ISSN: 2317-0840

Predição do Desempenho de Estudantes por Regressão Múltipla

Adenilson A. Novaes[1], Jean Piton-Gonçalves[2†]

[1] *Raccoon Publicidade Ltda.*
[2] *Departamento de Matemática/Universidade Federal de São Carlos (UFSCar).*

Resumo: *No contexto da Ciência de Dados Educacionais, a Predição de Desempenho Acadêmico de Estudantes pode seguir a Mineração de Dados Educacionais, que busca explicitar quantitativamente o desempenho estudantil, norteando professores e instituições de ensino. A Regressão Linear Múltipla é uma metodologia de predição que pode ser aplicada a dados educacionais, como é o caso de dados do Exame Nacional do Ensino Médio (ENEM). Partindo dos dados do ENEM edição 2019, esta pesquisa formulou, testou e analisou sete modelos de regressão múltipla partindo de uma amostra de 18.908 candidatos. Tais modelos consideraram os escores das provas de (i) Linguagens, Códigos e suas Tecnologias, (ii) Matemática e suas Tecnologias, (iii) Ciências da Natureza e suas Tecnologias e (iv) Ciências Humanas e suas Tecnologias e (v) Redação; e os dados pessoais (iv) idade, (v) sexo e (vi) se concluiu o Ensino Médio em escola pública ou privada. Seis modelos apresentaram independência, variância constante, ausência de outliers influentes e significativos, permitindo uma ótima capacidade preditiva do desempenho do estudante.*

Palavras-chave: Predição do desempenho, Regressão Linear Múltipla, Ciência de Dados, Mineração de Dados Educacionais, Exame Nacional do Ensino Médio.

Abstract: *In the context of Educational Data Science, Student Academic Performance Prediction can follow Educational Data Mining, which seeks to make student performance quantitative, guiding teachers and educational institutions. Multiple Linear Regression is a forecasting methodology that can be applied to educational data, as is the case of data from the Exame Nacional do Ensino Médio (ENEM). Based on data from the ENEM 2019 edition, this research proposed, tested and analyzed seven multiple regression models based on a sample of 18.908 candidates. Such models considered the scores of the tests of (i) Languages, Codes and their Technologies, (ii) Mathematics and their Technologies, (iii) Natural Sciences and their Technologies and (iv) Human Sciences and their Technologies and (v) Writing; and personal data (iv) age, (v) sex and (vi) completed high school in a public or private school. Six models showed independence, constant variance, absence of influential and significant outliers, allowing for an excellent predictive capacity of student performance.*

Keywords: Performance prediction, Multiple Linear Regression, Data Science, Educational Data Mining, Exame Nacional do Ensino Médio.

Introdução

No campo educacional, uma preocupação de professores e gestores é predizer o desempenho do estudante em uma próxima disciplina ou em um curso (IBRAHIM, RUSLI, 2007; SOULE, 2017; RAJALAXMI et al., 2019). No que se refere à avaliação em larga escala, o Exame Nacional do Ensino Médio (ENEM) avalia nacionalmente e anualmente cerca de 6 milhões de candidatos[†], proporcionando o ingresso no Ensino Superior. Segundo Justino (2019), é o segundo maior exame em larga escala do mundo, sendo o primeiro o Gaokao[†] (China).

[†]Autor correspondente: jpiton@ufscar.br.
[†]Média dos últimos 5 anos.
[†]O exame Gaokao é realizado na China e oferece apenas uma chance aos candidatos.

Estruturalmente, o exame é elaborado com base em competências e habilidades avaliadas. Quando recorremos à literatura, os trabalhos relacionados ao ENEM estão ligados à análises quantitativas ou qualitativas (TRAVITZKI, 2017; SILVA, SANTIAGO, SANTOS, 2013). No último caso, é visto como um instrumento pedagógico que auxilia na reflexão do professor e dos estudantes (WIEBUSCH, 2012). A literatura internacional tem demonstrado preocupação com a predição quantitativa do desempenho, uma vez que os modelos preditivos podem ser implantados em Sistemas Tutores Inteligentes e Gamificação, por exemplo.

No contexto da Ciência de Dados Educacionais (CDE) (SILVA et al., 2017), a Predição de Desempenho Acadêmico de Estudantes (do inglês *Predicting Student Academic Performance*) pode seguir a Mineração de Dados Educacionais (EDM), que possibilita a busca por aprimorar cada vez mais a qualidade educacional, variando desde a descoberta de problemas e falhas em sistemas educacionais (ou de ensino) e processos de ensino até explicitando, quantificando o desempenho de um estudante. Por exemplo, os trabalhos de Rajalaxmi et al. (2019), Abledu (2012) e Gadavi, Patel (2017) buscam modelos matemáticos preditivos através da regressão linear que auxiliem em políticas educacionais, professores, gestores e estudantes. Do ponto de vista metodológico, a predição pode contemplar as seguintes abordagens (BAKER, YACEF, 2009): (i) Classificação, (ii) Estimação da Densidade e (iii) Regressão.

Do ponto de vista da Avaliação Educacional em Larga Escala, o ENEM é uma avaliação de desempenho realizada, anualmente, pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), Autarquia Federal vinculada ao Ministério da Educação (MEC). Trata-se de um exame de aplicação nacional e valoriza a capacidade do participante de utilizar os conhecimentos e as habilidades, determinando assim seu desempenho das habilidades e competências da educação escolar. Os resultados de cada aplicação do ENEM geram um conjunto significativo de dados, que vão desde a nota (escore) do candidato até respostas de um questionário sócio-econômico.

No que se refere à análise de dados do ENEM sob a ótica da regressão, a revisão da literatura aponta para o trabalho de Silva Filho (2017), que determinou um modelo que possibilita a predição de desempenho de estudantes do Ensino Médio no ENEM através da regressão logística. Dias et al. (2020) também utilizam a regressão logística e identificaram possíveis variáveis que interferem na certificação ou não certificação de conclusão do Ensino Médio, utilizando resultados no ENEM da edição de 2015. Quanto à regressão linear múltipla, não foram encontrados artigos científicos, dissertações ou teses que apliquem tal regressão em dados do ENEM.

No contexto da regressão em aplicações educacionais é natural que a interpretação do conjunto de dados considere mais de uma variável independente. São exemplos: idade, ano escolar, notas de diferentes avaliações, frequência escolar, dentre outros. Nesse caso, a regressão em uma variável torna-se insuficiente e a solução é considerar a metodologia da análise de dados multidimensional (HOFFMANN, 2016).

Considerando que (i) a análise de dados multidimensional apontada por Hoffmann (2016) é adequada para propósitos educacionais e (ii) há uma lacuna de estudos no que se refere à aplicação da regressão múltipla em resultados de provas do ENEM, essa pesquisa responde à seguinte questão: "é possível propor e avaliar modelos de regressão múltipla com base em dados do ENEM observando correlações entre as diferentes áreas do conhecimento?" Para respondê-la, considerou-se uma amostra de 18.908 candidatos do referido exame de 2019, partindo da modelagem via regressão linear múltipla.

Predição de Desempenho

A Ciência de Dados surge da necessidade do processamento e da interpretação de um volume substancial de dados e é aplicada em diversos setores da economia, tais como o governamental, industrial, marketing, bancos, medicina, dentre outros (CAMILO, 2009). Segundo Porto, Ziviani (2014) sua base de formação advém da fusão entre ciências da computação e as modelagens

estatística e matemática. A ciência de dados é desafiada a resolver problemas de organização e gerenciamento de dados, análise de dados e de redes complexas. Nesse campo do conhecimento, o profissional deve ser capaz de transformar os dados em informação, uma vez que somos ricos em dados, contudo, pobres em informação (HAN, KAMBER, PEI, 2011).

Inserida na ciência de dados, a CDE permite apoiar sistemas educacionais e computacionais. Por exemplo, é possível identificar o desempenho dos estudantes da Educação Básica e, em seguida, ter-se ações quanto às taxas de reprovação e metas de ensino.

O conceito de "predição" esteve, no passado, associado às estruturas mágicas e ilógicas mas, atualmente, trata-se da identificação de padrões e relações. Nessa direção, Loh (2014) menciona que identificar padrões é algo antigo e surge da predição de variações do tempo, fases da Lua e seus eclipses, plantio e colheita são exemplos. Para o referido autor, um modelo pode ser utilizado para a predição de eventos futuros, uma vez que trata-se da generalização de uma realidade. Contudo, se o modelo for impreciso, poderá não corresponder a uma visão correta da realidade.

