ISSN: 2675-4185

Tendência de adoção de livros didáticos conforme características das obras e das escolas.

Trends in the adoption of textbooks based on the characteristics of literary works and schools

Área temática: Temas Transversais

RUIZ, Alanys Patricia Bronhara Affonso
e-mail: alanysruiz@hotmail.com
ARAÚJO, Elton Gean
e-mail: egarauj@yahoo.com.br

RESUMO

O Programa Nacional do Livro Didático (PNLD) representa um dos maiores setores do mercado editorial no Brasil, movimentando mais de 1 bilhão de reais por ano. Implementado pelo governo, atende todas as escolas públicas do país que optarem por participar, tendo essas a autonomia de selecionar as obras que desejam adquirir. Portanto, a capacidade das editoras de alcançar mais escolas durante o período de divulgação é crucial para maximizar as oportunidades de vendas. O objetivo geral do trabalho foi compreender as relações entre as características das instituições com o perfil das obras escolhidas, e gerar insights para otimização das estratégias de vendas. Foram coletados e tratados os dados do resultado das escolhas no programa PNLD 2023-Ef1-Objeto 1, disponibilizados pelo FNDE, o CENSO escolar 2022 e as notas do IDEB 2021 pelas escolas. Foram realizadas análises de correspondência entre as variáveis, e a construção de um modelo logístico binário. A partir dos resultados das análises de correspondência e o mapa perceptual foi possível observar que as categorias esfera estadual, região sul, localização rural, alunado pequeno e IDEB alto tiveram relações significativas com as coleções robustas. Já os livros simples tiveram uma relação significativa com as categorias esfera municipal, localização urbana, alunado grande, nota de IDEB médio e regiões sudeste e centro-oeste. O modelo logístico apresentou uma performance satisfatória, com oportunidade de melhorias. A análise deste reforçou a relação entre as obras simples e instituições com grande alunado e as robustas com escolas estaduais da região sul.

Palavras-chave: PNLD, ESCOLA, REGRESSÃO.

ABSTRACT

The National Textbook Program (PNLD) represents one of the largest sectors of the publishing market in Brazil, generating over 1 billion reais annually. Implemented by the government, it serves all public schools in the country that choose to participate, giving them the autonomy to select the works they wish to acquire. Therefore, the ability of publishers to reach more schools during the promotion period is crucial to maximizing sales opportunities. The primary objective of this study was to understand the relationships between the characteristics of institutions and the profiles of the chosen works, and to generate insights for optimizing sales strategies. Data were collected and processed from the PNLD 2023-Ef1-Objeto 1 program results provided by FNDE, the 2022 School CENSO, and the 2021 IDEB scores for schools. Correspondence analysis between variables and the construction of a binary logistic model were carried out. The results of the correspondence analysis and the perceptual map revealed that the categories of state sphere, southern region, rural location, small student body, and high IDEB scores were significantly associated with robust collections. In contrast, simpler books were significantly associated with the categories of municipal sphere, urban location, large student body, medium IDEB score, and the southeastern and central-western regions. The logistic model showed satisfactory

ISSN: 2675-4185

performance, with room for improvement. This analysis reinforced the relationship between simpler works and institutions with large student bodies and robust works with state schools in the southern region.

Keywords: PNLD, School, Regression.

1 INTRODUÇÃO

O Programa Nacional do Livro Didático (PNLD) tem como propósito fornecer obras didáticas, literárias, pedagógicas e materiais de apoio à educação, de todas as disciplinas da grade curricular de cada segmento, para escolas de educação básica e instituições de educação infantil das redes públicas (federal, estadual, municipal e conveniadas), sendo o maior programa de distribuição de material didático do país. Para este existe um sistema regulamentado e gratuito (MEC, 2023).

O setor de livros didáticos do Brasil é um dos mais concorridos e lucrativos, em escala global. É uma indústria muito concentrada, onde há poucas editoras grandes e muito espaço. Em relação a valores, no ano de 2016, por exemplo, houve um investimento do governo de 1 bilhão de reais em livros, alcançando cerca de 25 milhões de alunos e 160 mil escolas (FREISLEBEN; KAERCHER, 2022). Dados do FNDE (2023) dos anos de 2019/2020, mostram valores parecidos, mais de 1 bilhão investido, 120 mil escolas e 32 milhões de alunos atendidos. Entre 2012 e 2021 o valor total investido na aquisição de livros no programa foi de quase 11,6 bilhões (ABRELIVROS, 2024).

O programa atende as escolas públicas de todo Brasil, que de acordo com Censo Escolar 2023 (2024) totaliza 137.914 instituições e 35.032.185 alunos (considerando todos os segmentos). Por ano, 45 mil escolas em média participam desse processo de escolha e aquisição de novos materiais, alcançando cerca de 11 milhões de alunos.

As editoras que participam do programa têm um período médio de 3 meses para a divulgação dos seus materiais para as escolas, fazendo visitas e enviando amostras. Tendo em vista esse tempo, o número de escolas, a extensão territorial do país e o tamanho do mercado, observa-se que a eficiência do trabalho das editoras nesse período é de grande relevância.

Diante dessa necessidade, o propósito deste projeto consiste em identificar tendências nas escolhas realizadas de acordo com as características das escolas e os perfis das obras selecionadas. Esta análise busca oferecer insights para a otimização do processo em questão, uma vez que apresentado aos decisores a coleção com perfil mais provável de ser escolhido,

ISSN: 2675-4185

maior a chance de resultados positivos e maior economia de recursos. Para tanto, serão aplicados métodos de análise combinatória e o modelo de regressão logística.

2 REFERENCIAL TEÓRICO

O Programa Nacional do Livro Didático (PNLD) acontece em ciclos, de forma a alternar o atendimento aos segmentos de educação, o de educação infantil, ensino fundamental anos iniciais, ensino fundamental anos finais e ensino médio. Nos ciclos em que os segmentos não são atendidos é feita uma complementação correspondente a reposição necessária de livros e novas matrículas. (MEC, 2023).

O processo é controlado pelo órgão do Fundo Nacional de Desenvolvimento da Educação (FNDE), que é responsável pela compra, distribuição e logística da entrega das obras didáticas. Os materiais são avaliados e selecionados pelo Ministério da Educação, no âmbito da Secretaria de Educação Básica (SEB), sendo as escolas públicas inscritas no programa as principais responsáveis pela tomada de decisão no processo de escolha das obras (MEC, 2023).

Sobre o funcionamento do programa, durante seu ciclo há algumas etapas. Dentre elas: adesão (escolas demonstram interesse na participação do programa), editais (regras para as inscrições dos livros), inscrição das editoras (inscrição das obras pelas empresas), avaliação (valida se as obras inscritas estão de acordo com as regras do edital e aprova ou não para participação no programa), guia do livro (guia sobre as obras aprovadas disponibilizado às escolas pelo FNDE), escolha (período em que os professores escolhem as obras que serão utilizadas), pedido (formalização das escolhas e pedido de compra realizado pelas escolas), aquisição (processo de negociação e compra do FNDE com as editoras), produção (firmamento de contrato e produção dos materiais), análise de qualidade física, distribuição e recebimento (MEC, 2023).

Antes do período de escolha das obras, as editoras exercem a divulgação de seus livros. Práticas comumente realizadas são as visitas nas escolas, o contato direto com os professores e o envio de amostras dos materiais. Os perfis das escolas, como tamanho (em número de alunos) e localização, impactam nas estratégias. Quanto ao próprio produto, há características que são analisadas pelos professores (como a abordagem, autor, linguagem, diagramação, exercícios e orientação) que também são considerados durante a divulgação (VIDAL, 2016).

