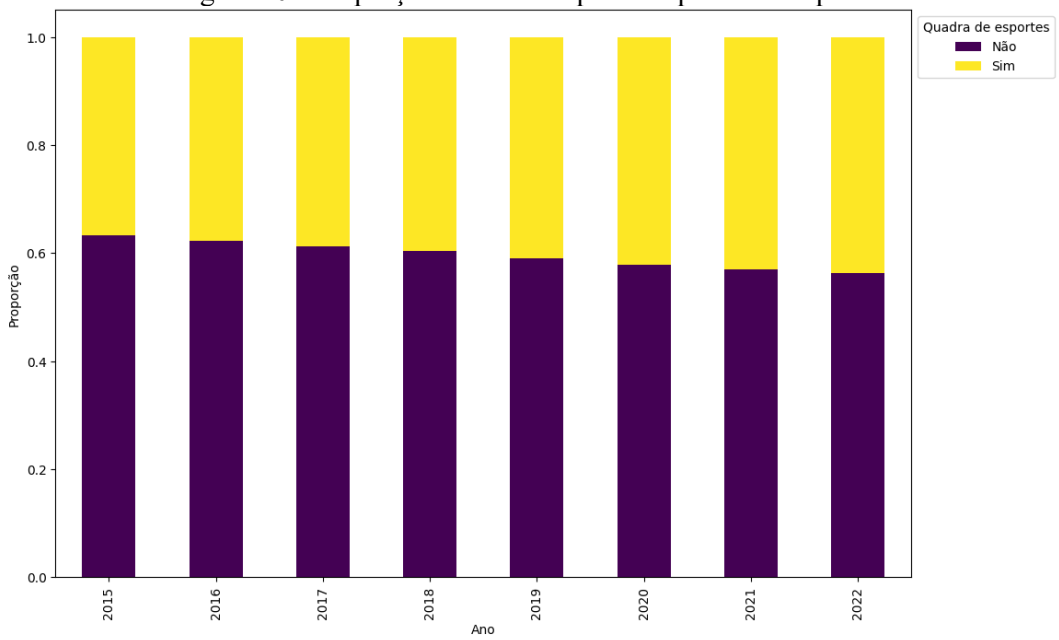


Fonte: O Autor (2024)

Quadras de esportes representam outro fator que despertou curiosidade neste trabalho. Enquanto os dados indicam que há uma política de expansão das escolas que desfrutam, na sua estrutura, de um espaço dedicado à prática esportiva de diversas modalidades, seus efeitos na taxa de aprovação escolar – se é que existem – parecem ser modestos. A Figura 20 mostra o

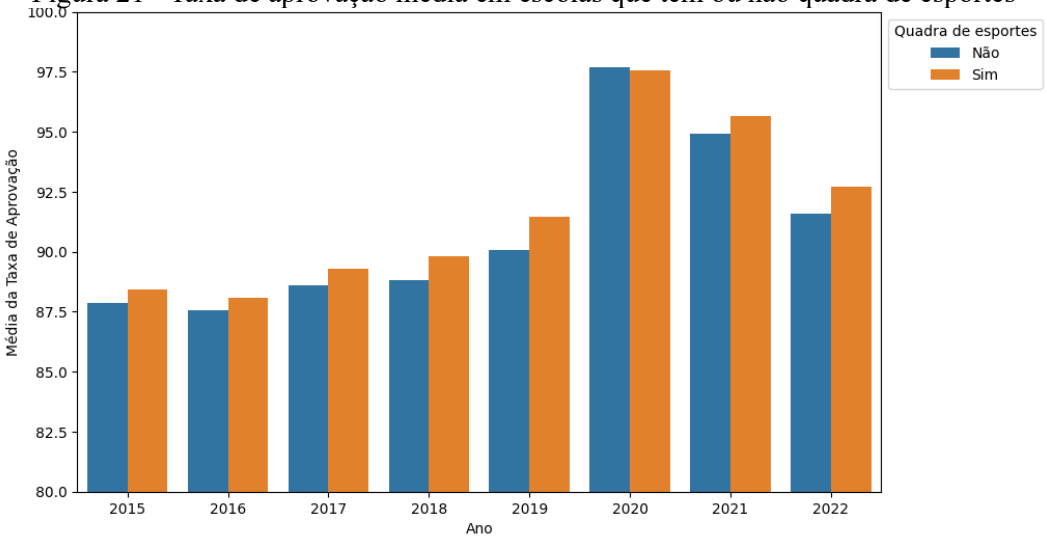
constante crescimento da proporção de escolas com quadra esportiva, passando de 36,66% no ano de 2015 para 43,59% em 2022. Em paralelo, as taxas de aprovação média desses dois grupos não apresentam grandes diferenças (Figura 21). Este trabalho, então, verificará se essa característica tem uma contribuição relevante.

Figura 20 - Proporção de escolas que têm quadra de esportes



Fonte: O Autor (2024)

Figura 21 - Taxa de aprovação média em escolas que têm ou não quadra de esportes

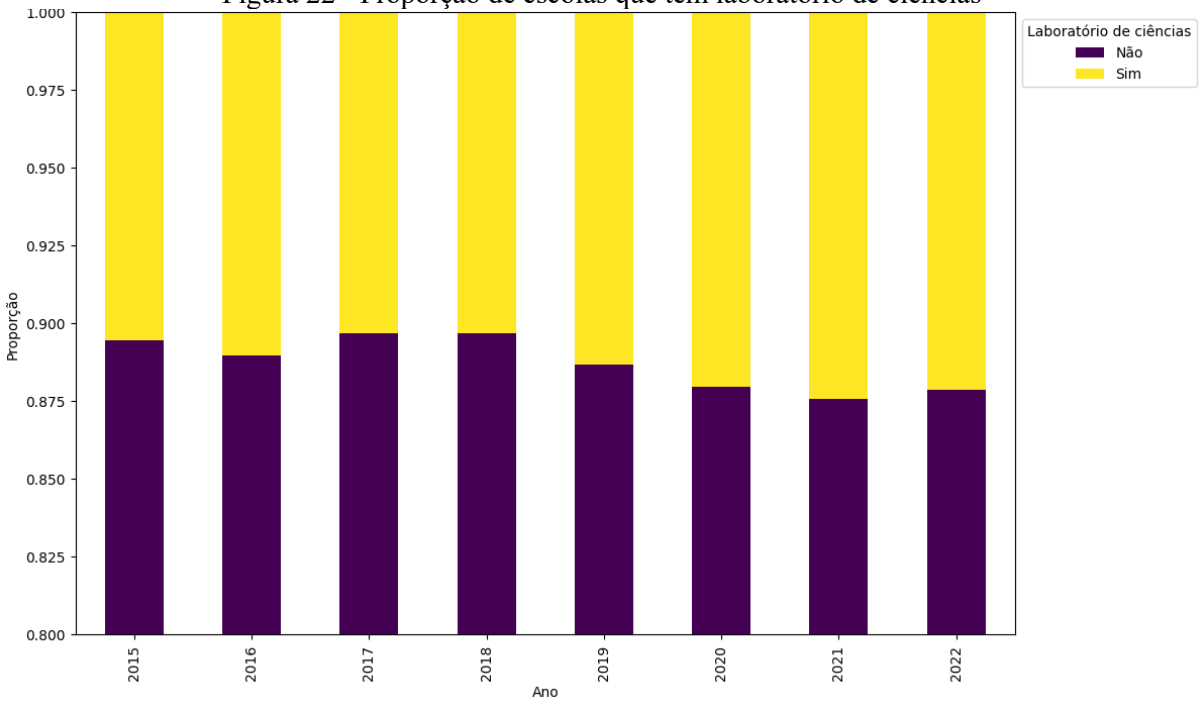


Fonte: O Autor (2024)

A proporção de escolas que possuem, na sua estrutura, disponível aos estudantes, um laboratório de ciências, tem aumentado, mesmo que de maneira discreta (Figura 22). Tal qual o caso das bibliotecas, as médias das taxas de aprovação anuais demonstram algo que não era o esperado: as escolas que possuem laboratório de ciências performam abaixo nas suas taxas de aprovação – ainda que a diferença venha diminuindo – daquelas que não têm esse equipamento

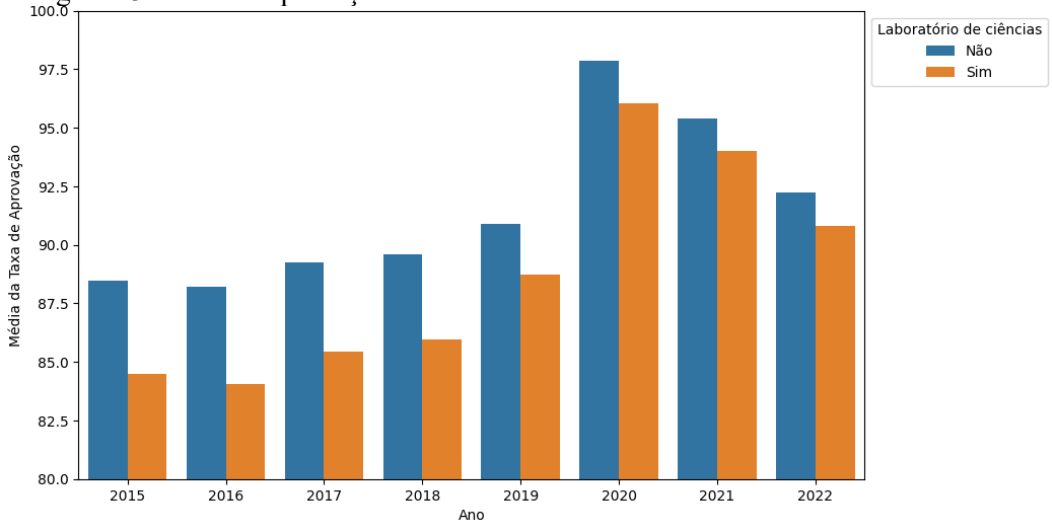
ao seu dispor (Figura 23). Esse atributo, portanto, também será utilizado no modelo para que seja verificada a sua contribuição para as taxas de aprovação das escolas estudadas. Já o caso dos laboratórios de EP não justifica sua inclusão no modelo, visto que, além de serem ínfimas as escolas que possuem o equipamento, elas foram registradas apenas no ano de 2022 (Figura 24).

Figura 22 - Proporção de escolas que têm laboratório de ciências



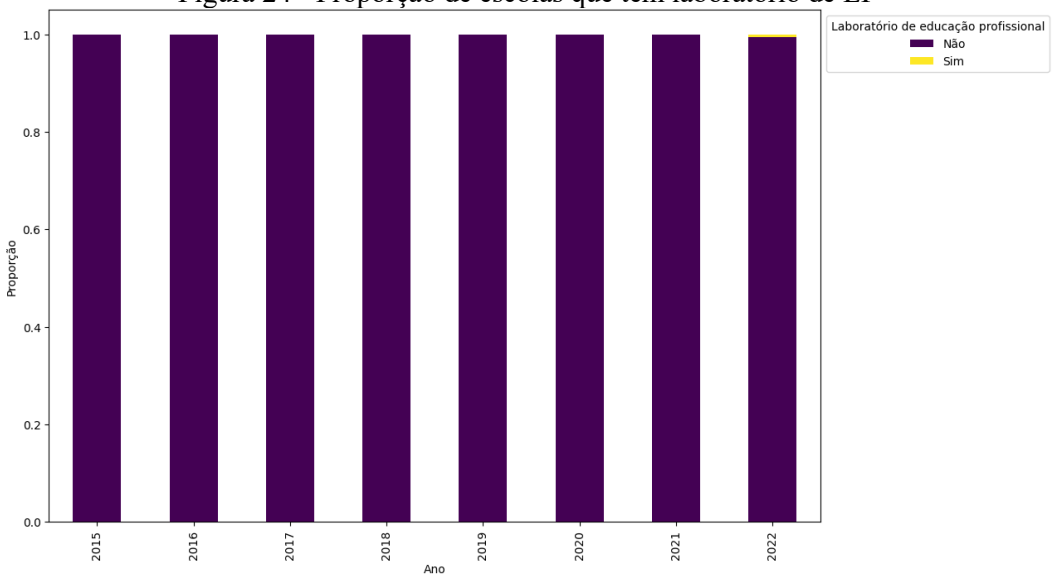
Fonte: O Autor (2024)

Figura 23 - Taxa de aprovação média em escolas com ou sem laboratório de ciências



Fonte: O Autor (2024)

Figura 24 - Proporção de escolas que têm laboratório de EP



Fonte: O Autor (2024)

Um problema a ser enfrentado por este trabalho foi a decomposição de alguns dos indicadores. É o caso do índice que trata da adequação da formação docente, que foi repartido em cinco grupos, tais quais cada um representa a parcela de docentes que se enquadram naquele grupo previamente definido. O mesmo ocorre com o índice de esforço docente, dividido em seis níveis distintos. Portanto, dado que esses dados são referentes ao EF e EM, são 22 variáveis que, se fizerem parte do modelo, exigirão um tempo e esforço computacional bastante maior. E não apenas esse problema seria englobado, mas também o fato de que, por serem, para o modelo, variáveis distintas, seria necessário realizar uma análise do que significa cada grupo ou nível e quais seriam faixas de percentuais aceitáveis para uma escola. Isso aumentaria, além do que foco deste trabalho, a sua complexidade.

Ainda, foi feita uma avaliação da correlação entre essas variáveis e a variável resposta do modelo proposto. Foi possível notar – como explicita o Apêndice B – que nenhuma delas tem uma correlação forte o suficiente para justificar entrarem no modelo, especialmente ponderando a complexidade e o esforço computacional citados anteriormente. Por conseguinte, foram excluídas do modelo, enxugando-o consideravelmente. No próximo tópico, serão apresentados atributos que foram analisados com o auxílio de testes estatísticos.

3.3 TESTES ESTATÍSTICOS

Este trabalho enfrentou a limitação de não ter acesso aos dados de aprovação individual de cada estudante, dificultando a avaliação de fatores como gênero, cor, nível socioeconômico etc. Porém, para que alguns desses fatores demográficos não fossem deixados de fora, foi feita

uma avaliação da proporção de matrículas do sexo feminino e masculino de cada escola, em cada ano, bem como de matrículas de pessoas brancas e não brancas, comparando-as com as respectivas proporções para todo o país. No cenário nacional, as matrículas variaram, ano a ano, em torno de 48% do sexo feminino e 52% do sexo masculino. Já as matrículas de pessoas não brancas representaram mais do que o triplo daquelas autodeclaradas brancas, como aponta a Tabela 1.

Tabela 1 - Proporção de matrículas por gênero e cor

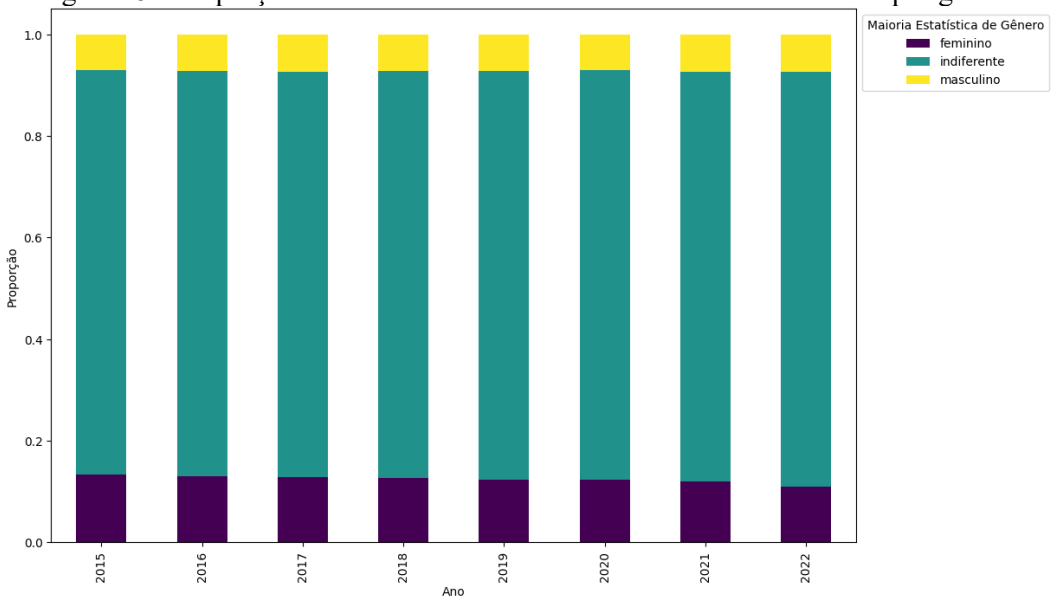
Ano	Feminino	Masculino	Branca	Não branca
2015	47,87%	52,13%	22,94	77,06
2016	47,83%	52,17%	23,37	76,63
2017	47,81%	52,19%	23,52	76,48
2018	47,93%	52,07%	23,73	76,27
2019	48%	52%	23,83	76,17
2020	48%	52%	23,89	76,11
2021	48,38%	51,62%	23,82	76,18
2022	48,38%	51,62%	23,64	76,36

Fonte: O Autor (2024)

Aquelas escolas cuja proporção de matrículas foi considerada diferente do parâmetro de maneira estatisticamente significante, considerando um índice de significância de 5%, foram classificadas de acordo. Nos casos em que não foi possível classificar a maioria significante, o atributo indicou que a proporção é indiferente. Para isso, foi aplicado um teste z para proporções (DOANE; SEWARD, 2016), comparando a parcela de matrículas por gênero – ou cor – de cada escola em cada ano com a correspondente nacional.

