Modelo de regressão logística aplicado na classificação do desempenho de alunos do 5º ano do ensino fundamental de escolas estaduais do Espírito Santo no Saeb

Logistic regression model applied to school performance classification of students from the 5th grade of primary education in SAEB in state schools in Espírito Santo

Thiago de Andrade Guedes Tecnólogo em Logística. Instituto Federal do Espírito Santo (Ifes) – Brasil. thiagoandradeoficial@gmail.com.

Katarina Rosa Lemos Tecnóloga em Logística. Instituto Federal do Espírito Santo (Ifes) – Brasil. katarinarosalemos@gmail.com.

Adonai José Lacruz Doutor em Administração. Instituto Federal do Espírito Santo (Ifes) e Programa de Pós-graduação em Administração da Universidade Federal do Espírito Santo (PPGAdm/UFES) – Brasil. adonai.lacruz@ifes.edu.br.

RESUMO

Este artigo tecnológico identifica, por meio de regressão logística, os indicadores educacionais que melhor diferenciam as notas de alunos do 5º ano do ensino fundamental de escolas estaduais do estado do Espírito Santo nas provas de Língua Portuguesa e de Matemática no Sistema de Avaliação da Educação Básica (Saeb) de 2017. A amostra foi composta 364 escolas estaduais do estado do Espírito Santo, classificadas como de melhor (aquelas com as 25% maiores notas) e de pior desempenho (aquelas com as 24% menores notas). Os resultados revelam que o conjunto ótimo de variáveis que melhor discrimina o desempenho em relação à língua portuguesa é composto pelas variáveis Nível socioeconômico dos discentes, Indicador de regularidade do corpo docente e Indicador de complexidade de gestão da escola. Já em relação à disciplina de matemática, o mesmo conjunto de variáveis, acrescido da variável Alunos por turma, se mostrou mais relevante. Dessa forma, os resultados podem ser vistos como um auxílio para as decisões de gestores públicos no que diz respeito a investimentos com base nos indicadores educacionais considerados prioritários, reforçando a importância do contexto escolar no desempenho dos alunos.

Palavras-chave: Regressão Logística. Gestão Escolar. Desempenho Escolar. Indicadores Educacionais. Políticas Educacionais.

ABSTRACT

This technological paper identifies the educational indicators that best differentiate school performance obtained by students from the 5th grade of primary education in state schools of Espírito Santo in the Math test and Portuguese Language test in the Saeb 2017 through logistic regression. The sample consisted of 364 schools of the public education network of Espírito Santo (Brazil), classified as the best performance and worst performance. The results showed that the Socioeconomic level of students, the Index of teacher regularity and the Indicator of school management complexity formed an optimal set of variables to discriminate the performance of schools in the Portuguese Language test. This same set of variables, and the variable Students per class, proved to be more relevant to evaluate the performance of schools in the Math test. These results can contribute to educational public policy decisions, reorganizing their investments based on priority educational indicators. The findings reinforce the importance of the school environment in school performance.

Keywords: Logistic Regression. Educational Management. School Performance. Educational indicators. Educational Policies.

Recebido em 13/10/2020. Aprovado em 25/11/2020. Avaliado pelo sistema *double blind peer review*. Publicado conforme normas da ABNT.
https://doi.org/10.22279/navus.2021.v11.p01-18.1444

1 INTRODUÇÃO

É latente a necessidade de atenção que o sistema educacional brasileiro requer. Dados do Índice de Desenvolvimento da Educação Básica (Ideb) de 2017, publicados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), revelam que apenas 30% das escolas estaduais do ensino fundamental alcançaram a meta de aprendizado em avaliações de larga escala de Língua Portuguesa e Matemática naquele ano. Alguns aspectos, agrupados como indicadores educacionais elencados pelo Inep, podem influenciar esses resultados (AMÉRICO; LACRUZ, 2017).

Trabalhos anteriores buscam mecanismos que possam contribuir para as decisões das autoridades no sistema educacional ao identificar com maior minúcia seus aspectos de carência institucional (LACRUZ; AMÉRICO; CARNIEL, 2019; LAUTERT; LAUTERT; ROLIM; LODER, 2011; CAMPELLO; LINS, 2008; NORONHA; CARVALHO; SANTOS, 2001).

Por se tratar de uma pesquisa relacionada aos estudos organizacionais, privilegiaram-se neste artigo tecnológico (MOTTA, 2017) os dados relativos à organização escolar; isto é, os indicadores educacionais (i.e. contextuais), a fim de contribuir para a prática gerencial, mais especificamente da administração pública. Nesse sentido, esta pesquisa busca explicar como e em qual grau se relacionam os indicadores educacionais com a chance de as escolas de ensino fundamental da rede pública estadual do Espírito Santo, considerando as séries iniciais, pertencerem ao grupo daquelas com maior e menor desempenho segundo o critério de corte aqui estabelecido.

A equipe gestora de uma escola é aquela determinada a organizar o ambiente escolar afim de assegurar os meios de alcançar os objetivos pedagógicos e as metas educacionais. Quando são obtidos dados processados e seus resultados avaliados por essa equipe, são postas em perspectiva as possibilidades de desfazer dificuldades e criar métodos de alcançar melhor desempenho. As avaliações educacionais de larga escala são tratadas nesta pesquisa como fonte de informações para construção de um modelo que permita fundamentar a tomada de decisões da liderança escolar no sentido de otimização do desempenho. No mesmo sentido, a construção deste modelo pode auxiliar os formadores de políticas públicas de educação. Dentro do contexto discutido, a secretaria de educação estadual, por exemplo, pode visualizar os indicadores que mais impactam os resultados e, assim, desenvolver algum nexo na realocação de recursos a fim de sanar ou amenizar os impactos negativos, já que o objetivo é ter melhor desempenho.

Por isso, para construir o modelo que viabiliza a análise dessa influência, esta pesquisa utiliza a técnica estatística Regressão Logística que proporciona uma análise multivariada adequada para as situações como a descrita, em que a variável dependente é categórica e assume resultados binários (FÁVERO *et al.*, 2009). A regressão logística permite a criação de um modelo matemático que estabeleça a probabilidade de uma observação pertencer a um determinado grupo em função do comportamento de um conjunto de variáveis independentes.

Em busca da estimativa da possível influência dos indicadores conjunturais das escolas no desempenho no Sistema de Avaliação da Educação Básica (Saeb), foram utilizados como variáveis independentes os indicadores de contexto das escolas que consideram a quantidade de alunos por turma (ATU), as horas-aula diárias (HAD), o indicador de regularidade do corpo docente (MIRD), o indicador de esforço docente (IED), o indicador de complexidade da gestão da escola (ICG), o nível socioeconômico dos discentes (NSE) e a adequação de formação docente (AFD). Como variável dependente foi utilizado, então, o grupo de escolas formado de acordo com a nota da prova (GNP).

Como esclarecem Américo e Lacruz (2017), o Saeb avalia estudantes de áreas rurais e urbanas dos anos iniciais e finais do ensino fundamental público em português e matemática, de forma censitária, a cada dois anos, inserindo-se no discursivamente entre as avaliações nacionais aplicadas em larga escala para mensurar a qualidade na educação. O Saeb pode ser entendido "[...] como uma estratégia que permite ao governo brasileiro avaliar a qualidade do ensino das escolas públicas e, ao mesmo tempo, responsabilizar cada uma dessas organizações, bem como dirigentes e professores, pelos resultados dos alunos" (AMÉRICO; LACRUZ, 2017, p. 855), cujo movimento no campo do *management* se conveniou chamar *accountability*.

O trabalho estará seccionado em uma revisão da literatura que engloba os estudos de modelos de regressão logística e suas aplicações, na descrição dos processos metodológicos que foram utilizados para o tratamento dos dados e a análise dos seus resultados, na discussão desses resultados e no debate de suas implicações e utilização do modelo proposto.