ISSN: 2675-4185

As editoras são informadas sobre as obras que poderão participar do programa após o resultado da avaliação final, e podem fazer a divulgação destas até um dia antes do primeiro dia de registro da escolha no sistema. Esse intervalo tem uma média de 2 a 3 meses (FNDE, 2023).

3 METODOLOGIA

3.1 Base de dados

Para obtenção dos dados foram utilizadas 3 principais fontes. A primeira foi o CENSO 2022, que mapeia as escolas públicas do Brasil, disponibilizados no site GOV.BR (2023), para coleta de informações gerais das escolas. A segunda foi a paletização das compras das obras didáticas de cada escola no PNLD 2023 Obras Didáticas Objeto 1, para ensino fundamental anos iniciais, disponibilizado pelo SIMAD no site do FNDE (2023). A terceira foi as notas das escolas no IDEB de 2021, divulgada pelo INEP no site GOV.BR (2023).

Consolidando as informações, a Tabela 1 descreve as variáveis utilizadas no presente trabalho.

Tabela 1. Descrição das variáveis utilizadas no estudo.

Descrição da Variável	**Tipo**	**Categoria/ Observação**
MEC	Categórica	Identificação as observações
REGIAO	Categórica	CentroOeste Nordeste Norte Sudeste Sul
LOCALIZACAO	Categórica	Urbana Rural
ESFERA	Categórica	Estadual Municipal
CLASSIFICACAO	Categórica	Simples Robusta
IDEB	Numérica	0 a 10
ALUNADO	Numérica	

Fonte: Dados originais da pesquisa.

A variável categórica “CLASSIFICACAO” foi estabelecida com base em um estudo realizado por uma equipe editorial de uma empresa sediada em São Paulo. O estudo analisou quatro coleções de matemática participantes do programa. As categorias ‘Robusta’ e ‘Simples’ foram determinadas com base na linguagem e abordagem adotadas nas obras. As obras que foram

ISSN: 2675-4185

classificadas como ‘Robustas’ apresentam uma linguagem mais formal, teorias mais extensas e exige maior protagonismo do aluno. Já as obras classificadas como ‘Simples’, adotam uma abordagem mais acessível, com uma linguagem menos formal.

A variável “MEC” é o código MEC, um código identificador único atribuído a cada escola no Brasil, definido pelo Ministério de Educação. “REGIAO” representa a região geográfica de cada escola dentro do país. “LOCALIZACAO” especifica se a escola está situada em uma área rural ou urbana. “ESFERA” indica a dependência administrativa da instituição, podendo ser estadual ou municipal. A variável “ALUNADO” representa o número de alunos matriculados na escola, para que seja possível mensurar seu tamanho. Essas características foram selecionadas para o estudo, uma vez que são fundamentais para compreender e comparar o perfil e funcionamento de diferentes escolas, e consequentemente podem impactar na escolha dos livros didáticos.

A variável “IDEB” é o Índice de Desenvolvimento da Educação Básica (IDEB), expresso em uma escala de 0 a 10. Esse índice tem como objetivo principal mensurar a qualidade da educação oferecida naquela unidade educacional, e é calculado considerando dois componentes principais: o desempenho dos alunos em avaliações padronizadas de língua portuguesa e matemática, e a taxa de aprovação escolar. A análise busca relacionar também esse dado com o perfil da obra didática, visando compreender possíveis correlações que possam influenciar o processo de aprendizado.

Para a base de dados utilizada no estudo considerou-se as escolas que participaram do programa do PNLD 2023 Obras Didáticas Objeto 1 que tiveram uma escolha fragmentada, ou seja, adotaram obras de forma independente, sem seguir a decisão do município. Além disso, os dados são referentes às escolhas apenas das coleções de matemática que foram categorizadas como Robusta ou Simples. O total de observações é de 7.061 escolas.

3.2 Pré-processamento

A primeira etapa consistiu na coleta e consolidação dos dados. Foi necessário realizar uma análise prévia dos resultados das adoções das coleções. Resumidamente, as redes de ensino têm a opção de adotar um modelo de escolha unificado, em que a decisão é tomada por votação e a coleção mais votada é utilizada por todas as escolas do grupo ou município definido, ou um modelo fragmentado, no qual cada escola recebe a coleção que escolheu (GECEB, 2024). Com base nesses resultados, foi realizada uma análise para identificar quais escolas seguiram quais

ISSN: 2675-4185

tendências de escolha. Conforme mencionado, apenas as escolas com escolha fragmentada foram consideradas no estudo, de modo que as características analisadas fossem individuais para cada instituição.

Em seguida, a partir do código MEC foi possível cruzar as informações fornecidas nas diferentes fontes já mencionadas e consolidar as informações em uma única base. Foi feita então uma análise descritiva das variáveis.

As etapas subsequentes visaram investigar as relações entre as características das escolas e o perfil das obras escolhidas. Para isso, foram aplicadas análises de correspondência simples e múltipla, além do modelo de Regressão Logística. Neste último, a variável "CLASSIFICACAO" foi considerada como variável dependente. Todas as análises foram realizadas por meio do Software R (CORE TEAM, 2024).

3.3 Análise de correspondência

A análise de correspondência simples (ANACOR) é uma metodologia estatística que analisa a associação e intensidade entre as categorias de duas variáveis categóricas. Essa é realizada através de uma tabela de contingência na qual as frequências absolutas são dispostas para cada par de categorias das variáveis em estudo (FÁVERO; BELFIORE, 2017).

Para análise são calculadas as frequências absolutas observadas e esperadas, os resíduos e os valores de χ^2. O teste do χ^2 possibilita analisar se a frequência das categorias de uma variável em relação às categorias da outra é aleatória ou se há uma associação significativa entre elas (FÁVERO; BELFIORE, 2017). Para o teste de associação do χ^2, foi considerado um nível de significância (p-valor) de 5% em que:

H0: as variáveis se associam de forma aleatória

H1: a associação entre as variáveis não se dá de forma aleatória.

Em sequência calculou-se os resíduos padronizados e os resíduos padronizados ajustados, que avaliam os padrões de cada categoria de uma variável com base na sua aparição ou ausência quando combinada com cada categoria da outra variável. Os resíduos padronizados ajustados foram utilizados para estipular a correspondência entre as categorias das variáveis. A um nível de significância de 5%, os valores superiores a 1,96 interpretam-se que há uma associação significativa entre as categorias (FÁVERO; BELFIORE, 2017).

Já na análise de correspondência múltipla (ACM) é possível avaliar a relação entre mais de duas variáveis e suas categorias, a partir de uma matriz binária. A matriz binária promove a

ISSN: 2675-4185

indicação da ocorrência (1) ou não (0) de dado evento para todas as combinações de categorias, gerando assim uma tabela de contingência. Cada linha da matriz representa uma dimensão, e possui um único valor 1, que será a inércia principal dessa dimensão. A partir das inércias, é possível calcular os autovalores e autovetores, e, por fim, as coordenadas que representam a relação entre as variáveis na ACM (FÁVERO; BELFIORE, 2017). Em resumo, as categorias das variáveis são representadas em um espaço multidimensional e a proximidade entre as categorias nesse espaço indica associações entre elas.

