UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Impactos e desafios da educação pré, durante e pós-pandemia - Um estudo com os dados do ENEM nos últimos anos através de modelos hierárquicos

Adriana de Fátima Lourençon Watanabe

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Adriana de Fátima Lourençon Watanabe

Impactos e desafios da educação pré, durante e pós-pandemia - Um estudo com os dados do ENEM nos últimos anos através de modelos hierárquicos

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Profa. Dra. Mariana Cúri

Versão original

São Carlos

2023

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m

Watanabe, Adriana de Fátima Lourençon
Impactos e desafios da educação pré, durante e pós-pandemia - Um estudo com os dados do ENEM nos últimos anos através de modelos hierárquicos / Adriana de Fátima Lourençon Watanabe ; orientadora Mariana Cúri. – São Carlos, 2023.
86 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Cúri, Mariana, orient. II. Título.

Adriana de Fátima Lourençon Watanabe

Impactos e desafios da educação pré, durante e pós-pandemia - Um estudo com os dados do ENEM nos últimos anos através de modelos hierárquicos

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Profa. Dra. Mariana Cúri

Original version

São Carlos

2023

Às minhas filhas Lana e Lara, pela compreensão, amor e inspiração.

AGRADECIMENTOS

Agradeço primeiramente a Deus, presença constante em minha vida.

Ao meu esposo Ailton, pelo apoio e amor incondicional em todos os momentos.

À professora e orientadora Prof[a]. Dr[a]. Mariana Cúri pelas reflexões e contribuições neste trabalho.

Aos professores e tutores do MBA - Inteligência Artificial e Big Data pelo aprendizado e contribuição na minha formação acadêmica.

A todos que direta ou indiretamente colaboraram no aprimoramento deste trabalho.

"Sucesso é o resultado da prática constante de fundamentos e ações vencedoras. Não há nada de milagroso no processo, nem sorte envolvida. Amadores aspiram, profissionais trabalham."

Bill Russel (um dos maiores jogadores da história da NBA)

Trecho extraído do livro: Transformando Suor em Ouro

RESUMO

Watanabe, A. F. L. **Impactos e desafios da educação pré, durante e pós-pandemia - Um estudo com os dados do ENEM nos últimos anos através de modelos hierárquicos**. 2023. 86p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Devido a relevância do ENEM no cenário educacional brasileiro, é importante investigar os mecanismos que afetam o desempenho dos estudantes e sua associação com diversos fatores, entre eles, o contexto social em que cada um vive. Nesse estudo, o ajuste dos Modelos de Regressão Hierárquicos considerando os microdados de 2018 a 2022, possibilitou verificar diferenças regionais significativas, principalmente com relação ao Tipo de Dependência Administrativa e o acesso à tecnologia, tão discutido e necessário nesse período com a chegada da pandemia do *Covid-19*. De fato, houve uma alteração no perfil dos interessados em se inscrever nesse exame nos últimos anos, porém não há indicativos de grandes alterações nos valores estimados para a nota geral. Desse modo, observou-se no ano de 2022, que o contexto social favoreceu o desempenho de cerca de 74% dos alunos Concluintes, porém, não foi determinante para outros 26%. Tais resultados contribuem para o debate acerca do desenvolvimento de políticas educacionais afirmativas e seus desafios a médio e longo prazo.

Palavras-chave: Modelos de Regressão Hierárquicos. Mineração de Dados Educacionais. Predição. ENEM.

ABSTRACT

Watanabe, A. F. L. **Impacts and challenges of education before, during and after the pandemic - A study with ENEM data in recent years using hierarchical models**. 2023. 86p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Given the relevance of ENEM (the Brazilian National High School Exam) to the Brazilian educational scenario, it is worth investigating the mechanisms that affect student's performance and their association with several conditions, including the student's social context. Here, the Hierarchical Regression Models fitting of data from 2018 to 2022 made suggested significant regional differences, mainly regarding the 'Type of Administrative Dependency' and access to technology, much discussed and necessary after the *Covid-19* pandemic. In fact, there has been a change in the profile of students interested on this exam in recent years, but there is no indication of major changes in the estimated values for the overall score. It was observed that the social context favored the performance of about 74% of students, in 2022; however, this was not determinant for the remaining 26%. Such results add to the debate on the development of affirmative educational policies and their challenges in the medium and long term.

Keywords: Hierarchical Regression Models. Educational Data Mining. Prediction. ENEM.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABEP	Associação Brasileira de Empresas de Pesquisa
CCEB	Critério de Classificação Econômica Brasil ou Critério Brasil
CCI	Coeficiente de Correlação Intraclasse
CE	Características Escolares
CETIC	Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação
CF	Características Familiares
CI	Características Individuais
ENEM	Exame Nacional do Ensino Médio
FIES	Fundo de Financiamento Estudantil
IBGE	Instituto Brasileiro de Geografia e Estatística
ICC	Indicador de Capital Cultural [IBGE] Instituto Brasileiro de Geografia e Estatística
IDEB	Índice de Desenvolvimento da Educação Básica
IDHM	Índice de Desenvolvimento Humano Municipal
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
INSE	Indicador de Nível Socioeconômico
INSE_A	Indicador de Nível Socioeconômico Ampliado
ITC	Indicador de Tecnologia e Conectividade
LGPD	Lei Geral de Proteção de Dados
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
PIB	Produto Interno Bruto

