

ANÁLISE COM MODELAGEM MULTINÍVEL: CONTRIBUIÇÃO PARA ESTUDO SOBRE AS VARIÁVEIS SIGNIFICANTES PARA AS NOTAS NO ENEM

Analysis with multi-level modeling: contribution to the study on significant variables for enem grades

Bárbara Castilho Máximo¹
Luiz Fernando Caldeira Ribeiro²

RESUMO

O Enem é uma das avaliações mais importantes para a educação básica do Brasil, reunindo estudantes de todos os tipos de escola, renda familiar e cor, à fim de estabelecer o ingresso ao ensino superior, seja ele em universidades públicas ou em particulares contando com bolsa ou financiamento estudantil. Contando com uma extensa base de dados sobre os inscritos do ano de 2019 e dados sobre as escolas que os mesmos estudaram em seu último ano escolar o presente trabalho teve o objetivo de testar e demonstrar quais variáveis ligadas ao aluno e à escola podem ser consideradas explicativas para a média das cinco notas que compõe a avaliação. Para esta demonstração foi desenvolvido um algoritmo supervisionado de “machine learning” conhecido como modelagem linear generalizado multinível, onde os dados referentes aos inscritos são aninhados aos dados referentes as escolas, podendo assim demonstrar que este tipo de técnica proporciona melhores resultados quando comparada a um modelo de regressão linear simples para este tipo de caso. O modelo foi construído com 36 variáveis, chegando a um valor de máxima verossimilhança restrita de 65.144,52. Além da utilidade preditiva do modelo, também permitiu uma análise socioeconômica do ingresso para o ensino superior, pois os resultados têm forte relação com dados como a raça/cor, renda, sexo e escolaridade dos pais dos alunos, também a dispositivos de acesso à informação nas escolas, como bibliotecas e computadores, permitindo assim o apontamento para fatores que podem ser aprimorados tanto no meio da educação básica quanto no social.

Palavras-chave: educação; modelo hierárquico linear; “machine learning”.

ABSTRACT

Enem is one of the most important assessments for basic education in Brazil, bringing together students from all types of school, family income and color, in order to establish admission to higher education, whether in public universities or in private ones with a scholarship. or student finance. With an extensive database on 2019 enrollments and data on the schools they studied in their last school year, this work aimed to test and demonstrate which variables linked to the student and the school can be considered explanatory. for the average of the five grades that make up the evaluation. For this demonstration, a supervised machine learning algorithm known as generalized multilevel linear modeling was developed,

¹ Universidade de São Paulo - Escola Superior de Agricultura "Luiz de Queiroz"

² Universidade do Estado de Mato Grosso

where the data referring to the subscribers are nested to the data referring to the schools, thus being able to demonstrate that this type of technique provides better results when compared to a model simple linear regression for this type of case. The model was built with 36 variables, reaching a restricted maximum likelihood value of 65,144.52. In addition to the predictive utility of the model, it also allowed a socioeconomic analysis of admission to higher education, as the results are strongly related to data such as race/color, income, gender and education of the students' parents, as well as information access devices in schools, such as libraries and computers, thus allowing the pointing to factors that can be improved both in the middle of basic and social education.

Keywords: education; linear hierarchical model; machine learning.

INTRODUÇÃO

Atualmente, o Brasil conta com uma importante avaliação educacional, o Exame Nacional do Ensino Médio (Enem), instituído em 1998, que tem o propósito de avaliar o desempenho de estudante ao fim da educação básica, tendo sua pontuação utilizada para o ingresso no ensino superior por meio de programas como o Sistema de Seleção Unificada (SISU) e o Programa Universidade para Todos (ProUni), bem como pleitear financiamento estudantil por meio de programas como o Fundo de Financiamento Estudantil (FIES) (INEP, 2021a).

O exame conta com 180 questões abrangendo quatro áreas do conhecimento, sendo elas: linguagens, códigos e suas tecnologias; ciências humanas e suas tecnologias; ciências da natureza e suas tecnologias; e matemática e suas tecnologias, podendo ser realizado por qualquer pessoa que tenha concluído o ensino médio ou esteja concluindo, nesse caso participando como treineiros (INEP, 2021a).

O Enem gera uma extensa base de dados abertos, onde é possível ter acesso aos aspectos sociais de seus participantes e suas respectivas pontuações. Além dessa base de dados, também é possível a consulta ao censo escolar, uma pesquisa estatística realizada anualmente, que é uma importante ferramenta para aplicação de políticas públicas para educação como a implantação do Fundo de Desenvolvimento do Ensino Fundamental e de Valorização do Magistério (Fundef) em 1998. Com o censo é possível ter acesso aos dados referentes a todas as escolas brasileiras, sendo públicas ou privadas (INEP, 2010).

Há um desafio para o país de melhorar a qualidade da educação básica, e por isso é importante obter um diagnóstico dos problemas relacionados para que sejam desenvolvidas estratégias e políticas de médio e longo prazo para uma na situação atual. Desde a década de

1990, houveram diversas reformas educacionais que acarretaram na importância do sistema avaliador como chave para estimular mudanças. Apesar de melhorias que aconteceram nas duas últimas décadas como baixos índices de repetência e de abandono, há o baixo desempenho dos alunos, trazendo a preocupação com a qualidade de ensino, chegando à conclusão que essa preocupação deve ser uma constante. (MALUSÁ et al. 2014)

Para ilustrar essa evolução das últimas décadas, o Índice de Desenvolvimento da Educação Básica (IDEB), que é uma das iniciativas do Inep para mensuração do desempenho do sistema educacional brasileiro (INEP, 2021b), demonstra um aumento de 55,26% no índice de 2005 a 2019 para os anos iniciais do ensino fundamental alcançando a meta de 5,9 pontos, aumento de 40% para os anos finais, ficando 0,3 pontos abaixo da meta de 5,2 e um aumento de 23,53% para o ensino médio, ficando 0,8 pontos abaixo da meta de 5,0.

A ciência de dados pode ser utilizada para extração de informações não óbvias de grandes bases de dados, a partir da definição de problemas, seu uso permite a mineração e limpeza da disponibilizada, bem como uso de aprendizado de máquina (“machine learning”) para extração de padrões desses dados, identificando agrupamentos e como esses dados se relacionam em conjunto, sendo úteis para a tomada de decisão na resolução de problemas (KALLEHER E TIERNEY, 2018). Chegando em uma hierarquia, onde a partir dos dados são geradas informações, que quando aplicadas chegam no conhecimento (FÁVERO E BELFIORE, 2017).

O crescimento do uso da ciência de dados nas últimas décadas se deu a grande capacidade de armazenamento de dados e o aumento da capacidade computacional disponível atualmente otimizando assim o uso do “machine learning” (KALLEHER E TIERNEY, 2018).

Em diversos trabalhos as variáveis que costumam ser mais examinadas para se chegar ao desempenho escolar são o nível de rendimento familiar, nível de escolaridade dos pais e a raça/cor do estudante e este padrão de distribuição tende a mostrar o padrão de desempenho escolar dos alunos (TRAVITZKI et al. 2016). Para ir além disso, tendo a possibilidade de se trabalhar com dados referentes a situação social daqueles que participam do Enem, bem como o censo escolar, surge o questionamento de quais fatores ligados a escola e sua estrutura, tem maior influência na pontuação destes alunos.