O número de dimensões é definido pela diferença do número de categorias totais e o número de variáveis categóricas. Cada dimensão terá uma proporção de variância explicada pelos seus pesos, que captura uma parte da variabilidade total dos dados, sendo as dimensões mais importantes aquelas com os maiores autovalores. Ao plotar o mapa perceptual as dimensões provenientes dos dois maiores autovalores são representadas nos eixos x e y, proporcionando uma visualização dos padrões de associação entre as categorias (EDISON BERTONCELO, 2022).

Nessa etapa, o primeiro passo consistiu em categorizar as variáveis quantitativas "IDEB" e "ALUNADO", a fim de viabilizar a análise de correspondência. A variável "IDEB" foi categorizada em "BAIXO", "MEDIO" e "ALTO", enquanto a variável "ALUNADO" foi dividida em "PEQUENA", "MEDIA" e "GRANDE". Para realizar essa transformação, foi utilizada a função "mutate" do pacote “dplyr” do R. A definição das categorias forma feitas a partir dos quartis do conjunto de dados. Essa classificação não representou o valor da informação em termos absolutos, mas sim em relação à distribuição dos dados. Em seguida, todas as variáveis qualitativas foram convertidas em factores.

Posteriormente, foi conduzida uma análise de correspondência simples entre a variável “CLASSIFICACAO” e todas as demais variáveis, com o intuito de avaliar se a associação entre elas era estatisticamente significativa. Além do mapa de calor dos resíduos padronizados ajustados para examinar as relações entre as categorias das variáveis.

Após verificar se havia associação estatisticamente significativa entre as variáveis foram feitas análises de correspondência múltipla, usando a função “dudi.acm” do pacote
“ade4”, e em seguida foram plotados os mapas perceptuais para visualizar as relações das categorias, consolidando as coordenadas-padrão obtidas por meio da matriz binária.

3.4 Regressão logística binária

ISSN: 2675-4185

A regressão logística binária é uma técnica estatística utilizada para estimar a probabilidade da ocorrência de um evento com base no comportamento de variáveis independentes. Para isso é utilizada a equação matemática conhecida por logito (Z), em que a variável dependente segue uma distribuição de Bernoulli (Figura 1). O objetivo é definir parâmetros que otimizam o desempenho do modelo, estimando-os por máxima verossimilhança (FÁVERO; BELFIORE, 2017).

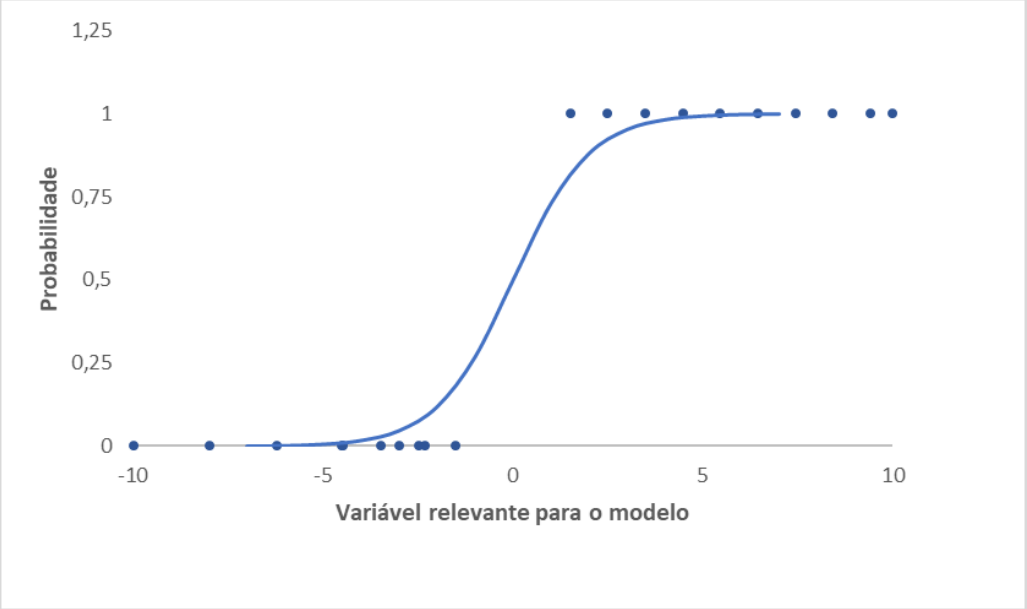


Figura 1. Gráfico função Regressão Logística P = f(Z).

Fonte: Autoria própria.

O logaritmo da função de verossimilhança, conhecido como log likelihood (LL), é uma métrica que buscamos maximizar, pois quanto maior a soma de seus valores, mais próximo de zero está a função, o que indica um melhor ajuste do modelo aos dados observados (FÁVERO; BELFIORE, 2017).

Para análise foram consideradas todas as variáveis categóricas. Como mencionado anteriormente, a variável “CLASSIFICACAO” foi definida como variável dependente. Nesse contexto a categoria “SIMPLES” foi considerada como evento e a “ROBUSTA” como não evento.

O processo inicial envolveu a conversão das variáveis qualitativas em variáveis dummy. Em seguida, foi aplicado o modelo utilizando a função glm, com o parâmetro family = "binomial", para ajustar o modelo de regressão logística.

ISSN: 2675-4185

Foi aplicado o procedimento stepwise, utilizando a função “step” do R. Esse procedimento exclui do modelo as variáveis que não se mostram significantes a um nível de significância de 95% (p-valor > 0,05), selecionando um conjunto que melhor explicam a variabilidade na variável de resposta (SICSÚ; SAMARTINI; BARTH, 2023).

Para avaliar a eficiência do modelo alguns parâmetros foram considerados. O LL é um deles, como explicado anteriormente, quanto maior seu valor, melhor o ajuste realizado ao modelo. Já os valores AIC (Equação 1) e BIC (Equação 2), sendo n é tamanho da amostra), quanto menores, indicam um melhor resultado de eficiência (SANTOS, 2017).

$$AIC = -2LL + 2p \qquad p=\text{número de parâmetros} \qquad (1)$$

$$BIC = -2LL + 2p \log(n) \qquad n=\text{número de observações} \qquad (2)$$

Outras métricas importantes são derivadas da matriz de confusão (Figura 2). Esta matriz resume e compara a classificação prevista pelo modelo para cada observação com a verdadeira ocorrência. O ponto de corte, ou cutoff, é valor predefinido (entre 0 e 1) que classifica as observações com base em suas probabilidades calculadas. Aplicando o modelo com o cutoff definido e classificando as observações é possível construir a matriz de confusão, na qual os valores são:

- Verdadeiro positivo (VP): evento classificado como evento (obra simples predita como obra simples)
- Falso negativo (FN): evento classificado como não evento (obra simples classificada como robusta)
- Falso positivo (FP): não evento classificado como evento (obra robusta classificada como simples)
- Verdadeiro negativo (VN): não evento classificado como não evento (obra robusta classificada como robusta).

		VALOR PREDITO	
		SIMPLES	ROBUSTO
REAL	SIMPLES	VP	FN
	ROBUSTO	FP	VN

Figura 2. Exemplificação da Matriz de confusão.

Fonte: Autoria Própria.