2 REGRESSÃO LOGÍSTICA

A técnica de regressão logística é uma das ferramentas mais utilizadas no meio acadêmico e organizacional, uma vez que estima o comportamento de diversas variáveis que influem sobre uma variável dependente categórica (FÁVERO *et al.*, 2009). Dessa forma, tem-se que objetivo principal da técnica está em analisar a relação entre variáveis independentes (ou explicativas) numéricas e/ou categóricas e uma variável dependente categórica (ou variável resposta) dicotômica/binária ou multinominal, fornecendo um modelo que prevê tal relação entre as variáveis.

Um modelo de regressão logística binária refere-se onde a variável resposta do modelo possui distribuição de Bernoulli (ou binomial), ou seja, quando assume apenas dois valores possíveis. Já a regressão logística multinominal pode ser vista como uma extensão do modelo logístico binário, em situações nas quais a variável resposta assume múltiplas categorias (HOSMER; LEMESHOW, 1989).

Em síntese, a regressão logística tem por função descrever a relação entre uma variável de resposta discreta (dicotomia é o caso mais comum) e uma (regressão logística simples) ou mais variáveis independentes (regressão logística múltipla), geralmente chamadas de preditoras, variáveis explicativas ou covariáveis.

Em um modelo de regressão logística, tem-se a variável resposta assumindo valores binários, ou seja, 1 ou 0, onde pode-se inferir situações dicotômicas de "sim ou não", "possui ou não possui", "sucesso ou insucesso" etc. Em todas as regressões a quantidade chave está no valor médio da variável binária ou dicotômica de acordo com o valor da variável independente. Esta quantidade é denominada valor médio condicional, podendo ser descrita como $E[Y=1 \mid X]=\pi(x)$, onde se assume que Y tem por função representar a variável resposta e X representar a variável explicativa (HOSMER; LEMESHOW, 1989).

O uso da técnica de regressão logística é apropriado para diversas situações, pois possibilita a análise do efeito de uma ou mais variáveis independentes (categóricas ou não) em relação a uma variável dependente dicotômica ou binária, informando com (1) para certa característica analisada e (0) para outra a ser observada ou para a falta desta (HOSMER; LEMESHOW, 1989).

Melhor dizendo, trata-se de uma técnica que tem o poder de descrever a relação entre diversas variáveis independentes (X_i) e uma variável dependente de caráter binário (Y), explicitada com 1 ou 0 (HOSMER; LEMESHOW, 1989). O modelo descreve o valor provável de Y através do modelo explicitado na equação 1.

$$E(Y) = \frac{1}{1+e^{-z}} \quad (1)$$

O foco da análise de regressão logística está em descrever a função matemática de Y em relação aos valores de X_i e de β_i. Desse modo, é possível ajustar os parâmetros do modelo (HOSMER; LEMESHOW, 1989). A expressão geral do modelo logístico é evidenciada pela equação 2.

$$f(z) = \frac{1}{1+e^{-z}} \quad (2)$$

De modo que pode ser escrita da forma completa como segue:

$$Z = \beta_0 + \sum_{i=1}^{K} \beta_i x_1 \quad (3)$$

Onde *z*, vetor logit, é tido como log *odds*, com seu valor variando de -∞ a +∞. Dessa forma, a função *f(z)*, função logit, normaliza a saída do modelo para o intervalo [0,1] e, então, fornece a probabilidade de ocorrência do evento.

O logit, função de regressão, pertence à categoria de modelos estatísticos, na qual as variáveis explicadas são qualitativas ou quantitativas, mas apenas características qualitativas podem ser observadas.

A transformação logit é considerada uma transformação chave na análise de modelos de regressão logística, tendo em vista que o seu intuito é aplicar um modelo de linearização logarítmica. Essa conversão, tomando como exemplo o caso da regressão logística simples, é definida como:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \quad (4)$$

De onde vem que:

$$g(x) = ln\left(\frac{\frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}}{1-\frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}}\right) \quad (5)$$

De modo que:

$$g(x) = ln\left(\frac{\frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}}{\frac{1}{1+e^{\beta_0+\beta_1 x_1}}}\right) \quad (6)$$

E, portanto, tem-se que:

$$g(x) = \ln\left(e^{\beta_0+\beta_1 x_1}\right) = \beta_0 + \beta_1 x_1 \quad (7)$$

Essa conversão é particularmente importante porque o modelo com essa conversão possui vários atributos do modelo de regressão linear, como a linearidade nos parâmetros, o fato de ser contínua e seus valores variando em $\mathbb{R}$.

Denomina-se a transformação como transformação logit de π(x). A razão $\frac{\pi(x)}{1-\pi(x)}$ denomina-se razão de chances ou *Odds Ratio* (OR). A OR contém a probabilidade de o evento ocorrer dividido pela probabilidade de que o evento não ocorra (HOSMER; LEMESHOW, 1989).

Dessa maneira, tem-se que π(x) é a probabilidade de ocorrência do evento e 1 – π(x) a não ocorrência deste mesmo evento.

De acordo com Hosmer e Lemeshow (1989), o modelo logístico teve forte aplicabilidade devido ao fato de ser uma técnica simples e com propriedade teórica, no entanto, isso se deve principalmente em relação a sua capacidade de fornecer o logaritmo da razão de chances. Ou seja, torna-se possível identificar a porcentagem em razão de chance de cada variável independente interferir na variável dependente. A base de interpretação da razão de chance é o valor 1, assim, valores acima deste representam valores incrementais; se o valor for menor que 1, são chamados fatores protetivos, onde o incremento de uma unidade na variável independente decrementa a chance desta em relação ao aumento na variável dependente; se igual a 1, significa que não possui associação entre as variáveis.

Como citado anteriormente, existem dois modelos de regressão logística: simples (univariado) e múltiplo (multivariado). Um modelo de regressão logística simples é utilizado para o caso de regressão com uma única variável explicativa. Dessa forma, busca-se demonstrar a influência da variável independente na variável dependente, a qual tem caráter binário ou dicotômico, tendo a possibilidade de assumir um dos dois valores binários possíveis.

Deste modo, neste modelo de regressão logística, segundo Hosmer e Lemeshow (1989), assume-se a equação 8:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}} \quad (8)$$

Onde: β_0 é o coeficiente do intercepto; β_1 é o coeficiente da variável independente, ou melhor, o parâmetro a ser estimado; x_1 é a variável explicativa ou independente.

Hosmer e Lemeshow (1989) apresentam uma generalização do modelo de regressão logística para o caso com mais de uma variável explicativa.

Anteriormente, foi considerado o modelo de regressão logística univariado, ou seja, para o caso em que tem-se apenas uma única variável explicativa. Agora, considera-se o modelo múltiplo. Neste tem-se um conjunto de k variáveis explicativas, onde assume-se o vetor $x^T \equiv (x_1, x_2, ..., x_k)$.

Neste caso, $\pi(x)$ tem a seguinte expressão:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}} \quad (9)$$

Onde: β_0 é o coeficiente do intercepto; $\beta_1, \beta_2, \ldots, \beta_k$ são os coeficientes de cada variável independente, ou melhor, os parâmetros a serem estimados; $X, X_2, \ldots, X_k$ são as variáveis explicativas ou independentes, onde pode-se considerar X_i, com i = 1, 2, ..., k.

Já o logit da Regressão logística múltipla é considerado da seguinte forma:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \quad (10)$$

2.1 Formas de seleção do modelo

De acordo com Fávero *et al.* (2009), o método mais utilizado para a seleção de variáveis que irão compor o modelo de regressão é o *stepwise*, onde se tem a inclusão das variáveis de forma individual de modo a se obter as melhores variáveis a serem utilizadas (HAIR *et al.*, 2005) e o método de *backward*, o qual inclui todas as variáveis explicativas no modelo inicial, sendo retiradas uma a uma de acordo com a significância estatísticas destas, deixando somente as mais significativas.

No método *stepwise*, ou seleção passo a passo, o processo começa com o passo à frente, porém, depois que a segunda variável é inserida no modelo, é realizado o teste da razão de verossimilhança de modo a verificar se a primeira variável ainda continua no modelo. Caso permaneça, é feita a seleção de uma terceira variável da mesma maneira que no procedimento passo à frente. Caso uma terceira variável seja inserida no modelo, é realizado o teste de modo a verificar se as duas ainda permanecem no modelo. É possível de que uma delas ou as duas sejam excluídas do modelo. Dessa forma, busca-se a inclusão de uma nova variável. Se esta é inserida, busca-se então a exclusão das demais que compõem o modelo. O fim do processo se dá quando não se consegue inserir e nem excluir mais variáveis no modelo.