Primeiramente, foram realizados testes para verificar a maioria estatística de acordo com o gênero e seus resultados foram compilados em um gráfico que mostra as parcelas das escolas que possuem maioria feminina, masculina e indiferente (Figura 25). A grande maioria das escolas possui diversidade de gênero condizente com a distribuição nacional daquele ano, sendo em torno de 80% das escolas em cada ano compostas por distribuições indiferentes de matrículas por gênero. Em seguida, as escolas com maioria de matrículas femininas foram de 13% em 2015 a 10% em 2022 e aquelas com maioria de matrículas masculinas se mantiveram na casa dos 7% (Tabela 2).

Figura 25 - Proporção de escolas de acordo com a maioria de matrículas por gênero



Fonte: O Autor (2024)

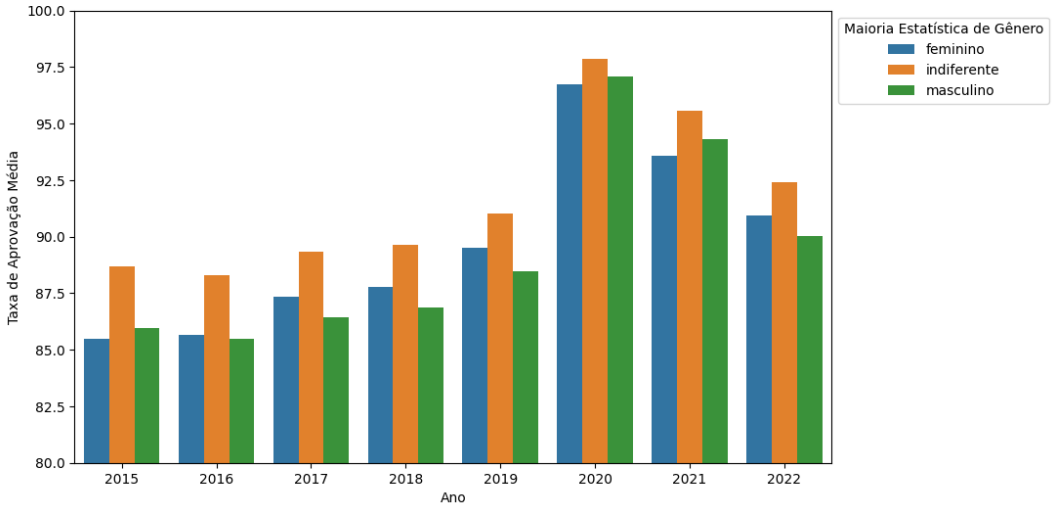
Tabela 2 - Proporção de escolas com maioria de matrículas por gênero

Ano	Feminino	Indiferente	Masculino
2015	13,37%	79,54%	07,08%
2016	13,05%	79,81%	07,14%
2017	12,94%	79,69%	07,37%
2018	12,62%	80,17%	07,21%
2019	12,39%	80,39%	07,22%
2020	12,3%	80,6%	07,1%
2021	12,02%	80,56%	07,42%
2022	10,95%	81,77%	07,28%

Fonte: O Autor (2024)

Quando se trata de avaliar as suas taxas de aprovação, é perceptível como as escolas cuja maioria de matrículas por gênero é estatisticamente indiferente figuram sempre acima daquelas cujas maiorias das matrículas são femininas ou masculinas (Figura 26). Quando comparadas escolas com maioria feminina ou masculina, há uma proximidade e alguma variação, apesar das primeiras estarem ligeiramente acima na maioria dos anos considerados. Portanto, este trabalho questiona se a diversidade de gênero condizente com a diversidade das matrículas nacionalmente – leia-se com maioria estatisticamente indiferente – é relevante para uma maior taxa de aprovação escolar.

Figura 26 - Taxa de aprovação média das escolas por maioria de gênero



Fonte: O Autor (2024)

O mesmo procedimento foi feito para avaliar as escolas baseado no critério de cor. Como pode ser verificado na Figura 27, mais da metade – de 52% a 55% – das escolas possui maioria estatisticamente significante de matrículas de pessoas não brancas. As escolas cuja maioria de matrículas é de pessoas brancas se manteve pouco acima de 30%, enquanto as demais – entre 17% e 13% – não possuem maioria estatisticamente significante, sendo classificadas como indiferente. Os resultados detalhados numericamente estão aglutinados na Tabela 3.

Tabela 3 - Proporção de escolas com maioria de matrículas por cor

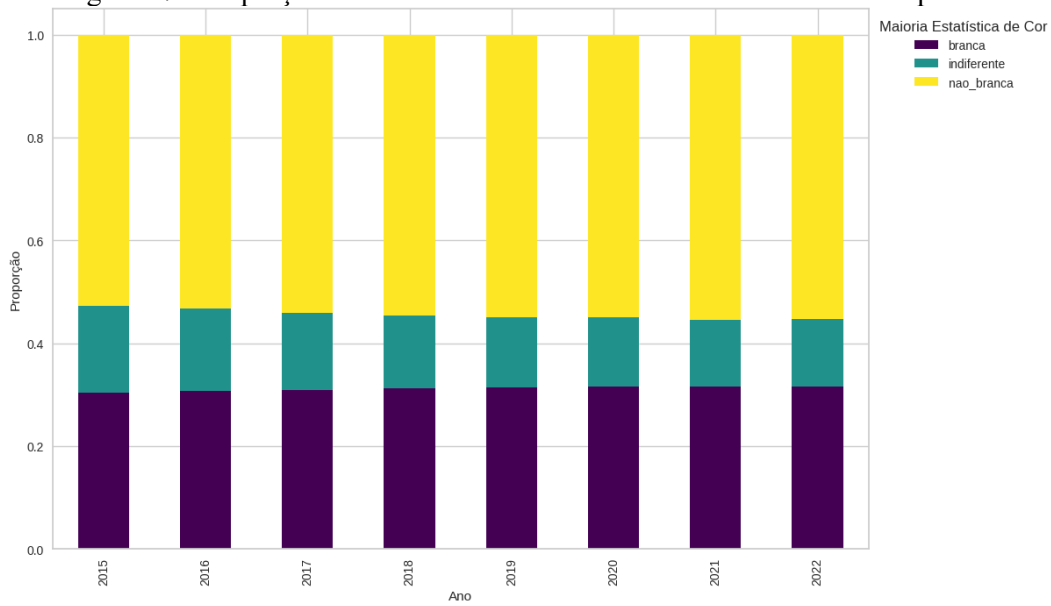
Ano	Branca	Indiferente	Não branca
2015	30,48%	16,85%	52,67%
2016	30,76%	16%	53,24%
2017	30,98%	14,96%	54,06%
2018	31,17%	14,22%	54,61%
2019	31,41%	13,68%	54,91%
2020	31,52%	13,46%	55,02%
2021	31,57%	12,97%	55,46%
2022	31,62%	13,05%	55,33%

Fonte: O Autor (2024)

Ao contrário da análise feita sobre as maiorias por gênero, na qual as escolas que não possuem maioria estatisticamente significante apresentam rendimento melhor às outras, no caso da avaliação por cor, há um padrão diferente. As escolas com maioria de matrícula de pessoas brancas rendem, constantemente, acima daquelas que não possuem maioria relevante e, com maior distanciamento, das escolas cuja maioria de matrículas é de pessoas não brancas (Figura 28). Dessa forma, será investigada a relevância desse fator para as taxas de aprovação escolar,

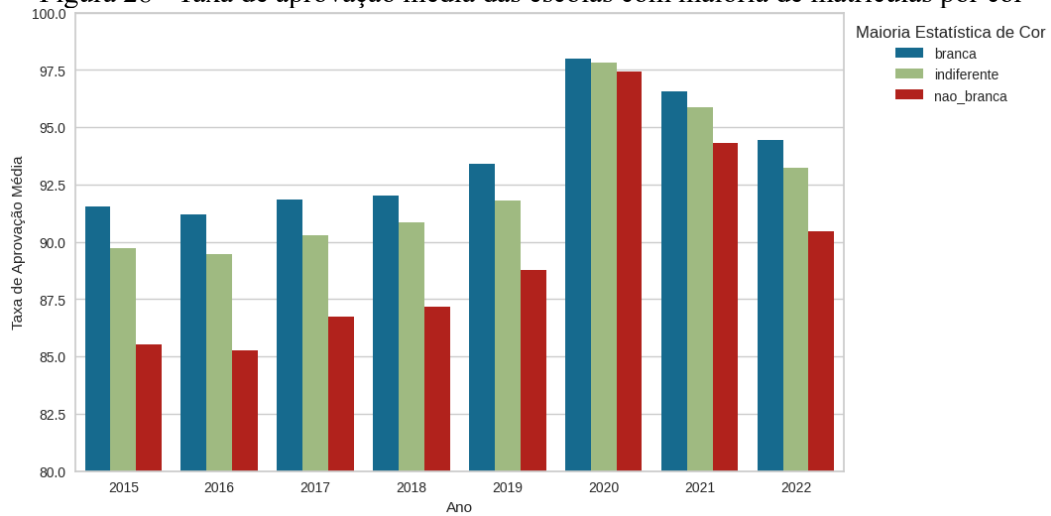
buscando compreender se escolas com maioria estatisticamente significante branca possuem os melhores resultados e quão relevante é essa tendência.

Figura 27 - Proporção de escolas de acordo com a maioria de matrículas por cor



Fonte: O Autor (2024)

Figura 28 - Taxa de aprovação média das escolas com maioria de matrículas por cor



Fonte: O Autor (2024)

Ainda, aglutinando a este tópico as análises estatísticas feitas no anterior, foram realizados testes com o intuito de inferir sobre a significância relativa às diferenças das taxas de aprovação escolar das estratificações apresentadas. Estas são a região geográfica, maioria estatística de cor e gênero, fornecimento de energia, abastecimento de esgoto, existência de biblioteca, quadra de esportes, laboratório de ciências e laboratório de informática.

Em todos os casos, este trabalho buscou avaliar se as diferenças aparentes nas taxas de aprovação escolar são estatisticamente significantes. Para tanto, estabeleceu em 5% o índice de significância para todos os testes realizados. como todas as amostras são suficientemente

grandes – na casa de milhares de elementos – a necessidade de um teste de normalidade foi dispensada, de acordo com o Teorema Limite Central (DOANE; SEWARD, 2016).

Assim sendo, foram utilizados dois testes distintos para verificar a significância estatística das diferenças entre as médias de aprovação amostrais. Nos casos em que as amostras são as cinco regiões geográficas do Brasil e as três classificações para maioria estatística das escolas, foi explorada a Análise de Variância (ANOVA) de um fator, por serem mais de duas amostras. Nos demais casos, os quais demandam a testagem de apenas duas amostras – representando a ausência ou presença do atributo –, o teste aplicado foi o teste t para amostras independentes (DOANE; SEWARD, 2016).

Em suma, as hipóteses formuladas seguiam a seguinte estrutura:

H_0: As médias das amostras são iguais;

H_1: Pelo menos uma das médias das amostras é diferente.

A Tabela 4 sintetiza os resultados dos testes ANOVA, enquanto a Tabela 5 reúne os resultados dos testes t para cada ano e reforça que as diferenças das taxas de aprovação escolar em todas as amostras e todos os anos são estatisticamente significantes para um índice de significância de 5%. Portanto, infere-se que estes fatores devem estar presentes no modelo a ser construído para prever as taxas de aprovação escolar.

Tabela 4 - Resultados de ANOVA para regiões, gênero e cor

Ano	Regiões		Gênero		Cor	
	Estatística F	p-valor	Estatística F	p-valor	Estatística F	p-valor
2015	2022,55	0,0	662,50	7,71e-281	3488,64	0,0
2016	1821,14	0,0	499,81	7,30e-211	3305,91	0,0
2017	1797,23	0,0	412,17	4,39e-173	2608,06	0,0
2018	1756,78	0,0	380,67	1,76e-159	2511,40	0,0
2019	1849,36	0,0	318,04	1,91e-132	2429,22	0,0
2020	124,43	3,74e-91	158,82	1,35e-63	73,98	7,85e-27
2021	1029,23	0,0	347,03	6,00e-145	796,80	0,0
2022	1648,31	0,0	662,50	7,71e-281	1801,63	0,0

Fonte: O Autor (2024)

Tabela 5 - Resultados dos testes t para as variáveis binárias (apenas p-valor)

Ano	Energia	Esgoto	Biblioteca	Quadra de esportes	Lab. de ciências	Lab. de informática
2015	0,0	2,19e-156	1,09e-45	1,09e+00	1,66e-288	4,74e-11
2016	0,0	2,17e-202	1,22e-48	9,90e+02	0,0	5,97e-10
2017	0,0	0,0	3,61e-41	1,02e-09	4,22e-269	3,16e-27
2018	0,0	7,81e-294	6,23e-49	2,51e-36	2,14e-247	5,38e-27
2019	0,0	0,0	4,7e-24	9,46e-87	7,99e-109	4,60e-74
2020	1,02e-52	3,93e-79	7,47e-53	2,9e-03	4,19e-159	1,32e-04
2021	9,45e-307	4.18e-263	9,97e-18	2,01e-25	4,2e-70	7,15e-11
2022	0,0	0,0	1,12e-15	2,80e-59	7,37e-50	3,88e-97

Fonte: O Autor (2024)

Para o desenvolvimento do modelo preditivo, foi necessário realizar algumas adaptações nos dados categóricos do conjunto de dados. As variáveis categóricas foram mapeadas para valores numéricos, facilitando o processamento pelos algoritmos de AM. Foi o caso da variável icg_nivel_complexidade_gestao_escola, que indica o nível de complexidade da gestão escolar. Ela foi transformada de uma escala nominal (no formato descrito no tópico de limpeza e tratamento dos dados) para uma escala ordinal de 1 a 6.

O mesmo foi feito para a variável que trata das regiões do país onde as escolas se inserem. A ordem estabelecida seguiu a análise das regiões que possuíram, mais frequentemente, as taxas de aprovação escolar maiores ou menores, portanto: Norte, Nordeste, Centro-Oeste, Sul e Sudeste, assumindo, respectivamente, os valores 1, 2, 3, 4 e 5. A variável que distingue entre escolas urbanas e rurais, foi mapeada com 'urbana' sendo 1 e 'rural' sendo 2. Por fim, a variável rede, que identifica a administração da escola, foi codificada com 'municipal' sendo 1, 'estadual' sendo 2 e 'federal' sendo 3.