Quando o propósito é educacional, os dados podem ser diversos: boletins escolares, notas de avaliações, questionários socioeconômicos, grau de escolaridade, tipo de escola, currículo escolar, etc. Partindo desses dados é possível predizer o desempenho do estudante, por exemplo, em um curso ou disciplina, permitindo uma intervenção que minimize a evasão escolar, que é um problema relevante no contexto educacional brasileiro (SANTOS, 2020).

Metodologicamente, a predição do desempenho do estudante pode seguir a Mineração de Dados Educacionais (EDM) que, de acordo com Baker, Yacef (2009), pode ser definida como a aplicação dos métodos e técnicas da mineração de dados na análise de dados educacionais. Para os autores a predição pode contemplar as abordagens de (i) Classificação, (ii) Estimação da Densidade e (iii) Regressão.

Outra abordagem são os Modelos Hierárquicos[iv], que consideram a estrutura do agrupamento dos dados organizados em níveis. Essa "hierarquia é uma propriedade intrínseca da população de interesse. Este tipo de estrutura é bastante comum em sistemas educacionais" (Natis, 2001). É importante frisar que em modelos hierárquicos não assume-se a independência para as observações. No que se refere aos dados do ENEM, o trabalho de Lobo, Cassuce e Cirino (2017) construiu um modelo hierárquico com o objetivo de identificar fatores determinantes do desempenho escolar em candidatos nordestinos que realizaram o ENEM edição 2013. Foram determinados dois níveis, em que o nível 1 considera os aspectos familiares do estudante e o nível 2 o desempenho na prova de matemática. O trabalho de Araújo (2019) investigou os determinantes que influenciam no desempenho de candidatos do ENEM na cidade de Viçosa e, para isso, utilizou um modelo hierárquico que obtêm a correlação entre os candidatos da mesma escola em três níveis: aluno, escola e o ano escolar.

Em termos de ENEM, Adeodato, Filho, Rodrigues (2014) aplicaram a Regressão Logística do ENEM 2011 combinada com Árvores de Decisão. Na mesma direção, Filho (2017) utilizou o R, Weka e Knime para construir e avaliar modelos de predição via Regressão Logística a partir de dados do ENEM de 2014. Recentemente, Alves, Cechinel, Queiroga (2018) aplicaram os dados do ENEM 2015 em um modelo preditivo do indicador de desempenho das notas da prova de Matemática e suas Tecnologias das escolas do Ensino Médio, de maneira que os modelos finais gerados foram treinados e testados por meio dos algoritmos Naive Bayes e J48 do Weka. Gomes et al. (2020) realizaram uma análise preditiva do desempenho em Matemática de candidatos ao ENEM de 2011, adotando o algoritmo CART (*Classification and Regression Trees*) em linguagem R, o que culminou em um modelo que explicou 29,97% da variância do desempenho da amostra.

No âmbito internacional, Abledu (2012) analisou problemas de ordem administrativa e técnica que influenciavam nas avaliações de cursos Politécnicos em Gana através da regressão linear múltipla. Gadhavi, Patel (2017), determinaram um modelo que proporcionou aos estu-

[iv] Também conhecidos como Modelos Lineares Multinível, Modelos de Efeitos Aleatórios e Modelos de Componentes da Variância.

dantes uma auto análise do desempenho, também com regressão linear. Um importante estudo foi o realizado por Ibrahim, Rusli (2007), em que compararam três modelos de desempenho, cujo objetivo foi predizer a métrica Média de Pontos Cumulativos (do inglês *Cumulative Grade Point Average* - CGPA), que é a média de pontuações acumuladas ao longo do curso, sendo a média geral ao final da graduação. Na Universidade do Sul de Illinois (Carbondale), o projeto de Soule (2017) predisse o sucesso estudantil por meio da regressão logística. Na mesma direção, Rajalaxmi et al. (2019) utilizaram informações do uso da internet por graduandos em Engenharia para predizer seus desempenhos acadêmicos através da regressão linear múltipla.

Em aplicações educacionais é possível que a interpretação do conjunto de dados considere mais de uma variável independente. São exemplos: idade, ano escolar, notas de diferentes avaliações, frequência escolar, dentre outros. Nesse caso, a regressão linear simples torna-se insuficiente e a solução é considerar a metodologia da análise de dados multidimensional (HOFFMANN, 2016). Para uma análise multivariada (ou multidimensional), a Regressão Linear Múltipla parte essencialmente da Equação 1:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_j, \tag{1}$$

em que Y é a variável dependente e X_j $(j = 1, 2, \ldots, k)$ na k-ésima variável independente. Em outras palavras, o modelo passa a considerar dimensões maiores ou iguais à três. Especificamente quando temos duas variáveis independentes, temos um plano de regressão. Para ilustrarmos a situação, tomemos a Equação 2.

$$Y = 73,887816 + 1,868802 \cdot X_1 + 0,024555 \cdot X_2, \tag{2}$$

A Figura 1 representa o modelo da Equação 2, determinada a partir do método dos mínimos quadrados para um caso multivariado composto por duas variáveis independentes (X_1, X_2).

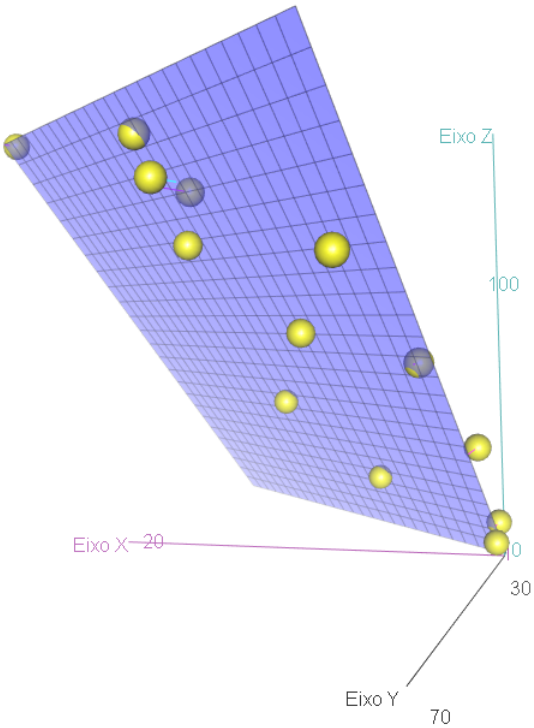

Figura 1: Plano Tridimensional da Equação (2).

Fonte: Autores.

Uma vez que os parâmetros do modelo são determinados, analisa-se a sua efetividade. Nesse contexto, o trabalho de Huang, Fang (2010) modelaram o desempenho acadêmico de estudantes de Engenharia na disciplina de Dinâmica[v], resultando em único modelo que foi testado quanto à sua efetividade. Para tal, os autores consideraram o teste F de significância e o Coeficiente de Determinação R^2 para casos multidimensionais. Este pode ser sintetizado como "a fração da variância total em Y explicada pelo modelo de regressão com a da variância atribuída aos resíduos" (FARIA, 2011, p. 14) e pode-se descrevê-lo como o quadrado do coeficiente de correlação de Pearson.

Uma vez que é determinado o coeficiente de determinação, é possível quantificar a capacidade explicativa do modelo, ou seja, quão uma variável resposta pode ser explicada pelas variáveis preditoras. Contudo, um problema com essa medida está relacionado à quantidade de variáveis, uma vez que, quanto mais variáveis forem adicionadas ao modelo, maior é o valor do coeficiente. Uma solução é utilizar o Coeficiente de Determinação Ajustado, que é uma medida que "penaliza" o número excessivo de variáveis do modelo. Uma vez que aumenta-se a dimensão, o valor do coeficiente tende a diminuir, possibilitando um ajuste ótimo ao modelo.

Após determinar a capacidade preditiva e a existência de variáveis explicativas, é necessário analisar os resíduos[vi]. A análise ocorre pela validação de alguns pressupostos da regressão simples univariada e é complementada por outros conceitos, tais como normalidade dos erros, independência dos erros, variância constante, existência de *outliers* e de colinearidade ou multicolinearidade entre as variáveis.