PNAD	Pesquisa Nacional por Amostra de Domicílio
PROUNI	Programa Universidade para Todos
RMSE	Root Mean Squared Error
SINAES	Sistema Nacional de Avaliação da Educação Superior
SISU	Sistema de Seleção Unificada
TICs	Tecnologias da Informação e Comunicação
TRI	Teoria de Resposta ao Item
UNESCO	United Nations Educational, Scientific, and Cultural Organization

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

Para ter acesso às melhores oportunidades de trabalho, é necessário ter uma boa base educacional. Anos de estudo podem não garantir os melhores salários, mas irá possibilitar uma chance maior de se adequar às exigências do mercado de trabalho que vem se tornando cada vez mais competitivo. Uma vez que este acesso é a garantia de melhores oportunidades no futuro, a falta dele, implica em piores perspectivas profissionais e consequentemente menores salários e chance de mobilidade social.

No Brasil, a desigualdade educacional pode ser vista em vários segmentos da sociedade, mas com o surgimento da pandemia, causada pelo novo coronavírus (WU *et al.*, 2020) e sua respectiva doença (*Covid-19*), tais problemas se agravaram (UNESCO, 2020). Diante de restrições e medidas preventivas impostas pela nova realidade, as instituições educacionais ficaram fechadas e obrigadas a uma adaptação.

Alunos que tiveram acesso a recursos como internet e computadores, bem como a adequação do ensino para o remoto e ambiente propício para o estudo, continuaram se preparando, apesar das dúvidas e incertezas do momento. Por outro lado, muitos alunos encararam outra realidade, impulsionada agora pela exclusão digital[1] (CRISTO, 2020) e (FRITOLI; POLATO, 2021).

De acordo com o Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (CETIC, 2019), que monitora a adoção das TICs no Brasil, em 2019, 27% dos domicílios, não possuiam nem computador nem internet, sendo este percentual muito distinto com relação ao tipo de área, urbana(24%) / rural(46%), região geográfica, Sudeste(24%) / Sul(25%) / Norte(25%) / Centro-Oeste(28%) / Nordeste(33%) e classe social A(0%) / B(3%) / C(17%) / DE(49%).

Já o Instituto Brasileiro de Geografia e Estatística (IBGE, 2019), através da Pesquisa Nacional por Amostra de Domícios (PNAD Contínua), divulgou naquele ano, que apenas 41% dos domicílios havia computador e que em apenas 83% dos domicílios havia utilização de internet. Tais números reforçam a desigualdade digital e tecnológica no qual estavam inseridos muitos brasileiros, mesmo antes da pandemia, e sinalizou os desafios quanto à eficácia de alcance igualitário e de qualidade do ensino com o fechamento das escolas (TODOS-PELA-EDUCAÇÃO, 2021).

[1] Definida como a exclusão ao acesso pleno às TICs (Tecnologias da Informação e Comunicação). Pode ser subdivida em: exclusão instrumental (falta de equipamentos necessários para acessar a internet, como computador, celular, ...), exclusão infraestrutural (a regiao não possuir acesso à internet), exclusão financeira (incapacidade de pagar pelo serviço de internet).

Além disso, os estudantes das classes mais vulneráveis tiveram que lidar com outras situações como: aumento de abuso doméstico, cuidar dos irmãos, pressão por buscar atividade laboral para auxiliar nas despesas do lar, insegurança alimentar, entre outros. Sem a perspectiva de continuar estudando, o problema se agravou ainda mais, dado a possibilidade da repetência e evasão escolar. Ambos indicadores refletem o desperdício de recursos, pois mantêm o sistema educacional inchado, seja para atender alunos que demoram a completar, ou sequer completam o ciclo do ensino da educação básica (Ver Anexo A).

Na rede pública, cada estado/município se articulou de uma forma diferente ao desafio de implementar medidas para a continuidade do ensino durante a pandemia. Nem todos conseguiram se estruturar com as plataformas online, disponibilizar equipamentos, fornecer ou subsidiar acesso à internet ou materiais didáticos de apoio, apesar dos esforços de professores e escolas em manter uma comunicação direta com os alunos via e-mail, telefone, redes sociais e aplicativos de mensagens. Já o setor privado, com mais recursos e uma gestão mais dinâmica e autônoma, não teve problema com escalabilidade e rapidamente se adaptou às novas exigências. Na prática, a diferença de desempenho dos alunos de um sistema de ensino e outro que já era grande, tendeu a aumentar, criando um abismo maior entre os alunos da rede pública e as universidades públicas.

O Censo Escolar realizado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), autarquia federal vinculada ao Ministério da Educação do Brasil (INEP, 2020c), já apontava em 2019 a presença de fragilidades com relação à infraestrutura das escolas, principalmente no âmbito dos recursos tecnológicos disponíveis. Entre as escolas de ensino médio cuja dependência administrativa era privada, cerca de 72% possuiam acesso à internet para ensino e aprendizagem. Por outro lado, as dependências administrativas municipal, estadual e federal, tais percentuais correspondiam respectivamente a 53%, 66% e 86%, sendo bem distintos de acordo com a região geográfica.