Para os dados referentes às matrículas, também foram determinados valores numéricos, nesses casos, de 1 a 3. Para a variável maioria_estatistica_cor, que classifica a predominância de cor entre os estudantes de cada escola, a transformação seguiu a seguinte abordagem: 'branca' recebeu o valor 1, 'não_branca' 2 e 'indiferente' 3. Similarmente, a transformação da variável maioria_estatistica_genero, que indica a predominância de gênero, foi codificada com 'feminino' recebendo 1, 'masculino' 2 e 'indiferente' 3.

Essas transformações foram realizadas utilizando a função map() do pacote pandas, que substitui os valores originais pelas suas correspondências numéricas conforme os dicionários acima especificados. A aplicação destas transformações garante a interpretação dos modelos, que irá avaliar a contribuição positiva ou negativa para valores maiores ou menores.

A manipulação, limpeza e tratamento inicial dos dados previamente narrada culminou no conjunto de variáveis que alimentarão o modelo a ser construído. O Apêndice C lista os fatores que foram utilizados para treinar e testar o modelo de previsão. Esse resultado engloba aqueles que não foram excluídos e aqueles que foram criados para viabilizar a análise. São 21 variáveis presentes na base de dados final, incluindo 20 explicativas e a variável dependente. No tópico a seguir, será descrita a utilização da biblioteca PyCaret como ferramenta de desenvolvimento do modelo.

3.4 CONSTRUÇÃO DO MODELO

O uso da biblioteca PyCaret para construção do modelo preditivo se deu pelo seu caráter facilitador da aplicação dos conceitos de AM a partir da linguagem de programação Python. Essa ferramenta visa democratizar o acesso a esse conhecimento a partir da automação das operações necessárias para o seu desenvolvimento com a escrita de poucas linhas de código. É necessário, antes, instalar o pacote com o comando !pip install pycaret[full] e, em seguida, importar as ferramentas de regressão. Para tal, foi utilizado o código from pycaret.regression import *.

Em seguida, foi configurado um experimento usando a função setup(), que recebeu como parâmetros a base de dados, a variável alvo, o código da sessão do experimento, o nome do experimento e o tamanho da base de teste. A Tabela 6 mostra o resultado do experimento e alguns dos parâmetros utilizados para gerá-lo. A variável resposta foi definida como a coluna taxa_aprovacao_escola, a semente que identifica a aleatoriedade do experimento é representada pelo valor 1914 e o tamanho da base de teste arbitrado foi 0,8.

O próximo passo compreende a comparação dos modelos. Basta a utilização da função compare_models() que a biblioteca entende que deve construir modelos preditivos a partir das informações do experimento anteriormente definido. Foram comparados 18 modelos distintos a partir de diversas métricas de performance (Tabela 7). Os indicadores de desempenho usados foram o Erro Absoluto Médio (MAE), Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE), Coeficiente de Determinação (R^2), Raíz do Erro Logarítmico Quadrático Médio (RMSLE) e Média Percentual Absoluta do Erro (MAPE).

Esses modelos são de diversos tipos: Random Forest Regressor, Decision Tree Regressor, Gradient Boosting Regressor, AdaBoost Regressor são baseados em árvores de decisão; Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Bayesian Ridge, Lasso Least Angle Regression e Orthogonal Matching Pursuit são modelos de regressão linear; CatBoost

Regressor, Extreme Gradient Boosting (XGBoost) e Light Gradient Boosting Machine (LightGBM) são modelos de boosting e ensemble; K Neighbors Regressor é um modelo baseados em vizinhança e similaridade; e Huber Regressor, Passive Aggressive Regressor, Least Angle Regression e Dummy Regressor são modelos de regressão alternativos (GARETH *et al.*, 2013; GÉRON, 2022; HASTIE *et al.*, 2009; ZHOU, 2012).

O modelo de regressão que obteve melhor desempenho em todos os parâmetros – como destacado na cor amarela – foi do tipo CatBoost (uma abreviação de *Categorical Boosting*), um algoritmo particularmente eficiente para lidar com dados categóricos e heterogêneos (HANCOCK; KHOSHGOFTAAR, 2020). Esse método trata atributos categóricos como números, sem que haja uma ponderação arbitrária entre os seus valores, evitando um extenso pré-processamento. O modelo, então, utiliza múltiplas árvores de decisão, cada uma treinada para corrigir os erros das anteriores e agrega esses resultados em uma previsão robusta. Esse resultado foi armazenado em uma variável denominada best_model para que as demais análises fossem realizadas.

Para realizar a previsão, foram obtidos dados do censo escolar das escolas públicas brasileiras e dos indicadores presentes no modelo (ver Apêndice C). O primeiro foi obtido pela organização Base dos Dados e os demais no site do INEP. O resultado foram 136.921 escolas aptas, ou seja, que possuem alguma das informações presentes no modelo preditivo e passaram pelos mesmos procedimentos de limpeza e tratamento da base que construiu o modelo preditivo. O próximo capítulo apresentará os resultados da previsão para o ano de 2023 e a análise da série histórica das informações obtidas em termos de fatores mais relevantes e sua influência positiva ou negativa às taxas de aprovação escolar.

Tabela 6 - Resultado da configuração do experimento usando PyCaret

DESCRIÇÃO	VALOR
Session id	1914
Target	taxa_aprovacao_escola
Target type	Regression
Original data shape	(882064, 21)
Transformed data shape	(882064, 37)
Transformed train set shape	(705651, 37)
Transformed test set shape	(176413, 37)
Numeric features	6
Date features	1
Categorical features	13
Rows with missing values	5,3%
Preprocess	True
Imputation type	simple
Numeric imputation	mean
Categorical imputation	mode
Maximum one-hot encoding	25
Encoding method	None
Fold Generator	KFold
Fold Number	10
CPU Jobs	-1
Use GPU	False
Log Experiment	False
Experiment Name	Dissertação_2024
USI	81b6

Fonte: O Autor (2024)

Tabela 7 - Comparação entre modelos

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
CatBoost Regressor	5.0537	63.0.309	79.391	0.4663	0.1506	0.0637	1.167.240
Extreme Gradient Boosting	5.0799	63.6.720	79.794	0.4608	0.1509	0.0640	190.360
Light Gradient Boosting Machine	5.1310	64.1.041	80.064	0.4572	0.1515	0.0647	254.080
Gradient Boosting Regressor	5.3082	66.7.557	81.703	0.4347	0.1535	0.0669	1.989.820
Random Forest Regressor	5.2275	67.4.301	82.115	0.4290	0.1519	0.0654	8.710.980
Linear Regression	6.1060	80.4.191	89.676	0.3190	0.1592	0.0763	122.500
Ridge Regression	6.1060	80.4.191	89.676	0.3190	0.1592	0.0763	95.080
Bayesian Ridge	6.1059	80.4.191	89.676	0.3190	0.1592	0.0763	101.380
K Neighbors Regressor	5.7559	81.0.038	90.001	0.3141	0.1593	0.0721	6.598.500
Elastic Net	6.1499	81.5.353	90.296	0.3096	0.1598	0.0769	84.540
Lasso Regression	6.1582	81.6.410	90.355	0.3087	0.1599	0.0770	82.380
Lasso Least Angle Regression	6.1583	81.6.411	90.355	0.3087	0.1599	0.0770	80.520
Huber Regressor	6.0578	84.9.650	92.176	0.2805	0.1618	0.0769	315.320
Orthogonal Matching Pursuit	6.4169	86.5.496	93.031	0.2671	0.1617	0.0800	75.360
AdaBoost Regressor	7.4399	98.9.894	99.275	0.1617	0.1638	0.0872	478.380
Passive Aggressive Regressor	7.7959	1.12.4.173	105.051	0.0480	0.1709	0.0960	109.400
Dummy Regressor	8.2098	1.18.0.954	108.671	-0.0000	0.1766	0.1025	87.920
Decision Tree Regressor	6.9800	1.34.4.208	115.939	-0.1383	0.2213	0.0863	214.540
Least Angle Regression	6457063.9760	407794900569935.0625	9031045.4820	3494897471569.5103	2.8043	72190.8642	7.4060

Fonte: O Autor (2024)

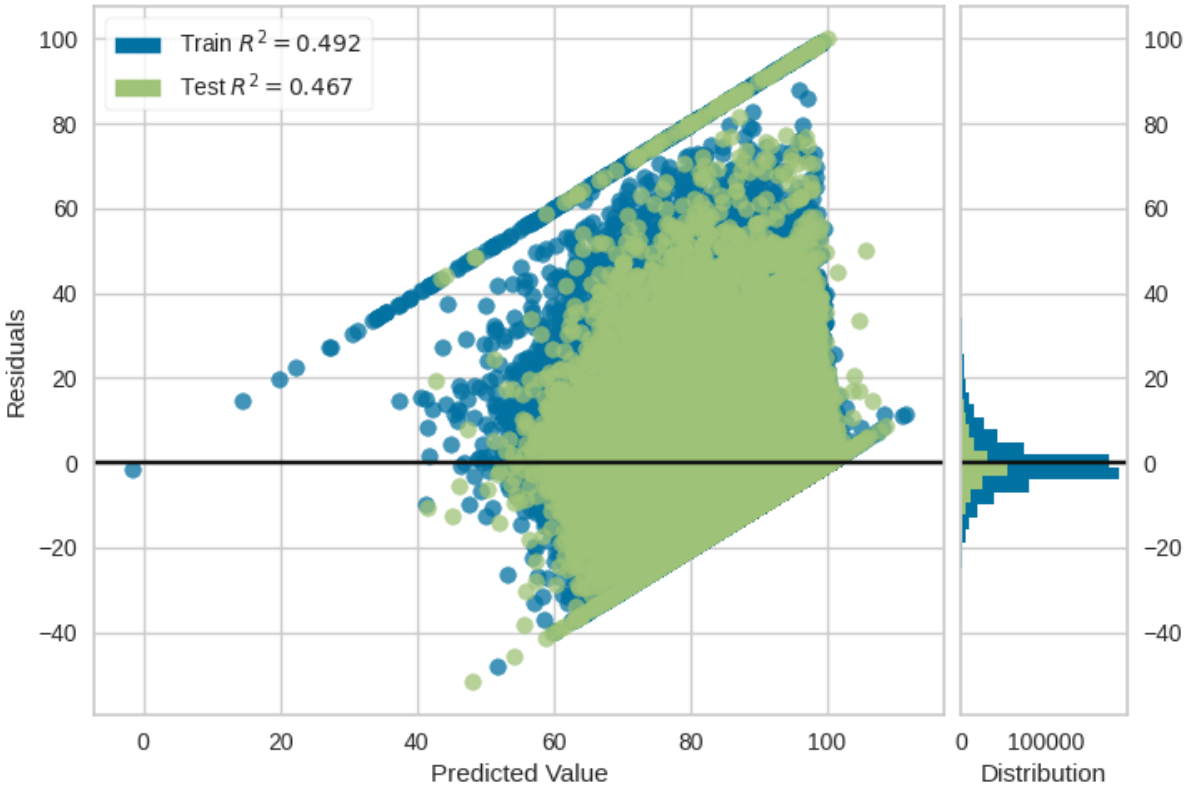
4 RESULTADOS

Este capítulo apresentará os resultados obtidos a partir da análise e aplicação do modelo preditivo. Este capítulo está dividido em três seções principais. Primeiramente, o modelo será avaliado pelo quanto ele explica a realidade com base nos dados de treino e de teste. Em seguida, será aplicado aos dados obtidos relativos ao ano de 2023 e seus resultados analisados à luz das estratificações feitas no desenvolvimento dessa pesquisa. Por fim, ele será interpretado pela lógica dos Valores de Shapley, avaliando como a variação dos fatores afeta positiva ou negativamente o modelo e exemplificando como essa contribuição se manifesta.

4.1 MODELO DE PREVISÃO

Como exposto no capítulo anterior, o algoritmo selecionado – por ter alcançado o melhor desempenho – para desenvolver o modelo preditivo foi a Regressão CatBoost. Para avaliá-lo, foi feita uma análise dos resíduos (Figura 29). O gráfico no traz, no eixo horizontal, os valores previstos pelo modelo, enquanto no vertical estão os resíduos, ou seja, as diferenças entre estes e os seus respectivos valores reais, de acordo com os dados de treino ou de teste. Ainda, o lado direito da imagem mostra a distribuição dos resíduos, proporcionando a visualização acerca da dispersão e tendência central dos resíduos.

Figura 29 - Resíduos do modelo de regressão CatBoost



Fonte: Esta pesquisa (2024)

Nota-se que existem resíduos tanto positivos quanto negativos e que a distribuição dos resíduos segue, visualmente, o formato de uma distribuição normal. É perceptível que há uma tendência central em torno do 0 e há uma relativa simetria da distribuição, o que indica não haver um viés significativo no modelo. Pela grande concentração dos resíduos perto do 0, o modelo se mostra eficaz ao prever, com pequenos erros, a maioria das entradas. Porém, a grande maioria dos resíduos se concentra entre 20 e -20, com alguns pontos extrapolando o limite superior. Essa grande dispersão em torno da média indica que ainda há erros consideráveis. Destacam-se alguns pontos nos quais o resíduo é muito acentuado.

Evidentemente, o cenário ideal é aquele com a menor dispersão possível – que demonstra a precisão do modelo. Entretanto, o erro menos prejudicial ao contexto estudado é aquele que subestima os valores. Resíduos positivos indicam que o valor real é maior que o previsto. Em uma situação na qual os gestores públicos buscam alocar esforços para a melhoria da taxa de aprovação de uma ou mais escolas para o ano seguinte, é de se esperar que aquelas cujas taxas previstas são menores recebam maior atenção e vice-versa. Logo, subestimar os valores acarretaria na sobre alocação de recursos, enquanto errar para mais ocasionaria menos recursos empregados do que o indicado. No contexto do melhoramento, o dano causado pela sobre alocação é menor, pois o potencial de melhoria tende a ser maior. Erros de supervalorização do indicador, na conjuntura citada, podem resultar na negligência de uma escola – ou rede – que

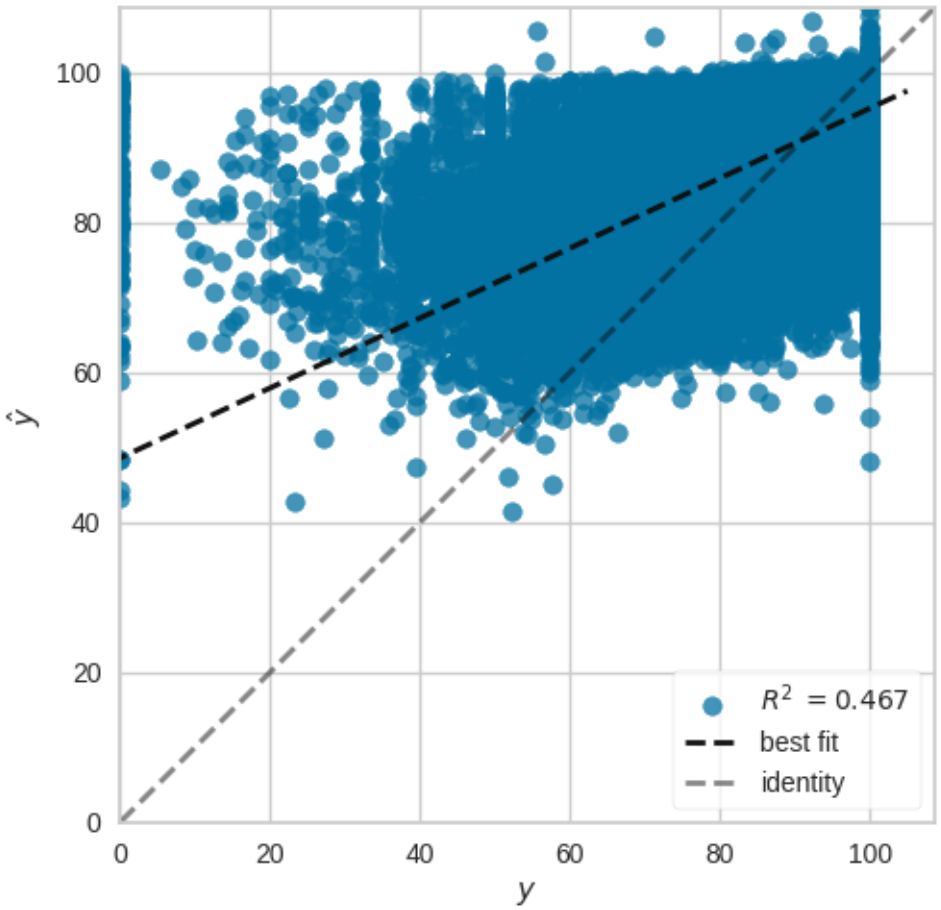
precise de maior atenção do que a que receberá. Em síntese, se não for observada simetria, é preferível uma assimetria positiva.