O objetivo do trabalho foi desenvolver um algoritmo de “machine learning” chamado de modelo multinível de regressão, pois de acordo com Fávero e Belfiore, (2017), este tipo de modelo leva em consideração o aninhamento dos dados entre os indivíduos, sendo os

participantes do Enem, e as escolas onde esses indivíduos se agrupam, podendo assim verificar se existem características de escola que explicam o desempenho no exame. Courgeau (2003) exemplifica que na educação existem níveis além da estrutura social e do processo de formação pessoal que não devem ser desprezados, como a classe deste aluno, as características da escola, sua família e o local onde vive, portanto é possível identificar diversos níveis, hierárquicos ou não, que exercem diversos efeitos ao comportamento do estudante.

MATERIAL E MÉTODOS

Nesta seção serão descritos os materiais e métodos para desenvolvimento dessa pesquisa, sendo definida como experimental pois se trata de um determinado objeto de estudo, no caso o desempenho de estudantes no ENEM, onde variáveis que são capazes de influenciá-lo serão selecionadas e avaliadas (GIL, 2017).

Banco de dados

Foram utilizadas duas bases de dados disponibilizadas pelo INEP, sendo os microdados referentes ao ENEM 2019, que foram extraídos em https://download.inep.gov.br/microdados/microdados_enem_2019.zip (INEP, 2019a) e os microdados do censo escolar de educação básica de 2019, extraídos em https://download.inep.gov.br/dados_abertos/microdados_censo_escolar_2019.zip (INEP, 2019b).

Os microdados do ENEM foram compostos por 136 variáveis e 5.095.270 observações. Para gerar uma amostra desta população, os dados foram filtrados em participantes que residem em 5 cidades, escolhidas por terem tamanho de população e Índice de Desenvolvimento Humano Municipal (IDHM) semelhantes, como pode ser observado na Tabela 1, sendo uma de cada região do Brasil, sudeste, nordeste, sul, centro-oeste e norte, respectivamente. A filtragem dos indivíduos foi feita levando em consideração apenas os que realizaram todas as etapas da avaliação e o preenchimento do estabelecimento que concluiu o ensino médio.

Tabela 1 - Cidades escolhida para amostra

Cidade	População estimada ¹	IDHM 2010 ¹	Inscritos ENEM ²	Inscritos filtrados ²
Sorocaba (SP)	695.328	0,798	12.919	2.924
Feira de Santana (BA)	624.107	0,712	25.200	2.815
Londrina (PR)	580.870	0,778	11.770	1.210
Aparecida de Goiânia (GO)	601.844	0,718	13.631	2.479
Porto Velho (RO)	548.952	0,736	24.575	3.339
Total			88.095	12.767

Fonte: Dados originais da pesquisa

Nota: ¹ Dados retirados do IBGE (2021); ² Dados filtrados a partir dos microdados do ENEM 2019 (INEP 2019a)

Entre as existentes na base, foram escolhidas para o desenvolvimento do algoritmo 32 variáveis, que estão apresentadas na Tabela 2, sendo algumas para a filtragem da base, referentes ao município de residência, estado e escola, as cinco notas do aluno para compor a variável dependente, bem como variáveis consideradas interessantes para influenciar este trabalho que podem apontar desigualdades entre as pessoas, sendo comuns a esse tipo de estudo como apontado por Travitzki et al. (2016), como raça/cor e renda familiar, grau de acesso à informação, como a posse de computador e acesso à internet, escolaridade de seus pais ou responsáveis.

Tabela 2 - Seleção de variáveis utilizadas no desenvolvimento da modelagem multinível retiradas do dicionário de dados dos microdados do ENEM 2019

Nome da variável	Descrição
CO_MUNICIPIO_RESIDENCIA	Código do município de residência
NU_IDADE	Idade
TP_SEXO	Sexo
TP_COR_RACA	Cor/raça
TP_NACIONALIDADE	Nacionalidade
TP_ST_CONCLUSAO	Situação de conclusão do Ensino Médio
TP_ENSINO	Tipo de instituição que concluiu ou concluirá o Ensino Médio
CO_ESCOLA	Código da Escola (Censo Escolar da Educação Básica)
TP_LOCALIZACAO_ESC	Localização (Escola)
IN_SEM_RECURSO	Indicador de inscrito que não requisitou nenhum recurso
NU_NOTA_CN	Nota da prova de Ciências da Natureza
NU_NOTA_CH	Nota da prova de Ciências Humanas
NU_NOTA_LC	Nota da prova de Linguagens e Códigos
NU_NOTA_MT	Nota da prova de Matemática
NU_NOTA_REDACAO	Nota da prova de redação
Q001	Até que série seu pai, ou o homem responsável por você, estudou?
Q002	Até que série sua mãe, ou a mulher responsável por você, estudou?
Q006	Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)
Q022	Na sua residência tem telefone celular?
Q024	Na sua residência tem computador?
Q025	Na sua residência tem acesso à Internet?

Fonte: INEP, 2019a

Os microdados referentes ao censo escolar contem 274 variáveis e 228.521 observações, a base foi filtrada pelas escolas presentes na amostra dos microdados do ENEM. Esta seleção, com 24 opções, listadas na Tabela 3, foi realizada com o objetivo de listar os dados relacionados as condições de saneamento, acesso à informação dentro da escola, como bibliotecas, laboratórios e computadores, tipo de administração e tamanho da instituição, sendo pontos que podem ser apontados para melhorias e mudanças de políticas públicas na área da educação.

Tabela 3 - Seleção de variáveis utilizadas no desenvolvimento da modelagem multinível retiradas do dicionário de dados dos microdados do censo escolar da educação básica 2019

Nome da variável	Descrição
CO_ENTIDADE	Código da Escola
TP_DEPENDENCIA	Dependência Administrativa
IN_ÁGUA_POTAVEL	Fornece água potável para o consumo humano
IN_ESGOTO_INEXISTENTE	Esgoto sanitário - Não há esgotamento sanitário
IN_BIBLIOTECA	Dependências físicas existentes e utilizadas na escola - Biblioteca
IN_LABORATORIO_CIENCIAS	Dependências físicas existentes e utilizadas na escola - Laboratório de ciências
IN_LABORATORIO_INFORMATICA	Dependências físicas existentes e utilizadas na escola - Laboratório de informática
IN_QUADRA_ESPORTES	Dependências físicas existentes e utilizadas na escola - Quadra de esportes coberta ou descoberta
IN_REFEITORIO	Dependências físicas existentes e utilizadas na escola - Refeitório
IN_ACESSIBILIDADE_INEXISTENTE	Recursos de acessibilidade para pessoas com deficiência ou mobilidade reduzida nas vias de circulação interna na escola - Nenhum dos recursos de acessibilidade listado
QT_SALAS_UTILIZADAS	Número de salas de aula utilizadas na escola (dentro e fora do prédio)
QT_SALAS_UTILIZA_CLIMATIZADAS	Condições das salas de aula utilizadas na escola (dentro e fora do prédio escolar) - Número de salas de aula climatizadas
QT_SALAS_UTILIZADAS_ACESSIVEIS	Condições das salas de aula utilizadas na escola (dentro e fora do prédio escolar) - Número de salas de aula com acessibilidade para pessoas com deficiência ou mobilidade reduzida.
IN_COMPUTADOR	Equipamentos existentes na escola - Computador
IN_DESKTOP_ALUNO	Computadores em uso pelos alunos (as) - Computador de mesa (desktop)
IN_COMP_PORTATIL_ALUNO	Computadores em uso pelos alunos (as) - Computador portátil
IN_TABLET_ALUNO	Computadores em uso pelos alunos (as) - Tablet
IN_INTERNET	Acesso à Internet
IN_ALIMENTACAO	Alimentação escolar para os alunos - PNAE/FNDE

Fonte: INEP, 2019b

Modelagem multinível de regressão

Um modelo multinível de regressão, também conhecido como modelo hierárquico linear ("hierarchical linear models (HLM)"), considerou o agrupamento natural dos dados, nessa base os fatores das escolas de educação básica não variam entre os alunos de uma mesma escola. Isso caracteriza o aninhamento das observações.