ISSN: 2675-4185

A partir desses critérios é possível gerar algumas métricas para avaliar a eficiência do modelo. A acurácia (EGM) (Equação 3) é a taxa global de acerto do modelo. A sensitividade (recall) (Equação 4) é a assertividade para o que foi previsto como evento. A especificidade (Equação 5) é a taxa de acerto para o que não foi evento. A precisão (Equação 6) é a taxa de acerto para o que foi classificado como evento.

$$Acurácia = \frac{VP+VN}{VP+VN+FP+FN} \quad (3)$$

$$Sensitividade = \frac{VP}{VP+FN} \quad (4)$$

$$Especificidade = \frac{VN}{VN+FP} \quad (5)$$

$$Precisão = \frac{VP}{VP+FP} \quad (6)$$

Outra métrica de avaliação resulta da curva ROC (Receiver Operating Characteristic), que é uma representação gráfica que mostra o desempenho do modelo de classificação em diferentes pontos de corte. Ela traça a taxa de verdadeiros positivos (sensitividade) em relação à taxa de falsos positivos (1 - especificidade) para todos os possíveis valores de corte. Assim é possível avaliar a eficácia do modelo, independente do cutoff. A proximidade da curva do canto superior esquerdo (coordenadas 1,1) indica melhor desempenho do modelo. O parâmetro AUC é a área abaixo da curva. Ou seja, quanto mais próximo de 1, melhor o desempenho do modelo (GIOLO, 2017).

Para a criação da matriz de confusão foi utilizada a função “confusionMatrix” e a curva ROC a partir da função “roc”.

Também é possível gerar insights analisando os coeficientes calculados para o modelo a partir do conceito odds ratio (OR). O OR indica a força da associação entre a variável preditora e a variável resposta, revelando quantas vezes mais provável é que o evento ocorra em uma categoria em relação a outra, e é determinado pela exponencial do coeficiente (FERNANDES; FIGUEIREDO FILHO; ROCHA; NASCIMENTO, 2020).

4 RESULTADOS E DISCUSSÕES

4.1 Análises descritivas

Os primeiros resultados observados foram as análises descritivas extraídas do banco de dados. Em relação às variáveis quantitativas foram construídos, para cada categoria de perfil

ISSN: 2675-4185

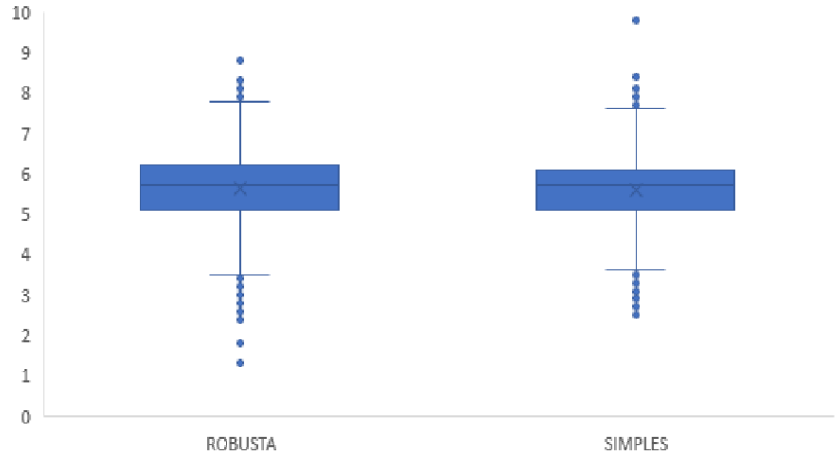
de obra (simples e robusta), um bloxplot relacionado às notas do IDEB das escolas e um relacionado ao número de alunos, representados na Figura 3. As métricas também foram demonstradas na Tabela 2.

Tabela 2. Estatísticas variáveis quantitativas.

	IDEB		ALUNADO	
	ROBUSTA	SIMPLES	ROBUSTA	SIMPLES
Mínimo	1,30	2,50	10,00	14,00
1° Quartil	5,10	5,10	113,00	137,00
Mediana	5,70	5,70	202,00	252,00
Média	5,62	5,60	248,37	284,56
3° Quartil	6,20	6,10	328,25	387,00
Máximo	8,80	9,80	1739,00	1467,00

Fonte: Dados originais da pesquisa.

A - Bloxplot - Notas IDEB de escolas que adotaram obras simples ou robusta



B - Bloxplot - Número de alunos de escolas que adotaram obras simples ou robusta

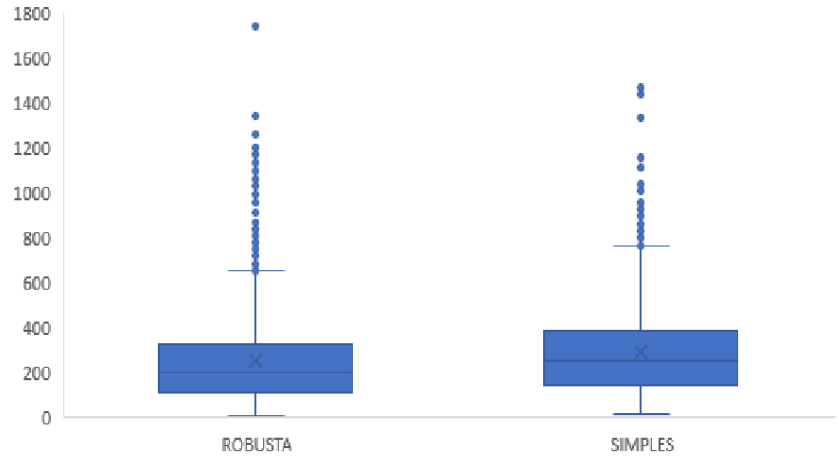


Figura 3. Bloxplot variáveis quantitativas.

ISSN: 2675-4185

Fonte: Dados originais da pesquisa.

Na figura 3A pôde-se notar que a distribuição da categoria robusta apresentou maior variabilidade nas notas do IDEB, com uma concentração maior entre o intervalo de notas 5,1 e 6,2. Já entre as escolas que adotaram obras simples a distribuição foi mais simétrica, apresentando uma leve tendência de escolas com IDEB mais alto preferirem. No entanto, seria esperado que as escolas com IDEB mais alto optassem pelas obras robustas, que geralmente oferecem conteúdo mais aprofundado. Uma possível explicação para esse resultado seria o ponto máximo muito maior.

Na figura 3B foi possível observar uma distribuição do número de alunos por escola, das escolas que aderiram obras robustas, positivamente assimétrica. A distribuição das escolas com os livros simples, teve uma distribuição mais simétrica quando comparada. Ambas apresentaram outliers, e no caso da categoria simples, a variabilidade foi maior.

Vale ressaltar que os outliers podem ter diversas origens, além das que estão sendo levadas em consideração nos gráficos. Como diferenças na qualidade de ensino, infraestrutura, condições socioeconômicas dos alunos, localização da escola, tamanho do município, perfil dos alunos, entre outros fatores.

Para as variáveis categóricas foram plotados gráficos (Figura 4) para visualizar a distribuição das categorias. A base de dados incluiu 7061 escolas participantes do programa, cada uma escolhendo uma coleção de matemática. Quanto à classificação do perfil das obras escolhidas, mais de 65% optaram pelas coleções robustas, o restante pelas obras simples (Figura 4A). Quanto às regiões geográficas do Brasil, quase 43% das escolas estão localizadas no Sudeste, sendo a região Centro-Oeste a menos abrangente (Figura 4B). Referente a localização, 84% das instituições encontram-se em áreas urbanas, enquanto 16% estão em áreas rurais (Figura 4C). Em relação a dependência administrativa, quase 70% são escolas municipais, e o restante estaduais (Figura 4D). Esses resultados evidenciam a diversidade e distribuição das características das escolas participantes do programa.