O gráfico também expõe o coeficiente de determinação, denotado por R^2, dos dados de treino e de teste utilizados. O valor de R^2 do treino aponta que o modelo explica 49.2% da variabilidade dos dados de treino, enquanto para os dados de teste esse valor é de 46.7%. A diferença entre os valores de R^2 de treino e teste, por ser pequena, sugere que o modelo não está sofrendo de um desequilíbrio no seu ajuste aos dados de treino e de teste. Ademais, este patamar de coeficiente de determinação é considerado bom (CHICCO; WARRENS; JURMAN, 2021), em especial num ambiente tão complexo quanto o educacional e com influência de tantos fatores externos. No capítulo a seguir, serão discutidas as implicações disso para os objetivos do trabalho.

O gráfico que foca no erro de previsão reforça a análise dos resíduos (Figura 30). Este, diferente do anterior, apresenta os valores reais da base de teste no eixo horizontal e aqueles que foram previstos pelo modelo no eixo vertical. A reta de referência, a identidade (*identity*), denotada pela linha tracejada cinza, indica o parâmetro do resultado ideal: as previsões são iguais aos valores que foram observados na realidade. Já a reta tracejada preta (*best fit*) indica a tendência geral das previsões quando comparadas às observações. Quanto mais próximas as duas retas, melhor.

Novamente observa-se uma grande dispersão dos erros em torno da reta identidade, o que demonstra a dificuldade do modelo em prever algumas situações. Em especial, os erros se concentram à medida que os valores reais aumentam (como verificado, também, na análise dos resíduos). Por outro lado, tais erros são bem menores se comparados àqueles relativos aos valores reais mais baixos. A reta de melhor ajuste dos erros sugere a sua tendência. Dado que ela é mais plana do que a reta identidade, conclui-se que o modelo tende a subestimar os valores altos e superestimar os valores baixos. O cruzamento ocorre próximo ao valor de 90% de valor observado. Como os erros se concentram nos valores altos, tal conclusão corrobora com a ideia de que a maioria dos erros levam à sobre alocação de esforços para melhoria.

Figura 30 - Erros de previsão para a regressão CatBoost

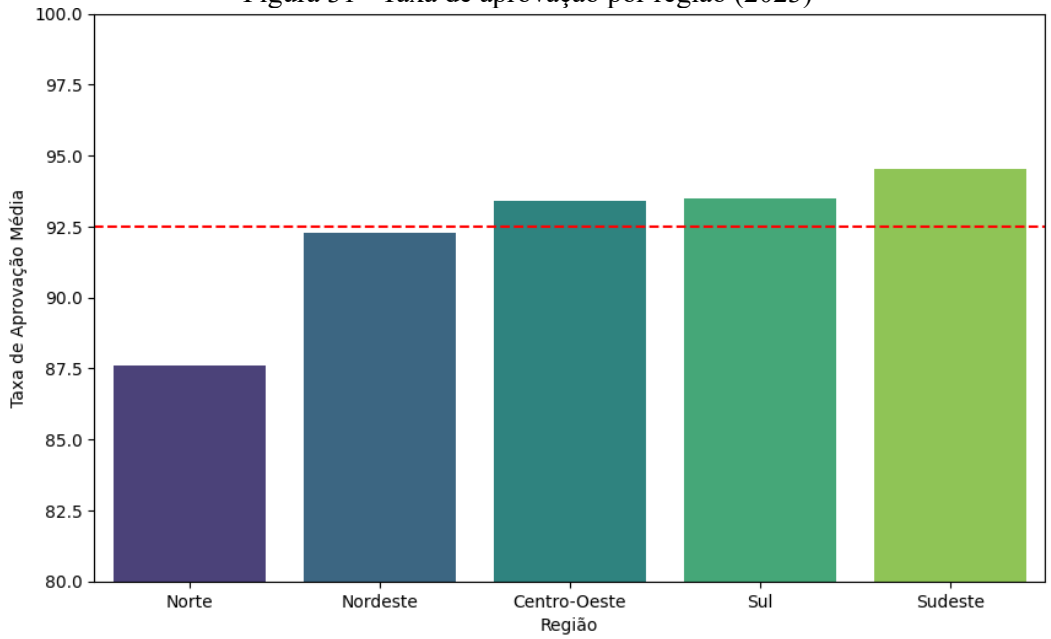


Fonte: Esta pesquisa (2024)

4.2 APLICAÇÃO DA PREVISÃO PARA 2023

O modelo construído foi aplicado ao conjunto de dados relativos ao ano de 2023, após a extração e manipulação previamente descrita. Os resultados da previsão seguiram o padrão que foi observado na análise estatística da base utilizada para construção do modelo. A Figura 31 mostra a taxa de aprovação média por região com o parâmetro da taxa de aprovação média de toda a base de dados. As médias variaram pouco em relação ao que foi observado anteriormente e as regiões, mantendo, em linhas gerais, os patamares de 2022. As maiores variações foram das regiões Centro-Oeste, com queda de 1% e Nordeste, com crescimento de 0,45%. Essas são seguidas de Norte (queda de 0,38%), Sul (aumento de 0,32%) e o Sudeste manteve, praticamente, a mesma taxa em relação à previsão (queda de 0,01%). A média nacional variou positivamente em, aproximadamente, 0,42%.

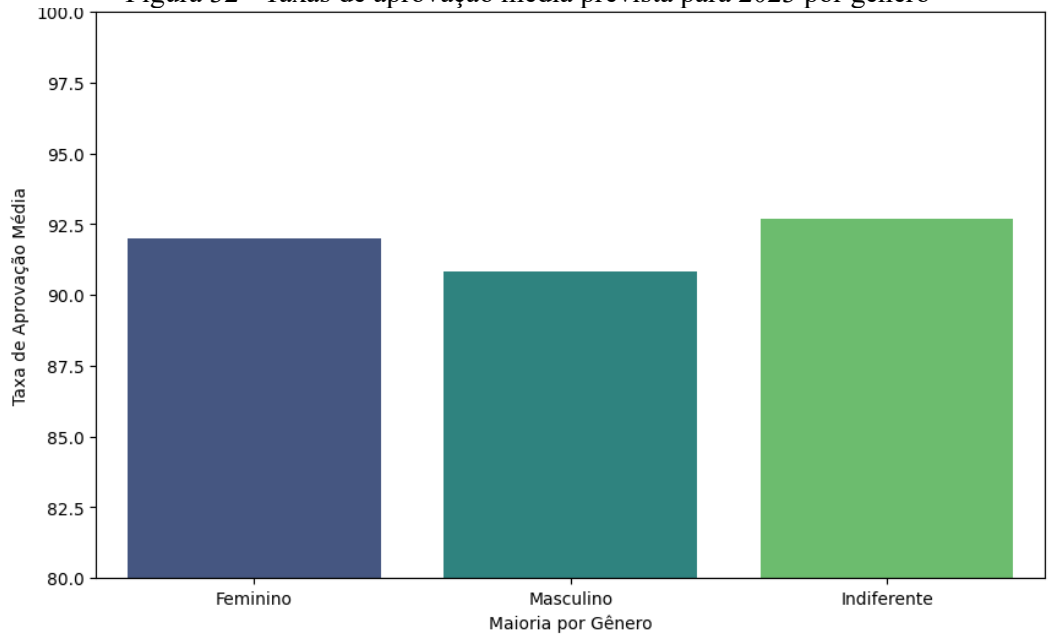
Figura 31 - Taxa de aprovação por região (2023)



Fonte: Esta pesquisa (2024)

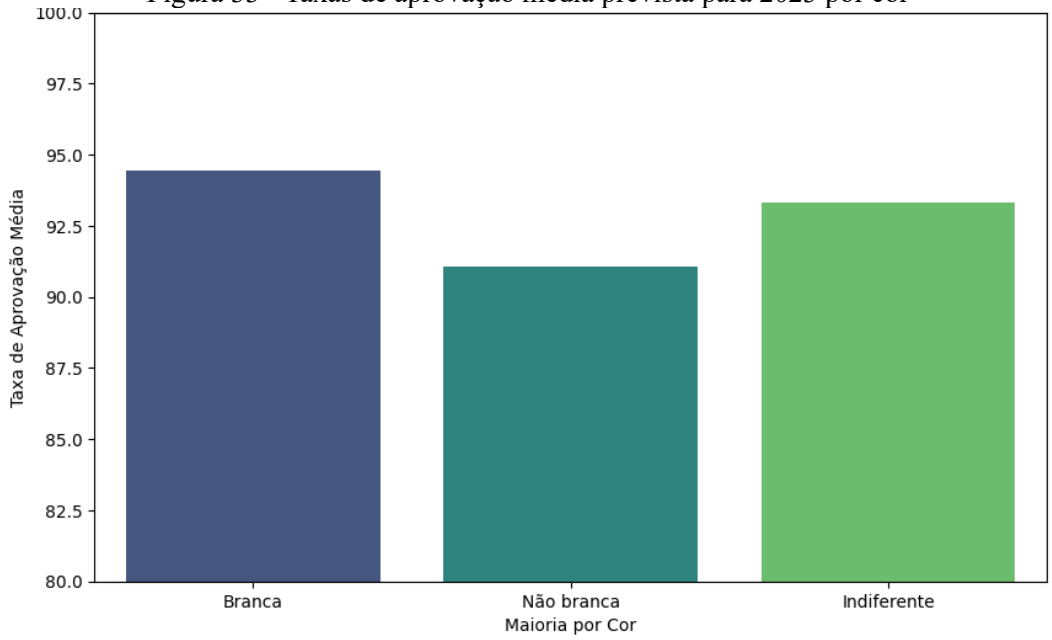
A avaliação feita por maioria de estudantes por fatores demográficos como gênero e cor segue a tendência de diminuição das diferenças, mesmo que mantendo o padrão que fora observado anteriormente. A Figura 32 detalha as médias das taxas de aprovação previstas para escolas com maioria estatisticamente significantes do gênero feminino, masculino e as cuja maioria não é significante. Já a Figura 33 mostra o mesmo para cor entre escolas cujas maiorias estatisticamente significantes de matrículas são brancas, não brancas ou não existe, denotado como “indiferente”.

Figura 32 - Taxas de aprovação média prevista para 2023 por gênero



Fonte: Esta pesquisa (2024)

Figura 33 - Taxas de aprovação média prevista para 2023 por cor

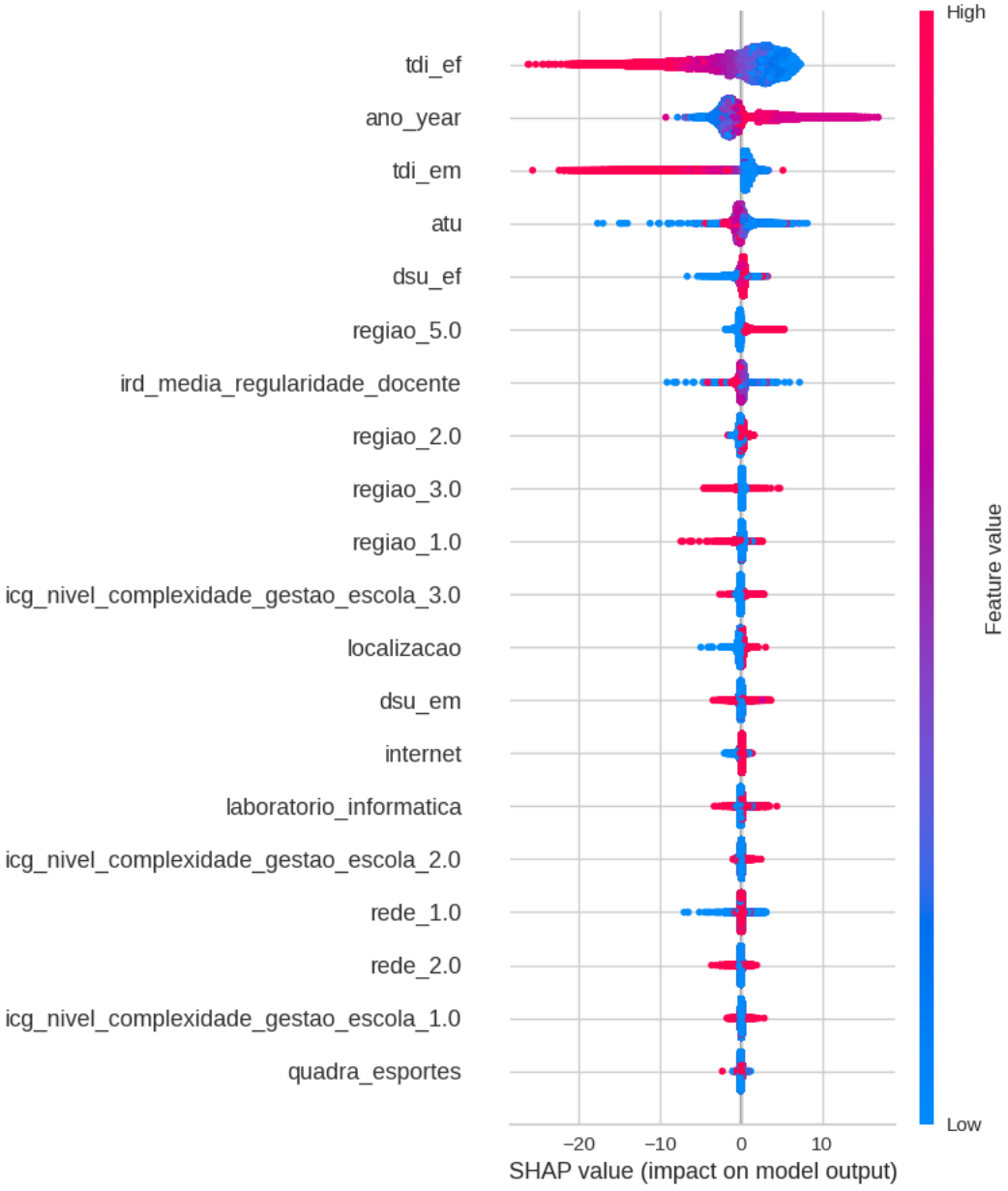


Fonte: Esta pesquisa (2024)

4.3 INTERPRETAÇÃO DO MODELO

Na biblioteca PyCaret, o método SHAP é explorado a partir da função interpret_model(), que recebeu apenas a variável best_model, criada para armazenar a escolha do melhor modelo de acordo com as métricas citadas (Tabela 7). O resultado dessa interpretação é ilustrado pela Figura 34 abaixo, que traz as variáveis à esquerda, a coloração dos pontos à direita e o corpo do gráfico mostra a linha base e como os diferentes pontos indicam Valores de Shapley positivos ou negativos. Quanto mais à esquerda, maior a influência negativa e, quanto mais à direita, maior a influência positiva. E os fatores que estão mais acima exercem maior contribuição ao modelo em comparação aos que estão mais abaixo. Para facilitar a leitura da figura, as variáveis apresentadas na interpretação do modelo estão descritas na Tabela 6.