Este tipo de modelo permite a análise das heterogeneidades individuais, bem como entre os agrupamentos em eles fazem parte e com isso permitindo especificar componentes aleatórios ao nível dos estudantes e das escolas (FÁVERO E BELFIORE, 2017). Os componentes aleatórios são interessantes para tirar conclusões de uma população a partir de uma amostra (SNIJDERS E BOSKER, 2012).

A variável dependente, ou seja, o fenômeno a ser estudado, foi definida a partir média das notas das cinco áreas da avaliação (NU_NOTA_CN, NU_NOTA_CH, NU_NOTA_LC, NU_NOTA_MT, NU_NOTA_REDACAO), essa média aritmética comum é utilizada como nota para o ProUni (MEC, 2021), já para o SISU foi utilizada a média aritmética ponderada, em que os pesos diferem para cada curso e são definidos por cada instituição de ensino superior (MEC, 2020). Demais variáveis a serem testadas no modelo foram denominadas de explicativas.

Para o desenvolvimento desse modelo, foi utilizado o software R v. 4.1.0 (R Foundation, 2021) em conjunto com o pacote nlme (PINHEIRO et al. 2021). O primeiro passo realizado foi a geração de um modelo nulo, utilizando o método de máxima verossimilhança restrita, ("restricted estimation of maximum likelihood (REML)"), estimando a variância dos erros, onde nenhuma variável explicativa é inserida, que pode ser representado pela equação matemática abaixo.

Nível 1:

$$notas_{ij} = b_{0j} + r_{ij} \quad (1)$$

Nível 2:

$$b_{0j} = \gamma_{00} + u_{0j} \quad (2)$$

Gerando por fim

$$notas_{ij} = \gamma_{00} + u_{0j} + r_{ij} \quad (3)$$

Onde, b_{0j} é um intercepto do segundo nível, somado aos termos de erro, r_{ij} e u_{0j} , de ambos os níveis.

A partir desse momento, foi possível realizar a comparação com um modelo de regressão linear nulo comum e verificar a partir de um teste de razão de verossimilhança, que pode ser chamado de “LR Test (Likelihood Ratio Test)”, sendo possível rejeitar a hipótese nula de que os interceptos forem iguais a zero, levando-se em consideração um nível de significância de 5% (FÁVERO e BELFIORE, 2017).

Após eliminação da hipótese nula, foram desenvolvidos modelos já considerando os interceptos aleatórios de variáveis explicativas de alunos, testando se são significantes para a ocorrência das notas finais, para isso, todas as variáveis que são do tipo qualitativa foram transformadas, onde cada categoria da mesma se torna uma variável binária (“dummy”), assumindo os valores de 0 ou 1, sendo 1 quando a categoria de interesse está presente e 0 quando não, sendo geradas novas variáveis, onde k é o número de categorias (FÁVERO e BELFIORE, 2017).

Por conta da demanda computacional para a geração dos modelos, a cada teste foi adicionada uma variável na lista de interceptos, podendo assim comparar seus p-valores, se tem nível de significância de 5% ($p\text{-valor} < 0,05$) ou não, ou se são alterados após a entrada de um novo intercepto.

Após chegar em uma listagem de interceptos aleatórios, verificou-se a inclinação das variáveis referentes aos alunos como de acordo com as escolas e foi realizado testes com as mesmas para incluir as inclinações aleatórias ao modelo, permitindo dessa forma chegar ao modelo final.

RESULTADOS E DISCUSSÃO

Visualização da base de dados

Para fim de análise da base de dados, é interessante observar a relação da média entre as variáveis escolhidas para estudo para compreender como as notas estão distribuídas entre as mesmas, podendo assim auxiliar na discussão do estudo.

A Figura 1 demonstra equilíbrio entre os quartis por cidade, apenas a cidade de Londrina (PR) que tem um maior destaque por conta do 3º quartil mais elevado.

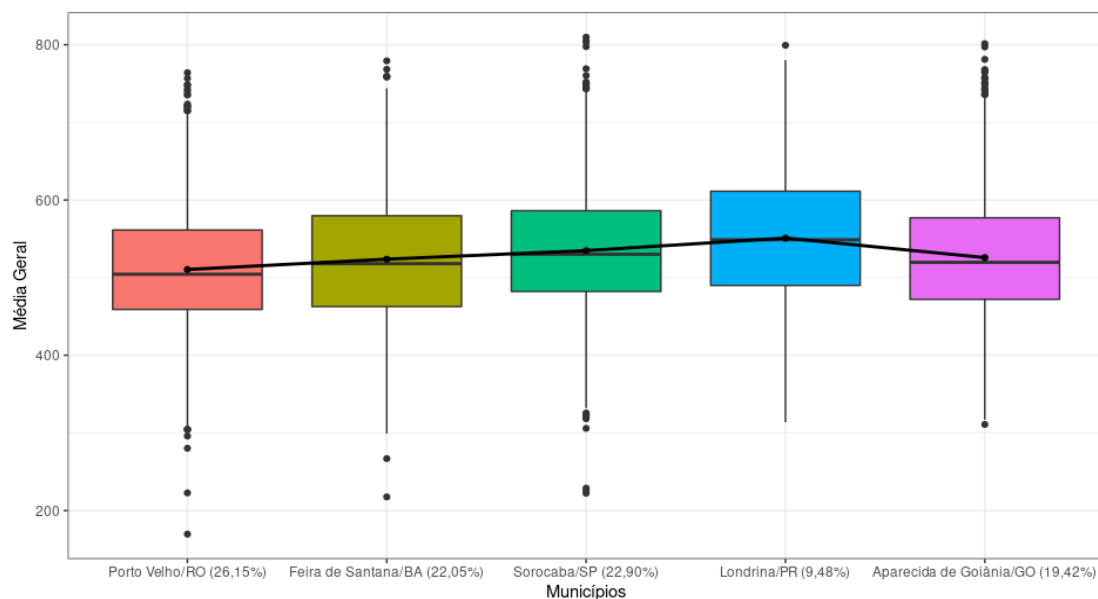


Figura 1 - Representação dos quartis da média geral por município

Fonte: Resultados originais da pesquisa

A Figura 2 demonstra a distribuição por sexo, mostrando a mediana mais elevada para o sexo masculino, com uma diferença de 17,75 pontos. Esse padrão se repete nas cinco cidades presentes no estudo, como pode ser visto na Figura 3. Araújo (2019) demonstrou mesmo padrão em seu trabalho com dados de 2015 a 2017 na cidade de Viçosa (MG) com as notas das provas objetivos, porém nas notas de redação são maiores entre as pessoas do sexo feminino.