No método *backward*, ou seleção para trás, o processo se inicia com o ajuste de todas as possíveis variáveis independentes a entrar no modelo. É feita a comparação do desvio do modelo logístico contendo todas as variáveis com os desvios dos modelos que ocasionam a eliminação das variáveis de forma individual. Se o nível descritivo encontrado a partir do teste da razão de verossimilhança tiver significância, a variável candidata entra no modelo e o procedimento é encerrado; caso contrário, ela é excluída do modelo. Das demais variáveis que sobraram no modelo, se escolhe novamente a de menor contribuição e verifica-se se foi ou não significante. Caso seja significante, ela é inserida no modelo e o procedimento tem seu fim, porém, caso não seja, ela é retirada do modelo e dá-se sequência ao processo. Ressalta-se que nesta técnica, uma vez que a variável é retirada do modelo ela não entra mais.

Outro método utilizado é o *Forward* ou seleção para frente. Este, basicamente, consiste em iniciar o modelo sem as variáveis independentes, efetuando, de forma a testar, passo a passo a adição de uma nova variável, utilizando os critérios de comparação para sua escolha (como o teste t ou F), acrescentando a variável que melhor aperfeiçoa o modelo e fazendo a repetição deste procedimento até quando não for mais possível elevar de forma significante a acurácia do modelo.

No caso multivariado, o método mais utilizado é o da Máxima Verossimilhança. Considerando a independência das observações, a função de verossimilhança é dada por:

$$L(\beta) = \ln[L(\beta)] = \Sigma\{y_i \ln[\pi(x)] (1 - y_i) \, ln(1 - \pi(x))\} \quad (11)$$

Podendo ser escrita da seguinte fora:

$$\sum_{i=1}^{n}[y_i\beta_0 + y_i\beta_1 x_1 + \cdots + y_i\beta_k x_k - \ln(1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k})] \quad (12)$$

A escolha do modelo e variáveis apropriadas, da mesma forma como o modo em que as variáveis são introduzidas no modelo, são missões deveras complexas e que devem ser realizadas de forma a tomar o cuidado de explorar ao máximo as inter-relações entre todas as variáveis (HENNEKENS; BURING, 1987).

2.2 Investigando a qualidade do modelo logístico

Em um modelo de regressão, espera-se encontrar o melhor modelo possível capaz de estimar os coeficientes que melhor explicam a relação das variáveis do modelo. É quase que descartada a hipótese de existência de um modelo que obtenha um Y real e por isso lança-se mão de modelos capazes de medir a qualidade do modelo logístico de modo a utilizar os melhores e mais eficientes modelos.

Existem várias formas de medir a qualidade do modelo, como o Akaike *Information Criteria* (AIC) e o *Deviance* (D). Medidas gráficas também são utilizadas para esta finalidade, como é o caso da curva ROC (*Receiver Operating Characteristic*).

O AIC, desenvolvido por Akaike (1974), daí o nome, pode ser explicado como sendo um método da minimização das variáveis independentes sem a necessidade de envolver testes estatísticos. Com tal critério, é possível se obter um modelo com parcimônia, o qual terá um menor número de variáveis e que possua uma boa qualidade do ajuste (PAULA, 2013). Com isso, depreende-se que o modelo mais indicado é o que representa um menor AIC. Neste critério, segundo trabalho de Akaike (1974), é admitido a existência de um modelo considerado "real", que por sua vez, descreve os dados que é desconhecido, tentando selecionar dentre um grupo de modelos avaliados, minimizando, assim, a divergência de Kullback-Leibler (K-L). A estimativa do AIC para um certo modelo é definida como: $\mathrm{AIC} = -2logL + 2P$, tendo que LogL é o valor do logaritmo da função de verossimilhança considerando-se as estimativas dos parâmetros e P sendo o número de parâmetros.

O método *Deviance*, também entendido como função desvio, é a distância do logaritmo relativo à função de log-verossimilhança do modelo de início, o qual contém k parâmetros, e do modelo final contendo um valor P de parâmetros. Segundo Pereira (2014), sendo a estatística menor que o Qui-quadrado de referência, o modelo é considerado adequado. Com isso, define-se que valores D pequenos influem que, para um menor número de parâmetros estimados, terá então um ajuste muito bom, assim como o ajuste do modelo inicial, considerando, então, um modelo adequado. A expressão da função desvio é a seguinte:

$$D = -2\log \frac{\text{VEROSSIMILHANÇA DO MODELO AJUSTADO}}{\text{VEROSSIMILHANÇA DO MODELO SATURADO}} \quad (13)$$

Os valores de log-verossimilhança dos modelos ajustado e saturado, assumindo $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = x_i^T\beta$, são dados, respectivamente, como segue:

$$VM_A = \sum_{i=1}^{n} \ln\left[p(y_i; \mu_i, \emptyset)\right] \quad (14)$$

$$VM_S = \sum_{i=1}^{n} \ln\left[p(y_i; \mu_i = y_i, \emptyset\ )\right] \quad (15)$$

Podendo simplificar a função desvio em:

$$D = 2(VM_S - VM_A) \quad (16)$$

Em relação à curva ROC, é uma medida gráfica utilizada para visualizar a qualidade do ajuste do modelo. A curva ROC obtida no modelo (figura 1) testa a relação do *cutoff* com a especificidade e a sensibilidade, de modo que quanto maior for o *cutoff* determinado, maior será a especificidade do teste e menor é a sensibilidade; ao passo que quanto menor o *cutoff*, maior será a sensibilidade e menor é a especificidade. Essa é uma característica discriminativa de uma espécie de teste diagnóstico que permite evidenciar valores (*cutoff)* em que é possível obter a melhor proporção entre falsos positivos e falsos negativos, otimizando a relação da sensibilidade com a especificidade (ZHUOFAN, 2011). Dessa forma, entende-se que quanto maior é a área abaixo da curva ROC, maior é a capacidade do modelo em questão de discriminar os grupos requeridos (o que interessa ao estudo) dos não requeridos.

De acordo com Zhuofan (2011), esta área abaixo da curva ROC é chamada de AUC (*Area Under The Curve*), sendo essa uma medida que compara modelos e quanto mais próxima de 1, melhor é a acurácia do modelo. Essa área é uma métrica invariante em escala que qualifica a precisão das previsões do modelo proposto e independe de classificação.

2.3 Medidas de ajustamento

Em um modelo de regressão logística, são utilizadas diversas medidas para que o resultado seja o mais correto possível, ou seja, para que o modelo seja o mais ajustado possível. Dessa forma, têm-se alguns testes importantes ao utilizar um modelo de regressão, como os testes de McFadden, Cox e Snell, Nagelkerke, Hosmer e Lemeshow e Kolgomorov-Smirnov.

O Pseudo R-quadrado de McFadden é uma medida utilizada para realizar uma mensuração da qualidade do ajuste dos modelos estimados, ou seja, mede a qualidade do ajuste (Pseudo *Goodness Of Fit*) dos modelos estimados.

O cálculo do R-quadrado de McFadden, segundo Fávero *et al.* (2009), é dado por:

$$R^2{}_{logit} = \frac{-2LL_0-(-2LL_\beta)}{-2LL_0} \quad (17)$$

Em relação ao R-quadrado de Cox e Snell, entende-se que é um mecanismo comparável ao R-quadrado da regressão linear, porém, este teste não tem como finalidade indicar a proporção da variação constatada em relação à variável dependente e as independentes, tendo em vista que funções de probabilidade não lidam com esses tipos de variações. No entanto, é um mecanismo que pode ser empregado para comparar o desempenho de modelos concorrentes; entre duas equações logísticas consideradas igualmente válidas, prefere-se a que apresenta o maior Cox-Snell R^2.