Figura 34 - Interpretação do modelo por variável pelos Valores de Shapley



Fonte: Esta pesquisa (2024)

Tabela 8 - Descrição das variáveis interpretadas por meio dos Valores de Shapley

Variável	**Descrição**
tdi_ef	Taxa de distorção idade-série no Ensino Fundamental.
ano_year	Ano.
tdi_em	Taxa de distorção idade-série no Ensino Médio.
atu	Média de alunos por turma da escola.
dsu_ef	Taxa de docentes com Ensino Superior no EF.
regiao_5.0	Região Sudeste.
ird_media_regularidade_docente	Índice de regularidade docente, indicando a estabilidade dos professores na escola.
regiao_2.0	Região Nordeste.
regiao_3.0	Região Centro-Oeste.
regiao_1.0	Região Norte.
icg_nivel_complexidade_gestao_escola_3.0	Nível 3 de complexidade na gestão da escola.
localizacao	Localização da escola (urbana ou rural).
dsu_em	Taxa de docentes com Ensino Superior no EF.
internet	Acesso à internet na escola.
laboratorio_informatica	Disponibilidade de laboratório de informática.
icg_nivel_complexidade_gestao_escola_2.0	Nível 2 de complexidade na gestão da escola.
rede_1.0	Rede de ensino municipal.
rede_2.0	Rede de ensino estadual.
icg_nivel_complexidade_gestao_escola_1.0	Nível 1 de complexidade na gestão da escola.
quadra_esportes	Disponibilidade de quadra de esportes.

Fonte: Esta pesquisa (2024)

Portanto, a variável mais influente para o modelo é a taxa de distorção idade-série do EF. Para valores mais baixos da variável, a contribuição ao modelo se concentra em Valores de Shapley positivos até 8 pontos percentuais. Já valores mais altos exercem uma interferência muito mais danosa, chegando a contribuições negativas de mais de 20 pontos percentuais na previsão do modelo. Em seguida a variável ano foi exposta apenas para refletir o cenário em que as taxas de aprovação foram evoluindo ao longo do tempo, tendo seu pico nos dois anos de pandemia, quando se adotou uma política de avaliação por ciclos.

A variável de decisão seguinte é a taxa de distorção idade-série no EM nesse caso, valores mais baixos não exercem influência significativamente positiva. Entretanto, uma alta parcela de estudantes em situação de defasagem sugere causar efeitos desastrosos para a aprovação da escola. Pode-se dizer, então, que a distorção idade-série é um indicador que não traz benefícios tão significativos quanto os malefícios. Antes o contrário: no caso dessa taxa no EM, valores baixos levam, geralmente, a cenários nos quais não haverá perdas, enquanto valores altos aumentam muito as chances de reprovação ou abandono.

A quantidade média de alunos por turma na escola possui uma contribuição relevante para a previsão do modelo. Entretanto, a relação não é trivial. A maioria dos valores menores estão presentes tanto na faixa de influência positiva, mas há uma quantidade considerável de pontos na faixa negativa. Para valores altos, há uma prevalência de pontos na área negativa, mas muito próximos ao zero. Isso faz crer que menos estudantes por turma tendem a ter uma influência positiva na taxa de aprovação, porém, em alguns casos, isso pode representar um problema. Já os valores maiores estão mais concentrados na faixa de influência negativa, mas a distância do zero é muito pequena. A interpretação da variável indica que as escolas devem manter a atual quantidade média de estudantes por turma – em torno de 20 – e, se conveniente, diminuir com cautela.

A quantidade de docentes com ensino superior no EF surge como o próximo fator mais relevante para as taxas de aprovação. Seu comportamento aponta para um efeito nulo para as escolas com altas parcelas dos profissionais graduados. Por outro lado, escolas com baixas taxas de professores com graduação na área resultam em contribuições negativas para a previsão da aprovação escolar. Esse resultado indica que a maior quantidade possível de docentes com grau de formação superior é fator basilar para as taxas de aprovação das escolas. Ou seja, a fartura não é um diferencial positivo, mas a carência é sensivelmente prejudicial.

Os atributos seguintes oferecem uma contribuição modesta ao algoritmo, portanto este trabalho irá focar nas respostas às questões levantadas na fase anterior e em aspectos que interessaram ao objetivo do trabalho. A contribuição da região geográfica na qual se encontra a escola é analisada caso a caso. O modelo avaliou as regiões como variáveis binárias em que 0 (pontos em azul) indica que a escola não se localiza naquela região e 1 (pontos em vermelho) significa que ela está naquela região. A região Sudeste (região_5.0) figurou mais alto no ranking. Estar nesta região contribui positivamente para a taxa de aprovação escolar. Caso contrário, não há influência significativa, apesar de haver pontos com Valor de Shapley ligeiramente negativo.

Apenas outras três regiões aparecem no ranking: Nordeste, Centro-Oeste e Norte, em ordem de relevância. Na primeira, tanto pontos vermelhos quanto azuis se concentram em torno do meio, mas há uma leve tendência positiva às escolas nordestinas em detrimento das outras regiões. No segundo caso, há influência tanto positiva quanto negativa no fato da escola ser do Centro-Oeste. Já as escolas do Norte apresentam a mais clara influência das três: a maioria dos pontos vermelhos está no campo negativo da escala.

A influência da não existência de serviço de internet é, basicamente, negativa. Nos casos em que a escola dispõe do fornecimento desse serviço, os resultados de aprovação não são afetados de forma significativa. Isso indica que a internet pode ser interpretada como item básico para uma escola de tal sorte que a sua presença não traz benefícios, mas a sua ausência é sentida. A presença de uma quadra de esportes na escola é o último fator dos que foram listados na análise estatística a aparecer na interpretação do modelo. Nesse caso, não aparenta ser categorizado como uma influência negativa ou positiva significativa.

Chama atenção o comportamento da variável que versa sobre a localização (se rural ou urbana) da escola. Escolas urbanas apresentaram contribuição negativa, enquanto o oposto foi observado para escolas rurais, o que pode ser contraintuitivo para o senso comum. A presença de um laboratório de informática não teve contribuição constantemente positiva, e pode ser, também, uma informação que vá de encontro à crescente importância que as tecnologias digitais têm na vida da sociedade. Os outros fatores listados, a saber, fornecimento de energia, cobertura da rede de esgoto, presença de biblioteca, existência de laboratório de ciências, maioria de gênero e de cor, não foram contemplados no ranking de contribuição. Portanto, podemos considerar irrelevante suas contribuições para o modelo.

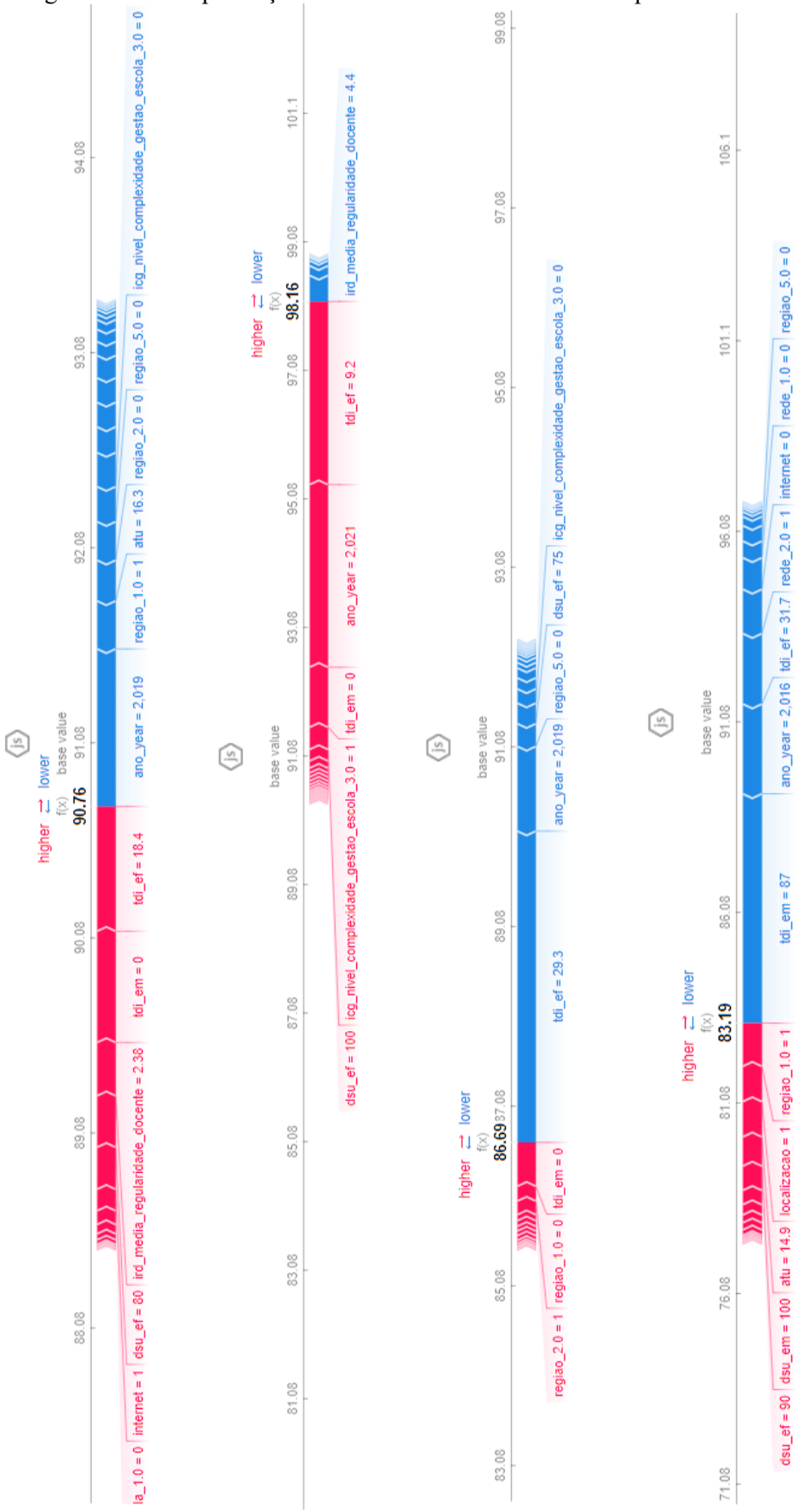
A Figura 35 exemplifica como essas variáveis exercem influência na previsão do modelo, individualmente. O modelo encontrou um valor base de 91,08%, que corresponde ao valor médio de todas as previsões. As faixas em azul indicam as variáveis que contribuíram negativamente para a previsão e o tamanho da sua contribuição. Em vermelho, aquelas cuja contribuição foi positiva. Para gerar os gráficos, foi utilizado o comando interpret_model(best_model, plot='reason', observation=).

No caso mais à esquerda (observation=23), foi prevista uma taxa de aprovação de 90,76%, bem próximo ao valor base. A escola hipotética faz parte da região Norte e exemplifica como a região tem importância. O fato de ser desta região tem efeito negativo – o maior, depois da variável “ano”. Uma quantidade relativamente baixa de alunos por turma possui sua parcela de depreciação da variável resposta. Logo em seguida, o fato de não ser das regiões Nordeste e Sudeste também exercem um efeito de encolhimento na taxa de aprovação prevista. As baixas

taxas de distorção idade-série, tanto no EF quanto no EM, são o principal fator que elevou a previsão ao patamar do valor base.

A segunda previsão (observation=7) mostra uma taxa bem maior que a base: 98,16%. Os principais fatores, excetuando o ano de 2021, foram, novamente, as baixas taxas de distorção idade-série no EM e EM. Isso corrobora com a interpretação do modelo, que mostra essas variáveis entre as que possuem os maiores Valores de Shapley. O mesmo pode ser notado nas outras duas previsões (observation=10 e observation=1914). Com valores encontrados na casa dos 80% – respectivamente, 86,69% e 83,19% –, altos valores de distorção idade-série, seja no EF ou EM, levam o modelo a predizer valores abaixo do valor base. No capítulo seguinte, serão discutidas as conclusões que esses resultados sugerem a este trabalho e conjecturados possíveis trabalhos futuros que esta dissertação pode inspirar.

Figura 35 - Exemplificações da influência das variáveis nas previsões do modelo



Fonte: Esta pesquisa (2024)

5 CONCLUSÕES

Este trabalho teve como objetivo desenvolver e aplicar um modelo de previsão acerca da taxa de aprovação escolar nas escolas públicas brasileiras e interpretá-lo de acordo com as suas características e suas respectivas contribuições para o resultado da previsão. A utilização da biblioteca PyCaret foi um fator primordial para comparação entre diferentes modelos de AM de maneira automatizada, com baixa exigência de escrita de código. O algoritmo que apresentou melhor desempenho nas métricas avaliadas foi a regressão CatBoost.

O modelo desenvolvido apresentou bons resultados, levando em consideração o nível de complexidade e o caráter inovador que o estudo se propôs. Contudo, há espaço para melhorias. O valor do R^2 próximo de 0,5 indica um bom começo para a exploração desse tipo de modelo e metodologia, mas não é suficiente para a implementação como tomada de decisões importantes como no campo da educação pública. A magnitude dos erros do modelo pode ser abreviada com a inclusão de outras variáveis e a utilização de uma capacidade computacional mais robusta para lidar com mais dados. Outros desafios a serem enfrentados estão associados à subjetividade dos objetivos educacionais e a sua incorporação aos modelos de AM, tornando possível a conversão das suas previsões em intervenções educacionais (LIU *et al.*, 2023)

Além disso, foi aplicado o modelo para previsão das taxas de aprovação de 2023, ainda não divulgadas até o momento da escrita desta dissertação. Os resultados seguiram, na média, a tendência de crescimento ao longo dos anos avaliados, excetuando dos dois anos nos quais houve uma mudança na política de avaliação. Por fim, também foi utilizado o método SHAP para interpretação das contribuições de cada fator dentro do modelo e como eles o afetam, o que conferiu uma maior transparência aos seus resultados.

Através da interpretação do modelo pelos Valores de Shapley, a ampla contribuição das taxas de distorção idade-série de uma escola foi evidenciada, seja no EF ou EM. Essa realidade é praticamente irreversível, representando um problema que pode ser, na melhor das hipóteses, mitigado. Uma sugestão para lidar com essa situação é o desenho e implementação de uma política de reforço específica para lidar com as deficiências de conteúdo curricular voltada a esses estudantes, bem como para reduzir a taxa de abandono associada aos fatores externos à escola. Dessa forma, espera-se que, nos anos em que essa política esteja vigente, seja possível superar as dificuldades geradas por esta sucessão de reprovações e abandonos e romper com tal tendência.

Como medida para evitar que novos estudantes sejam afetados por esse problema, sugere-se a que a política de aprovação em ciclos seja reimplementada, o que pode oferecer uma abordagem mais flexível e adaptativa para lidar com as diferentes velocidades de aprendizagem dos alunos. Além disso, a avaliação dos estudantes em ciclos possibilita a correção do seu direcionamento se, por exemplo, no primeiro ano do ciclo, for de maus resultados. Isso pode evitar o sentimento de insucesso que antecede à desmotivação e tem poder de ocasionar outra reprovação e até a desistência do ano letivo. Essa política também pode prover o tempo que profissionais de educação necessitam para lidar com o contexto específico de cada estudante, mitigando os efeitos de um mau ano letivo e os riscos de novos insucessos.

A análise das taxas de aprovação também revelou tanto o benefício obtido por uma escola pelo fato de estar localizada em uma região quanto as perdas associadas ao fato de não ser parte de outra. Esta última informação inclui o benefício e malefício combinados pela possibilidade de ser de qualquer outra região. Por exemplo, há um prejuízo associado à não presença de uma escola no Sudeste, o que inclui as contribuições positivas e negativas das outras regiões, somadas. O tipo de localização da escola também teve um impacto nas previsões. Escolas urbanas apresentaram uma tendência a taxas de aprovação mais baixas, enquanto escolas rurais mostraram contribuição tanto positiva quanto negativa.

Estas duas últimas observações frutos da pesquisa podem indicar que escolas em regiões e locais mais pobres são as mais afetadas negativamente, enquanto aquelas que se localizam nas mais ricas tendem a se beneficiarem. O que pode ser reflexo da má distribuição de recursos é um apontamento já observado como correção no FUNDEB. Essa política, apesar de importante para o financiamento da educação básica, gerou desigualdades de orçamentos entre as cidades. Municípios pobres localizados em estados ricos recebiam menos repasses do Governo Federal em comparação àqueles mais ricos que estão localizados em estados pobres e o novo FUNDEB, aprovado em 2020, tende a corrigir essas assimetrias (IBSA, 2022).