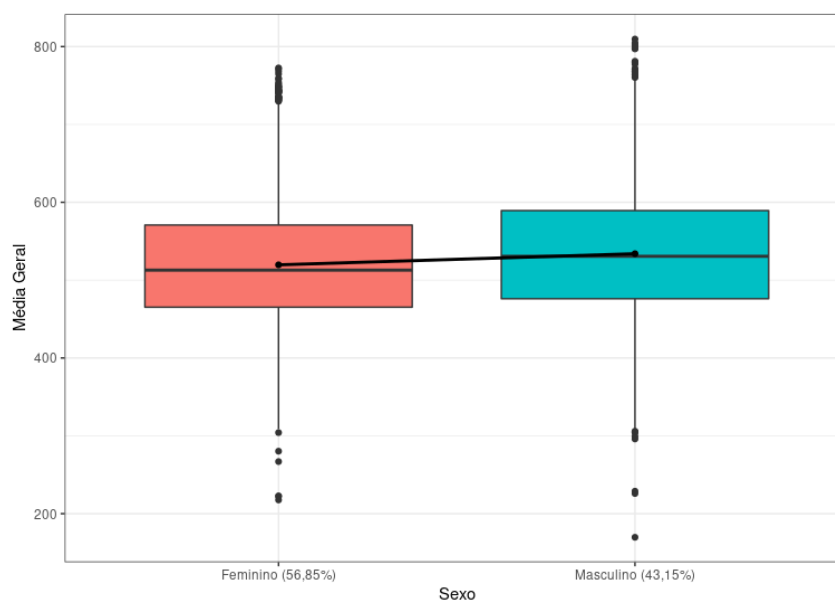


Figura 2 - Representação dos quartis da média geral por sexo

Fonte: Resultados originais da pesquisa

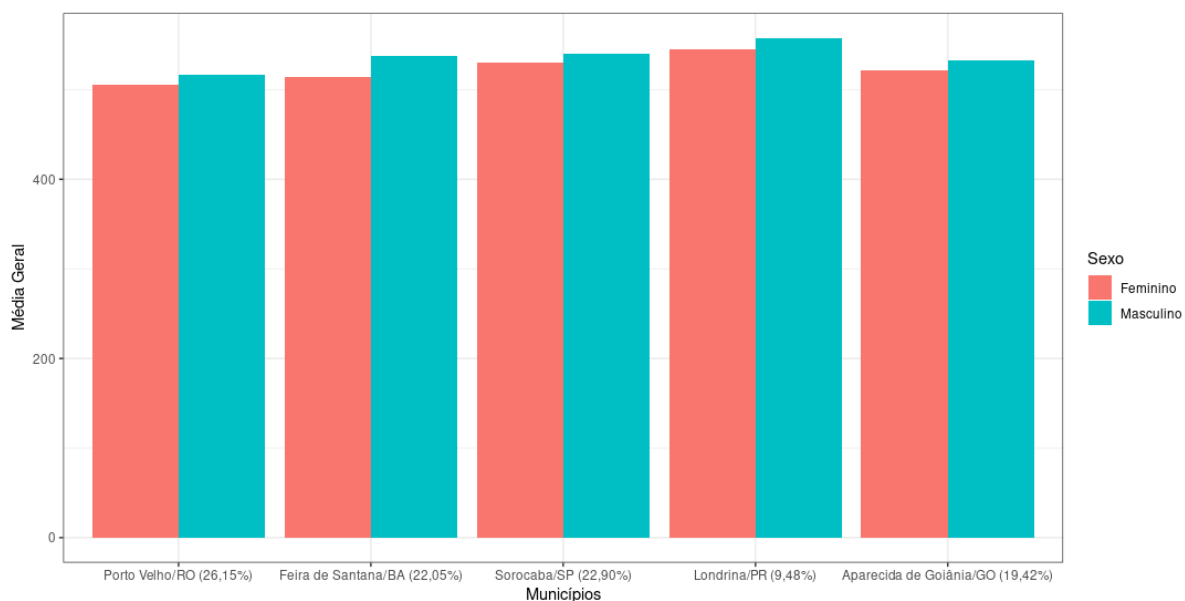


Figura 3 - Distribuição das médias das notas por sexo entre os municípios

Fonte: Resultados originais da pesquisa

Na Figura 4 pode se observar o maior destaque para as notas vindas de inscritos que se declararam como brancos, onde os valores de 3º quartis dos declarados como pardos e negros ficam próximos do valor de sua mediana

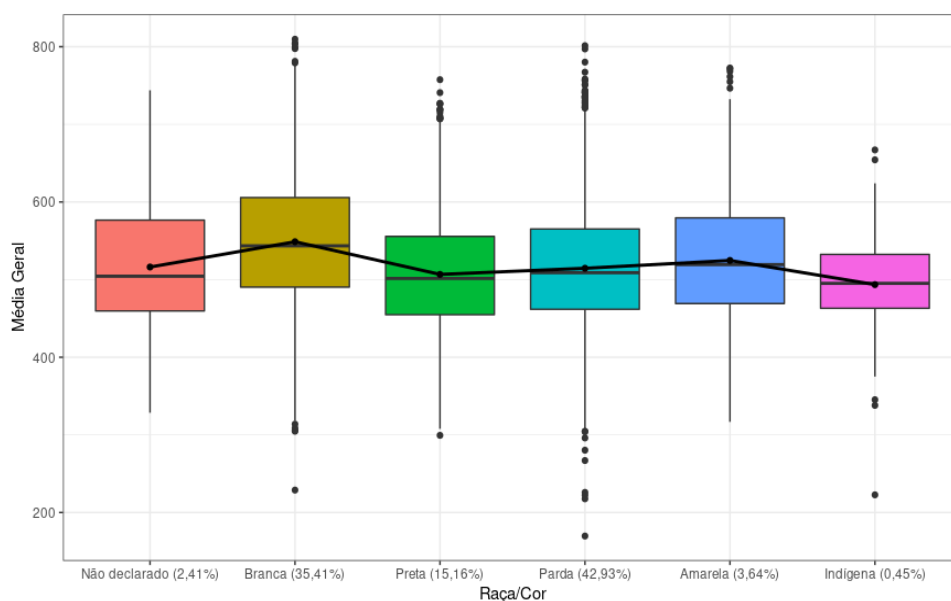


Figura 4 - Representação dos quartis da média geral por raça/cor

Fonte: Resultados originais da pesquisa

A Figura 5 demonstra que este padrão está presente nas cidades Porto Velho (RO), Feira de Santana (BA) e Aparecida de Goiânia (GO), nas demais os indivíduos declarados amarelos tem as maiores médias.

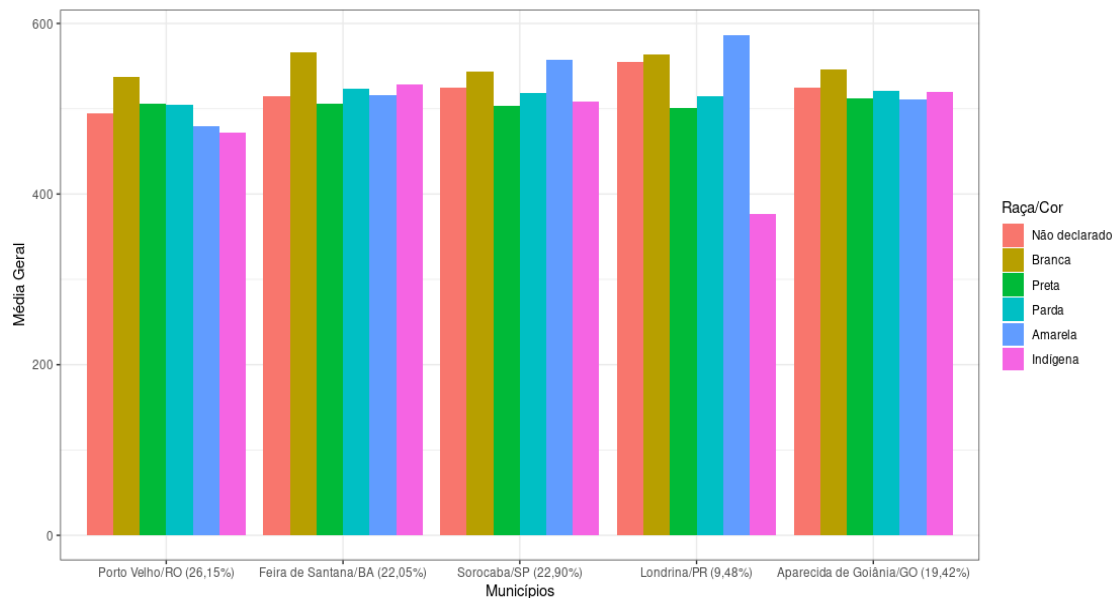


Figura 5 - Distribuição das médias das notas por raça/cor entre os municípios
Fonte: Resultados originais da pesquisa

De acordo com o IBGE (2010) a cidade de Londrina (PR) é a que tem a maior população de pessoas amarelas entre as listadas, como pode ser visto na Figura 6, contando com 17.448 habitantes em 2010, apesar de ter mais de 70% de sua população como branca, a baixa porcentagem de indígenas realizando a prova vai de acordo com os habitantes das cidades, que variam de cerca de 500 a 1400.