Além destes segmentos que apresentam disparidade de acesso e oportunidades de aprendizagem, outros fatores também já indicavam grande influência no desempenho do estudante. Entre eles, destacam-se o nível educacional dos pais e a renda familiar como determinantes positivos nesse desempenho, pelo fato de estarem correlacionados com o nível de recursos disponíveis para serem gastos com educação, bens e serviços de maior qualidade para seus filhos influenciando direta e indiretamente no seu aprendizado (MORAES; PERES; PEDREIRA, 2021).

Uma vez que o ensino médio compreende os anos finais da formação básica educacional e é o período onde se amadurece a escolha pela carreira, direcionando o estudante para o ensino superior ou mercado de trabalho, tal segmento se mostrou mais impactado com o início da pandemia dada a preparação para os vestibulares e provas de larga escala, como o ENEM.

1.2 O ENEM

O ENEM, Exame Nacional do Ensino Médio, é uma prova de avaliação de desempenho individual realizado pelo INEP muito concorrida no país. Criado em 1998, tinha por objetivo avaliar a qualidade do ensino médio brasileiro, mas, com o passar do tempo, se tornou uma forma de acesso às universidades por meio dos programas como o SISU (Sistema de Seleção Unificada)[2], PROUNI (Programa Universidade para Todos)[3] e FIES (Fundo de Financiamento Estudantil)[4] representando a ampliação e a democratização do acesso à educação superior (INEP, 2020b).

O ENEM é uma prova diferente dos vestibulares tradicionais aplicados pelas próprias universidades, pois é elaborado com base em competências, habilidades e na transdisciplinaridade das questões. Aqui, o desafio é fazer com que o aluno saiba interpretar e desenvolver o senso crítico, dedicando mais esforços à multidisciplinaridade dos conhecimentos do que somente à memorização dos conteúdos (JUNIOR, 2021).

Atualmente, o modelo de prova do ENEM consiste em 180 questões objetivas de múltipla escolha estruturadas por competências: Linguagens, Códigos e suas Tecnologias (LC), Ciências Humanas e suas Tecnologias (CH), Ciências da Natureza e suas Tecnologias (CN), Matemática e suas Tecnologias (MT), além de uma proposta de Redação (RE). Os participantes também respondem a um questionário com questões sobre seu nível socioeconômico, família, educação e trabalho.

As questões são elaboradas e corrigidas de acordo com a Teoria de Resposta ao Item (TRI) [5].

---

[2] SISU: Sistema eletrônico gerido pelo MEC, reúne as vagas ofertadas por instituições públicas de ensino superior de todo o Brasil, sendo a maioria delas ofertadas por instituições federais. O processo de seleção dos estudantes é autônomo em relação àqueles realizados no âmbito das demais instituições de ensino superior, e é efetuado com base nos resultados do ENEM de acordo com a Lei de Cotas e outras políticas de ações afirmativas que podem ser adotadas pelas instituições públicas de ensino superior. <https://acessounico.mec.gov.br/sisu>

[3] PROUNI: Programa que tem por objetivo destinar bolsas de estudos em instituições de ensino superior privadas, em cursos de graduação e sequênciais de formação específica. É direcionado aos estudantes de baixa renda originários do ensino médio público ou do ensino privado, neste caso, na condição de bolsistas. Parte de seus recursos, são destinados aos negros, indígenas e portadores de deficiência. <http://prouniportal.mec.gov.br/tire-suas-duvidas-pesquisa/o-prouni/47-como-funciona-o-prouni>

[4] FIES: Programa que tem como objetivo conceder financiamento aos estudantes em cursos superiores não gratuitos, com avaliação positiva pelo Sistema Nacional de Avaliação da Educação Superior (SINAES) e por instituições de educação superior não gratuitas aderentes ao programa. É direcionado aos estudantes de baixa renda. O valor financiado poderá ser pago num prazo posterior a sua formação. <http://portalfies.mec.gov.br/?pagina=home>

[5] Conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar uma certa resposta a um item como função dos parâmetros do item e da habilidade do respondente. Assim, quanto maior a habilidade, maior a probabilidade de acerto no item.

Desse modo, não só o total de acertos são contabilizados, mas também as características das questões e o padrão de respostas dos estudantes para o cálculo das notas. Assim, duas pessoas com a mesma quantidade de acertos na prova são avaliadas de forma distintas, pois o resultado depende de quais itens foram respondidos corretamente e quais não foram, o que expressa habilidades diferentes de cada um dos indivíduos. Além disso, as provas são comparáveis ano a ano. Isso permite a elaboração de provas diferentes com o mesmo grau de dificuldade (ANDRADE; TAVARES; VALLE, 2000).