A quantidade média de alunos por turma apresentou um comportamento de difícil análise. Enquanto a sua contribuição para o modelo foi uma das maiores, foi observada uma influência predominantemente positiva em um menor número de estudantes por turma. Por outro lado, há uma quantia razoável de pontos azuis (associados a valores menores) exercendo uma força negativa na previsão. Esse fato é pouco conclusivo acerca de que políticas devem ser adotadas para melhoria das taxas de aprovação e exige uma investigação mais detalhada e aprofundada, avaliando o rendimento de turmas diferentes com quantidade de estudantes diferentes na mesma escola. Pela interpretação do modelo, recomenda-se a diminuição cautelosa da média de alunos por turma.

A quantidade de docentes com ensino superior representa um fator cuja contribuição é nula, quanto maior for a parcela. Todavia, menores proporções exercem influência negativa no modelo, indicando que a busca pela totalidade de docentes com ensino superior em uma escola é algo basilar. Ou seja, não traz efeitos positivos, mas evita os negativos. Isso pode parecer uma conclusão óbvia, mas uma investigação mais cuidadosa revela que, nos últimos anos, essa exigência por uma formação adequada tem sido flexibilizada. Desde os anos 1990 o espaço de formação docente vem sendo aberto para outros ambientes que não as Instituições de Ensino Superior. O chamado Novo Ensino Médio, implementado em 2017 e discutido em 2024, coloca como requisito para dar aula na educação básica apenas o "reconhecimento do notório saber" (MACEDO, 2017). O modelo indica, portanto, que essas políticas sejam descontinuadas e, no lugar, o foco seja direcionado para a formação dentro do ambiente acadêmico. Parcerias entre prefeituras, estados e os centros de formação docente nas universidades locais são sugeridas.

O fato de uma escola ter ou não acesso à internet não representou uma força significativa. Em geral, o acesso não traz efeitos positivos, mas a sua ausência gera efeitos ligeiramente negativos. Isso indica que é positiva uma política de ampliação – e até universalização – de escolas com internet, mas não é uma das medidas prioritárias na busca pela melhoria da taxa de aprovação escolar.

Em síntese, a interpretação do modelo a partir do método SHAP, utilizando a biblioteca PyCaret, para calcular os Valores de Shapley viabilizou a compreensão dos fatores prioritários que contribuem às taxas de aprovação escolar a partir dos dados explorados. A distorção idade-série obteve a maior contribuição, seguida da média de alunos por turma, docentes com ensino superior no EF e outras características com menor importância. O trabalho destacou, entre elas, as regiões, localização e internet para realizar um debate para além do ranking de contribuição. Essas informações são de grande valor, em especial para o debate que está sendo feito em 2024 para determinar o novo PNE, que estará vigente para os próximos 10 anos.

5.1 IMPACTOS

Os resultados deste trabalho têm o potencial de influenciar de diversas formas na sociedade. Esta pesquisa informa gestores públicos acerca de como melhorarem suas as taxas de aprovação escolar. Isso pode reduzir a evasão escolar, visto que reprovações contribuem para a desistência da vida escolar. O aumento da escolarização, por sua vez, possui relevantes efeitos socioeconômicos. Pessoas com níveis mais altos de escolaridade tendem a ter mais e melhores oportunidades de emprego e rendas mais elevadas. Ou seja, elas possuem uma melhor qualidade

de vida e têm a capacidade de ativar a economia pelo consumo. Além disso, esse fator, quando em alta, também tem relação com baixos índices de criminalidade e melhores condições de saúde.

Políticas com o intuito de – no curto prazo – controlar e – no longo – reverter os índices altos de distorção idade-série, como um reforço sistemático a esses estudantes e a avaliação em ciclos, pelo que foi concluído por esta dissertação, podem resultar em efeitos bastante positivos nas taxas de aprovação e nesses indicadores sociais. Ainda, a pesquisa apresenta um conjunto de informações qualificadas para guiar o melhor uso dos recursos públicos. Uma política que visa investir na universalização do acesso à internet, apesar de não ter evidências de efeito puramente benéfico, tende a nivelar as escolas e evitar menores taxas de aprovação. Por outro lado, escolhas de investimento como a construção de quadras de esportes, bibliotecas, laboratório de ciências e laboratório de informática não apresentaram motivos para serem priorizadas.

Por fim há evidência de que a escolarização forma agentes ambientais cujas ações são mais responsáveis e amistosas ao meio ambiente, adotando práticas sustentáveis e apoiando políticas de conservação (CHANKRAJANG; MUTTARAK, 2017). Logo, a educação contribui não apenas para o desenvolvimento econômico e social, mas também para a sustentabilidade ambiental. Esta pesquisa demonstrou que há efeitos negativos em escolas urbanas em relação às rurais. Dessa forma, é de se esperar que políticas que visem melhorar as taxas de aprovação das escolas urbanas contribuirão para as questões ambientais desses agrupamentos. Destacam-se exemplos como limpeza de rios, ruas, arborização das áreas urbanas, defesa de áreas de proteção ambiental etc.

5.2 LIMITAÇÕES

Apesar dos resultados promissores alcançados com o modelo de previsão, variadas são as limitações que precisam ser reconhecidas para proporcionar um entendimento claro dos desafios enfrentados e dos pontos a serem aprimorados. A primeira dificuldade enfrentada foi sobre a disponibilidade dos dados. No início desta pesquisa, a maior parte das bases disponíveis pelo Governo Federal foram retiradas do ar a pretexto de adequá-las à Lei Geral de Proteção de Dados. Além de ter atrasado o trabalho, impossibilitou a análise individualizada dos resultados estudantis. Fatores como gênero, raça, nível socioeconômico e outras questões relativas aos indivíduos não puderam ser explorados de maneira completa e com nível de detalhamento que exigem.

Ainda em relação aos dados, é necessário considerar que, apesar da fonte pública dos dados conferir credibilidade à sua apresentação, a qualidade, em alguns casos, é um fator negativo. Em especial, os dados relativos ao ano de 2023, em muitos casos, ausentes ou incompletos. Possivelmente o processo de coleta não foi realizado a partir das melhores práticas. como observado em algumas variáveis, podem afetar a precisão do modelo e das previsões subsequentes.

A tarefa de tentar representar a complexidade de fatores que influenciam à educação é, em si, uma limitação. Nenhum modelo é capaz de capturar, em sua totalidade, os elementos relativos aos contextos socioeconômicos que se relacionam ao desempenho escolar. Outras áreas, especialmente das ciências sociais, possuem papel importante no complemento dessa simplificação para torná-la tão próxima da realidade quanto possível. Infraestrutura das escolas, formação dos professores, condições socioeconômicas das famílias, questões psicossociais dos estudantes e dos profissionais não foram integradas à análise. Muitas vezes, faltam dados detalhados para representá-los.

O trabalho também se limitou às escolas públicas brasileiras que possuem dados relativos às taxas de aprovação dos seus estudantes de EF e EM. Apesar da abrangência, as características e particularidades regionais e locais foram suprimidas em proveito de um modelo para o país como um todo. É possível que algumas questões discutidas neste trabalho não sejam aplicáveis a realidades específicas, seja por aspectos culturais ou físicos. Um modelo mais específico pode ser construído para abarcar as questões características de cada subconjunto dos sistemas de educação brasileiro. O modelo atual também depende da coleta e divulgação de dados acerca das características das escolas brasileiras por ano letivo, o que ocorre apenas no ano seguinte. Portanto, o modelo serve apenas para informar políticas futuras e não corrigir o ano vigente.

5.3 TRABALHOS FUTUROS

Devido ao caráter inovador do trabalho, diversas possibilidades de pesquisa surgem a partir desse. A união conceitos de AM e Teoria dos Jogos para avaliar as taxas de aprovação escolar nas instituições públicas brasileiras de educação básica, usufruindo, para isso, da praticidade da biblioteca PyCaret, gerou resultados que indicam esses caminhos de continuação. Algumas dessas direções para pesquisas futuras serão discutidas neste tópico.

O primeiro caso – e mais óbvio – é a validação no modelo a partir da divulgação das taxas de aprovação reais para o ano de 2023 e comparação com as previsões obtidas. Como discutido, há espaço pra melhorias no modelo, o que foi concluído a partir de dados de teste. A partir de

dados desconhecidos, a análise da precisão do modelo tende a ser mais assertiva. Isso pode dar às futuras pesquisas um arcabouço argumentativo maior para a melhoria do modelo.

Um aspecto que não foi investigado por esta dissertação foi a contribuição de tecnologias digitais para as taxas de aprovação escolar. Como já discutido, o fato de não haver dados para toda a linha temporal estudada, levou a estes fatores serem excluídos do modelo construído. Porém, há uma lacuna para investigação apenas desses fatores para melhor compreensão de quais dessas tecnologias contribuem de maneira mais intensa para os resultados escolares. Por uma outra perspectiva, essas tecnologias podem ser incorporadas ao modelo construído, usando como base temporal os anos em que há dados acerca dessas variáveis. Essa investigação tem importância primordial com a crescente digitalização das relações cotidianas.

É necessária, também, uma futura investigação mais aprofundada para entender o perfil das escolas urbanas que contribuem para índices de aprovação abaixo da média e daquelas que contribuem positivamente. A partir dessas informações, outros dados podem ser incorporados a um modelo para antever os resultados e agir preventivamente. O mesmo se aplica aos casos em que escolas de algumas regiões algumas regiões sofrem mais com baixa aprovação e outras apresentam melhores resultados.

De maneira geral, o aprimoramento contínuo do modelo é essencial para acompanhar as mudanças e dinâmicas no cenário educacional. A inclusão de novas variáveis, a utilização de técnicas avançadas de análise de dados e a adaptação a novos contextos podem contribuir para o desenvolvimento de um modelo mais preciso e eficiente. O desenvolvimento de um modelo mais complexo, a partir de uma maior capacidade computacional, pode oferecer melhores resultados na previsão das taxas de aprovação. Este modelo deve ser capaz de capturar as nuances e interações entre as diferentes variáveis, proporcionando uma análise mais profunda e detalhada.

A utilização de outros conceitos e ferramentas da Teoria dos Jogos também é enxergada como uma pesquisa de alto potencial. A ideia de jogos cooperativos associada à eficiência na alocação de recursos abre horizontes para intercâmbio entre diferentes modelos matemáticos, seja na educação ou em quaisquer áreas de interesse público.

Em conclusão, este trabalho oferece uma base sólida para a análise e previsão das taxas de aprovação escolar, mas também aponta para a necessidade de avanços contínuos no desenvolvimento de soluções nessa área. Os avanços em um modelo de previsão com alto R^2 e altos índices de acerto pode vir a ser um produto de interesse público para o desenvolvimento social e econômico do país. A educação é um campo dinâmico e desafiador, bem como possui importância central para uma sociedade. A aplicação de modelos preditivos pode contribuir

significativamente para a melhoria das políticas educacionais, o sucesso acadêmico dos estudantes e, por consequência, para uma sociedade menos desigual e socialmente mais justa.

REFERÊNCIAS

AAS, K.; JULLUM, Martin; LØLAND, A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. **ArXiv**, [*s. l.*], v. abs/1903.10464, 2019.

ABRAHA, Abraha Tesfay. Analyzing spatial and non-spatial factors that influence educational quality of primary schools in emerging regions of Ethiopia: Evidence from geospatial analysis and administrative time series data. **Journal of Geography and Regional Planning**, [*s. l.*], v. 12, n. 1, p. 10–19, 2019.

AGÊNCIA BRASIL. **Estudo mostra que pandemia intensificou uso das tecnologias digitais**. [*S. l.*], 2021. Disponível em: https://agenciabrasil.ebc.com.br/geral/noticia/2021-11/estudo-mostra-que-pandemia-intensificou-uso-das-tecnologias-digitais#:~:text=Ao%20todo%2C%2081%25%20da%20popula%C3%A7%C3%A3o,urbanos%20e%2029%25%20nos%20rurais. Acesso em: 16 jun. 2024.

AL MUTAWA, Abdullah; SRUTHI, Sai. Enhancing Human-Computer Interaction in Online Education: A Machine Learning Approach to Predicting Student Emotion and Satisfaction. **INTERNATIONAL JOURNAL OF HUMAN-COMPUTER INTERACTION**, [*s. l.*], 2023.

ALFONSO J. GIL, María Luz Cacheiro-González, Ana María Antelm-Lanzat; PÉREZ-NAVÍO, Eufrasio. School dropout factors: a teacher and school manager perspective. **Educational Studies**, [*s. l.*], v. 45, n. 6, p. 756–770, 2019.

ALI, Moez. PyCaret: An open source, low-code machine learning library in Python. **PyCaret version**, [*s. l.*], v. 2, 2020.

BARROS, R. P. **Políticas públicas para redução do abandono e evasão escolar de jovens**. [*S. l.*]: Fundação Brava, Instituto Unibanco, Insper, Instituto Ayrton Senna, 2017.

BOREL, Emile. La théorie du jeu et les équations intégralesa noyau symétrique. **Comptes rendus de l'Académie des Sciences**, [*s. l.*], v. 173, n. 1304–1308, p. 58, 1921.

BORJA GAMBAU, Juan G. Rodríguez, Juan C. Palomino; SEBASTIAN, Raquel. COVID-19 restrictions in the US: wage vulnerability by education, race and gender. **Applied Economics**, [*s. l.*], v. 54, n. 25, p. 2900–2915, 2022.

BRASIL. **Lei de Diretrizes e Bases da Educação Nacional.** 20 dez. 1996. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9394.htm.

BRASIL, Secretaria do Tesouro Nacional. **Ministério da Educação**. [*S. l.: s. n.*], 2024. Disponível em: https://portaldatransparencia.gov.br/orgaos-superiores/26000-ministerio-da-educacao. Acesso em: 13 jun. 2024.

BRASIL, Agência. **Rendimento domiciliar per capita se recupera em 2022, informa o IBGE**. [*S. l.*], 2023. Disponível em: https://agenciabrasil.ebc.com.br/economia/noticia/2023-05/rendimento-domiciliar-capita-se-recupera-em-2022-informa-o-ibge. .

BRASIL, CONSTITUIÇÃO. **Constituição da República Federativa do Brasil**. Brasília: [*s. n.*], 1988.

BRASIL, CONSTITUIÇÃO. **Planejando a próxima década: conhecendo as 20 metas do Plano Nacional de Educação**. Brasília: Ministério da Educação, 2014.

BRUIN, Boudewijn de. Game theory in philosophy. **Topoi**, [*s. l.*], v. 24, n. 2, p. 197–208, 2005.

BRYK, Anthony S. *et al.* **Learning to Improve: How America's Schools Can Get Better at Getting Better**. [*S. l.*]: Harvard Education Press, 2015.

CAMACHO, Diogo M. *et al.* Next-Generation Machine Learning for Biological Networks. **Cell**, [*s. l.*], v. 173, p. 1581–1592, 2018.

CARDONA, T *et al.* Data Mining and Machine Learning Retention Models in Higher Education. **JOURNAL OF COLLEGE STUDENT RETENTION-RESEARCH THEORY & PRACTICE**, [*s. l.*], v. 25, n. 1, p. 51–75, 2023.

CASTRO, Marcia C *et al.* Brazil's unified health system: the first 30 years and prospects for the future. **The lancet**, [*s. l.*], v. 394, n. 10195, p. 345–356, 2019.

CHANKRAJANG, Thanyaporn; MUTTARAK, Raya. Green Returns to Education: Does Schooling Contribute to Pro-Environmental Behaviours? Evidence from Thailand. **Ecological Economics**, [*s. l.*], v. 131, p. 434–448, 2017.

CHANTREUIL, Frederic *et al.* **Magnitude and evolution of gender and race contributions to earnings inequality across US regions**. **RESEARCH IN ECONOMICS**THE BOULEVARD, LANGFORD LANE, KIDLINGTON, OXFORD OX5 1GB, OXON, ENGLANDELSEVIER SCI LTD, , 2021.

CHICCO, Davide; WARRENS, Matthijs J.; JURMAN, Giuseppe. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. **PeerJ Computer Science**, [*s. l.*], v. 7, p. e623, 2021.

COPPIN, Ben. **Inteligência artificial**. [*S. l.*]: Grupo Gen-LTC, 2015.