Comparando os dados das Figuras 5 e 6 também é notável que pessoas declaradas brancas não são a maioria nas cidades onde tiveram as maiores pontuações. Os estudos realizados por Araújo (2019) demonstram correlação positiva entre as notas e os autodeclarados brancos, colaborando também para demonstrar o contexto de exclusão racial presente no país (MELO et al. 2021). Outro fator a ser observado é que as duas cidades com maior porcentagem de pessoas brancas também têm os melhores desempenhos (Figura 1) como também é apresentado na pesquisa de Melo et al. (2021) com os dados do Enem realizado em 2018.

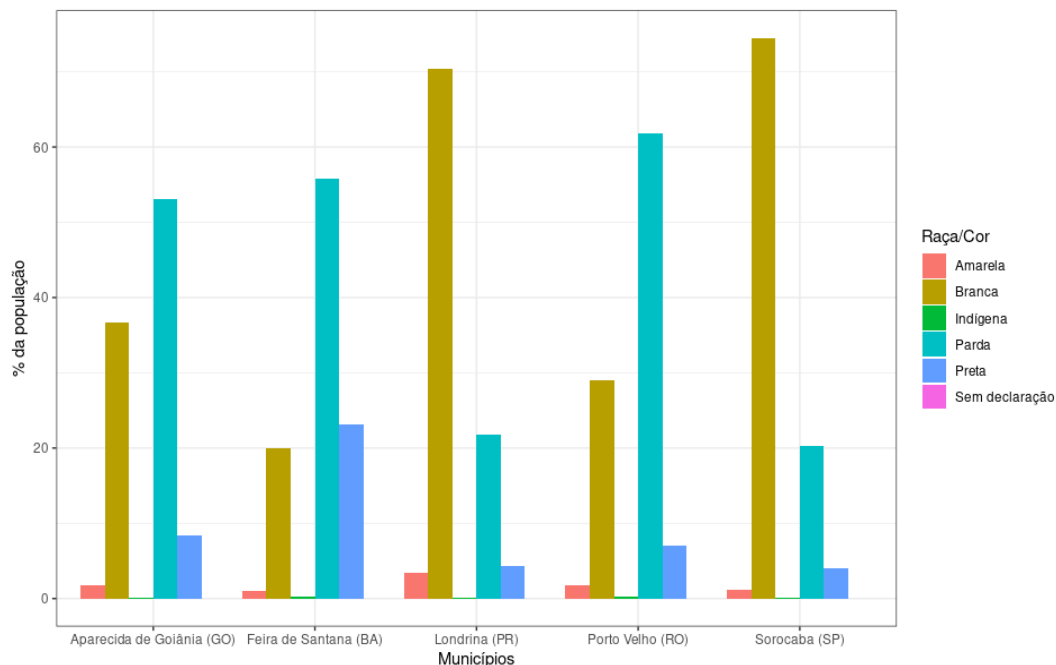


Figura 6 - Porcentagem da população total distribuída por raça/cor por município
Fonte: Dados retirados do IBGE (2010)

A renda dos participantes demonstra ser crucial para os resultados, a Figura 7 apresenta que os quartis das notas aumentam de acordo com as faixas. De acordo com Osorio (2009) há uma relação entre as classes sociais e raça e a conclusão do ensino médio, onde a porcentagem de negros concluintes é bem menor do que de pessoas brancas, e a diferença aumenta conforme a diminuição de renda, a mesma relação também é demonstrada entre os que já faziam parte do ensino superior, demonstrando assim que o acesso ao ensino superior não é equitativo no ponto de vista de classe ou raça/cor.

Esses dados reforçam a necessidade de políticas no sistema educativo do país que vão além da educação (TRAVITZKI et al. 2016).

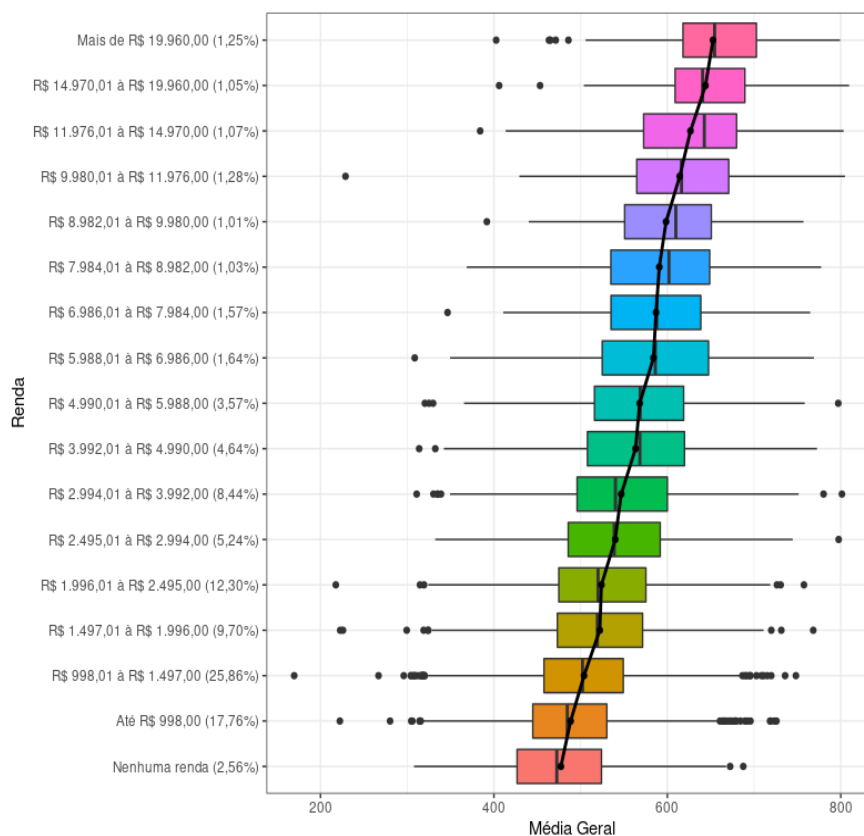


Figura 7 - Representação dos quartis da média geral por renda mensal familiar
 Fonte: Resultados originais da pesquisa

Uma maior diferença entre os quartis está apresentada na Figura 8, onde a mediana das escolas com administração privada já se encontra acima da média de 600 pontos e está quartil acima de todos os pontos das escolas com administração estadual e municipal.