ISSN: 2675-4185

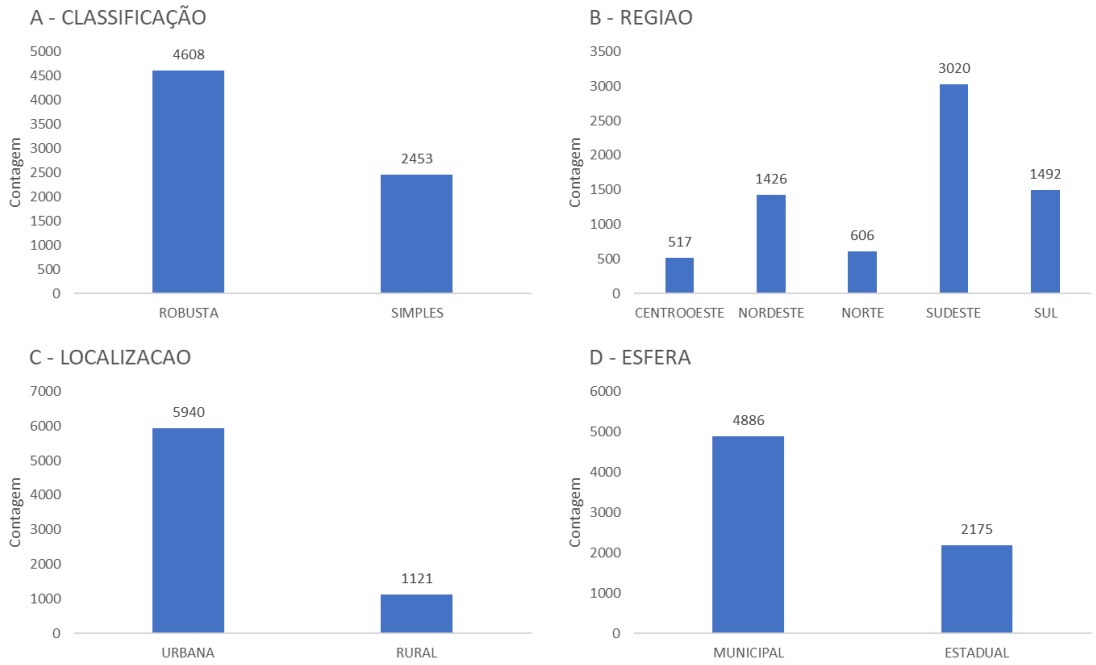


Figura 4. Análise descritiva do conjunto de dados.

Fonte: Dados originais da pesquisa.

4.2 Análises de correspondência simples entre a variável dependente e as variáveis categóricas

Os resultados das análises de correspondência simples foram os seguintes:

"CLASSIFICACAO" x "REGIAO":

Tabela 3. Estatísticas: "CLASSIFICACAO" x "REGIAO".

χ^2 Total	65,27
p-valor	< 0,01

Fonte: Dados originais da pesquisa.

Como descrito na Tabela 3, como p-valor < 0,05 considerou-se H1 como verdadeiro, ou seja, a associação das variáveis "CLASSIFICACAO" e "REGIAO" não se dá de forma aleatória.

Tabela 4. Tabela de correlação "CLASSIFICACAO" x "REGIAO".

		CLASSIFICACAO	
		SIMPLES	**ROBUSTA**
REGIAO	**Centro-Oeste**	3,20	-3,20
	Nordeste	-0,58	0,58
	Norte	0,31	-0,31
	Sudeste	4,69	-4,69

ISSN: 2675-4185

	Sul	-7,37	7,37

Fonte: Dados originais da pesquisa.

Pelo resultado apresentado na Tabela 4 foi possível identificar uma associação estatisticamente significativa entre as categorias região "Sul" e coleções classificadas como "Robusta", as regiões "Sudeste" e "CentroOeste" e as coleções classificadas como "Simples". Ou seja, indica que há uma tendência das escolas da região Sul do país a adotarem obras com perfil robusto, e escolas das regiões Sudeste e Centre-Oeste têm uma tendência de adotarem livros com perfil Simples.

"CLASSIFICACAO" x "LOCALIZACAO":

Tabela 5. Estatísticas: "CLASSIFICACAO" x "LOCALIZACAO".

χ^2 Total	9,24
p-valor	0,002

Fonte: Dados originais da pesquisa.

Como observado na Tabela 5 o p-valor < 0,05 considerou-se H1 como verdadeiro, ou seja, a associação das variáveis "CLASSIFICACAO" e "LOCALIZACAO" não se dá de forma aleatória.

Tabela 6. Tabela de correlação "CLASSIFICACAO" x "LOCALIZACAO".

		CLASSIFICACAO	
		SIMPLES	**ROBUSTA**
LOCALIZACAO	**RURAL**	-3,04	3,04
	URBANA	3,04	-3,04

Fonte: Dados originais da pesquisa.

Pelo resultado apresentado na Tabela 6 foi possível identificar uma tendência de escolha das escolas rurais em livros mais robustos e escolas urbanas em coleções mais simples.

"CLASSIFICACAO" x "ESFERA":

Tabela 7. Estatísticas: "CALSSIFICACAO" X "ESFERA".

χ^2 Total	29,66
p-valor	< 0,01

Fonte: Dados originais da pesquisa.

Com o p-valor < 0,05 (Tabela 7) considerou-se H1 como verdadeiro, ou seja, a associação das variáveis "CLASSIFICACAO" e "ESFERA" não se dá de forma aleatória.

ISSN: 2675-4185

Tabela 8. Tabela de correlação “CALSSIFICACAO” X “ESFERA”.

		CLASSIFICACAO	
		SIMPLES	**ROBUSTA**
ESFERA	**MUNICIPAL**	5,45	-5,45
	ESTADUAL	-5,45	5,45

Fonte: Dados originais da pesquisa.

Pelo resultado apresentado na Tabela 8 foi possível identificar uma relação significativa entre as escolas municipais e a escolha de coleções simples e as escolas estaduais e as obras mais robustas.

“CLASSIFICACAO” x “IDEB”:

Tabela 9. Estatísticas: “CLASSIFICACAO” x “IDEB”.

χ^2 Total	10,04
p-valor	0,007

Fonte: Dados originais da pesquisa.

Na Tabela 9 podemos observar também um p-valor < 0,05, o que significa que existe uma associação significativa entre as categorias das variáveis “CLASSIFICACAO” e “IDEB”.

Tabela 10. Tabela de correlação “CALSSIFAICACAO” x “IDEB”.

		CLASSIFICACAO	
		SIMPLES	**ROBUSTA**
IDEB	**ALTO**	-2,71	2,71
	MEDIO	2,85	-2,85
	BAIXO	-0,65	0,65

Fonte: Dados originais da pesquisa.

A Tabela 10 mostra que há uma associação estatisticamente significativa entre a utilização de coleções robustas e o alcance de IDEB alto, bem como entre a utilização de coleções simples e o alcance de IDEB médio.

“CLASSIFICACAO” x “ALUNADO”:

Tabela 11. Estatísticas: “CLASSIFICACAO” x “ALUNADO”.

χ^2 Total	78,08
p-valor	< 0,01

Fonte: Dados originais da pesquisa.

Entre as variáveis “CLASSIFICACAO” e “ALUNADO” a associação não acontece de forma aleatória, tendo em vista o p-valor < 0,05 observado na Tabela 11, rejeitando-se H0.