COURNOT, Augustin. **Recherches sur les Principes Mathématiques de la Théorie des Richesses**. English edition, 1897ed. London: Macmillan, 1838.

CURIEL, Imma. **Cooperative game theory and applications: cooperative games arising from combinatorial optimization problems**. [*S. l.*]: Springer Science & Business Media, 2013. v. 16

DAHIS, Ricardo *et al.* Data Basis (Base dos Dados): Universalizing Access to High-Quality Data. **Available at SSRN 4157813**, [*s. l.*], 2022.

DANIEL, Ben Kei. Big Data and data science: A critical review of issues for educational research. **British Journal of Educational Technology**, [*s. l.*], v. 50, n. 1, p. 101–113, 2019.

DAY, Christopher; GU, Qing; SAMMONS, Pam. The Impact of Leadership on Student Outcomes: How Successful School Leaders Use Transformational and Instructional Strategies

to Make a Difference. **Educational Administration Quarterly**, [*s. l.*], v. 52, n. 2, p. 221–258, 2016.

DIETTERICH, Thomas G. Machine learning. **ACM Comput. Surv.**, [*s. l.*], v. 28, p. 3, 1996.

DOANE, David P; SEWARD, Lori Welte. **Applied statistics in business and economics**. [*S. l.*]: Mcgraw-Hill, 2016.

DUFWENBERG, M. Game theory. **Wiley interdisciplinary reviews. Cognitive science**, [*s. l.*], v. 2 2, p. 167–173, 2011.

DUPÉRÉ, V. *et al.* Stressors and Turning Points in High School and Dropout. **Review of Educational Research**, [*s. l.*], v. 85, p. 591–629, 2015.

DURYEA, S.; LAM, D.; LEVISON, D. Effects of Economic Shocks on Children's Employment and Schooling in Brazil. **Journal of development economics**, [*s. l.*], v. 84 1, p. 188–214, 2007.

EDDY, W. Large Data Sets in Statistical Computing. [*s. l.*], p. 8382–8386, 2001.

EKER, Remzi; ALKIS, Kamber Can; AYDIN, Abdurrahim. Assessment of large-scale multiple forest disturbance susceptibilities with AutoML framework: an Izmir Regional Forest Directorate case. **JOURNAL OF FORESTRY RESEARCH**, [*s. l.*], v. 35, n. 1, 2024.

ENGBOM, Niklas; MOSER, Christian. Earnings inequality and the minimum wage: Evidence from Brazil. **American Economic Review**, [*s. l.*], v. 112, n. 12, p. 3803–3847, 2022.

FAHD, Kiran *et al.* Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. **EDUCATION AND INFORMATION TECHNOLOGIES**, [*s. l.*], v. 27, n. 3, p. 3743–3775, 2022.

FIANI, Ronaldo. **Teoria dos jogos: para cursos de administração e economia**. [*S. l.*]: Elsevier Brasil, 2015.

FILHO, João Cardoso Palma. Políticas públicas de financiamento da educação no Brasil. **EccoS – Revista Científica**, [*s. l.*], v. 8, n. 2, p. 291–312, 2008.

FREIRE, Paulo. **Pedagogia do oprimido**. [*S. l.: s. n.*], 1971.

FROHLICH-WYDER, Marie-Therese; BACHMANN, Hans-Peter; SCHMIDT, Remo S. Classification of cheese varieties from Switzerland using machine learning methods: Free volatile carboxylic acids. **LWT-FOOD SCIENCE AND TECHNOLOGY**, [*s. l.*], v. 184, 2023.

GARETH, James *et al.* **An introduction to statistical learning: with applications in R**. [*S. l.*]: Spinger, 2013.

GARG, Mausam Kumar; CHOWDHURY, Poulomi; KANCHAN, S. K. Md Illias. **An overview of educational inequality in India: The role of social and demographic factors**. **FRONTIERS IN EDUCATION**AVENUE DU TRIBUNAL FEDERAL 34, LAUSANNE, CH-1015, SWITZERLANDFRONTIERS MEDIA SA, , 2022.

GÉRON, Aurélien. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow**. [*S. l.*]: O'Reilly Media, Inc., 2022.

GIBBONS, Robert S. **Game theory for applied economists**. [*S. l.*]: Princeton University Press, 1992.

GONZALEZ ALEU, Fernando *et al.* Increasing service quality at a university: a continuous improvement project. **Quality Assurance in Education**, [*s. l.*], v. 29, n. 2/3, p. 209–224, 2021.

GUBBELS, J.; PUT, C. E. van der; ASSINK, M. Risk Factors for School Absenteeism and Dropout: A Meta-Analytic Review. **Journal of Youth and Adolescence**, [*s. l.*], v. 48, p. 1637–1667, 2019.

GUPTA, Rahul *et al.* Long term estimation of global horizontal irradiance using machine learning algorithms. **OPTIK**, [*s. l.*], v. 283, 2023.

HANCOCK, John T.; KHOSHGOFTAAR, Taghi M. CatBoost for big data: an interdisciplinary review. **Journal of Big Data**, [*s. l.*], v. 7, n. 1, p. 94, 2020.

HANUSHEK, Eric A.; WOESSMANN, Ludger. Education, knowledge capital, and economic growth. *In*: THE ECONOMICS OF EDUCATION. [*S. l.*]: Elsevier, 2020. p. 171–182. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/B9780128153918000148. Acesso em: 22 out. 2020.

HASTIE, Trevor *et al.* **The elements of statistical learning: data mining, inference, and prediction**. [*S. l.*]: Springer, 2009. v. 2

HOSSAIN, SM *et al.* Impact Prediction of Online Education During COVID-19 Using Machine Learning: A Case Study. *In*: INTELLIGENT SUSTAINABLE SYSTEMS, WORLDS4 2022, VOL 2, 2023. (AK Nagar et al., Org.)**Anais [...]**. [*S. l.: s. n.*], 2023. p. 567–582.

IBGE. **IBGE divulga o Rendimento Domiciliar per Capita e o Coeficiente de Desequilíbrio Regional de 2022**. [*S. l.: s. n.*], 2023. Disponível em: https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/37023-ibge-divulga-o-rendimento-domiciliar-per-capita-e-o-coeficiente-de-desequilibrio-regional-de-2022#:
:text=Em%202022%2C%20o%20rendimento%20nominal,um%20CDR%20de%200%2C63.

INEP. **Distorção idade-série é maior entre os meninos**. [*S. l.*], 2021. Disponível em: https://www.gov.br/inep/pt-br/assuntos/noticias/censo-escolar/distorcao-idade-serie-e-maior-entre-os-meninos. Acesso em: 2 set. 2023.

INEP. **Pesquisa revela resposta educacional à pandemia em 2021**. [*S. l.: s. n.*], 2023a. Disponível em: https://www.gov.br/inep/pt-br/assuntos/noticias/censo-escolar/pesquisa-revela-resposta-educacional-a-pandemia-em-2021.

INEP. **Resumo Técnico Censo Escolar 2023**. [*S. l.: s. n.*], 2023b. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/resumo_tecnico_censo_escolar_2023.pdf.

INSTITUTO BRASILEIRO DE SOCIOLOGIA APLICADA. O novo FUNDEB e seus impactos para os estados e municípios em 2022. [*s. l.*], 2022. Disponível em: https://ibsa.org.br/o-novo-fundeb-e-seus-impactos-para-os-estados-e-municipios-em-2022/.

IPARRAGUIRRE-VILLANUEVA, Orlando *et al.* Comparison of Predictive Machine Learning Models to Predict the Level of Adaptability of Students in Online Education. **INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS**, [*s. l.*], v. 14, n. 4, p. 494–503, 2023.

JANIESCH, Christian; ZSCHECH, Patrick; HEINRICH, K. Machine learning and deep learning. **Electronic Markets**, [*s. l.*], v. 31, p. 685–695, 2021.

JORDAN, Michael I.; MITCHELL, T. Machine learning: Trends, perspectives, and prospects. **Science**, [*s. l.*], v. 349, p. 255–260, 2015.

JOSE, Rejath *et al.* Cardiovascular Health Management in Diabetic Patients with Machine-Learning-Driven Predictions and Interventions. **APPLIED SCIENCES-BASEL**, [*s. l.*], v. 14, n. 5, 2024.

KAR, SP *et al.* Assessment of learning parameters for students' adaptability in online education using machine learning and explainable AI. **EDUCATION AND INFORMATION TECHNOLOGIES**, [*s. l.*], 2023.

KILIC, K. *et al.* Soft ground tunnel lithology classification using resampling and supervised learning. *In*: , 2023. (A Benardos, G Anagnostou, & VP Marinos, Org.)**PROCEEDINGS OF THE ITA-AITES WORLD TUNNEL CONGRESS 2023, WTC 2023: Expanding Underground-Knowledge and Passion to Make a Positive Impact on the World**. [*S. l.*]: Int Tunnelling & Underground Space Assoc, 2023. p. 2733–2741.

LIU, Lydia T. *et al.* Reimagining the machine learning life cycle to improve educational outcomes of students. **Proceedings of the National Academy of Sciences**, [*s. l.*], v. 120, n. 9, 2023. Disponível em: https://dx.doi.org/10.1073/pnas.2204781120.

LUAN, H; TSAI, CC. A Review of Using Machine Learning Approaches for Precision Education. **EDUCATIONAL TECHNOLOGY & SOCIETY**, [*s. l.*], v. 24, n. 1, p. 250–266, 2021.

MACEDO, Jussara Marques de. Reconhecimento do notório saber e a inclusão excludente do professor na educação básica: qual o lugar da universidade na formação?. [*s. l.*], v. 21, p. 1239–1259, 2017.

MCINTYRE, Nora A. **Access to online learning: Machine learning analysis from a social justice perspective**. **EDUCATION AND INFORMATION TECHNOLOGIES**ONE NEW YORK PLAZA, SUITE 4600, NEW YORK, NY, UNITED STATESSPRINGER, , 2023.

MENDONÇA, Yuri V. S.; NARANJO, Paola G. Vinueza; PINTO, Diego Costa. The Role of Technology in the Learning Process: A Decision Tree-Based Model Using Machine Learning. **Emerging Science Journal**, [*s. l.*], v. 6, p. 280–295, 2023.

MUHAMEDYEV, R. *et al.* The use of machine learning "black boxes" explanation systems to improve the quality of school education. **COGENT ENGINEERING**, [*s. l.*], v. 7, n. 1, 2020.

MUNIR, Hussan; VOGEL, Bahtijar; JACOBSSON, Andreas. Artificial Intelligence and Machine Learning Approaches in Digital Education: A Systematic Revision. **INFORMATION**, [*s. l.*], v. 13, n. 4, 2022.

MUNN, Michael; PITMAN, David. **Explainable AI for Practitioners**. [*S. l.*]: O'Reilly Media, Inc., 2022.

MYERSON, Roger B. **Game Theory: Analysis of Conflict**. Cambridge, Massachusetts: Harvard University Press, 1991.

NASH, John F. NON-COOPERATIVE GAMES. **Classics in Game Theory**, [*s. l.*], 1951.

NO, Fata; TANIGUCHI, K.; HIRAKAWA, Yukiko. School dropout at the basic education level in rural Cambodia: Identifying its causes through longitudinal survival analysis. **International Journal of Educational Development**, [*s. l.*], v. 49, p. 215–224, 2016.

PALACIOS, Carlos A. *et al.* Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile. **ENTROPY**, [*s. l.*], v. 23, n. 4, 2021.

PARK, Young-Eun. Uncovering trend-based research insights on teaching and learning in big data. **Journal of Big Data**, [*s. l.*], v. 7, n. 1, p. 93, 2020.

PEKTAS, ST. A Systematic Analysis of Machine Learning Studies in Education. *In*: PROCEEDINGS OF THE 15TH INTERNATIONAL CONFERENCE ON EDUCATION TECHNOLOGY AND COMPUTERS, ICETC 2023, 2023. (ACM, Org.)**Anais [...]**. [*S. l.: s. n.*], 2023. p. 451–455.

POUDYAL, Sujan; MOHAMMADI-ARAGH, Mahnas J.; BALL, John E. Prediction of Student Academic Performance Using a Hybrid 2D CNN Model. **Electronics**, [*s. l.*], v. 11, n. 7, p. 1005, 2022.

QEDU. **Distorção idade-série no Brasil**. São Paulo: Fundação Lemann, 2023. Disponível em: https://qedu.org.br/brasil/distorcao-idade-serie?ano=2020&dependencia_id=5&localizacao_id=0&ciclo_id=AF.

RADHAKRISHNAN, Menaka; BORUAH, Swagata; RAMAMURTHY, Karthik. EEG-Based Anamoly Detection for Autistic Kids - A Pilot Study. **TRAITEMENT DU SIGNAL**, [*s. l.*], v. 39, n. 3, p. 1005–1012, 2022.

RAHUL; KATARYA, Rahul. A Systematic Review on Predicting the Performance of Students in Higher Education in Offline Mode Using Machine Learning Techniques. **WIRELESS PERSONAL COMMUNICATIONS**, [*s. l.*], v. 133, n. 3, p. 1517–1546, 2023.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. **Nature Machine Intelligence**, [*s. l.*], v. 1, p. 206–215, 2018.

RUSSELL, Stuart J; NORVIG, Peter. **Artificial intelligence: a modern approach**. [*S. l.*]: Pearson, 2016.

SANTOS, Mariana Cristina Silva *et al.* The Bolsa Familia Program and educational indicators of children, adolescents, and schools in Brazil: a systematic review/Programa Bolsa Familia e

indicadores educacionais em criancas, adolescentes e escolas no Brasil: revisao sistematica. **Ciencia & saude coletiva**, [*s. l.*], v. 24, n. 6, p. 2233–2248, 2019.

SHAPLEY, Lloyd S. A value for n-person games. [*s. l.*], 1953.

SMITH, C. A. B.; NEUMANN, John Von; MORGENSTERN, O. Theory of Games and Economic Behavior. **Journal of the American Statistical Association**, [*s. l.*], v. 40, p. 263, 1945.

SUNDARARAJAN, Mukund; NAJMI, Amir. The Many Shapley Values for Model Explanation. *In*: , 2020. (Hal Daumé III & Aarti Singh, Org.)**Proceedings of the 37th International Conference on Machine Learning**. [*S. l.*]: PMLR, 2020. p. 9269–9278. Disponível em: https://proceedings.mlr.press/v119/sundararajan20b.html.

TABACARU, Gigi *et al.* A Robust Machine Learning Model for Diabetic Retinopathy Classification. **JOURNAL OF IMAGING**, [*s. l.*], v. 10, n. 1, 2024.

TODOS PELA EDUCAÇÃO. **Anuário Brasileiro da Educação Básica 2021**. São Paulo: Editora Moderna, 2022.

TSAI, Eline R. *et al.* Turnaround time prediction for clinical chemistry samples using machine learning. **CLINICAL CHEMISTRY AND LABORATORY MEDICINE**, [*s. l.*], v. 60, n. 12, p. 1902–1910, 2022.

UBS; CREDIT SUISSE. **Global Wealth Report 2023: leading perspectives to navigate the future**. Zurich: UBS; Credit Suisse, 2023.

UNESCO. **The hidden crisis: Armed conflict and education**. Paris: UNESCO, 2011.

UNICEF. **Panorama da distorção idade-série no Brasil**. [*S. l.: s. n.*], 2018. Disponível em: https://www.unicef.org/brazil/media/461/file/Panorama_da_distorcao_idade-serie_no_Brasil.pdf.

UNICEF; IPEC. **Educação brasileira em 2022 – a voz de adolescentes**. São Paulo: UNICEF; Ipec, 2022.

VAN DE SCHOOT, Rens *et al.* An open source machine learning framework for efficient and transparent systematic reviews. **Nature machine intelligence**, [*s. l.*], v. 3, n. 2, p. 125–133, 2021.

VILLEGAS-CH, William; GOVEA, Jaime; REVELO-TAPIA, Solange. Improving Student Retention in Institutions of Higher Education through Machine Learning: A Sustainable Approach. **SUSTAINABILITY**, [*s. l.*], v. 15, n. 19, 2023.