1.3 Panorama Geral do ENEM nos últimos anos

Ao longo dos anos, a aplicação do ENEM no país se consolidou, investindo em acessibilidade, isenção da taxa de inscrição, ampliação de municipios para a aplicação do exame, além da sua utilização para políticas públicas pontuais e estruturais para melhoria do ensino brasileiro. Em números, a quantidade de inscritos que era de aproximadamente 200 M (mil) em 1998 aumentou, chegando a quase 9 MM (milhões) de inscritos em 2014 e 2016, decaindo posteriormente para um patamar menor que 7 MM de inscritos (INEP, 2019) e (INEP, 2020b) (Figura 1).

Com o início da pandemia e a necessidade de garantir condições sanitárias adequadas para a realização desse exame, mudanças significativas ocorreram. Normalmente era realizado entre os meses de outubro e novembro, mas, o ENEM 2020, ocorreu em janeiro de 2021, após muita discussão sobre o seu adiamento. Iniciou o ENEM Digital, um modelo opcional de exame realizado em ambiente virtual. Foi necessário a reaplicação do exame no

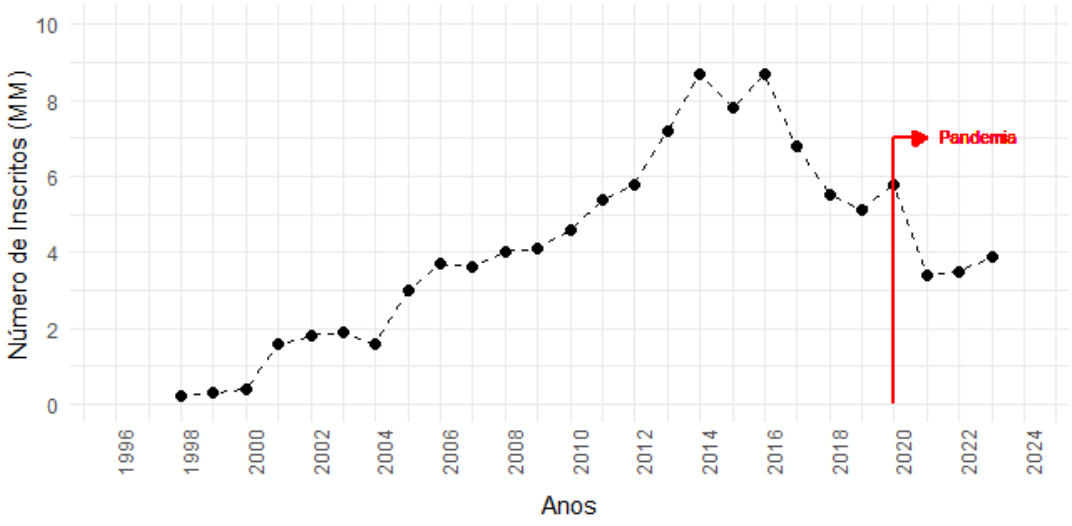


Figura 1 – Número de Inscritos no ENEM desde sua primeira edição em 1998. Fonte: (INEP, 2019) e (INEP, 2020b).

estado do Amazonas e dois municípios de Rondônia que não fizeram as provas nas datas regulares devido a decretos locais. Mesmo assim, observou um significativo percentual de ausentes na versão escrita (52%), digital (68%) e na reaplicação (72%) dos inscritos (INEP, 2020a).

Já o ENEM 2021, voltou a ser realizado em novembro. Apesar da diminuição do percentual de ausentes nos dois dias da prova, notou-se uma diminuição expressiva no número de inscritos. Enquanto que no ENEM 2020 havia aproximadamente 5.8 MM de inscritos, nas realizações posteriores, esse número ficou abaixo dos 4 MM.

Um levantamento realizado pelo INEP (2023) entre os participantes do ENEM 2022 (aproximadamente 1MM de respostas), apontou os principais hábitos de estudo durante o segundo ano da pandemia. Entre os resultados, observou-se que 6% interromperam os estudos, 77% disseram que aprendiam mais estudando de forma presencial e 78% vivenciaram algum tipo de problema para estudar ou manter-se informado. Já os principais motivos apontados por não conseguir se dedicar aos estudos foram: 34% teve dificuldade de compreender o conteúdo por falta de explicação de um professor ou outra pessoa em tempo real e 31% se sentiu desestimulado por não ter colegas com quem interagir sobre o que estava estudando. Além disso, 14% tiveram que trabalhar ou trabalhavam em serviços essenciais e 8% precisaram cuidar dos irmãos ou ajudar alguém doente (Figura 2).