Materiais e Métodos

Seguindo a modelagem do desempenho do estudante via regressão linear múltipla, esta pesquisa muniu-se de uma amostra de candidatos que realizaram o ENEM da Edição 2019, que culminaram em quatro experimentos descritos e discutidos nas próximas seções.

Materiais

Os microdados[vii] do ENEM 2019 contemplam diversos dados sobre os candidatos, que vão desde os dados pessoais (não identificáveis) até o escore[viii] do candidato em cada prova, em um total de 3.2GB de dados em texto plano.

Adota-se um estudo amostral em que a chave primária é que o candidato resida na Região Metropolitana de Ribeirão Preto do Estado de São Paulo, região composta por 34 municípios em uma população de aproximadamente 1,7 milhão de habitantes[ix]. Considera-se apenas os candidatos que responderam às cinco provas e desconsiderados os treineiros, ausentes[x] e desclassificados[xi]. Aplicando os referidos critérios, a amostra culminou em 18.908 candidatos válidos.

Quanto à estrutura do exame, a parte objetiva da prova é composta por um total de 180 itens[xii] que avaliam o candidato nas áreas de (i) Linguagens, Códigos e suas Tecnologias, (ii)

[v] Composta por conteúdos relacionados à mecânica estrutural, dinâmica e controle do sistema e projetos de máquinas.

[vi] Diferença entre os valores projetados pelo modelo e os valores reais, também compreendida como erro.

[vii] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

[viii] É a pontuação ou nota do candidato padronizada na escala de 0 a 1000

[ix] Dados do Instituto Brasileiro de Geografia e Estatística (IBGE) no ano 2018. Fonte https://emplasa.sp.gov.br/RMRP

[x] Existem casos em que o candidato compareceu em uma das provas, mas se ausentou em outra.

[xi] Por apresentarem irregularidades e/ou infringirem as regras do exame.

[xii] Para Osterlind (1998), os itens de um teste não podem ser chamados de questões, pois um item pode assumir outros formatos, que não são necessariamente interrogativos. Além disso existem parâmetros psicométricos associados, tais como dificuldade e discriminação.

Matemática e suas Tecnologias, (iii) Ciências da Natureza e suas Tecnologias e (iv) Ciências Humanas e suas Tecnologias, sendo complementado com uma (v) Redação de caráter dissertativo-argumentativo. São 45 itens objetivos de múltipla escolha por área na parte objetiva. De acordo com o número de acertos e erros (em cada uma das quatro provas) é atribuído um escore por área.

O escore é estimado por meio da Teoria de Resposta ao Item (TRI) que, de acordo com Omitido (2020, p. 670) "propõe uma modelagem estatístico-matemática para as características latentes do examinado e modela a probabilidade de um indivíduo responder corretamente a um item em função do seu traço latente que é psicometricamente mapeado para um estimador θ". Detalhes da TRI são encontrados em Wainer, Mislevy (1990) e Lord (1980).

Sistematicamente, o conjunto de dados considerados para a geração dos modelos estão descritos na Tabela 1.

Tabela 1: Dados considerados para os testes.

Variável	Descrição	Variação	Categórica	Tamanho	Tipo
`idade`	Idade	-	-	3	numérica
`sexo`	Sexo	M/F	Masculino/Feminino	1	alfanumérica
`tp_escola`	Tipo de escola do Ensino Médio	1 a 4	Pública/Privada	1	numérica
`CN`	Escore de Ciências da Natureza	-	-	9	numérica
`CH`	Escore de Ciências Humanas	-	-	9	numérica
`LC`	Escore de Linguagens e Códigos	-	-	9	numérica
`MT`	Escore de Matemática	-	-	9	numérica
`RD`	Escore da Redação	-	-	9	numérica

Fonte: Microdados do ENEM 2019 (Adaptado).

Do ponto de vista computacional, esta pesquisa implementa algoritmos nas linguagens C++ e R[xiii]. A amostra foi coletada a partir de algoritmos desenvolvidos em C++, visto que havia uma capacidade computacional de processamento e memória limitada e o R apresenta extensa lentidão e uso exponencial da memória RAM.

Para os experimentos apresentados utiliza-se a linguagem R. Enquanto destaques, o pacote `car`[xiv] calculou as tabelas de análise de variância para objetos produzidos pelas sub-rotinas `lm` (para ajuste linear), e `glm` (para ajuste de modelos lineares generalizados) e forneceu estatísticas do teste F para modelos lineares multivariados produzidos por `lm`. Outro pacote importante foi o `vif`[xv], responsável pelo teste de colinearidade das variáveis.

Métodos

A pesquisa inicia-se com a definição da quantidade e do tipo das variáveis, assim como os pressupostos quanto à dependência ou independência. Para isso considera-se os dados quantitativos que possibilitasse a realização da análise e alguns binários (utilizados como variáveis *dummy*[xvi]). De outra forma, a escolha dos valores dependentes e independentes se deu devido à área de atuação dos autores, gerando insumos para discussões não apenas da comunidade, mas também próprias. São realizados testes e análises em cada experimento, composto por *testes*, os seguintes modelos de regressão linear múltiplos:

Experimento 1
Teste 1.0: $\texttt{MT} = \alpha + \beta_1 \cdot \texttt{CN} + \beta_2 \cdot \texttt{CH} + \beta_3 \cdot \texttt{LC} + \beta_4 \cdot \texttt{RD}$

[xiii] https://cran.r-project.org
[xiv] https://cran.r-project.org/web/packages/car/index.html
[xv] https://cran.r-project.org/web/packages/VIF/index.html
[xvi] https://medium.com/data-hackers/vari%C3%A1veis-dummy-o-que-%C3%A9-quando-usar-e-como-usar-78de66cfcca9

Experimento 2
Teste 2.0: $\texttt{MT} = \alpha + \beta_1 \cdot \texttt{CN} + \beta_2 \cdot \texttt{CH} + \beta_3 \cdot \texttt{LC} + \beta_4 \cdot \texttt{RD} + \beta_5 \cdot \texttt{tp\_escola}$

Experimento 3
Teste 3.0: $\texttt{MT} = \alpha + \beta_1 \cdot \texttt{idade} + \beta_2 \cdot \texttt{sexo} + \beta_3 \cdot \texttt{tp\_escola}$

Experimento 4
Teste 4.0: $\texttt{MT} = \alpha + \beta_1 \cdot \texttt{CN} + \beta_2 \cdot \texttt{CH} + \beta_3 \cdot \texttt{LC} + \beta_4 \cdot \texttt{RD} + \beta_5 \cdot \texttt{idade} + \beta_6 \cdot \texttt{sexo}$
Teste 4.1: $\texttt{LC} = \alpha + \beta_1 \cdot \texttt{CN} + \beta_2 \cdot \texttt{CH} + \beta_3 \cdot \texttt{MT} + \beta_4 \cdot \texttt{RD} + \beta_5 \cdot \texttt{idade} + \beta_6 \cdot \texttt{sexo}$
Teste 4.2: $\texttt{CN} = \alpha + \beta_1 \cdot \texttt{MT} + \beta_2 \cdot \texttt{CH} + \beta_3 \cdot \texttt{LC} + \beta_4 \cdot \texttt{RD} + \beta_5 \cdot \texttt{idade} + \beta_6 \cdot \texttt{sexo}$
Teste 4.3: $\texttt{CH} = \alpha + \beta_1 \cdot \texttt{CN} + \beta_2 \cdot \texttt{MT} + \beta_3 \cdot \texttt{LC} + \beta_4 \cdot \texttt{RD} + \beta_5 \cdot \texttt{idade} + \beta_6 \cdot \texttt{sexo}$

Metodologicamente, cada teste verifica se os parâmetros via regressão linear múltipla são significativos quanto à confiança de 5% e se fornecem bom poder preditivo. Uma vez confirmados, segue a análise de resíduos. Em suma, os experimentos contemplam as variáveis, os parâmetros estimados, os valores do teste F de significância e o valor que indica o poder preditivo, o R^2 ajustado. Devido ao número de variáveis ser maior do que três, não é possível expressar graficamente o hiperplano dos modelos.