VITORINO DE ARRUDA, Danilo; RAMOS, Francisco. APROVACÃO ESCOLAR À LA SHAPLEY: as desigualdades da educação pública brasileira. *In*: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 2023, São José dos Campos - SP, BR. **Anais do Simpósio Brasileiro de Pesquisa Operacional**. São José dos Campos - SP, BR: Galoa, 2023. Disponível em: https://proceedings.science/sbpo/sbpo-2023/trabalhos/aprovacao-escolar-a-la-shapley-as-desigualdades-da-educacao-publica-brasileira?lang=pt-br. Acesso em: 28 fev. 2023.

WHIG, Pawan *et al.* A novel method for diabetes classification and prediction with Pycaret. **MICROSYSTEM TECHNOLOGIES-MICRO-AND NANOSYSTEMS-INFORMATION STORAGE AND PROCESSING SYSTEMS**, [*s. l.*], v. 29, n. 10, SI, p. 1479–1487, 2023.

WORLD BANK. **World Development Report 2018: Learning to Realize Education's Promise**. [*S. l.*]: World Bank, 2018. Disponível em: https://openknowledge.worldbank.org/handle/10986/28340.

XIN, Lei *et al.* POPs identification using simple low-code machine learning. **SCIENCE OF THE TOTAL ENVIRONMENT**, [*s. l.*], v. 921, 2024.

XU, QA; DENG, H. Research on education management system based on machine learning and multidimensional data modeling. **APPLIED MATHEMATICS AND NONLINEAR SCIENCES**, [*s. l.*], 2023.

YOUSAFZAI, Bashir Khan; HAYAT, Maqsood; AFZAL, Sher. Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. **EDUCATION AND INFORMATION TECHNOLOGIES**, [*s. l.*], v. 25, n. 6, p. 4677–4697, 2020.

ZAMAN, Bushra *et al.* Modeling education impact: a machine learning-based approach for improving the quality of school education. **JOURNAL OF COMPUTERS IN EDUCATION**, [*s. l.*], 2023.

ZHOU, Zhi-Hua. **Ensemble methods: foundations and algorithms**. [*S. l.*]: CRC press, 2012.

APÊNDICE A – Características das variáveis dos dados baixados

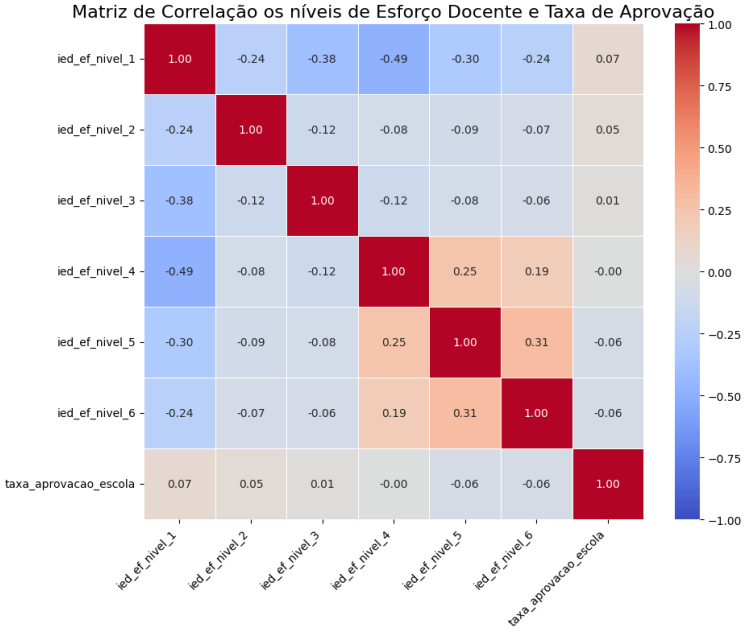
#	Nome da coluna	Descrição	Não-ausentes	Tipo
1	ano	Ano do registro	1231837	Int64
2	id_municipio	Código do município	1231837	object
3	id_escola	Código da escola	1231837	object
4	localizacao	Escola rural/urbana	1231837	object
5	rede	Dependência administrativa	1231837	object
6	atu_ef	# de alunos por turma EF	923967	float64
7	atu_em	# de alunos por turma EM	183458	float64
8	atu_ef_turmas_unif_multi_fluxo	# de alunos por turma EF turmas unificadas	376827	float64
9	had_ef	# alunos por turma EF	727437	float64
10	had_em	# alunos por turma EM	183457	float64
11	had_em_nao_seriado	# hora-aula docente EM não seriado	5169	float64
12	tdi_ef	% de distorção idade-série EF	923967	float64
13	tdi_em	% de distorção idade-série EM	182011	float64
14	taxa_aprovacao_ef	% de aprovação EF	828399	float64
15	taxa_aprovacao_em	% de aprovação EM	162568	float64
16	taxa_aprovacao_em_nao_seriado	% de aprovação EM não seriado	5242	float64
17	taxa_reprovacao_ef	% de reprovação EF	828399	float64
18	taxa_reprovacao_em	% de reprovação EM	162568	float64
19	taxa_reprovacao_em_nao_seriado	% de reprovação EM não seriado	5242	float64
20	taxa_abandono_ef	% de abandono EF	828399	float64
21	taxa_abandono_em	% de abandono EM	162568	float64
22	taxa_abandono_em_nao_seriado	% de abandono EM não seriado	5242	float64
23	dsu_ef	% de docentes com ensino superior EF	923958	float64
24	dsu_em	% de docentes com ensino superior EM	183456	float64
25	dsu_ep	% de docentes com ensino superior EP	18594	float64
26	afd_ef_grupo_1	% de docentes EF com formação adequada - grupo 1	925571	float64
27	afd_ef_grupo_2	% de docentes EF com formação adequada - grupo 2	925571	float64
28	afd_ef_grupo_3	% de docentes EF com formação adequada - grupo 3	925571	float64
29	afd_ef_grupo_4	% de docentes EF com formação adequada - grupo 4	925571	float64
30	afd_ef_grupo_5	% de docentes EF com formação adequada - grupo 5	925571	float64
31	afd_em_grupo_1	% de docentes EM com formação adequada - grupo 1	183448	float64
32	afd_em_grupo_2	% de docentes EM com formação adequada - grupo 2	183448	float64
33	afd_em_grupo_3	% de docentes EM com formação adequada - grupo 3	183448	float64
34	afd_em_grupo_4	% de docentes EM com formação adequada - grupo 4	183448	float64
35	afd_em_grupo_5	% de docentes EM com formação adequada - grupo 5	183448	float64
36	ird_media_regularidade_docente	Índice médio de regularidade docente	1180390	float64
37	ied_ef_nivel_1	% de docentes EF com índice de esforço - nível 1	926254	float64

38	ied_ef_nivel_2	% de docentes EF com índice de esforço - nível 2	926254	float64
39	ied_ef_nivel_3	% de docentes EF com índice de esforço - nível 3	926254	float64
40	ied_ef_nivel_4	% de docentes EF com índice de esforço - nível 4	926254	float64
41	ied_ef_nivel_5	% de docentes EF com índice de esforço - nível 5	926254	float64
42	ied_ef_nivel_6	% de docentes EF com índice de esforço - nível 6	926254	float64
43	ied_em_nivel_1	% de docentes EM com índice de esforço - nível 1	183564	float64
44	ied_em_nivel_2	% de docentes EM com índice de esforço - nível 2	183564	float64
45	ied_em_nivel_3	% de docentes EM com índice de esforço - nível 3	183564	float64
46	ied_em_nivel_4	% de docentes EM com índice de esforço - nível 4	183564	float64
47	ied_em_nivel_5	% de docentes EM com índice de esforço - nível 5	183564	float64
48	ied_em_nivel_6	% de docentes EM com índice de esforço - nível 6	183564	float64
49	icg_nivel_complexidade_gestao_escola	Índice de complexidade da gestão a escola (de 1 a 6)	1231802	object
50	regiao	Região geográfica da escola	1231837	object
51	sigla_uf	Sigla da Unidade Federativa	1231837	object
52	tipo_localizacao_diferenciada	Inclui território indígena, de quilombo, entre outros	1231837	object
53	quantidade_matricula_feminino	Quantidade de matrículas do sexo feminino	1231546	Int64
54	quantidade_matricula_masculino	Quantidade de matrículas do sexo masculinio	1231546	Int64
55	quantidade_matricula_nao_declarada	Quantidade de matrículas do de raça/cor não declarada	1231546	Int64
56	quantidade_matricula_branca	Quantidade de matrículas do de raça/cor branca	1231546	Int64
57	quantidade_matricula_preta	Quantidade de matrículas do de raça/cor preta	1231546	Int64
58	quantidade_matricula_parda	Quantidade de matrículas do de raça/cor parda	1231546	Int64
59	quantidade_matricula_amarela	Quantidade de matrículas do de raça/cor amarela	1231546	Int64
60	quantidade_matricula_indigena	Quantidade de matrículas do de raça/cor indígena	1231546	Int64
61	quantidade_matricula_fundamental	Quantidade de matrículas EF	1231546	Int64
62	quantidade_matricula_medio	Quantidade de matrículas EM	1231546	Int64
63	agua_inexistente	Abastecimento de água (0 - existente; 1 - inexistente)	1231837	Int64
64	energia_inexistente	Fornecimento de energia (0 - existente; 1 - inexistente)	1231837	Int64
65	esgoto_inexistente	Atendimento rede de esgoto (0 - existente; 1 - inexistente)	1231837	Int64
66	banheiro	Escola tem banheiro (1 - sim; 0 - não)	653039	Int64
67	biblioteca	Escola tem biblioteca (1 - sim; 0 - não)	1231837	Int64
68	alimentacao	Escola tem alimentação na escola (1 - sim; 0 - não)	1231837	Int64
69	laboratorio_ciencias	Escola tem laboratório de ciências (1 - sim; 0 - não)	1231837	Int64
70	laboratorio_informatica	Escola tem laboratório de informática (1 - sim; 0 - não)	1231837	Int64
71	laboratorio_educacao_profissional	Escola tem laboratório EP (1 - sim; 0 - não)	238683	Int64

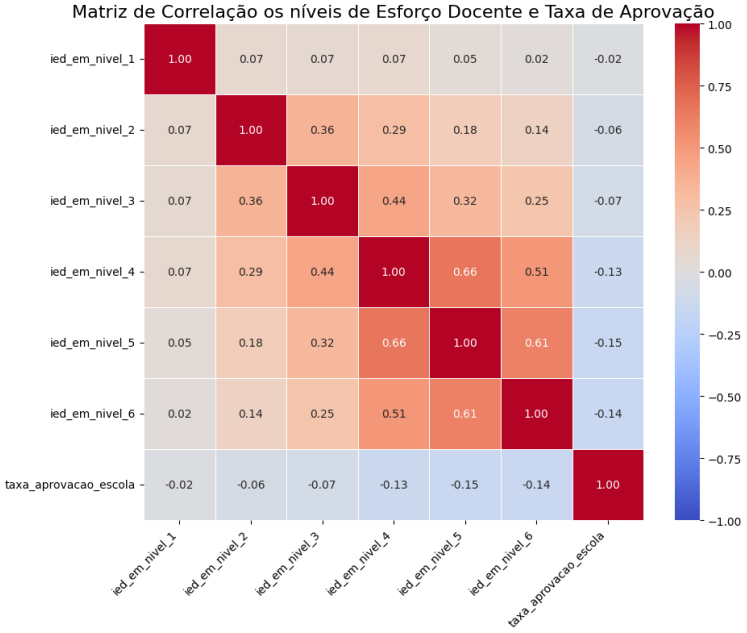
72	quadra_esportes	Escola tem quadra de esportes (1 - sim; 0 - não)	1231837	Int64
73	sala_multiuso	Escola tem sala multiuso (1 - sim; 0 - não)	653039	Int64
74	desktop_aluno	Escola tem computador de mesa para aluno (1 - sim; 0 - não)	653039	Int64
75	computador_portatil_aluno	Escola tem computador portátil para aluno (1 - sim; 0 - não)	653039	Int64
76	tablet_aluno	Escola tem tablet para aluno (1 - sim; 0 - não)	653039	Int64
77	internet	Escola tem acesso à internet (1 - sim; 0 - não)	1231837	Int64

APÊNDICE B – Matrizes de correlação

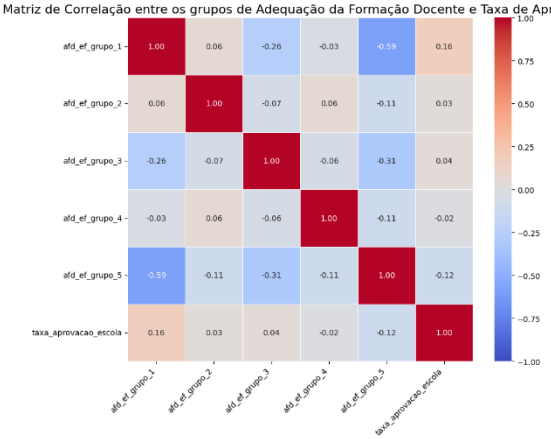
Matriz de Correlação os níveis de Esforço Docente e Taxa de Aprovação



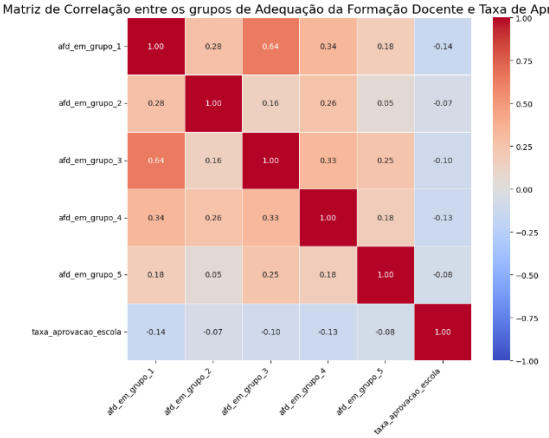
Matriz de Correlação os níveis de Esforço Docente e Taxa de Aprovação



Matriz de Correlação entre os grupos de Adequação da Formação Docente e Taxa de Aprovação



Matriz de Correlação entre os grupos de Adequação da Formação Docente e Taxa de Aprovação



APÊNDICE C – Variáveis para treino e teste do modelo final

#	Nome da coluna	Descrição	Não-ausentes	Tipo
1	ano	Ano do registro	882064	datetime64[ns]
2	atu	Média de alunos por turma	882064	float64
3	taxa_aprovacao_escola	Taxa de aprovação média da escola	882064	float64
4	localizacao	Escola rural/urbana	882064	object
5	rede	Dependência administrativa	882064	object
6	tdi_ef	% de distorção idade-série EF	882064	float64
7	tdi_em	% de distorção idade-série EM	882064	float64
8	ird_media_regularidade_docente	Índice médio de regularidade docente	882064	float64
9	icg_nivel_complexidade_gestao_escola	Índice de complexidade da gestão a escola (de 1 a 6)	882064	object
10	regiao	Região geográfica da escola	882064	object
11	biblioteca	Escola tem biblioteca (1 - sim; 0 - não)	882064	object
12	laboratorio_ciencias	Escola tem laboratório de ciências (1 - sim; 0 - não)	882064	object
13	quadra_esportes	Escola tem quadra de esportes (1 - sim; 0 - não)	882064	object
14	internet	Escola tem acesso à internet (1 - sim; 0 - não)	882064	object
15	maioria_estatistica_genero	Maioria estatística de gênero (1 - feminino, 2 - indiferente, 3 - masculino)	882064	object
16	maioria_estatistica_cor	Maioria estatística de cor (1 - branca, 2 - indiferente, 3 - não branca)	882064	object
17	laboratorio_informatica	Escola tem laboratório de informática (1 - sim; 0 - não)	882064	object
18	energia_inexistente	Fornecimento de energia (0 - existente; 1 - inexistente)	882064	object
19	esgoto_inexistente	Atendimento rede de esgoto (0 - existente; 1 - inexistente)	882064	object
20	dsu_ef	% de docentes com ensino superior EF	882064	float64
21	dsu_em	% de docentes com ensino superior EM	882064	float64