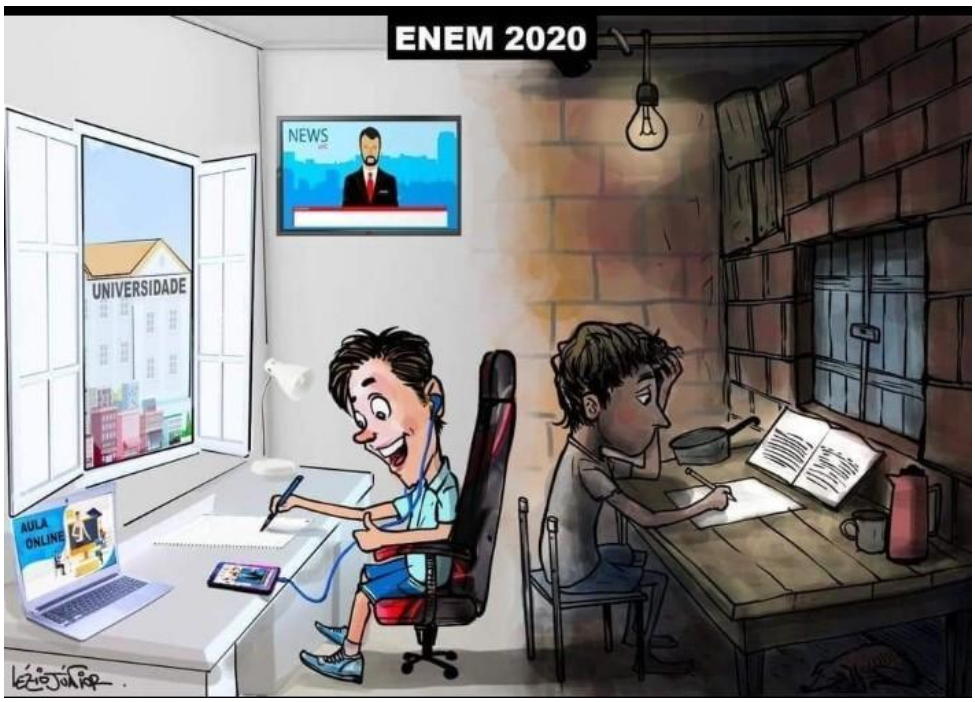


Figura 2 – Ilustração baseada na perspectiva da preparação dos alunos para o ENEM 2020. Disponível em: https://brainly.com.br/tarefa/27396432. Acesso em: 25 fev. 2023.

Com relação a infraestrutura para estudar ou manter-se informado, 44% disseram ter tido algum tipo de dificuldade, entre elas, 59% devido a conexão da internet, 27% por causa do equipamento (computador ou notebook) pouco disponível por ser compartilhado com outras pessoas ou com configuração insuficiente. Já os tipos de ajuda que recebeu com mais frequência foram: para acessar a internet (40%), com explicações de conteúdo (24%) e com a alimentação (14%).

Dada a importância que o ENEM tem no cenário educacional brasileiro, muitos estudos já foram realizados com o intuito de verificar quais fatores interferem no desempenho dos estudantes ao longo dos anos. Fatores estes relacionados principalmente com as características individuais, familiares, escolares, regionais e de práticas/qualificação docente, conforme será descrito no Capítulo 3.

É com o intuito de dar continuidade no acompanhamento dos fatores mais relevantes associados com o desempenho do estudante nesse exame que o presente estudo pretende analisar dados mais recentes através de técnicas de mineração de dados educacionais.

De fato, tais informações foram profundamente afetadas pelo momento caótico e incerto vivido com a pandemia e podem não refletir o perfil de todos os possíveis candidatos que poderiam ter prestado esse exame além do alto percentual de candidatos ausentes.

Porém, indicadores observados ano a ano, poderão ser úteis em verificar se as ações afirmativas propostas pelas políticas públicas anteriores e atuais conseguem mitigar o problema de acessibilidade ao ensino superior via ENEM.

2 OBJETIVOS

2.1 Objetivos Gerais e Específicos

2.1.1 Objetivo Geral

Investigar o perfil socioeconômico dos estudantes que se inscreveram no ENEM nos últimos anos através de técnicas de mineração de dados, verificando sua relação com o desempenho no exame e a alteração desse comportamento principalmente após início da pandemia.

2.1.2 Objetivos Específicos

- Descrever o perfil socioeconômico dos estudantes que se inscreveram no ENEM no período de 2018 a 2022, de acordo com o desempenho geral neste exame;
- Propor métodos preditivos que permitam identificar quais fatores socioeconômicos estão mais associados a um melhor desempenho em cada ano de realização do exame;
- Comparar metodologias, observando sua performance, vantagens e desvantagens;
- Verificar a capacidade preditiva de um determinado método em diferentes anos de realização do exame.

2.2 Justificativa e Motivação

Dada a importância do ENEM no cenário das políticas públicas relacionadas à educação básica e superior no país, pretende-se analisar e quantificar o quão a pandemia alterou o perfil socioeconômico dos estudantes inscritos nesse exame, apresentando um panorama geral nos últimos anos. Tais informações poderão contribuir para um mapeamento mais específico dos segmentos mais atingidos pela pandemia em termos de déficits educacionais possibilitando o contínuo monitoramento da eficácia das ações afirmativas de políticas públicas no país.

2.3 Organização do Trabalho

Este estudo será estruturado em mais seis capítulos, além desta introdução, organizados da seguinte forma:

- **Capítulo 3**: revisão da literatura e trabalhos relacionados mais recentes destacando as edições do ENEM e a metodologia utilizada para análise das informações em cada um deles;
- **Capítulo 4**: fundamentação teórica, conceitos e métodos aplicados neste trabalho;
- **Capítulo 5**: metodologia utilizada para coleta, análise, pré-processamento e processamento das informações;
- **Capítulo 6**: resultados obtidos referentes às comparações ano a ano e metodológicas;
- **Capítulo 7**: discussão dos resultados;
- **Capítulo 8**: considerações finais, bem como as limitações do trabalho.