Descritivamente, os Experimentos 1, 2 e 3 predizem o desempenho do estudante em Matemática e suas Tecnologias, sendo que: (i) o Teste 1 considera o escore das outras três provas e redação, (ii) o Teste 2 acrescenta o tipo da escola e o (iii) Teste 3 considera somente a idade, sexo e tipo de escola. O Experimento 4 é composto por quatro modelos de regressão, que buscam correlacionar o escore das quatro provas com idade e sexo.

Resultados e Discussão

Após a definição dos materiais e métodos para a realização da pesquisa foi possível iniciar os experimentos através do *software* R. Os quatro experimentos passam pela análise de regressão linear múltipla para definição de seus coeficientes e, posteriormente, pela validação estatística dos mesmos. Os resultados e discussões podem ser acompanhados nas próximas seções.

Experimento 1

Neste experimento utiliza-se dados de 18.908 candidatos. A Tabela 2 explicita os coeficientes determinados e o p-valor calculado foi de $2e^{-16}$ para todas as variáveis.

Tabela 2: Parâmetros (valores) encontrados para os coeficientes do Teste 1.

Variável	Coeficiente	Valor
Independente	α	-75,622298
CN	β_1	0,568728
CH	β_2	0,250989
LC	β_3	0,284213
RD	β_4	0,094078

Fonte: Autores.

Além desses valores individuais, o modelo é significativo pelo Teste F e tem poder de predição de 55,99% (moderado) que é considerado como adequado, uma vez que o universo de dados do ENEM que poderiam explicar o desempenho é grande, além do modelo de avaliação (são dois dias apenas, o que não permite isolar situações imprevisíveis como o estresse e ansiedade do candidato). A análise de resíduos (Figura 2) mostra quatro gráficos que permitem identificar a independência dos erros e variância, normalidade e existência de *outliers* influentes. O primeiro

(*residuals vs fitted*) e o terceiro gráfico (*scale-location*) auxiliam na análise da independência e variância. Nesse teste é possível afirmar que os erros são independentes e apresentam variância constante, uma vez que estão distribuídos aleatoriamente em torno de zero (mais visível no terceiro gráfico).

Inspecionando o segundo gráfico (*normal Q-Q*) da Figura 2, é possível afirmar que há normalidade dos erros, visto que os valores estão próximos da linha. Pelo quarto gráfico (*Cook's distance*) verifica-se que não há *outliers* influentes, visto que os valores mais altos são bem distantes de 1,0; apresentando no máximo 0,04.

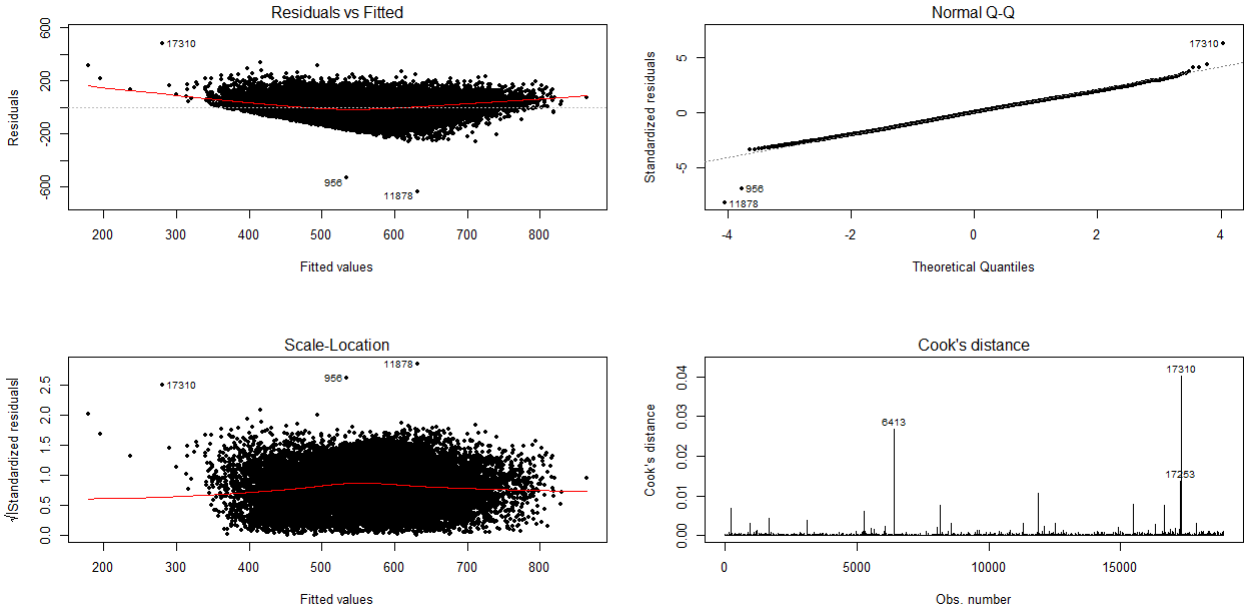

Figura 2: Análise de resíduos do Teste 1.

Fonte: Autores.

Para a análise de colinearidade utiliza-se a *Variance Inflation Factor* (VIF) que apresenta os resultados descritos na Tabela 3. Como não resulta em valores superiores a cinco, é possível afirmar que não há colinearidade entre as variáveis e, consequentemente, o modelo é considerado, enquanto preditivo, para modelar o desempenho em Matemática.

Tabela 3: Verificação da colinearidade do Teste 1 - VIF.

CN	CH	LC	RD
2,355071	2,963565	2,759483	1,763712

Fonte: Autores.

Experimento 2

Diferentemente do Experimento 1, são considerados 8.925 candidatos (47,2% do total) nesse experimento, uma vez que a variável `tp_escola` foi filtrada (no banco) apenas para os candidatos concluintes do Ensino Médio, reduzindo o conjunto de dados. Neste caso, a Tabela 4 apresenta os resultados obtidos após a aplicação da regressão múltipla, em que o p-valor da variável independente foi de $3,29e^{-3}$, enquanto as demais $2,0e^{-16}$.

De acordo com a Tabela 4 as variáveis são significativas e o modelo é significativo, demonstrando 56,26% de poder preditivo. Comparado ao Teste 1, este apresenta pouca melhora

Tabela 4: Parâmetros (valores) encontrados para os coeficientes do Teste 2.

Variável	**Coeficiente**	**Valor**
Independente	α	-23,559799
`CN`	β_1	0,522887
`CH`	β_2	0,249763
`LC`	β_3	0,240958
`RD`	β_4	0,076005
`tp_escola`	β_5	30,571795

Fonte: Autores.

quando é adicionada a dimensão `tp_escola`. Nota-se uma diferença de 30 pontos no escore final de Matemática e suas Tecnologias entre a escola pública e a particular, sendo favorável à escola particular. Esta diferença é significativa, uma vez que os resultados são utilizados em grande maioria para ingresso em uma universidade ou faculdade e, através desta visão pode-se assumir que um candidato de escola particular tem vantagem sobre o candidato da escola pública para esta amostra de dados.

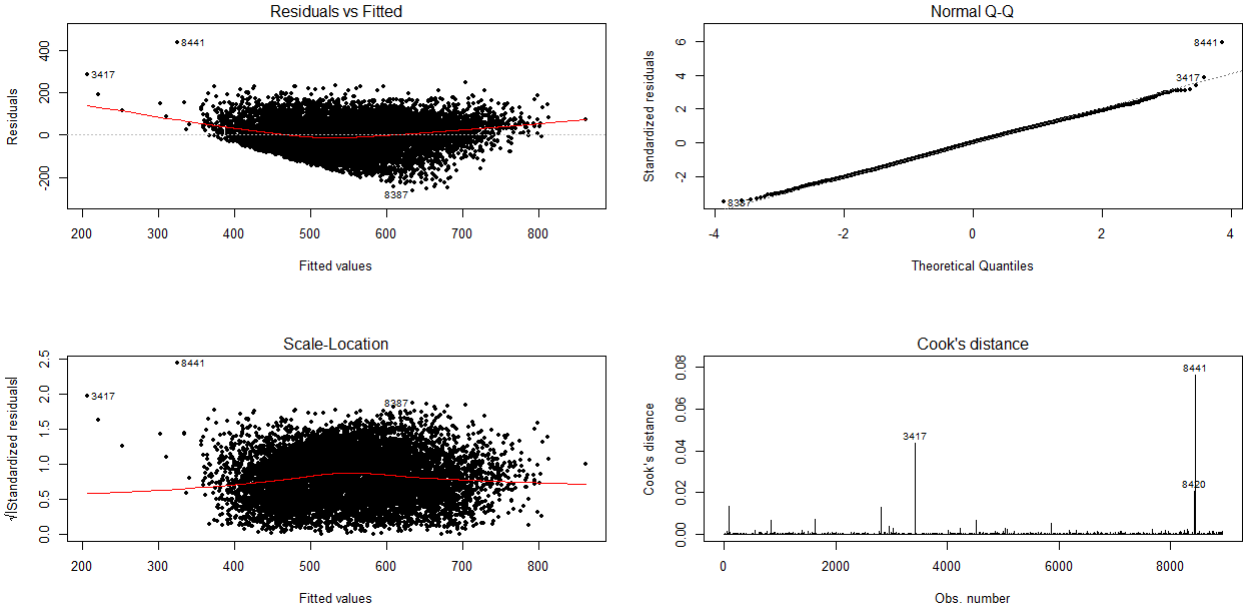

Figura 3: Análise de resíduos do Teste 2.