Os cálculos do R-quadrado de Cox e Snell, segundo Fávero *et al.* (2009), são dados pelos modelos:

$$R^2{}_{CS} = 1 - (\frac{L_0}{L_\beta})^{\frac{1}{N}} \quad (18)$$

$$R^2{}_{CS_{máx}} = 1 - (L_0)^{\frac{2}{N}} \quad (19)$$

Já o R-quadrado de Nagelkerke faz uma correção do R-quadrado de Cox e Snell, sobre-estimando-o de forma a poder alcançar seu valor máximo 1. Com isso, o valor do R-quadrado de Nagelkerke será sempre maior que o de Cox e Snell. A diferença do teste Nagelkerke R^2 para o Cox-Snell R^2 é se fazer mais compreensível que o de Cox e Snell (HAIR *et al.*, 2005), porém, a utilização dos dois em conjunto é importante para reforçar os achados e melhor compreendê-los.

Dessa forma, segundo Fávero *et al.* (2009), o cálculo do R-quadrado de Nagelkerke é dado pela equação 20:

$$\tilde{R}^2{}_N = \frac{R^2{}_{CS}}{R^2{}_{CS_{máx}}} \quad (20)$$

O teste de Hosmer e Lemeshow, por sua vez, avalia o modelo ajustado comparando as frequências observadas e as esperadas. Assim, o intuito deste teste é verificar se existem diferenças significativas entre as classificações realizadas pelo modelo de referência e a realidade observada. Desse modo, a hipótese a ser testada nesse tipo de teste é: H_0: "O modelo explica bem os dados".

Em relação ao teste de Kolmogorov-Smirnov (KS), este concede o parâmetro valor de prova (*p-value*), o qual deve ser interpretado como a medida do grau de concordância dos dados, e a hipótese nula (H_0), onde H_0 corresponde à consideração de que o modelo é adequado. Desta maneira, quanto menor for a significância, menor é a consistência entre os dados e a hipótese nula. Assim, para que o modelo seja adequado, o *p-value* deve ser maior que α, assegurando a não rejeição da hipótese nula.

2.4 Regressão logística com dados educacionais

Não é possível depreender com precisão informações acerca das primeiras realizações da técnica de regressão logística, porém, segundo McLachlan (1992), foram em estudos prospectivos relacionados à fatores de doenças coronárias. De acordo com Hosmer e Lemeshow (1989), a técnica teve retorno após o estudo de Truett, Cornfield e Kannel (1967), onde os autores estudaram o risco de doença coronária em um projeto denominado *Framingham Heart Study*. Com isso, o estudo possui grande relevância e ganhou notoriedade, sendo até os dias de hoje uma obra de suma importância na área da saúde.

A partir daí a regressão logística se tornou um mecanismo padrão para analisar dados de caráter dicotômicos e binários, sobretudo na área de ciências médicas, conforme Hosmer e Lemeshow (1989). Esta técnica apresentou uma elevação em seu uso muito rápida, sendo sua utilização difundida para diversas outras áreas, como por exemplo, a economia, administração, educação e outras.

Na área educacional, sua aplicação é bem vista pelos pesquisadores. Não se sabe ao certo o primeiro uso da regressão logística em dados educacionais, porém, o uso da regressão logística possibilitou diversos estudos ligados à educação, sendo sua utilização bastante diversificada.

Cunha, Cornachione Júnior e Martins (2008) utilizaram a regressão logística para verificar a existência de variáveis que influenciam a probabilidade associada à ocorrência de reprovação em disciplinas conforme as diversas percepções dos alunos do curso noturno de graduação em ciências contábeis de uma faculdade particular localizada em Belo Horizonte – MG. Com o uso da regressão, os autores identificaram que a reprovação está associada positivamente com os fatores de preferência dos alunos por detalhes, por ouvir, quando se trata de aulas expositivas e por altas expectativas em relação às suas notas, revelando importante contribuição para o meio educacional.

Vitelli, Rocha e Fritsch (2010) trouxeram a regressão logística para estudar a evasão de alunos nos cursos de graduação em uma instituição de ensino superior privada, identificando os fatores contribuintes. Os autores encontraram que a evasão está associada com a indefinição na escolha profissional, desempenho acadêmico e condição financeira dos discentes. Os resultados são importantes, pois com eles torna-se possível focar no problema e, dessa forma, buscar soluções com base nos achados.

Brunozi Júnior *et al.* (2011) estudaram, por meio da regressão logística, os efeitos das transferências intergovernamentais e arrecadação tributária sobre os indicadores sociais da saúde e educação em Minas Gerais. Os resultados do estudo dos autores trouxeram uma importante contribuição referente à educação ao identificar que a arrecadação tributária assim como a distribuição da quota-parte referentes ao ICMS Saúde da Família (Lei Robin Hood) são fatores que revelam discriminação na importância das receitas públicas citadas como fonte de fomento para a aplicação das demandas sociais básicas da população, no caso saúde e educação, revelando que a gestão dos governos municipais, levando em consideração os critérios de produto interno bruto gerado, transferências sociais e seguridade social, não compreende as condições necessárias para assegurar as determinações mínimas de bem-estar da sociedade na educação e saúde e, por consequente, a promoção do desenvolvimento econômico e social em âmbito dos municípios.

Dados mais recentes mostram importantes estudos com o uso da regressão logística aplicada à educação, tanto no ensino básico quanto o superior (SILVA *et al.*, 2020; ROCHA; TOLEDO JÚNIOR, 2020). Silva *et al.* (2020) trazem a aplicação da regressão logística no ensino básico, buscando identificar as prováveis relações entre a rede pessoal e o desempenho escolar dos alunos investigados, revelando que notas baixas no processo de admissão (alunos calouros) e baixo desempenho acadêmico no primeiro semestre (alunos com notas de aprovação) foram fatores preditivos de fracasso acadêmico. Já o estudo de Rocha e Toledo Júnior (2020) tem o foco no ensino superior e traz o uso da regressão logística para identificar fatores preditivos de insucesso acadêmico em um curso de medicina no Brasil; seus resultados mostram que alunos com redes densas tendem a tirar notas baixas na disciplina de Português, já os alunos com redes com muitos cliques apresentam certa tendência a tirarem notas baixas na disciplina de Matemática.

Segundo Niu (2018), em seu estudo *A review of the application of logistic regression in educational research: common issues, implications, and suggestions*, embora a aplicação da regressão logística seja bem

vista e bastante utilizada no campo da educação, ainda existem problemas geradores de preocupação na área educacional, sendo a fraqueza na compreensão conceitual, a diferença para o modelo de regressão linear e o significado e limitação da razão de chance.

3 PROCEDIMENTOS METODOLÓGICOS

Nesta seção descrevem-se os procedimentos metodológicos deste artigo tecnológico (MOTTA, 2017). A fim de favorecer a transparência na pesquisa, o *dataset* e o *script* para o *software* Stata utilizados foram disponibilizados no repositório Harvard Dataverse[1].

3.1 *Design* e método de pesquisa

Para a realização da pesquisa lançou-se mão de uma pesquisa descritiva, possuindo corte transversal e abordagem quantitativa.

Buscou-se, através da Regressão Logística, descrever por meio do modelo logístico binário, as variáveis que melhor discriminam as performances das escolas estaduais capixabas no Saeb de acordo com a inclusão de variáveis contextuais no modelo. Esclarece-se que o estado do Espírito Santo, como observam Gobbi *et al.* (2020), oferece uma boa perspectiva para análise do fenômeno investigado, pois contribui para a construção de uma amostra controlável e representativa de um contexto regional, com apenas 493 escolas estaduais.

3.2 Composição da amostra

A população de pesquisa do estudo é composta por 726 escolas estaduais urbanas e rurais dos 78 municípios do estado do Espírito Santo para as quais se dispunha de todas as variáveis independentes levantadas para o modelo. A amostra de pesquisa utilizada é composta por 364 escolas. Ressalta-se que a redução se deu por conta de o estudo considerar apenas as notas referentes aos quartis 1 e 3, ou seja, as 25% inferiores e as 25% superiores.