3 TRABALHOS RELACIONADOS

Uma revisão sistemática dos estudos publicados nos últimos dez anos sobre a influência de características socioeconômicas e escolares no desempenho dos estudantes considerando às realizações do ENEM de 1998 a 2019 foi realizada por Dutra, Júnior e Fernandes (2023). De maneira geral, identificaram que os fatores mais relevantes nessa associação foram: renda familiar, nível de escolaridade dos pais, idade, sexo e raça (entre os itens referentes às questões socioeconômicas) e a caracterização das escolas de origem dos estudantes (tanto com relação à estrutura física quanto pedagógica). A Figura 3 apresenta uma nuvem de palavras como resultado da frequência de ocorrências de termos correspondentes aos atributos dessa pesquisa.

Para obter tais resultados, observaram que diferentes métodos de mineração de dados foram aplicados, sejam estes com foco preditivo ou descritivo. Para o primeiro caso, destacaram primeiramente a utilização das Árvores de Decisão e os métodos Ensemble, seguido dos Modelos de Regressão Logística e Regressão Linear, como os principais algoritmos de Aprendizado Supervisionado voltados para as tarefas de classificação e regressão, ora utilizando como rótulo o nível de desempenho alto/baixo do estudante ora mensurando o desempenho médio geral (ou de uma competência específica). No segundo caso, destacaram o método Apriori, seguido pelo k-means, como os principais algoritmos de Aprendizado Não Supervisionado, sejam estes voltados para as tarefas de associação ou de agrupamento.

Figura 3 – Nuvem de palavras gerada na revisão sistemática da literatura proposta por Dutra, Júnior e Fernandes (2023)

Dados mais recentes do ENEM envolvendo o período de início da pandemia, foram explorados nos estudos de Cruz *et al.* (2022) (ENEM 2019 a 2020), através de Modelos de Regressão Linear, em Noguera e Aguiar (2023) (ENEM 2017 a 2020), através de Árvores de Decisão, Algorítimos SVM e Redes Neurais e em NETO (2023) (ENEM 2017 a 2021) através de Análises de Cluster.

O primeiro estudo comparou os anos estudados e constatou pequena queda no desempenho médio dos estudantes em 2020. O segundo estudo, focou no desenvolvimento de uma arquitetura baseada em *data warehousing* e a partir desta comparou diferentes algoritmos para predição do desempenho bom ou ruim dos participantes, sendo seus resultados muito semelhantes com relação aos fatores relevantes encontrados entre os anos estudados. Já o terceiro estudo, observou o comportamento de três grupos ao longo dos anos e constatou uma alteração no perfil dos estudantes, com o aumento dos grupos considerados mais favorecidos.

Uma outra abordagem muita utilizada em dados educacionais e mais especificamente nos dados do ENEM é o desenvolvimento de Modelos de Regressão Hierárquicos. Essa classe de modelos é uma extensão dos modelos de regressão tradicionais, cuja estrutura de agrupamento intrínseca aos dados pode ser observada e ajustada em níveis. Note que a estrutura de um sistema educacional é naturalmente organizado em níveis, já que um conjunto de alunos forma uma turma, um conjunto de turmas forma uma escola, escolas são agrupadas por município e assim por diante. Esse tipo de estrutura permite verificar o quão cada nível está associado com a variável resposta através da decomposição da variância estimada pelo modelo.

Entre os estudos mais recentes utilizando essa abordagem, estão os realizados por Jaloto e Primi (2021) (ENEM 2018), Duarte (2020) (ENEM 2019) e Brito e Pedroso (2023) (ENEM 2018 a 2021), cada qual considerou as informações referentes aos estudantes e as escolas ou municípios como níveis para avaliar o desempenho do estudante no exame.

Em Jaloto e Primi (2021), tal ajuste foi realizado em cada uma das competências, onde observou-se não só a influência de fatores socieconômicos no desempenho do estudante no exame, mas também o respectivo percentual da variabilidade atribuído aos tipos de administração escolar (municipal, estadual, federal e privada).

Em Duarte (2020), o ajuste foi realizado em duas competências, Linguagens e Códigos e Matemática, com o intuito de analisar o desempenho dos estudantes com e sem deficiência e o efeito dos que tiveram e não tiveram atendimento especializado. De maneira geral, além dos determinantes associados com o desempenho do estudante, observou-se que o efeito escola (seja ela pública ou privada) era maior entre os estudantes com deficiencia do que os estudantes sem deficiência em ambas as competências além de ressaltar a importância do atendimento especializado no diferencial da nota.

Em Brito e Pedroso (2023), o ajuste foi realizado considerando a média geral das notas, cujo objetivo era verificar o impacto da pandemia no desempenho dos estudantes do estado do Paraná. Para isso, dois períodos foram comparados: 2018 a 2019 (pré pandemia) e 2020 a 2021 (pós pandemia). Desse modo, além de verificar os fatores mais relevantes em cada biênio, observou o impacto do nível município na variabilidade das notas e a importância de incluir variáveis contextuais no ajuste, no caso, um indicador de desenvolvimento econômico e social municipal.