Fonte: Autores.

Partindo do terceiro gráfico (*Scale-location*) da Figura 3, pode-se afirmar que os erros são independentes e de variância constante, além de serem normais (*Normal Q-Q*) e não apresentarem *outliers* influentes (*Cook's distance*). De acordo com os resultados da Tabela 5, conclui-se que não há colinearidade entre as variáveis e o modelo é considerado, enquanto preditivo, apto para modelar o desempenho em Matemática quando a variável `tp_escola` é considerada.

Experimento 3

Assim como o Experimento 2, `tp_escola` é uma das variáveis, portanto, são considerados 8.925 candidatos (47,2% do total). Os parâmetros do modelo estão expressos na Tabela 6. Ressalta-se que todos os p-valores calculados são iguais a $2e^{-16}$.

Tabela 5: Verificação da colinearidade do Teste 2 - VIF.

CN	CH	LC	RD	tp_escola
2,352174	2,866952	2,717174	1,821912	1,294095

Fonte: Autores.

Tabela 6: Parâmetros (valores) encontrados para os coeficientes do Teste 3.

Variável	**Coeficiente**	**Valor**
Independente	α	599,6624
idade	β_1	-5,3508
sexo	β_2	42,9434
tp_escola	β_3	106,4465

Fonte: Autores.

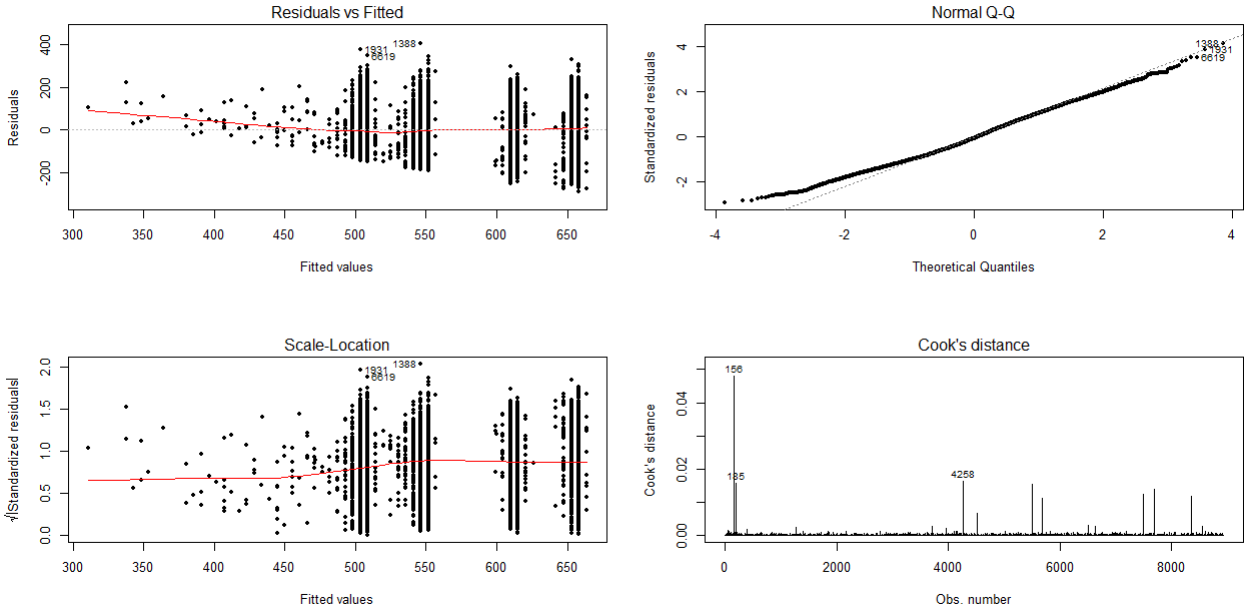

Figura 4: Análise de resíduos para Teste 3.

Fonte: Autores.

O modelo é significativo assim como as variáveis presentes na Tabela 6. Por outro lado, o poder preditivo calculado é de 23,79%. Através da análise de resíduos da Figura 4, nota-se que no gráfico *Scale-location* os erros não são independentes, pois estão distribuídos, em sua maioria, em faixas verticais. Diante disso, o modelo é desconsiderado para modelar o desempenho em Matemática. Como já não cumpre esse pressuposto, o modelo é definido como inadequado.

Apesar de inadequado, observa-se duas informações interessantes. A primeira sobre a idade e o valor de parâmetro negativo. Nesse caso, pode-se dizer que quanto mais distante do ano de conclusão do Ensino Médio, menor o desempenho alcançado na prova. A segunda quanto ao sexo que apresentou uma diferença de aproximadamente 43 pontos, significativa para os valores finais dos escores. Estas informações explicam pouco os resultados, devido ao baixo poder de predição, porém podem ser exploradas em futuros estudos.

Experimento 4

O último experimento é subdividido em quatro diferentes testes, de maneira que o objetivo é predizer o desempenho nas quatro áreas das provas objetivas. Resultados mostram que todos os modelos são significativos. Os poderes de predição para `MT`, `LC`, `CN` e `CH` são, respectivamente, 58,27%, 64,63%, 63,45% e 67,40%. A Tabela 7 traz os parâmetros determinados pela regressão linear múltipla, com o objetivo de predizer o desempenho do candidato em Matemática nas quatro áreas do exame. Os valores ausentes são as respostas para cada modelo. Por exemplo, no Teste 4.0 a resposta esperada é o escore em Matemática, portanto seu valor é ausente, uma vez que não é uma variável preditora mas sim a resposta.

Tabela 7: Parâmetros (valores) encontrados para os coeficientes do Experimento 4.

Variável	**Teste 4.0** `MT`	**Teste 4.1** `LC`	**Teste 4.2** `CN`	**Teste 4.3** `CH`
Independente	-52,692516	238,287321	36,711309	-29,726665
`MT`	-	0,062710	0,222632	0,089250
`LC`	0,302378	-	0,269171	0,625477
`CN`	0,519944	0,130371	-	0,234977
`CH`	0,241535	0,351049	0,272288	-
`RD`	0,105737	0.036979	0,066941	0,057272
`idade`	-1,219398	-0,211429	0,402815	0,882615
`sexo`	33,639099	-6,219731	8,202540	3,264952

Fonte: Autores.

De acordo com a Tabela 8, as variáveis menos explicativas e influentes (`RD`, `idade` e `sexo`) apresentam um número inferior a 2,0. Por outro lado, as variáveis que são mais explicativas e influentes são superiores a 2,0, indicando que os valores estão na tolerância admitida. O resultado igual a 3,002447 (`MT` na variável `CH`) é um exemplo de variável mais próxima da colinearidade, uma vez que o limite é 5,0.

Tabela 8: Coeficientes VIF para o Experimento 4.

Variável	**VIF** `MT`	**VIF** `LC`	**VIF** `CN`	**VIF** `CH`
`MT`	-	2,351933	2,119879	2,345711
`LC`	2,774597	-	2,728977	2,207223
`CN`	2,419838	2,640586	-	2,561526
`CH`	3,002447	2,394816	2,872264	-
`RD`	1,871728	1,897334	1,875720	1,884588
`idade`	1,042142	1,049129	1,048231	1,038728
`sexo`	1,055973	1,096807	1,098077	1,103620

Fonte: Autores.