A coleta de dados foi realizada por meio do site do Inep no período entre março e abril de 2020, destacando que os dados mais recentes no período da coleta eram referentes ao ano de 2017.

Para a realização do trabalho, foi necessária a construção de 2 bases de dados distintas com os dados levantados, sendo referentes às notas das disciplinas de língua portuguesa e matemática dos alunos dos anos iniciais no Saeb.

Foram utilizadas 7 variáveis contextuais (NSE, MIRD, IED, ICG, HAD, AFD e ATU) como independentes e a variável GNP como variável dependente. A descrição das variáveis é dada no quadro 1.

Quadro 1 – Definição das Variáveis

Variáveis		Descrição
Dependente	GNP	Grupo de acordo com a nota da prova
Independentes	NSE	Nível socioeconômico dos discentes
	MIRD	Média do indicador de regularidade do corpo docente
	IED	Indicador de esforço docente
	ICG	Indicador de complexidade de gestão da escola
	HAD	Horas-aula diária
	AFD	Adequação de formação docente
	ATU	Alunos por turma

Fonte: Elaboração própria.

[1] (cf. https://doi.org/10.7910/DVN/RCLAFX)

A caracterização das variáveis é dada a seguir:

a) GNP: o grupo é formado de acordo com a média da proficiência em língua portuguesa e matemática padronizada para um intervalo de 0 a 10. A nota da escola corresponde à média das notas dos discentes;
b) NSE: é a média aritmética simples da medida de NSE dos alunos, num intervalo definido com base em bens domésticos, renda, contratação de serviços e nível de escolaridade da família, a fim de possibilitar a visão geral do padrão de vida dos alunos;
c) MIRD: Observação da permanência dos professores nas escolas. Quanto mais próximo de 0, mais irregular é o vínculo do professor com a escola; quanto mais perto de 5, mais regular;
d) IED: Percentual de docentes por nível de esforço. As características consideradas para mensurar o nível de esforço são: número de escolas, turnos, alunos e etapas que o professor leciona. As escolas são classificadas do nível 1 (até 25 alunos, um único turno, escola e etapa) ao nível 6 (mais de 400 alunos, nos três turnos, em duas/três escolas e etapas). Utiliza-se como variável explicativa a soma dos percentuais dos níveis 4, 5 e 6;
e) ICG: Características consideradas para mensurar o ICG: porte das escolas; número de turnos, etapas e modalidades oferecidas. O nível 1 representa a menor complexidade de gestão escolar, enquanto o nível 6, a maior;
f) HAD: Número médio de horas-aula diária;
g) AFD: Vai do grupo 1 (licenciatura ou bacharelado com complementação pedagógica na disciplina que leciona) ao grupo 5 (não possui curso superior). Considera-se como variável explicativa o percentual de disciplinas, em cada etapa, ministradas por professores do nível 1;
h) ATU: Número médio que corresponde à divisão do número de matrículas pelo número de turmas da escola.

3.3 Tratamento dos dados

O primeiro passo necessário foi levantar os valores binários para a regressão e, dessa forma, utilizou-se os quartis 1 e 3 para a definição, definindo como 0 e 1, respectivamente. Ou seja, dividiu-se as amostras da seguinte forma: as notas referentes ao 1º quartil (25% inferiores) ficaram com o valor 0 e as notas referentes ao 3º quartil (25% superiores) ficaram com o valor 1. Fez-se isso para todos os modelos de regressão.

Assim, a variável dependente binária GNP assumiu os valores:

$$GNP = \begin{cases} 0: Nota \leq Quartil\ 1 \\ 1: Nota \geq Quartil\ 3 \end{cases}$$

A regressão logística foi processada para se obter as variáveis significativas do modelo bruto com a amostra de treino e, logo em seguida, o modelo ajustado, de modo a identificar a significância das variáveis inseridas no modelo final. Utilizou-se o método *Stepwise Forward* para verificar a significância das variáveis antes destas adentrarem o modelo ajustado.

Foi revelada a razão de chance das variáveis contextuais (variáveis independentes) do modelo ajustado, de forma a melhor compreender a chance de aumento e possível influência na nota da prova (variável dependente).

Para medir a qualidade da amostra foi utilizado o teste de McFadden e para avaliar a qualidade do modelo, utilizou-se a estatística de Hosmer e Lemeshow.

Para o teste de Hosmer e Lemeshow, assumiu-se a seguinte hipótese nula: *H_0: Não há diferenças significativas entre os resultados preditos pelo modelo e os observados.*

Foi realizado também o teste do Qui-quadrado, de modo a constatar se existe associação entre as variáveis do modelo (dependente e independentes) ou se a relação é somente de casualidade. Deste modo, assumiu-se a seguinte hipótese nula para este teste: *H_0: Não há associação entre as variáveis.*

Para observar os percentuais de acerto do modelo, estabeleceu-se uma tabela de classificação.

Como medida para verificar a acurácia estabelecida no modelo, utilizou-se a medida gráfica da curva ROC e o valor AUC de referência para cada disciplina.

Para o modelo foi utilizada a técnica de reamostragem Jackknife (TUKEY, 1958) para validação dos resultados.

Foi utilizado o *software* STATA para todos os processos estatísticos realizados no estudo. Foi adotado um nível de significância de 0,05 para o estudo.

4 ANÁLISE DOS RESULTADOS

De modo a analisar quais variáveis contribuem no pertencimento das escolas aos grupos de maior e menor desempenho no Saeb, foram realizadas duas regressões logísticas; um modelo para cada área de referência (língua portuguesa e matemática) de acordo com as séries iniciais do ensino fundamental (5º anos). Dessa maneira, foram dois modelos de regressão: modelo das notas de língua portuguesa e de matemática dos anos iniciais do ensino fundamental.

Para todos os modelos logísticos foi mantido o intercepto, pois é assumido que mesmo que todas as variáveis analisadas tenham o seu valor como zero, as notas das provas não seriam zeradas, pois outros fatores interferem nos resultados obtidos pelos alunos nas provas.

4.1 Modelo das notas de língua portuguesa dos anos iniciais

Após a divisão do modelo binário, foram quantificadas 182 observações (50%) para as 25% inferiores e 182 (50%) para as 25% superiores, totalizando 364 notas, configurando o *cutoff* do modelo.

O resultado da regressão com o modelo completo (tabela 1) mostram que as variáveis NSE, MIRD e ICG foram significantes ao nível de 5% (*p-value* < 0,05). O valor do Pseudo R^2 com todas as variáveis foi de 0,2016.

Tabela 1 – Modelo Logístico Completo – Saeb Língua Portuguesa (5º ano)

Variáveis	Coeficiente	Intervalo de confiança (95%)	Significância
NSE	1,565	(1,108 — 2,021)	0,000
MIRD	0,921	(0,461 — 1,381)	0,000
IED	0,000	(-0,005 — 0,005)	0,945
ICG	-0,325	(-0,527 — -0,124)	0,002
HAD	0,152	(-0,260 — 0,564)	0,471
AFD	0,000	(-0,006 — 0,006)	0,957
ATU	-0,001	- (-0,075 — 0,072)	0,969

Fonte: Elaboração própria.

Como nem todas as variáveis se mostraram estatisticamente significantes, foi processada regressão logística *Stepwise Forward*, utilizando como critério de saída o nível de significância de 0,10 e de entrada o nível de 0,05, as variáveis significativas para entrarem no modelo foram as mesmas. Desse modo, as variáveis que compuseram o modelo final foram: NSE, MIRD e ICG (Tabela 2). As demais variáveis foram descartadas por não apresentarem significância para o modelo.

Tabela 2 – Modelo Logístico Ajustado – Saeb Língua Portuguesa (5º ano)

Variáveis	OR[a]	Estimativa (IC 95%)[b]	Significância
NSE	4,766	1,561(1,120 — 2,002)	0,000
MIRD	2,455	0,898(0,457 — 1,339)	0,000
ICG	0,720	-0,328(-0,521 — -0,134)	0,001

[a]*Odds ratios* [b] Intervalo de confiança de 95%.

Fonte: Elaboração própria.