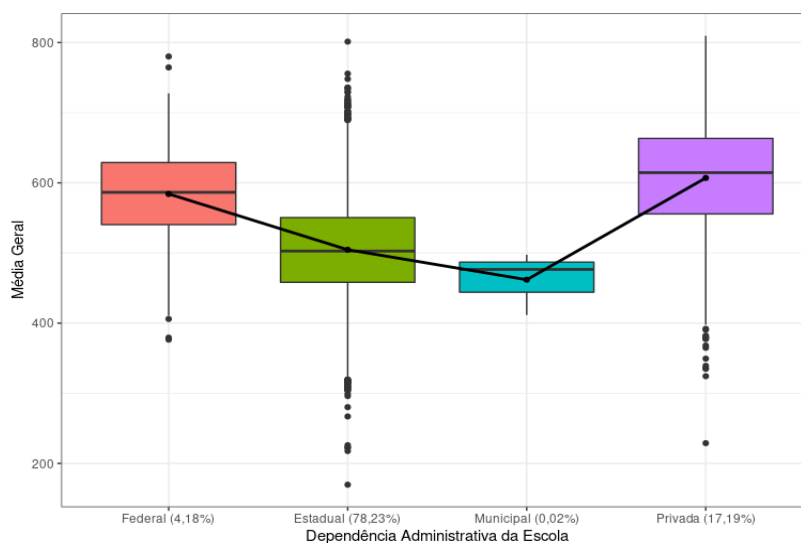


Figura 8 - Representação dos quartis da média geral por dependência administrativa da escola
 Fonte: Resultados originais da pesquisa

Quando visualizado por cidade, na Figura 9, as médias de escolas federais e privadas se mantêm bem próximas, porém as primeiras não estão presentes em todos os locais. A modelagem de regressão hierárquica de Araújo (2019) demonstrou que a dependência administrativa da escola foi o fator de maior relevância no desempenho dos alunos, porém nesse caso as federais têm o maior desempenho, seguido das privadas e estaduais.

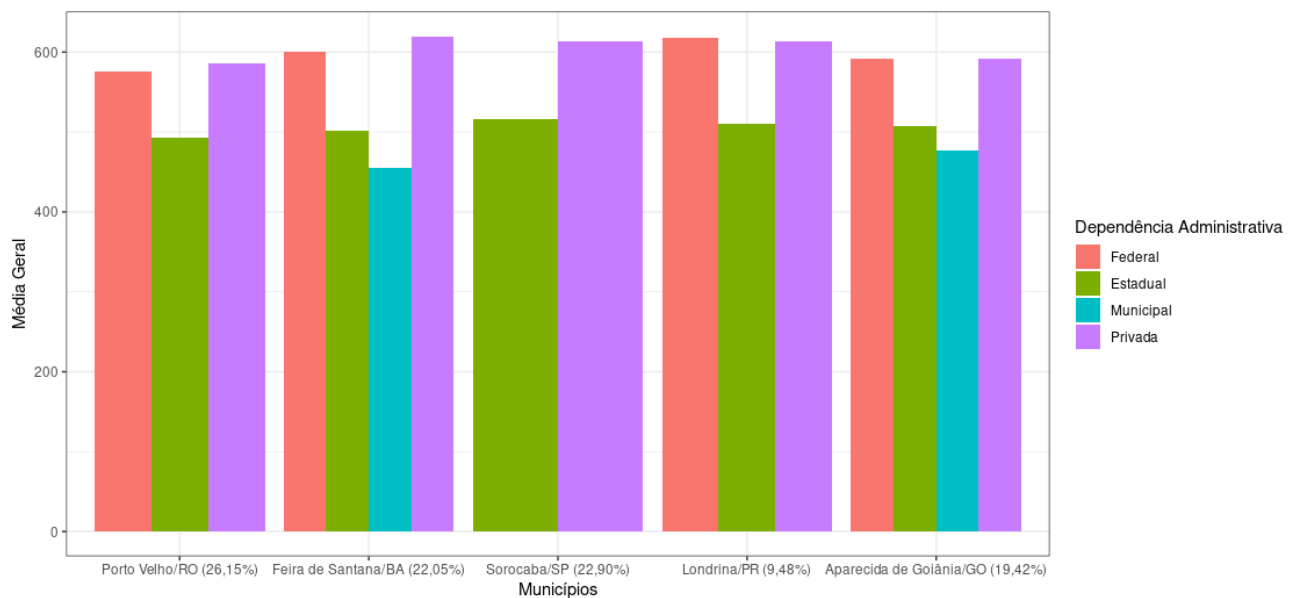


Figura 9 - Distribuição das médias das notas por dependência administrativa da escola entre os municípios

Fonte: Resultados originais da pesquisa

Modelagem

O primeiro teste a ser realizado foi a comparação entre os modelos nulos de HLM e de regressão linear comum, onde o LR Test apontou um resultado de -71.621 para o primeiro e -74.348 para o segundo, o HLM teve o maior resultado, podendo assim comprovar que os interceptos são diferentes de zero, rejeitando assim a hipótese nula.

Após testar todas as variáveis listadas nas Tabelas 2 e 3, sendo possível aumentar o valor de máxima verossimilhança restrita para $-65.146,56$. Na tabela 4 pode-se observar quais variáveis foram consideradas no modelo com interceptos aleatórios, dando destaque para o maior valor de intercepto representado pela TP_COR_RAÇA_1 com a categoria Branco e a TP_DEPENDENCIA_2, tendo o maior valor negativo, com a categoria Estadual.

Tabela 4 - Variáveis testadas como interceptos que tiveram p-valor aceito considerando nível de confiança de 95%

Variável	Valor do Intercepto	Tipo	Descrição da categoria representada pela <i>dummy</i>
TP_SEXO_M	4,917	<i>Dummy</i>	Masculino
TP_COR_RACA_1	144,752	<i>Dummy</i>	Branco
TP_NACIONALIDAD E_1	16,527	<i>Dummy</i>	Brasileiro(a)
Q001_E	6,578	<i>Dummy</i>	Completo o Ensino Médio, mas não completou a Faculdade.
Q001_F	9,728	<i>Dummy</i>	Completo a Faculdade, mas não completou a Pós-graduação.
Q001_G	8,273	<i>Dummy</i>	Completo a Pós-graduação.
Q002_E	3,408	<i>Dummy</i>	Completo o Ensino Médio, mas não completou a Faculdade.
Q002_F	9,172	<i>Dummy</i>	Completo a Faculdade, mas não completou a Pós-graduação.
Q006_C	5,431	<i>Dummy</i>	De R\$ 998,01 até R\$ 1.497,00.
Q006_D	11,494	<i>Dummy</i>	De R\$ 1.497,01 até R\$ 1.996,00.
Q006_E	8,911	<i>Dummy</i>	De R\$ 1.996,01 até R\$ 2.495,00.
Q006_F	15,394	<i>Dummy</i>	De R\$ 2.495,01 até R\$ 2.994,00.
Q006_G	16,780	<i>Dummy</i>	De R\$ 2.994,01 até R\$ 3.992,00.
Q006_H	17,750	<i>Dummy</i>	De R\$ 3.992,01 até R\$ 4.990,00.
Q006_I	20,630	<i>Dummy</i>	De R\$ 4.990,01 até R\$ 5.988,00.
Q006_J	23,658	<i>Dummy</i>	De R\$ 5.988,01 até R\$ 6.986,00.
Q006_K	20,691	<i>Dummy</i>	De R\$ 6.986,01 até R\$ 7.984,00.
Q006_L	18,568	<i>Dummy</i>	De R\$ 7.984,01 até R\$ 8.982,00.
Q006_M	18,886	<i>Dummy</i>	De R\$ 8.982,01 até R\$ 9.980,00.
Q006_N	20,036	<i>Dummy</i>	De R\$ 9.980,01 até R\$ 11.976,00.
Q006_O	26,380	<i>Dummy</i>	De R\$ 11.976,01 até R\$ 14.970,00.