Os resultados do Teste 4.0 (Figura 5) mostram que há independência e variância constante assim como a normalidade e a ausência de *outliers* influentes, sendo complementado pelo VIF da Tabela 8 que está adequado ao limite. Portanto, o modelo é considerado para modelar o desempenho em Matemática e suas Tecnologias.

Os resultados do Teste 4.1 (Figura 6) mostram que há independência e variância constante assim como a normalidade e a ausência de *outliers* influentes, com exceção de uma maior dispersão nas extremidades da reta do gráfico *Normal Q-Q*, o que não prejudica em totalidade o modelo. Portanto, o modelo é considerado para modelar o desempenho em Linguagens, Códigos e suas Tecnologias.

Analisando os Testes 4.2 (Figura 7) e 4.3 (Figura 8), verifica-se similaridades com os resultados do Teste 4.1, com destaque a uma maior dispersão nas extremidades da reta de normalização

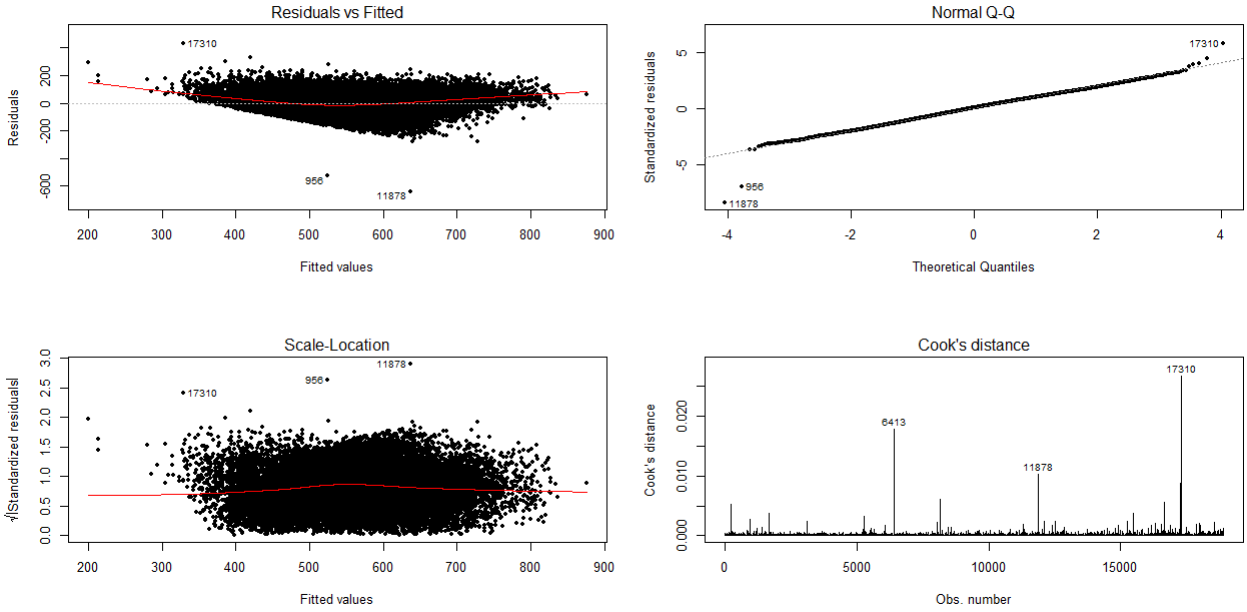

Figura 5: Análise de resíduos para Teste 4.0 - `MT`.

Fonte: Autores.

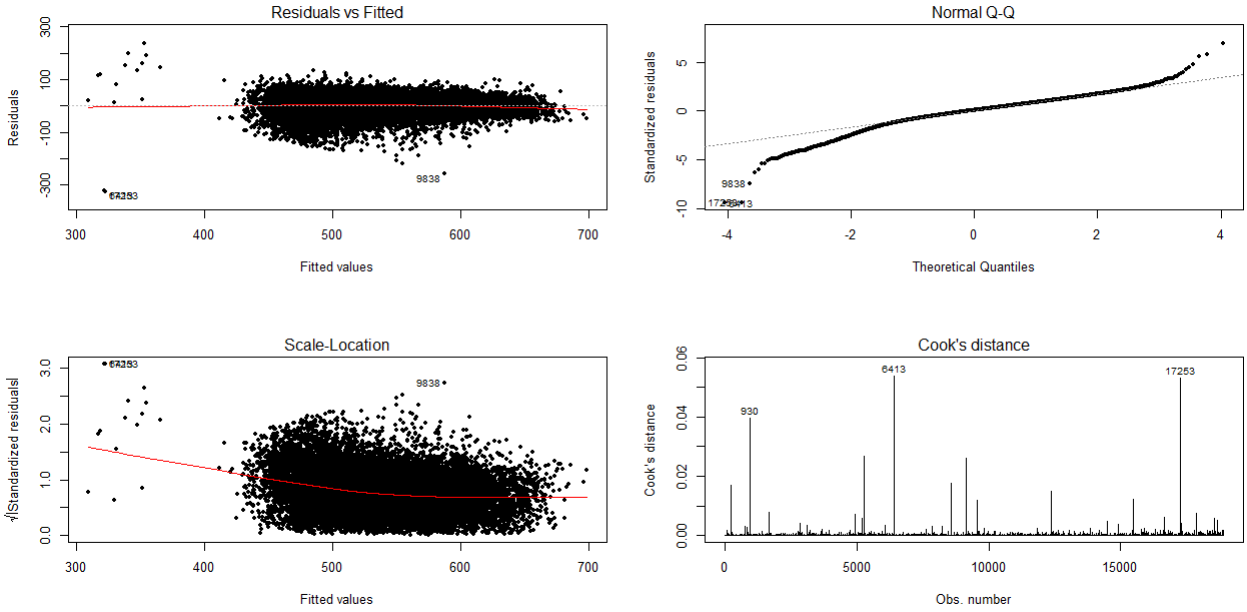

Figura 6: Análise de resíduos para Teste 4.1 - `LC`.

Fonte: Autores.

dos dados, com maior expressividade para `CN`. Ainda assim, como os critérios de independência, variância constante, ausência de *outliers* influentes e colinearidade são satisfatórios.

Com isso, conclui-se que os modelos dos Testes 4.2 e 4.3 modelam adequadamente, respectivamente, o desempenho dos candidatos em Ciências da Natureza e suas Tecnologias e em Ciências Humanas e suas Tecnologias.

De maneira geral, o poder de predição foi moderado na maior parte dos modelos, sendo

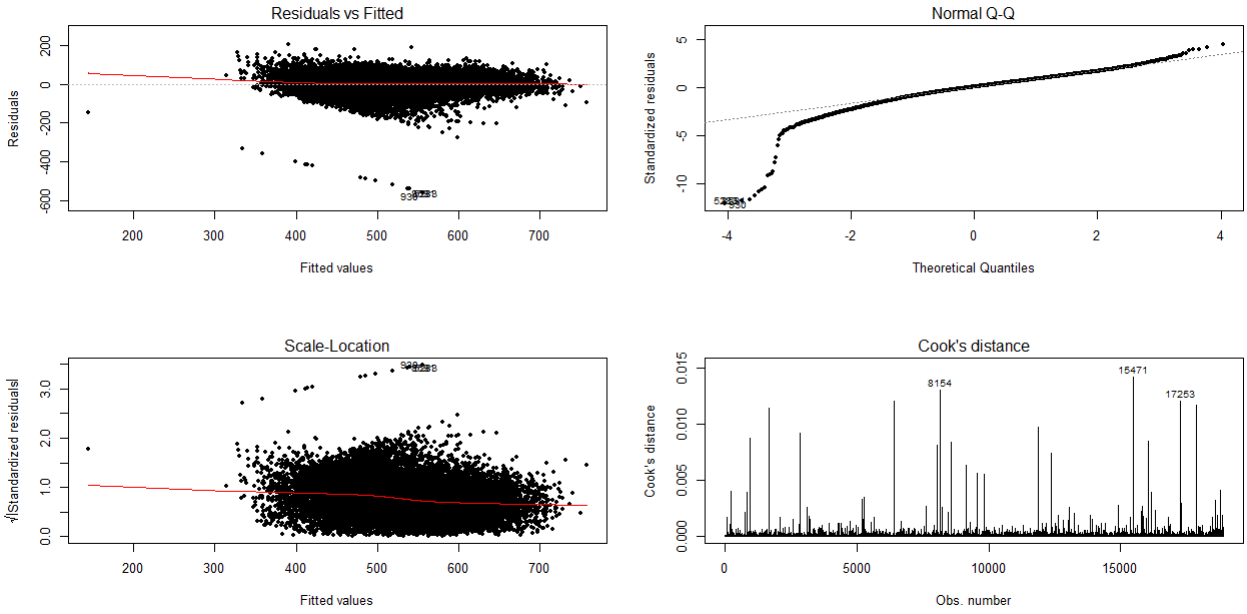

Figura 7: Análise de resíduos para Teste 4.2 - `CN`.