Após a análise do modelo ajustado (tabela 2), verifica-se que as três variáveis continuam associadas ao modelo (*p-value* < 0,05). Os valores da razão de chances (OR) encontrados, mostram dois desfechos diferentes para as variáveis. Antes de tudo, vale ressaltar que o valor base para a interpretação das OR é 1 (HOSMER; LEMESHOW, 1989). Deste modo, o valor acima de 1 indica um valor incremental, ou seja, sendo 4,766 o valor da OR da variável NSE, tem-se 3,766 de chance de aumento para cada 1 unidade de acréscimo no valor da variável GNP. Ele se dá para a variável MIRD, seu valor de OR de 2,455 indica que a chance de aumento é de 1,455 para cada 1 unidade acrescentada. Já o valor abaixo de 1 (ICG), indica que a tendência é diminuir a chance de aumento, ou seja, sendo 0,720 o valor da OR, tem-se que a chance de aumento na variável GNP diminui em 0,280 para cada 1 unidade adicionada ao valor da variável ICG. Dessa forma, tem-se que com o aumento no valor das variáveis NSE e MIRD, a chance de o valor da variável GNP crescer, aumenta. No caso do aumento no valor da variável ICG, a chance de o valor da variável GNP ter um crescimento é diminuída.

Os sinais positivos das estimativas das variáveis NSE e MIRD refletem que quanto maior o nível socioeconômico dos discentes e o indicador de regularidade do corpo docente, maior é a força de pertencimento ao grupo de maior desempenho. Já o sinal negativo da estimativa do coeficiente de regressão logística do modelo ajustado descreve que quanto maior o indicador de complexidade de gestão da escola, maior é a força de pertencimento ao grupo de menor desempenho (significativo para $\alpha = 5\%$).

Acrescenta-se que o teste de razão de verossimilhança falhou em rejeitar a hipótese (*p-value* = 0,97) que a exclusão de variáveis do modelo (Tabela 2) tenha reduzido a capacidade explicativa do modelo completo (Tabela 1).

Por sua vez, o teste de significância do conjunto de coeficientes do modelo feito por meio da distribuição Qui-quadrado, demonstra a significância do modelo completo em relação ao modelo nulo, ou seja, apenas com o intercepto (*p-value* < 0,001).

O resultado obtido do Pseudo R^2 de McFadden de 0,200 para o modelo ajustado é considerado razoável. McFadden sugeriu que valores de Pseudo R^2 entre 0,2 a 0,4 devem ser considerados para representar um ajuste muito bom do modelo (LOUVIERE; HENSHER; SWAIT, 2000), dessa forma, valores dentro deste limite e não tão distantes são considerados aceitáveis. Assim, mesmo devido ao pequeno número de variáveis do modelo, o valor é considerado satisfatório.

A medida de Hosmer e Lemeshow (Tabela 3) indica que não existe diferença significativa na distribuição de valores dependentes efetivos e previstos. O resultado do teste de Hosmer e Lemeshow mede a correspondência dos valores efetivos e previstos da variável dependente. Neste caso, o melhor ajuste do modelo é indicado por uma diferença menor na classificação observada e prevista. Um bom ajuste do modelo é indicado por um valor Chi-Square não significante (HAIR *et al.*, 2009).

Tabela 3 – Teste de Hosmer Lemeshow – Saeb Língua Portuguesa (5º ano)

Qui-quadrado	Graus de liberdade	Significância
7,89	8	0,444

Fonte: Elaboração própria.

A tabela de classificação apresentada na tabela 4, mostra que a taxa de acerto do modelo com as 3 variáveis foi alta. A taxa de acerto global é de 72,53%. Nenhuma taxa de acerto de grupos individuais se mostrou baixa, sendo 69,90% para o grupo com as notas inferiores e 75,95% para as superiores.

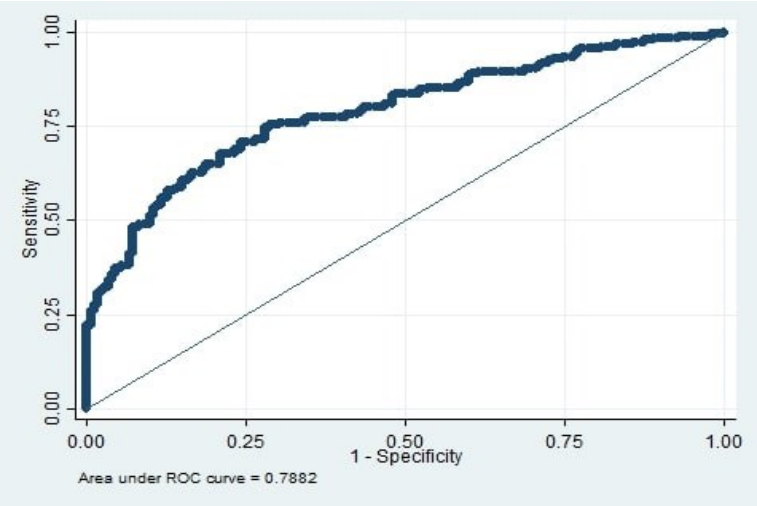
Tabela 4 – Tabela de Classificação – Saeb Língua Portuguesa (5º ano)

Classificados corretamente	
Grupo inferior	Grupo superior
0,699	0,7595
Global	
0,7253	

Fonte: Elaboração própria.

A curva ROC deste modelo (figura 1) revela uma boa acurácia do modelo quando traz o AUC com valor igual a 0,7882. Poder discriminante considerado aceitável pela gradação de Hosmer e Lemeshow (1989).

Figura 1 – Curva ROC – Saeb Língua Portuguesa (5º ano)



Fonte: Elaboração própria.

Como critério de validação dos resultados, foi realizado o procedimento de reamostragem de Jackknife, no qual se estima o valor de todas as observações retirando uma a uma da amostra. Os resultados permitem considerar que os resultados obtidos não sejam específicos da amostra utilizada na estimação.

Por fim, registra-se que não foram identificados problemas de multicolinearidade entre os preditores (VIF < 10) nem a presença de *outlier* na amostra (|DFBETA| < 1).

4.2 Modelo das notas de matemática dos anos iniciais

Seguindo para o modelo das notas de matemática dos anos iniciais, foi quantificado 182 observações (50%) para as 25% inferiores e 182 (50%) para as 25% superiores, tendo o total também de 364 notas, configurando o *cutoff* do modelo.

O resultado da regressão com o modelo completo para as notas de matemática dos alunos dos quintos anos (tabela 5) mostra que as variáveis significantes ao nível de 5% (*p-value*<0,05) foram a IRD, MIRD, ICG e ATU. O valor do Pseudo R^2 com todas as variáveis foi de 0,2424.

Tabela 5 – Modelo Logístico Completo – Saeb Matemática (5º ano)

Variáveis	Coeficiente	Intervalo de confiança (95%)	Significância
NSE	1,666	(1,180 — 2,152)	0,000
MIRD	1,310	(0,818 — 1,801)	0,000
IED	0,004	(-0,001 — 0,010)	0,144
ICG	-0,262	(-0,478 — -0,045)	0,018
HAD	0,082	(-0,333 — 0,496)	0,699
AFD	-0,002	(-0,009 — 0,004)	0,421
ATU	-0,089	(-0,163 — -0,014)	0,019

Fonte: Elaboração própria.

Da mesma forma que no caso anterior, como nem todas as variáveis foram estatisticamente significantes, processou-se a regressão logística *Stepwise Forward*, utilizando como critério de saída o nível de significância de 0,10 e de entrada o nível de 0,05, as variáveis significativas para entrarem no modelo ajustado

foram as mesmas significativas do modelo completo. Desse modo, as variáveis que compuseram o modelo final foram: NSE, MIRD, ICG e ATU (Tabela 6). As demais variáveis foram descartadas por não apresentarem significância para o modelo.

Tabela 6 – Modelo Logístico Ajustado – Saeb Matemática (5º ano)

Variáveis	OR[a]	Estimativa (IC 95%)[b]	Significância
NSE	5,206	1,650(1,168 — 2,132)	0,000
MIRD	3,923	1,367(0,894 — 1,840)	0,000
ICG	0,915	-0,286(-0,495 — -0,076)	0,007
ATU	0,002	-0,089(-0,163 — -0,0154)	0,018

[a]*Odds ratios* [b] Intervalo de confiança de 95%.
Fonte: Elaboração própria.