ISSN: 2675-4185

Tabela 12. Tabela de correlação “CLASSIFICACAO” x “ALUNADO”.

		CLASSIFICACAO	
		SIMPLES	**ROBUSTA**
ALUNADO	**GRANDE**	7,49	-7,49
	MEDIA	-0,49	-0,49
	PEQUENA	-6,91	6,91

Fonte: Dados originais da pesquisa.

Sobre as categorias das variáveis classificação e alunado é observado, na Tabela 12, que existe uma relação significativa entre “simples” e “grande”, bem como “robusta” e “pequena”.

4.3 Análises de correspondência múltipla entre a variável dependente e as variáveis categóricas

Após realizar uma análise de correspondência simples entre todas as variáveis em relação à variável "CLASSIFICACAO", observou-se que todas as variáveis apresentaram um p-valor inferior a 0,05. Esse resultado sugere uma associação estatisticamente significativa entre as variáveis, indicando que a relação entre elas não ocorre aleatoriamente. Dessa forma, conclui-se que todas as variáveis estão de alguma forma relacionadas com a classificação das obras. Portanto, não há necessidade de descartar nenhuma variável para conduzir uma análise de correspondência múltipla, que examina as inter-relações entre diversas variáveis simultaneamente.

Em uma análise com todas as 6 variáveis, em que no total existem 17 categorias, teremos 11 dimensões, as quais apresentaram variâncias variando entre 18,52% (maior) e 3,83% (menor). Na plotagem do mapa perceptual, em que apenas as 2 dimensões com maiores autovalores serão representadas nos eixos X e Y, a variabilidade total foi de 31,07%.

ISSN: 2675-4185

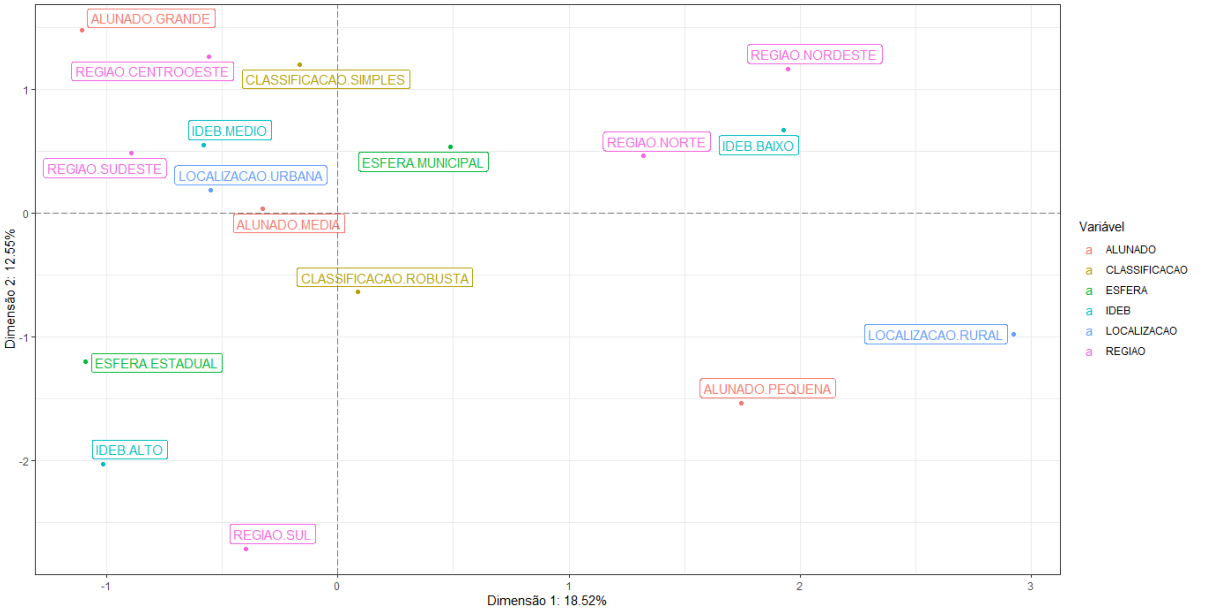


Figura 5. Mapa perceptual resultante da ACM.

Fonte: Dados originais da pesquisa.

A partir do mapa (Figura 5) é possível tirar algumas informações importantes. Pensando que a variável principal é a "CLASSIFICACAO" foi possível observar que:

A categoria "Robusta" apresentou relação positiva com as categorias LOCALIZACAO-RURAL e ALUNADO-PEQUENA em ambas as coordenadas, reforçando a inter-relação entre esses elementos em diferentes dimensões. Essas correlações também foram significativas na análise de correspondência simples (Tabelas 6 e 12). Além disso, foi observada também, uma relação positiva com as categorias ESFERA-ESTADUAL, IDEB -ALTO e REGIAO-SUL na coordenada Y.

Observando a categoria "Simples", identificou-se uma relação positiva com as categorias "REGIAO-CENTROOESTE", "REGIAO-SUDESTE", "LOCALIZACAO-URBANA", "IDEB-MEDIO" e "ALUNADO-GRANDE" em ambas as dimensões. Essas correlações também foram significativas na análise de correspondência simples (Tabela 4, 6, 10 e 12). Na dimensão Y, a relação foi positiva com as categorias "ESFERA-MUNICIPAL", "REGIAO-NORTE", "REGIAO-NORDESTE" e "IDEB-BAIXO".

Vale ressaltar que as dimensões plotadas apresentam apenas 31,07% da variabilidade, sendo assim, o mapa pode auxiliar na identificação de padrões e tendências gerais nos dados, mas não reflete todo o contexto.

Foi encontrado em alguns trabalhos análises das relações de algumas variáveis que foram estudadas. Colucci (2014) analisou o desempenho de algumas escolas municipais a partir da

ISSN: 2675-4185

nota do IDEB, relacionando com características como tamanho do município, investimento médio do governo por aluno e municipalização (razão do número de escolas públicas municipais pela soma do número de escolas públicas municipais e estaduais). Em uma das análises o resultado obtido foi de que as escolas das regiões sul e sudeste apresentaram um rendimento melhor, como justificativa de que são regiões socioeconomicamente mais desenvolvidas e apresentam maior média de investimento. Quanto à municipalização, foi um fator não explicativo para essa relação.

Caprara (2020) realizou um estudo sobre as condições de classe e o desempenho educacional. Em análises de regressão linear múltipla, os fatores semelhantes aos avaliados no presente estudo que tiveram maior relevância foram, dependência administrativa, que abrangia também as escolas particulares e públicas federais, e a localização (urbana ou rural). A região geográfica não apresentou tanta influência. Outros critérios que tiveram grande impacto foram variáveis que indicam posse de capitais e classe social.

Nas análises de correspondência múltipla de Caprara (2020) as notas altas do IDEB tiveram relação significativa com escolas de rede privada e federal, as notas médias com as instituições da rede estadual, e notas muito baixa com escolas de rede municipal, com uma variabilidade das dimensões de 83,6%.

Tanto Caprara (2020) quanto Colucci (2014) identificaram variáveis socioeconômicas relevantes para compreender o desempenho educacional nas escolas. Algumas dessas variáveis também poderiam ser utilizadas para examinar a conexão entre o perfil das obras didáticas e os critérios de escolha adotados pelas instituições, como os recursos disponíveis em sala de aula ou o investimento médio por aluno.