Aqui, foram destacados apenas os estudos mais recentes com foco preditivo no desempenho do estudante. É importante ressaltar que há muitos outros estudos envolvendo os dados do ENEM, sejam estes qualitativos, descritivos, descritos na visão docente ou agrupados por escola que não foram considerados aqui, mas que poderiam ser discutidos em outro momento [1].

[1] Principais fontes de pesquisa:
- Google Scholar: <https://scholar.google.com/>
- Scientific Eletronic Library Online - SciELO: <https://www.scielo.br/>
- Science Direct: <https://www.sciencedirect.com/>

4 FUNDAMENTAÇÃO TEÓRICA

Um dos objetivos da área de Ciência de Dados aplicada a dados educacionais, é aprimorar a qualidade dos sistemas e processos de ensino e aprendizagem tanto em ambientes individuais quanto em ambientes colaborativos.

De forma mais específica, Baker, Isotani e Carvalho (2011) e Costa *et al.* (2013), citam a necessidade e a importância de métodos e aplicações voltados para a identificação de problemas que afetem o desempenho dos alunos, meios de melhorar a qualidade do material didático, desenvolvimento de metodologias pedagógicas eficazes e de ferramentas que viabilizem a personalização dos ambientes educacionais de tal forma que contribuam para a melhoria das práticas em educação. Baker, Isotani e Carvalho (2011) citam ainda oportunidades na área educacional brasileira e o desafio de analisar e compreender o comportamento dos estudantes, principalmente pela diversidade populacional, cultural e econômica.

Uma vez que as informações analisadas aumentam em tamanho e dimensionalidade, pois não se restringem apenas à interação do aluno com um sistema educacional, e sim podem incluir dados administrativos (da escola, do professor, ...), dados demográficos (sexo, idade, ...), dados sobre motivação, estado emocional, habilidades em geral (SCHEUER, 2011), recomenda-se utilizar, na dinâmica de análise de dados, processos KDD (do inglês *Knowledge-Discovery in Databases*) que auxiliam de forma sequencial e padronizada, na extração de conhecimento em busca de padrões existentes a fim de gerar hipóteses e reflexões a cerca deste objetivo.

4.1 Etapas de um processo KDD

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), um processo KDD é interativo e iterativo e envolve algumas etapas que podem ser resumidas em:

- Objetivo: definição e entendimento do problema;
- Coleta: seleção das informações disponíveis e de interesse;
- Pré-Processamento: verificação, formatação e exploração das informações coletadas (checagem de inconsistências, dados incompletos, redundantes, outliers, entre outras análises que fizerem necessárias para a elaboração de uma base de dados apropriada);
- Transformação: principais tratativas nas informações de interesse que permitam a redução de dimensionalidade e/ou a agregação dos dados;

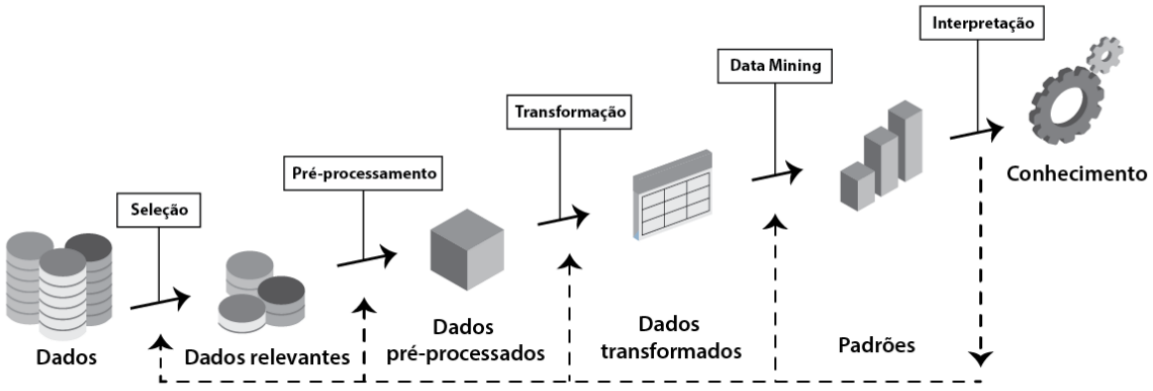


Figura 4 – Visão Geral das Etapas de um Processo KDD (Adaptação inspirada do trabalho de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)).

- Mineração de Dados: análise exploratória, escolha e avaliação do método/algoritmo de acordo com o objetivo proposto;
- Avaliação: validação e interpretação das informações obtidas;
- Implantação: utilização das informações para a tomada de decisão.

4.2 Mineração de Dados

É uma etapa do processo KDD que envolve a aplicação de métodos e algoritmos em busca de padrões consistentes através da detecção de relacionamentos sistemáticos entre as variáveis de interesse (AGARWAL, 2013).

A escolha deste método/algoritmo, irá depender do objetivo proposto e das variáveis disponíveis, sejam elas quantitativas discretas ou contínuas. De maneira geral, são divididos em problemas de predição e de descrição. No primeiro caso, algoritmos de aprendizagem supervisionadas são utilizados tanto para tarefas de classificação quanto regressão. Já no segundo caso, algoritmos de aprendizagem não supervisionados são utilizados para tarefas de agrupamento ou de associação de informações. [1].