Após a análise do modelo ajustado (tabela 6), verifica-se que as quatro variáveis ainda possuem associação significativa ao modelo (*p-value*<0,05). Os valores da razão de chances (OR) encontrados mostram que duas variáveis são incrementais e duas decrementais em relação a chance de aumento no valor da variável GNP. O valor de 5,206 da NSE indica que a chance de aumento na variável dependente é de 4,206 para cada 1 unidade de acréscimo no valor da variável independente. Em relação à variável MIRD, o valor de chance de incremento no valor da GNP é de 2,923, tendo considerando seu valor de OR de 3,923. Em relação ao valor abaixo de 1 da variável ICG, mais precisamente 0,915, indica que a chance de aumento no valor da variável resposta com o acréscimo de 1 unidade na variável explicativa, diminui 0,085. Para o valor de 0,002 de OR da variável ATU, a chance de aumento cai 0,998 para cada 1 unidade acrescentada na variável independente. Com isso, tem-se que com o aumento no valor das variáveis NSE e MIRD, a chance de o valor da variável GNP crescer, aumenta. No caso do aumento no valor das variáveis ICG e ATU, a chance de o valor da variável GNP ter um crescimento é diminuída.

As duas primeiras estimativas são positivas, ou seja, indicam que quanto maior os valores do nível socioeconômico dos discentes e do indicador de regularidade do corpo docente, maior é a força de pertencimento ao grupo de maior desempenho. Já os dois valores negativos das duas últimas estimativas, indicam que quanto maior o valor do índice de complexidade de gestão da escola e o número de alunos por turma, maior é a força de pertencimento ao grupo de menor desempenho (significativo para $\alpha = 5\%$).

Acrescenta-se também para o modelo de matemática que o teste de razão de verossimilhança falhou em rejeitar a hipótese (*p-value* = 0,99) que a exclusão de variáveis do modelo (Tabela 6) tenha reduzido a capacidade explicativa do modelo completo (Tabela 5).

Por sua vez, o teste de significância do conjunto de coeficientes do modelo feito por meio da distribuição Qui-quadrado, demonstra a significância do modelo completo em relação ao modelo nulo, ou seja, apenas com o intercepto (*p-value* < 0,001).

O resultado obtido do Pseudo R^2 de McFadden de 0,237 é considerado bom. Assim, seguindo a mesma ótica do que foi citado no modelo anterior, mesmo com o modelo contendo poucas variáveis, o valor é considerado satisfatório.

A medida de Hosmer e Lemeshow (Tabela 7), assim como no modelo anterior, indica que não existe diferença significativa na distribuição de valores dependentes efetivos e previstos (*p-value* = 0,186 > α=0,05).

Tabela 7 – Teste de Hosmer e Lemeshow – Saeb Matemática (5º ano)

Qui-quadrado	Graus de liberdade	Significância
11,28	8	0,186

Fonte: Elaboração própria.

Em relação à tabela de classificação do modelo, conforme visto na Tabela 8, mostra que a taxa de acerto do modelo com as 4 variáveis foi alta. A taxa de acerto global é de 73,63%. As duas taxas de acerto dos

grupos individuais se mostraram altas, sendo 72,63% para o grupo com as notas inferiores e 74,71% para as superiores.

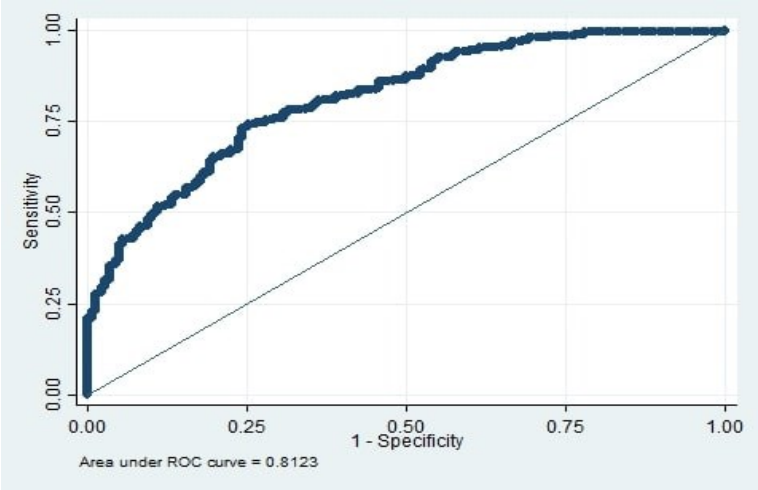
Tabela 8 – Tabela de Classificação – Saeb Matemática (5º ano)

Classificados corretamente	
Grupo inferior	Grupo superior
0,7263	0,7471
Global	
0,7363	

Fonte: Elaboração própria.

Na aplicação referente às notas de matemática, a curva ROC (Figura 2) revela uma boa acurácia do modelo quando traz o AUC com valor igual a 0,8123. Poder discriminante considerado aceitável, de acordo com a gradação de Hosmer e Lemeshow (1989).

Figura 2 – Curva ROC – Saeb Matemática (5º ano)



Fonte: Elaboração própria.

Como critério de validação dos resultados, foi realizado o procedimento de reamostragem de Jackknife também para o modelo de matemática. Os resultados, assim como no modelo anterior, permitem considerar que os resultados obtidos não sejam específicos da amostra utilizada na estimação.
Finalmente, registra-se que não foram identificados problemas de multicolinearidade entre os preditores (VIF < 10) nem a presença de *outlier* na amostra (|DFBETA| < 1).

5 CONSIDERAÇÕES FINAIS

A regressão logística é um método que pode ser usado para avaliar a influência de indicadores educacionais nas notas do Saeb. Neste artigo tecnológico (MOTTA, 2017), objetivamente, foi desenvolvido modelo aplicado às escolas da rede pública estadual no Espírito Santo. Frente a boa acurácia alcançada, os impactos revelados podem ser utilizados na tomada de decisão por parte das autoridades educacionais no sentido de rearranjar seus focos de investimento, por exemplo.

A relação do Índice de Complexidade de Gestão da Escola (ICG) com o desempenho na prova é inversa, de modo que quão maior o ICG, maior a chance de que a escola faça parte do grupo de escolas de pior desempenho. Tomando como nota que a complexidade da escola envolve aspectos como a quantidade de etapas ofertadas para alunos com idade mais elevada, o porte da escola, o número de etapas (séries ou níveis)

ofertados e o número de turnos ofertados, os resultados podem auxiliar questões de gestão escolar que envolvam futuras mudanças.

Os indicadores que impactam positivamente no desempenho das escolas envolvem fatores como a renda das famílias dos alunos (NSE) e o tempo de permanência dos professores nas escolas (MIRD). Dessas relações positivas, pode-se entender que o nível socioeconômico sugere um melhor rendimento escolar, já que as condições desses alunos permitem melhor acesso em geral; e que o tempo de permanência do corpo docente pode auxiliar o processo de aprendizagem dos alunos.

A pesquisa mostra, ainda, que quanto maior o índice de alunos matriculados em uma mesma turma da escola, menores são as notas da prova, dando margem a concluir que a qualidade do ensino pode ser prejudicada quando um professor precisa distribuir a atenção a muitos alunos.

Os indicadores revelados como impactantes permitem trazer questionamentos quanto aos aspectos de gestão escolar, partindo do princípio de que a liderança tenha meios de gerir as perspectivas pedagógicas de acordo com a manipulação organizacional que lhe é de alcance, constituindo-se na proposição de uma agenda de pesquisa decorrente dos resultados deste estudo. Limitar a quantidade de alunos por turma, controlar as horas de dedicação dos professores e oferecer melhores condições de convívio escolar àqueles que abrigam alguma carência social são fatores palpáveis a um gestor da rede estadual? Quais dos enfoques que compõem o ICG estão diretamente ligados à qualidade da gestão escolar?