4.4 Modelo de regressão logística binária

Gerando o modelo logístico obteve-se os parâmetros descritos na (Tabela 13). As variáveis “LOCALIZACAO” e “IDEB” não se mostraram estatisticamente significante, assim como as categorias “REGIAO-CENTROOESTE” e “REGIAO-SUDESTE”, essas foram excluídas do modelo em que se aplicou o processo de stepwise (Tabela 14).

Tabela 13. Resultado Modelo Logístico

	Coeficiente	Erro Padrão	p-valor
Intercept	-0,6685	0,1863	<0,001
LOCALIZACAO-URBANA	0,1116	0,0798	>0,05

ISSN: 2675-4185

ESFERA-ESTADUAL	-0,3528	0,0589	<0,001
IDEB	-0,0186	0,0354	>0,05
ALUNADO	0,0008	0,0001	<0,001
REGIAO-NORDESTE	-0,1217	0,1037	>0,05
REGIAO-CENTROOESTE	0,1904	0,1267	>0,05
REGIAOSUDESTE	0,063	0,0988	>0,05
REGIAOSUL	-0,3567	0,1123	<0,01
AIC	8997,8297		
BIC	9059,5908		

Fonte: Dados originais da pesquisa.

Tabela 14. Resultado Modelo Logístico Stepwise.

	Coeficiente	Erro Padrão	p-valor	OR
Intercept	-0,6283	0,0559	<0,001	0,533497977
ESFERA-ESTADUAL	-0,3503	0,0579	<0,001	0,704476715
ALUNADO	0,0009	0,0001	<0,001	1,000900405
REGIAO-NORDESTE	-0,1942	0,0679	<0,01	0,823493189
REGIAOSUL	-0,4284	0,0686	<0,001	0,651550742
AIC	8994,2895			
BIC	9028,6012			

Fonte: Dados originais da pesquisa.

Como já mencionado, a aplicação do processo stepwise tem o objetivo de melhorar a eficiência do modelo. Ao aplicar esse método, observou-se uma redução significativa nos valores de AIC e BIC. O modelo resultante apresentou um AIC de 8994,28 e um BIC de 9028,60, em comparação com o modelo sem a aplicação do método, que registrou valores de AIC 8997,83 e BIC 9059,59, respectivamente. Essa diminuição nos critérios sugere uma melhoria na qualidade do modelo após a implementação do processo.

Com a definição dos parâmetros após aplicação do procedimento de stepwise, a Equação (5) foi construída para o modelo de regressão do estudo, para probabilidade de CLASSIFICACAO = “SIMPLES”:

$$Log\left(\frac{\hat{P}}{1-\hat{P}}\right) = -0{,}6283 - 0{,}3503(ESFERA - ESTADUAL) + 0{,}0009(ALUNADO) - 0{,}1942(REGIAO - NORDESTE) - 0{,}4284(REGIAO - SUL) \quad (5)$$

A partir dos coeficientes pode-se observar os efeitos das variáveis na probabilidade de ocorrência do evento. As categorias “ESFERA-ESTADUAL”, “REGIÃO-NORDESTE” e “REGIÃO-SUL” apresentaram um sinal negativo, o que indica que quando for atribuída a presença dessas variáveis ocorre uma diminuição na probabilidade do evento em 0,3503,

ISSN: 2675-4185

0,1942 e 0,4284, respectivamente. Em outras palavras, a probabilidade de uma escola escolher uma obra simples, sendo de esfera estadual, ou localizada nas regiões nordeste ou sul do país, é menor em comparação às outras condições. Já o coeficiente da variável “ALUNADO” foi positivo, indicando que a probabilidade da ocorrência do evento aumenta em 0,0009 para cada unidade de aluno. Essas relações entre as categorias também foram notadas nos resultados de análise de correspondência.

Com as razões de chance apresentadas (OR), observa-se que a chance de uma escola estadual escolher uma obra simples é de 29,55% menor do que uma escola municipal. Em relação a variável “REGIÃO”, uma escola na região Sul tem 34,84% menos chance de escolher uma obra simples do que as regiões com coeficiente 0 (Sudeste, Centro-Oeste e Norte), enquanto no Nordeste a chance é 17,65% menor. As três categorias analisadas ("ESFERA-ESTADUAL", "REGIAO-SUL" e "REGIÃO-NORDESTE") apresentaram um OR menor que 1, indicando que sua presença diminui a probabilidade de escolha de obras simples. Já a variável “ALUNADO”, com OR maior que 1, aumenta a chance em 0,09% para cada 1 aluno. A partir do modelo foram construídas duas matrizes de confusão, uma com cutoff 0,50 (Tabela 15) e outra com cutoff 0,35 (Tabela 16). As estatísticas do resultado do estão descritas na Tabela 17.

Tabela 15. Matriz de confusão cutoff 0,50.

	SIMPLES	ROBUSTO
SIMPLES	29	45
ROBUSTO	2424	4563

Fonte: Dados originais da pesquisa.

Tabela 16. Matriz de confusão cutff 0,35.

	SIMPLES	ROBUSTO
SIMPLES	1415	2105
ROBUSTO	1038	2503

Fonte: Dados originais da pesquisa.

Tabela 17. Estatísticas resultantes das matrizes de confusão de cutoff 0,5 e 0,35.

Cutoff	0,5	0,35
Acurácia	0,6503	0,5549
Sensitividade	0,0118	0,5768
Especificidade	0,9902	0,5432
p-value	<2e-16	<2e-16

ISSN: 2675-4185

Fonte: Dados originais da pesquisa.

Com o cutoff 0,5 o modelo teve uma acurácia melhor, de 65,03%, do que aplicando um cutoff 0,35 que foi de 55,49%. Porém, é importante notar que a sensitividade foi extremamente baixa no primeiro caso, de apenas 1,18%, enquanto a especificidade foi de 99,02%. Isso indica uma falha significativa do modelo em identificar casos positivos, gerando muitos falsos negativos. Ou seja, muitas escolas que optaram pela obra simples, pelo modelo, foram erroneamente classificadas com as obras robustas. Já no segundo houve um equilíbrio mais adequado entre a sensitividade e a especificidade, com valores de 57,68% e 54,32%, respectivamente. Isso sugere que o modelo foi capaz de identificar uma proporção maior dos casos positivos sem comprometer drasticamente a identificação dos casos negativos.

Para avaliar o desempenho do modelo, independentemente do cutoff, foi gerada a curva ROC (Figura 6). A eficiência tende a ser maior quando mais próxima de 1 for a área da curva. Nesse caso, obteve-se um valor de 58,49%, o que indica que o modelo não tem um alto poder de discriminação, mas também não chega a ter um desempenho aleatório.