Q006_P	36,614	<i>Dummy</i>	De R\$ 14.970,01 até R\$ 19.960,00.
Q006_Q	34,936	<i>Dummy</i>	Mais de R\$ 19.960,00.
Q022_A	-12,858	<i>Dummy</i>	Não.
Q022_B	-6,513	<i>Dummy</i>	Sim, um.
Q025_B	5,716	<i>Dummy</i>	Não.
Q024_A	-9,235	<i>Dummy</i>	Sim, dois.
Q024_C	10,999	<i>Dummy</i>	Sim, três.
Q024_D	9,347	<i>Dummy</i>	Sim, quatro ou mais.
Q024_E	17,172	<i>Dummy</i>	Sim.
TP_ENSINO_2	-35,477	<i>Dummy</i>	Educação Especial - Modalidade Substitutiva
TP_LOCALIZACAO_E SC_2	-23,046	<i>Dummy</i>	Rural
TP_DEPENDENCIA_2	-76,309	<i>Dummy</i>	Estadual
IN_BIBLIOTECA	10,355	<i>Dummy</i>	-
IN_COMPUTADOR	66,799	<i>Dummy</i>	-
QT_SALAS_UTILIZA DAS	0,581	Númeri ca	-

Fonte: Resultados originais da pesquisa

A variável TP_COR_RACA_1 foi incluída como inclinação aleatória do modelo, permitindo que o valor de máxima verossimilhança restrita chagasse à -65.144,52.

Na Tabela 5 estão representados os erros-padrão das variâncias no componente de efeitos aleatórios do modelo produzido, o modelo demonstrou ter variação significativa a 95% de confiança no desempenho das notas em decorrência dos interceptos aleatórios do modelo ($\text{Var}(v_{0j})$), porém não obteve significância nas inclinações aleatórias ($\text{Var}(v_{1j})$), onde 14,18% da variância total do desempenho é devido ao interceptos e 1,23% devido a inclinação.

Tabela 5 - Erros-padrão dos efeitos aleatórios

Componentes	Estimativas	Erros-padrão	z	P-valor
Var(v_{0j})	630,11078	71,96624	8,755644	0,000
Var(v_{1j})	54,97221	37,40965	1,469466	0,142
Var(e)	3759,82718	50,42784	74,55856	0,000

Fonte: Resultados originais da pesquisa

Apesar das inclinações não terem atingido o P-valor, foram mantidas nos testes de predição por conta do aumento do valor de REML. A Tabela 6 compara as medias de notas por cidade com as médias estimadas a partir do modelo final.

Tabela 6 - Valores estimados pelo modelo por município

Município	Média Geral	Mediana	Média Estimada	Geral	Mediana Estimada
Aparecida de Goiânia (GO)	530,361	525,140	528,714		516,833
Feira de Santana (BA)	527,234	520,470	526,852		513,031
Londrina (PR)	553,950	550,100	555,573		536,043
Porto Velho (RO)	521,376	516,060	522,904		516,544
Sorocaba (SP)	536,022	531,320	535,500		518,836

Fonte: Resultados originais da pesquisa

A distribuição dos valores estimados pelos reais pode ser visualizada na Figura 10, é possível notar que o modelo tem tendência linear, porém ainda pode ser melhor ajustado continuando a bateria de testes no modelo com diferentes combinações de variáveis, possibilitando assim a diminuição dos termos de erro.

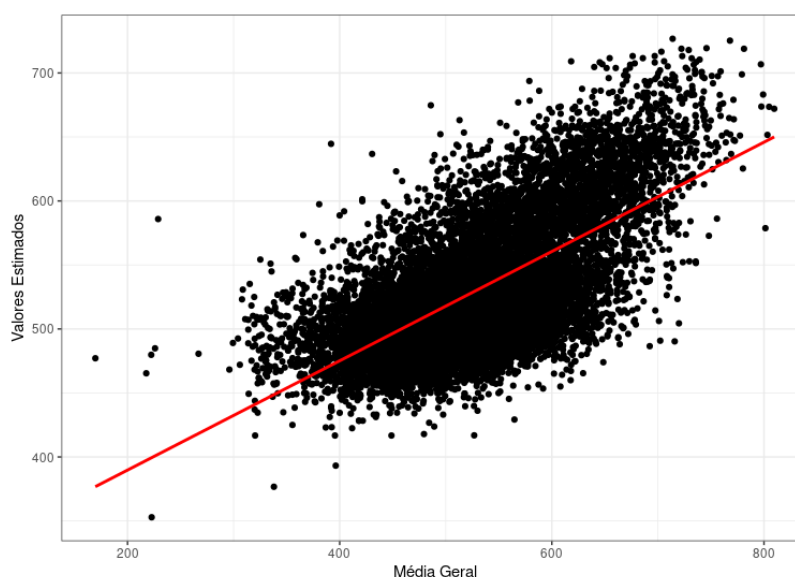


Figura 10 - Distribuição linear dos valores estimados a partir do modelo

Fonte: Resultados originais da pesquisa

Na Figura 11 demonstrando o aninhamento dos resultados entre três escolas presentes no estudo.

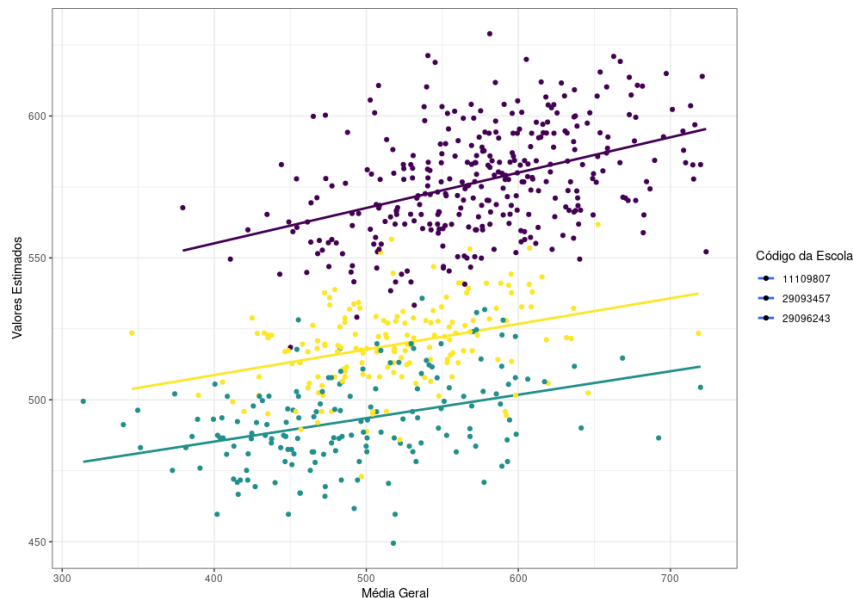


Figura 11 - Distribuição linear dos valores estimados a partir do modelo por escola
Fonte: Resultados originais da pesquisa

Em comparação com outros trabalhos, pode se notar semelhanças nas variáveis que foram adotadas, Travitzki et al. (2016) apresenta um modelo hierárquico com 3 níveis (aluno, escola e município) onde teve melhores resultados do que com 2 níveis, com dados do Enem de 2009 à 2012 sendo utilizado os fatores de renda, educação dos pais e raça/cor, demonstra também que o “efeito-escola” no Brasil ainda atenua desigualdades no país, contrariando resultados publicados em 1990, porem Travitzki (2013) também conclui que o Enem não é adequado para explicar o quanto as escolas contribuem para a democracia do país, sendo imprudente a supervalorização de seus resultados. Araújo (2019) também demonstra melhores resultados com modelos de 3 níveis, onde o terceiro no caso é o ano de realização da prova, o que não se aplicaria a este estudo.