Nesse contexto, Mesquita (2012) discute sobre como as publicações dos resultados sobre o desempenho das escolas pode fazer parte da construção de uma política nacional de educação que desperte reflexões acerca da comparação entre o que a sociedade espera do serviço público de educação prestado e daquele de fato observado.

Assim, a pesquisa buscou contribuir para decisões técnicas de emprego de recursos públicos e solução de problemas de gestão educacional no sentido de trazer uma análise de dados que proponha um mecanismo de comparação em que a realidade revelada pela nota de cada uma das escolas estaduais da rede pública do estado do Espírito Santo é interpretada de acordo com os indicadores contextuais do INEP.

REFERÊNCIAS

AMÉRICO, Bruno Luiz; LACRUZ, Adonai José. Contexto e desempenho escolar: análise das notas na Prova Brasil das escolas capixabas por meio de regressão linear múltipla. **Revista de Administração Pública**, v. 51, n. 5, p. 854-878, 2017.

CAMPELLO, Antonio de Vasconcellos Carneiro; LINS, Luciano Nadler. Metodologia de análise e tratamento da evasão e retenção em cursos de graduação de instituições federais de ensino superior. *In*: ENCONTRO NACIONAL DE ENGENHARIA DE PRODUCÃO, 28., 2008, Rio de Janeiro. **Anais eletrônicos** [...]. Rio de Janeiro: ABEPRO, 2008. Disponível em: http://www.abepro.org.br/biblioteca/enegep2008_tn_sto_078_545_11614.pdf. Acesso em: 23 set. 2020.

CUNHA, Jacqueline Veneroso Alves; CORNACHIONE JÚNIOR, Edgard B.; MARTINS, Gilberto de Andrade. Uma aplicação da regressão logística no inventário de estilos de aprendizagem de canfield (lsi) sob a ótica das reprovações acadêmicas. **Revista de Contabilidade e Organizações**, v. 3, n. 2, p. 100-112, 2008.

SILVA, Bruno Lopes da *et al*. Redes pessoais e desempenho escolar no ensino básico: um estudo sobre os alunos concluintes do Ensino Médio da Escola Estadual Santos Dumont, Parnamirim no Rio Grande do Norte, Brasil. **Revista Electrónica Educare**, v. 24, n. 1, p. 1-19, 2020.

FÁVERO, Luiz Paulo Lopes *et al*. **Análise de dados**: modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

GOBBI, Beatriz Christo *et al*. Uma boa gestão melhora o desempenho da escola, mas o que sabemos acerca do efeito da complexidade da gestão nessa relação? **Ensaio: avaliação e políticas públicas em educação**, v. 28, n. 106, p. 198-220, 2020.

HAIR, Joseph F. *et al*. **Análise multivariada de dados**. Porto Alegre: Bookman, 2009.

HAIR, Joseph F. *et al*. **Fundamentos de métodos de pesquisa em administração**. Porto Alegre: Bookman, 2005.

HENNEKENS, Charles H.; BURING, Julie E. **Epidemiology in Medicine.** Boston: Little, Brown and Company, 1987.

HOSMER, David. W.; LEMESHOW, Stanley. **Applied logistic regression**. New York: John Wiley & Sons, 1989.

BRUNOZI JÚNIOR, *Antônio Carlos et al*. Efeitos das transferências intergovernamentais e arrecadação tributária sobre os indicadores sociais da saúde e educação em minas gerais. **Revista de Informação Contábil**, v. 5, n. 2, p. 99-121, 2011.

LACRUZ, Adonai José; AMÉRICO, Bruno Luiz; CARNIEL, Fagner. Indicadores de qualidade na educação: análise discriminante dos desempenhos na Prova Brasil. **Revista brasileira de educação,** v. 24, e240002, p. 1-26, 2019.

LAUTERT, Lisandra Veiga dos Santos; ROLIM, Matheus; LODER, Liane Ludwig. Investigando processos de retenção no âmbito de um curso de engenharia elétrica. *In*: CONGRESSO BRASILEIRO DE EDUCAÇÃO EM ENGENHARIA, 39., 2011, Blumenau. **Anais eletrônicos** [...]. Brasília: ABENGE, 2011. Disponível em: http://www.abenge.org.br/cobenge/arquivos/8/sessoestec/art2094.pdf. Acesso em: 23 set. 2020.

LOUVIERE Jordan J.; HENSHER David A; SWAIT Joffre D. **Stated choice methods**. New York: Cambridge University Press, 2000.

McLACHLAN, Geoffrey J. **Discriminant Analysis and Statistical Pattern Recognition**. New York: John Wiley & Sons, 1992.

MOTTA, Gustavo da Silva. Como escrever um bom artigo tecnológico? **Revista de Administração Contemporânea**, vol. 21, n. 5, p. 4-8, 2017.

NORONHA, Adriana Backx; CARVALHO, Beatriz Montiani; SANTOS, Fabrício Fernando Foganhole dos. **Estudo do perfil dos alunos evadidos da faculdade de economia, administração e contabilidade,** campus Ribeirão Preto, e avaliação do tempo de titulação dos alunos atualmente matriculados. Ribeirão Preto: FEA/USP, 2001. (Texto para Discussão. Série Administração).

NIU, Lian. A review of the application of logistic regression in educational research: common issues, implications, and suggestions. **Educational Review**, v. 72, n. 1, p. 41-67, 2018.

MESQUITA, Silvana. Os resultados do Ideb no cotidiano escolar. **Ensaio: Avaliação e Políticas Públicas em Educação**, v. 20, n. 76, p. 587-606, 2012.

PAULA, Gilberto A. **Modelos de regressão com apoio computacional**. São Paulo: IME/USP, 2013.

PEREIRA, Natalia Herculano. **Modelo preditivo para intervenção coronária percutânea em pacientes com infarto agudo do miocárdio com supradesnivelamento do seguimento ST**. 2014. 94f. Dissertação (Mestrado em Modelos de decisão e saúde) – Programa de Pós-Graduação em Modelos de Decisão e Saúde, Universidade Federal da Paraíba, João Pessoa, 2014.

ROCHA, Bárbara Aparecida da Silva Rego; TOLEDO JÚNIOR, Antonio. Predictive Factors of Graduation Delay in a Medical Program: a Retrospective Cohort Study in Brazil, 2010-2016. **Revista brasileira de educação médica**, v. 44, n. 1, e001, p. 1-6, 2020.

TRUETT, Jeanne; CORNFIELD, Jerome; KANNEL, William. A multivariate analysis of the risk of coronary heart disease in Framingham. **J. chron. Dis.**, v. 20, n. 7, p. 511-524, 1967.

VITELLI, Ricardo Ferreira; ROCHA, Cleonice Silveira; FRITSCH, Rosangela. **Estudo sobre evasão nos cursos de graduação de uma instituição de ensino superior privada**: aplicação de regressão logística. Programa de Observatório de Educação INEP/CAPES, Núcleo em Rede, Projeto nº 44, Indicadores de Qualidade e Gestão Democrática. 2010. Disponível em: https://anpae.org.br/simposio2011/cdrom2011/PDFs/trabalhosCompletos/comunicacoesRelatos/0456.pdf. Acesso: 23 set. 2020.

TUKEY, John Wilder. Bias and confidence in not quite large samples. **Annals of Mathematical Statistics,** v. 29, n. 2, p. 614-623, 1958.

ZHUOFAN, Wu. **Proposta de um modelo de regressão binária com resposta contínua aplicado à análise dos dados do SINASC**: identificação de fatores de risco para o baixo peso ao nascer. 2011. 76f. Dissertação (Mestrado em Saúde na Comunidade) – Programa de Pós-Graduação em Saúde na Comunidade, Universidade de São Paulo, Ribeirão Preto, 2011.

AGRADECIMENTO

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (Fapes), Edital n. 21/2018 Universal, e dos Programas Institucionais de Bolsas de Iniciação Científica (Pibic) e de Voluntariado de Iniciação Científica (Pivic) do Instituto Federal do Espírito Santo (Ifes), Edital PRPPG 02/2019 Pibic/Pivic.