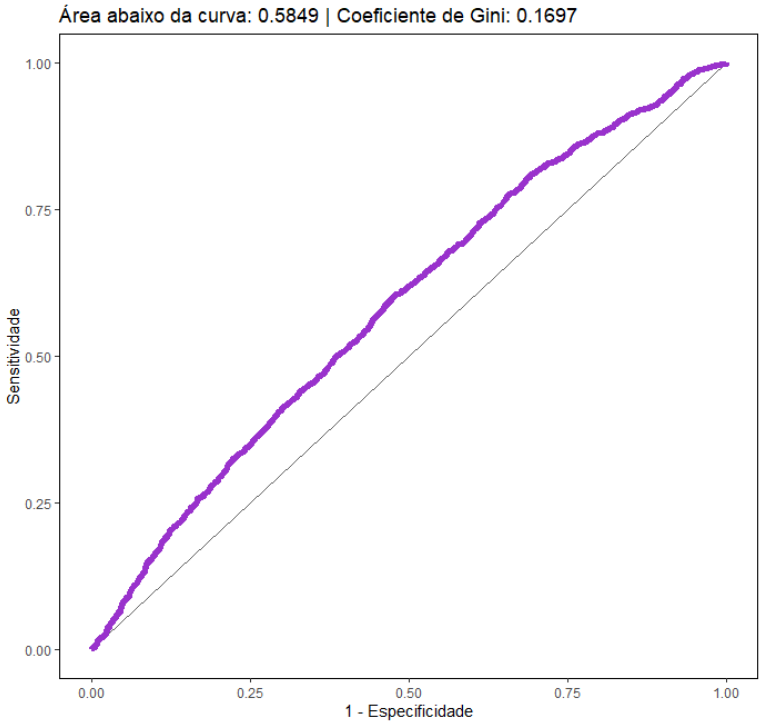


Figura 6. Curva ROC Modelo de Classificação das Obras com aplicação de stepwise.

Fonte: Dados originais da pesquisa.

5 CONSIDERAÇÕES FINAIS

ISSN: 2675-4185

A partir das análises de correspondência realizadas foi possível identificar uma relação estatisticamente significativa entre as obras simples com as categorias “ALUNADO-GRANDE”, “REGIÃO-SUDESTE”, “REGIAO-CENTROOESTE”, “ESFERA-MUNICIPAL”, “LOCALIZACAO-URBANA” e “IDEB-MEDIO”. Já as obras robustas apresentaram relações com as categorias “ALUNADO-PEQUENA”, “REGIAO-SUL”, “ESFERA-ESTADUAL”, “LOCALIZACAO-RURAL”, “IDEB-ALTO”. Na regressão logística o coeficiente da categoria “REGIAO-SUL” apresentou maior relevância para a não ocorrência do evento (não escolha da obra simples), seguido da categoria “ESFERA-ESTADUAL” e “REGIAO-NORDESTE”, com pesos negativos. O número de alunos teve um coeficiente positivo para a ocorrência do evento. Embora o modelo logístico construído tenha demonstrado um resultado satisfatório, há margem para aprimoramento. Uma das limitações do estudo foi a ausência de significância estatística de algumas das variáveis analisadas. A inserção de novas variáveis, que não sejam excluídas do procedimento stepwise, pode potencializar o desempenho do modelo. Além disso, a expansão do banco de dados, possivelmente de outros anos programa, pode enriquecer as análises e fortalecer ainda mais o modelo logístico. Essas características podem ser melhor exploradas para estratégias de divulgação das editoras no PNLD, que podem direcionar seus esforços para obras com maior probabilidade de serem selecionadas.

REFERÊNCIAS

ABRELIVROS. PNLD - Valores de aquisição. Disponível em: https://www.abrelivros.org.br/anuario/pnld-valores-de-aquisicao.html. Acesso em: 08 mar. 2024.

CAPRARA, Bernardo Mattes. Condição de Classe e Desempenho Educacional no Brasil. Educação & Realidade, Porto Alegre, v. 45, n. 4, p. 0-28, dez. 2020. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/2175-623693008.

COLUCCI, Lucas. O impacto na ponderação do peso da Prova Brasil e do indicador de rendimento no perfil das escolas municipais do ensino fundamental consideradas eficientes pela técnica DEA em transformar investimento financeiro em desempenho no IDEB em 2011. 2014. 129 f. Dissertação (Mestrado) - Curso de Administração de Organizações, Universidade de São Paulo, Ribeirão Preto, 2014.

EDISON BERTONCELO. Construindo espaços relacionais com a análise de correspondências múltiplas: aplicações nas ciências sociais. Brasília: Enap, 2022

ISSN: 2675-4185

FÁVERO, Luiz Paulo; BELFIORE, Patrícia. MANUAL DE ANÁLISE DE DADOS Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®. Rio de Janeiro: Editora GEN LTC, 2017.

FERNANDES, Antônio Alves Tôrres; FIGUEIREDO FILHO, Dalson Britto; ROCHA, Enivaldo Carvalho da; NASCIMENTO, Willber da Silva. Leia este artigo se você quiser aprender regressão logística. Revista de Sociologia e Política, Recife, v. 28, n. 74, p. 0-20, maio 2020. Disponível em: https://doi.org/10.1590/1678-987320287406en. Acesso em: 20 mar. 2024.

FNDE - Fundo Nacional de Desenvolvimento da Educação. Dados estatísticos. Disponível em: https://www.fnde.gov.br/index.php/programas/programas-do-livro/pnld/dados-estatisticos. Acesso em: 26 mar. 2023.

FNDE - Fundo Nacional de Desenvolvimento da Educação. Programa do Livro Disponível em: https://www.gov.br/fnde/pt-br/acesso-a-informacao/acoes-e-programas/programas/programas-do-livro.. Acesso em: 25 mar. 2023.

FNDE. SIMAD - Sistema do Material Didático. Disponível em: https://www.fnde.gov.br/distribuicaosimadnet. Acesso em: 26 mar. 2023.

FREISLEBEN, Alcimar Paulo; KAERCHER, Nestor André. O PNLD E O MERCADO DE LIVROS DIDÁTICOS NO BRASIL. Ciência Geográfica, Bauru, v. 26, n. 26, p. 391-404, jan. 2022.

GECEB - GERÊNCIA DE CURRÍCULO E EDUCAÇÃO BÁSICA (org.). Programa Nacional do Livro e do Material Didático (PNLD). Disponível em: https://curriculo.sedu.es.gov.br/curriculo/livrodidatico/. Acesso em: 20 fev. 2024

GIOLO, S. R. Introdução à Análise de Dados Categóricos com Aplicações. Editora: Blucher - Projeto Fisher ABE. 2017.

MEC - Ministério da Educação. PNLD. Disponível em: http://portal.mec.gov.br/busca-geral/318-programas-e-acoes-1921564125/pnld-439702797/12391-pnld#:~:text=O%20Programa%20Nacional%20do%20Livro,redes%20federal%2C%20estaduais%2C%20municipais%20e. Acesso em: 25 mar. 2023.

GOV.BR. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira | Inep. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar/resultados/2022. Acesso em: 26 mar. 2023.

INEP. CENSO ESCOLAR. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar/resultados. Acesso em: 10 mar. 2024.

INEP. Ideb - Resultados. Disponível em: https//www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados. Acesso em: 26 mar. 2023.

ISSN: 2675-4185

LIMA, Victória da Silva. Análise de regressão logística aplicada à educação online durante a pandemia da COVID-19. 2020. 45 f. Tese (Doutorado) - Curso de Estatística, Universidade de Brasília, Brasilia, 2021.

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

SANTOS, Damião Flávio. Modelo de regressão log-logístico discreto com fração de cura para dados de sobrevivência. Dissertação (Mestrado) — Universidade de Brasília, Brasília-DF, Brasil, 2017.

SICSÚ, Abraham Laredo (org.); SAMARTINI, André; BARTH, Nelson Lerner. TÉCNICAS DE MACHINE LEARNING. São Paulo: Blucher, 2023.

VIDAL, Cynthia dos Santos. O PROCESSO DE ESCOLHA DOS LIVROS DIDÁTICOS, NUMA ESCOLA PÚBLICA. 2016. 63 f. TCC (Graduação) - Curso de Licenciatura em Matemática, Departamento Acadêmico de Matemática, Universidade Tecnológica Federal do Paraná, Curitiba, 2016.