Silva Filho (2017) apresenta um modelo de regressão logística, com dados do Enem de 2014, onde existem semelhanças entre as variáveis explicativas do nível escola, em seu foi analisada a probabilidade de sucesso ou não sucesso dos alunos, porem o estudo contou apenas com escolas de administração federal, que tem bom desempenho, ficando abaixo apenas das particulares (Figura 8), sua modelagem também teve foco na titulação dos professores e fatores ligados a estrutura da escola, em contrapartida utilizando o mesmo tipo de regressão.

Adeodato (2016) calcula probabilidade da escola se boa ou ruim, com base no quartil superior das notas, seu estudo trata apenas de escolas particulares e dá destaque para variáveis como a formação de professores, laboratórios de ciências e informática entre outras coisas em sua maioria baseadas em condições econômicas e financeiras das instituições de ensino, o mesmo da destaque para o fato que parte das estruturas das escolas, como os computadores, podem não ser utilizados para fins educacionais, por isso precisam ser interpretadas com cuidado.

A utilização de variáveis referentes a educação dos pais é reforçada por Araújo (2019), em sua modelagem a escolaridade da mãe afeta mais o modelo do que a escolaridade do pai, o uso desse dado também é reforçado por Travitzki (2013).

Pode se notar certos padrões como a utilização de variáveis semelhantes, permitindo o levantamento de discussões sobre a desigualdade social ainda tão latente, mesmo com anos de diferença entre as demais bases de dados. Os resultados obtidos ainda podem ser melhorados caso se opte em utilizar outros tipos de modelagem em conjunto com o HLM, como por exemplo uso de modelos não supervisionados para diminuição da dimensionalidade da base de dados, bem como a inclusão de mais níveis.

CONCLUSÃO

O trabalho demonstrou a efetividade de um modelo linear generalizado multinível com dados aninhados.

Os resultados apresentaram uma importante relação entre diversos dados socioeconômicos dos alunos e também a influência de bens que fornecem acesso à informação, dados que estão relacionados com a escola também acabam seguindo essa linha, com a relação entre existência de computadores e bibliotecas nas escolas.

O estudo colaborou em apontar que a desigualdade social e racial no Brasil ainda, em todas as suas cinco regiões, é de grande impacto para o ingresso no ensino superior.

As maiores dificuldades encontradas para desenvolvimento deste projeto foram relacionadas a capacidade computacional disponível para testes, fazendo com os mesmos demorassem bastante para serem realizados, limitando assim o tempo disponível para realizar comparações de resultados com outros tipos de modelos.

REFERÊNCIAS

Adeodato, P. J. L. 2016. Data mining solution for assessing brazilian becondary school quality based on ENEM and census data. In: 13th CONTECSI - International Conference on Information Systems & Technology Management, 2016, São Paulo, SP, Brasil. Anais... p. 1112-1124.

Araújo, D.L. 2019. Determinantes do desempenho no ENEM dos concluintes do ensino médio no município de Viçosa-MG. Dissertação de Mestrado Profissional em Administração Pública em Rede Nacional. Universidade Federal de Viçosa, Viçosa, MG, Brasil.

Courgeau, D. 2003. Methodology and epistemology of multilevel analysis. 1ed. Kluwer Academic Publishers, London, UK.

Fávero, L. P.; Belfiore, P. 2017. Manual de análise de dados. 1ed. Elsevier, Rio de Janeiro, RJ, Brasil.

Gil, A. C. 2017. Como elaborar projetos de pesquisa. 6ed. Atlas, Rio de Janeiro, RJ, Brasil.
Instituto Brasileiro de Geografia e Estatística (IBGE). 2010. Censo demografico. Disponível em <<https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografico-2010.html?edicao=10503&t=resultados>>. Acesso em: 01 mai. 2022.

Instituto Brasileiro de Geografia e Estatística (IBGE). 2021. Cidades. Disponível em <<https://cidades.ibge.gov.br>>. Acesso em: 02 nov. 2021.

Instituto Nacional De Estudos E Pesquisas Educacionais Anísio Teixeira (INEP). 2010. A importância do Censo Escolar. Disponível em <<https://www.gov.br/inep/pt-br/assuntos/noticias/censo-escolar/a-importancia-do-censo-escolar>>. Acesso em: 16 out. 2021.

Instituto Nacional De Estudos E Pesquisas Educacionais Anísio Teixeira (INEP). 2019a. Microdados do Enem 2019. Disponível em <https://download.inep.gov.br/microdados/microdados_enem_2019.zip>. Acesso em: 30 out. 2021.

Instituto Nacional De Estudos E Pesquisas Educacionais Anísio Teixeira (INEP). 2019b. Microdados do censo escolar da educação básica 2019. Disponível em <https://download.inep.gov.br/dados_abertos/microdados_censo_escolar_2019.zip>. Acesso em: 14 mai. 2022.

Instituto Nacional De Estudos E Pesquisas Educacionais Anísio Teixeira (INEP). 2021a. Enem. Disponível em <<https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>>. Acesso em: 16 out. 2021.

Instituto Nacional De Estudos E Pesquisas Educacionais Anísio Teixeira (INEP). 2021b. Resultados do índice de desenvolvimento da educação básica | 2019: Resumo técnico. Disponível em <https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/resultados>

[_indice_desenvolvimento_educacao_basica_2019_resumo_tecnico.pdf](#)>. Acesso em: 17 out. 2021.

Kelleher, J.D.; Tierney, B. 2018. Data science. 1ed. The MIT Press, Cambridge, MA, USA.

Malusá, S.; Ordones, L.L.M.; Ribeiro, E. 2014. Enem: Pontos positivos para a educação brasileira. Revista Educação e Políticas em Debate. 3: 358-382.

Melo, R.O.; Freitas, A.C; Francisco, E.R.; Motokane, M. T. 2021. Impacto das variáveis socioeconômicas no desempenho do Enem: uma análise espacial e sociológica. Revista de Administração Pública (RAP). 55 (6): 1271-1294.

Ministério da Educação (MEC). 2020. SISU: Notas do ENEM. Disponível em <https://sisu.mec.gov.br/#/duvidas#nota_enem>. Acesso em: 01 nov. 2021.

Ministério da Educação (MEC). 2021. Tire suas dúvidas: ProUni e ENEM. Disponível em <<http://prouniportal.mec.gov.br/tire-suas-duvidas-pesquisa/prouni-e-enem>>. Acesso em: 01 nov. 2021.

Osorio, R. G. 2009. Classe, raça e acesso ao ensino superior no Brasil. Cadernos de Pesquisa. 39: 867-880.

Pinheiro, J.; Bates, D.; DebRoy, S.; Sarkar D.; R Core Team. 2021. nlme: Linear and Nonlinear Mixed Effects Models. Disponível em <<https://CRAN.R-project.org/package=nlme>>. Acesso em: 03 nov. 2021.

R Foundation. 2021. R Project. Disponível em <<https://www.r-project.org/>>. Acesso em: 03 nov. 2021.

Silva Filho, R. L. C. 2017. Modelo de análise e predição do desempenho dos alunos dos Institutos Federais de Educação usando o ENEM como indicador de qualidade escolar. Dissertação de Mestrado Profissional em Ciência da Computação. Universidade Federal de Pernambuco, Recife, PE, Brasil.

Snijders, T. A. B.; Bosker, R. J. 2012. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. 2ed. SAGE Publications, City Road, London, UK.

Travitzki, R. 2013. ENEM: limites e possibilidades do Exame Nacional do Ensino Médio enquanto indicador de qualidade escolar. Tese de Doutorado em Educação. Universidade de São Paulo, São Paulo, SP, Brasil.

Travitzki, R.; Couto, A. P.; Ferrão, M. E. 2016. Desigualdades educacionais e socioeconômicas na população brasileira pré-universitária: Uma visão a partir da análise de dados do ENEM. Arquivos Analíticos de Políticas Educativas. 24: 1-32.