Fonte: Autores.

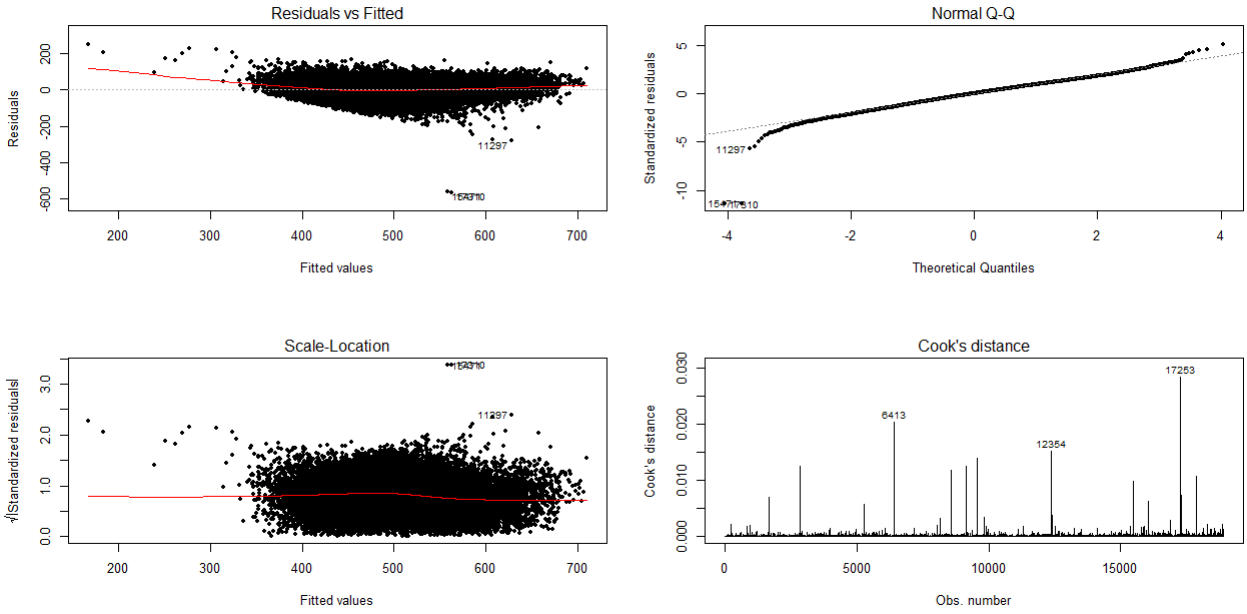

Figura 8: Análise de resíduos para Teste 4.3 - `CH`.

Fonte: Autores.

o valor mais alto para Ciências Humanas (67,40%), dificultando uma interpretação adequada do desempenho do estudante. Uma provável causa é a não consideração de outras variáveis e/ou dimensões. O trabalho de Figueirêdo, Nogueira, Santana (2014) aponta alguns fatores influenciadores, tais como renda familiar e escolaridade dos pais. Os referidos autores afirmam que "(...) um indivíduo com baixo *background* precisa esforçar em torno de 99,38% a mais do que um indivíduo com alto *background* para estar entre os 5% com melhores notas" (FIGUEIREDO

et al., 2014, p. 389-390).

No que se refere ao Experimento 1, este demonstra uma diferença significativa entre os valores dos coeficientes - Ciências da Natureza versus Ciências Humanas e Linguagens *versus* Redação - alinhando a correlação individual dessas áreas do conhecimento, sendo o maior coeficiente para Ciências da Natureza e o menor para Redação. Para elucidar este raciocínio utiliza-se o gráfico de Draftman presente na Figura 9 que representa a força da correlação de Pearson (de acordo com o tamanho da fonte do gráfico).

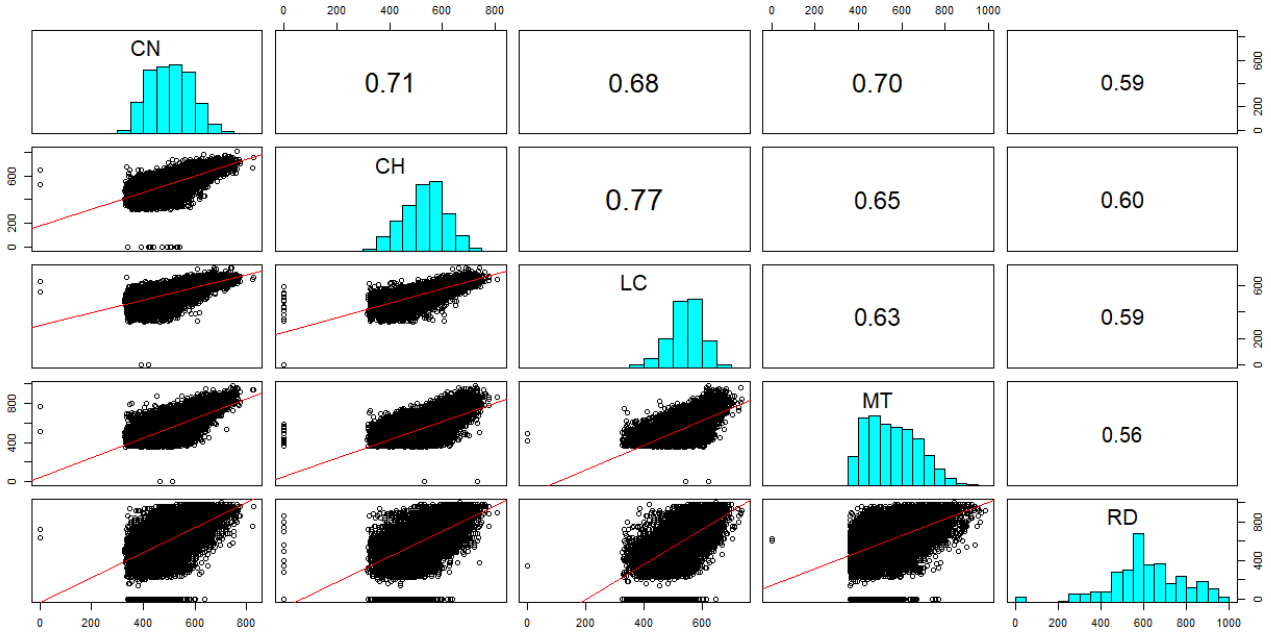

Figura 9: Gráfico de Draftman das correlações entre as notas (escores).

Fonte: Autores.

As análises finais inferem que, no Experimento 4, é observada que a variação dos pontos para a variável `sexo` é mais acentuada quando se compara Matemática com Linguagens e Códigos. Por outro lado, há simetria entre Matemática e Ciências Humanas no que se refere a idade. Enquanto na Matemática o desempenho diminui conforme a idade, em Ciências Humanas é o inverso.

Apesar do poder de predição ser menor no Teste 3, o tipo de escola influencia no desempenho do estudante. Nesse caminho, Sampaio, Guimarães(2009) analisaram o componente atribuído ao tipo de instituição de ensino e concluíram que as escolas públicas são menos eficientes do que as privadas, além de uma diferença de desempenho significativa entre as escolas federais e estaduais, favorável às federais. O trabalho de Moraes, Belluzzo (2014) analisa o diferencial de desempenho escolar através da decomposição de quantis, mostrando a favorabilidade para as escolas particulares em relação às públicas.

Conclusão

A presente pesquisa mostra que é possível, quando consideramos independência entre as observações, propor modelos de desempenho de estudantes a partir da regressão linear múltipla a partir de dados educacionais, como é o caso do ENEM. Essa instanciação mostra que é possível se obter modelos de predição de desempenho que possam apoiar na tomada de decisão de professores, gestores e estudantes. A revisão da literatura internacional mostra que o uso de

regressão linear múltipla ocorre há alguns anos, porém no Brasil é ainda incipiente quando se trata de uma aplicação com dados reais do ENEM.

Essa aplicação metodológica da regressão em dados educacionais proporciona e provê uma abordagem que poderá apoiar sistemas de ensino e computacionais, tais como (i) os Sistemas Tutores Inteligentes[xvii] que, dentre seus componentes, no módulo Modelo do Estudante representa o estado atual do desempenho do estudante em determinado domínio de conhecimento, que pode ser representado por um modelo via regressão linear múltipla; e (ii) o suporte à Gamificação[xviii] enquanto estratégia de aprendizagem mediado por jogos. Por exemplo, o trabalho de Gawel (2019) utilizou a gamificação como uma ferramenta automatizada que aumentou a eficácia do processo educacional, o que proporcionou um maior engajamento e motivação dos estudantes na disciplina de microeconomia. O nível de envolvimento nas atividades gamificadas e questionários avaliativos estimaram, por meio da regressão linear simples, a influência de elementos baseados em jogos sobre os resultados de aprendizagem da microeconomia.

Enquanto trabalho futuro, almeja-se desenvolver um sistema computacional que apoiará instituições de Ensino Superior a traçar o diagnóstico de calouros a partir dos resultados do ENEM.

References

ABLEDU, G. K. Multiple Regression Analysis of Assessment of Academic Performance of Students in the Ghanaian Polytechnics. *Research on Humanities and Social Sciences*, v. 2, n. 9, p. 15-25, 2012.

ADEODATO, P. J. L.; FILHO, M. M. S.; RODRIGUES, R. L. Predição de desempenho de escolas privadas usando o ENEM como indicador de qualidade escolar. *In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. 2014. p. 891.

ALVES, R. D.; CECHINEL, C.; QUEIROGA, E. Predição do desempenho de Matemática e Suas Tecnologias do ENEM utilizando técnicas de Mineração De Dados. *In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. 2018. p. 469.

ARAÚJO, D. L. de, Determinantes do desempenho no ENEM dos concluintes do ensino médio no município de Viçosa MG. Dissertação de Mestrado. Universidade Federal de Viçosa, Viçosa. 2019.

BAKER, R. S.; YACEF, K. The state of educational data mining in 2009: A review and future visions. *JEDM— Journal of Educational Data Mining*, v. 1, n. 1, p. 3-17, 2009.

CAMILO, C. O.; SILVA, J. C. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. Universidade Federal de Goiás (UFC), v. 29, 2009.

DIAS, G. V. et al. Um estudo sobre a certificação pelo Enem usando regressão logística. Revista Sítio Novo, v. 4, n. 1, p. 41-52, 2020.

FARIA, B. F. P. Teste F na regressão linear múltipla para dados temporais em correlação serial. 2011. Tese de Doutorado.

[xvii] De acordo com Omitido (2004), um Sistema Tutor Inteligente (STI) aplicado à Educação pode utilizar técnicas de Inteligência Artificial para mediar os processos de ensino/aprendizagem.

[xviii] Essencialmente, é a aplicação de elementos de jogos apoiados pela tecnologia em algum contexto específico.

FIGUEIRÊDO, E.; NOGUEIRA, L.; SANTANA, F. L. Igualdade de Oportunidades: Analisando o papel das circunstâncias no desempenho do ENEM. *Revista Brasileira de Economia*, v. 68, n. 3, p. 373-392, 2014.

FILHO, R. L. C. S. Modelo de análise e predição do desempenho dos alunos dos Institutos Federais de Educação usando o ENEM como indicador de qualidade escolar. Tese de Doutorado. Universidade Federal de Pernambuco, Recife. 2017.

GADHAVI, M.; PATEL, C. Student final grade prediction based on linear regression. *Indian J. Comput. Sci. Eng*, v. 8, n. 3, p. 274-279, 2017.

GAWEL, A. The Influence of Gamification Activities on Learning Outcomes in Higher Education. Pietrzykowski M. (Ed.): Bogucki Wyd. Nauk., Poznan. 2019.

GOMES, C. M. A.; SOUZA FLEITH, D.; MARIA, C. Preditores do Desempenho em Matemática de Estudantes do Ensino Médio. *Psicologia: Teoria e Pesquisa*, v. 36, p. e3638, 2020.

HAN, J.; KAMBER, M.; PEI, J. Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, v. 5, n. 4, p. 83-124, 2011.

HOFFMANN, R. Análise de regressão: uma introdução à econometria. São Paulo, 2016. 393p.

HUANG, S.; FANG, N. Regression models of predicting student academic performance in an engineering dynamics course. *In: American Society for Engineering Education. American Society for Engineering Education*, 2010.

IBRAHIM, Z.; RUSLI, D. Predicting students academic performance: comparing artificial neural network, decision tree and linear regression. *In: 21st Annual SAS Malaysia Forum*, 5th September. 2007.

JUSTINO, R. Estudantes universitários brasileiros e chineses: um estudo comparado dos exames Enem e Gaokao. 2019.

LOBO, G. D., CASSUCE, F. C. C., CIRINO, J. F. Avaliação do desempenho escolar dos estudantes da região nordeste que realizaram o ENEM: uma análise com modelos hierárquicos. *Revista Espacios*, v. 38, n. 5, p. 12, 2017.

LORD, F. M. Application of item response theory to practical testing problems. First ed. 1980.

MORAES, A. G. E.; BELLUZZO, W. O diferencial de desempenho escolar entre escolas públicas e privadas no Brasil. *Nova economia*, v. 24, n. 2, p. 409-430, 2014.

NATIS, L. Modelos hierárquicos lineares. *Estudos em Avaliação Educacional*, n. 23, p. 3-29, 2001.

OSTERLIND, S. F. Constructiong Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats. *Kluwer Academic Publishers - New York, Boston, Dordrecht, London, Moscow.* 1998.

OMITIDO (para revisão cega). 2020.

OMITIDO (para revisão cega). 2004.

PORTO, F.; ZIVIANI, A. Ciência de dados. *III Seminário de Grandes Desafios da Computação no Brasil*, Rio de Janeiro, RJ, 2014.

RAJALAXMI, R. R. et al. Regression Model for Predicting Engineering Students Academic Performance. *International Journal of Recent Technology and Engineering*, v. 7, p. 71-75, 2019.

SAMPAIO, B.; GUIMARÃES, J. Diferenças de eficiência entre ensino público e privado no Brasil. *Economia Aplicada*, v. 13, n. 1, p. 45-68, 2009.

SANTOS, J. A. Reflexões sobre a evasão escolar: uma problemática na educação brasileira. *Revista Teias*, v. 21, p. 260-270, 2020.

SILVA, F. A. .; SANTIAGO, M. L.; DOS SANTOS, M. C. Análise de itens da prova de matemática e suas tecnologias do ENEM que envolvem o conceito de números racionais à luz dos seus significados e representações. *Revista Eletrônica de Educação Matemática*, v. 8, p. 190-208, 2013.

SILVA, L. A. et al. Ciência de Dados Educacionais: definições e convergências entre as áreas de pesquisa. *In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. 2017. p. 764.

SILVA FILHO, R. L. C. Modelo de análise e predição do desempenho dos alunos dos Institutos Federais de Educação usando o ENEM como indicador de qualidade escolar. 2017.

SOULE, P. Predicting student success: A logistic regression analysis of data from multiple siu-c courses. Carbondale. OpenSIUC, 2017.

TRAVITZKI, R. Avaliação da qualidade do Enem 2009 e 2011 com técnicas psicométricas. Estudos em Avaliação Educacional, v. 28, n. 67, p. 256-288, 2017.

WAINER, H.; MISLEVY, R. J. Item response theory, item calibration, and proficiency estimation. 1990.

WIENBUSCH, E. M. Avaliação em larga escala: uma possibilidade para a melhoria da aprendizagem. in In: Anais do IX ANPED Sul, GT05: Estado e Política Educacional. Caxias do Sul/RS: ANPED Sul, 2012.