Na área educacional, Costa *et al.* (2013) afirmam a importância da adequação dos algoritmos de mineração de dados existentes para lidar com especificidades inerentes ao contexto educacional, tais como a não independência estatística e a hierarquia dos dados.

[1] Há outros tipos de algoritmos como o aprendizado semi-supervionado, os por reforço, entre outros, bem como outros tipos de tarefas como detecção de anomalias.

4.3 Aprendizado Supervisionado

É uma classe de algoritmos de aprendizado de máquina utilizada quando temos uma váriavel resposta de interesse (há informações sobre os rótulos). Neste caso, o objetivo consiste em encontrar uma relação (através de uma função) entre as variáveis independentes/atributos e a variável resposta/target de tal maneira que auxilie na previsão de novas informações. Isso é realizado através da construção/treinamento de um modelo/algoritmo capaz de mapear as relações contidas no conjunto de dados buscando poder de generalização suficiente para observações não vistas anteriormente.

Podem ser utilizados tanto em problemas de regressão (quando a variável resposta de interesse é quantitativa contínua) ou de classificação (quando a variável resposta de interesse é quantitativa discreta).

4.4 Modelos de Regressão Lineares

São amplamente utilizados em diversas áreas. Na área educacional, podemos escrever a relação entre as variáveis socioeconômicas (variáveis independentes) e o desempenho do estudante (variável dependente) como:

$$y_i = \beta_0 + \beta_1 * x_i + \varepsilon_i \tag{4.1}$$

Onde:

y_i: variável dependente/resposta, com i = 1,2,...,n
β_0: intercepto geral
β_1: coeficiente de inclinação
x_i: vetor com uma determinada variável independente
ε_i: resíduo associado ao indivíduo
que segue uma distribuição Normal com média 0 e variância σ^2

Os termos β_0 e β_1 são considerados parâmetros fixos para os quais valores específicos serão estimados. Para que o modelo seja considerado válido, parte-se do pressuposto que: as informações (e resíduos) sejam independentes, que exista homocedasticidade (igualdade de variâncias) e que as variáveis independentes não sejam correlacionadas (ausência de multicolinearidade).

4.5 Modelos de Regressão Lineares Hierárquicos

Também conhecidos como Modelos Multiníveis, é uma classe de modelos que leva em consideração a estrutura de agrupamento intrínseca aos dados. Tal estrutura pode ser organizada em níveis, uma vez que os individuos são organizados naturalmente em grupos, sofre influência dos mesmos, apresentando características semelhantes entre si e distintas entre os demais (efeito contextual) (BROWN, 2021).

Na área educacional, essa estrutura pode ser organizada considerando como nível 1 o individuo e como nível 2 a escola ou município. Dessa forma, podemos escrever essa relação de acordo com a notação de Antonakis, Bastardoz e Rönkkö (2021) por:

$$y_{ij} = \beta_{0j} + \beta_{1j} * x_{ij} + \varepsilon_{ij} \tag{4.2}$$

Onde:

y_{ij}: variável dependente/resposta, com i = 1,2,...,n_j, j = 1,2,...,J e n > J

n_j: número de indivíduos (nível 1) associado ao j-ésimo município

J: número de municípios (nível 2)

β_{0j}: intercepto associado ao j-ésimo município

β_{1j}: coeficiente de inclinação associado ao j-ésimo município

x_{ij}: vetor com uma determinada variável independente associado ao i-ésimo individuo

ε_{ij}: resíduo associado ao i-ésimo indivíduo

Sendo:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{4.3}$$

Onde:

γ_{00}: é o intercepto geral ou valor médio

u_{0j}: é o efeito aleatório associado ao j-ésimo município

$$\beta_{1j} = \gamma_{10} + u_{1j} \tag{4.4}$$

Onde:

γ_{10}: é o coeficiente associado ao j-ésimo município

u_{1j}: é o efeito aleatório associado ao j-ésimo município

Pressupõe-se aqui que os termos ε_{ij}, u_{0j} e u_{1j} referentes aos nível 1 (individuo) e nível 2 (município), são independentes e identicamente distribuídos seguindo uma distribuição Normal com média 0 e determinada variância.

A equação 4.2 mostra que o valor esperado da variável dependente y depende linearmente do valor observado de x (nível 1). Já as equações 4.3 e 4.4 mostram que a dependência linear apresentada no nível 1 pode variar entre as unidades do nível 2 de modo que cada município terá um intercepto e uma inclinação exclusivos.

Os termos β_{0j} e β_{1j} são considerados parâmetros aleatórios, para os quais estimamos uma distribuição. Desse modo, diferente dos modelos de regressão lineares clássicos, os modelos lineares hierárquicos trazem maior flexibilidade no ajuste dos parãmetros, uma vez que distintas retas de regressão podem ser ajustadas além de levar em consideração a estrutura de dependência entre as observações. Destaca-se também a precisão dos erros-padrão associados às estimativas dos parâmetros.

De maneira simplificada, podemos reescrever o modelo acima como:

$$y_{ij} = \gamma_{00} + \gamma_{10} * x_{ij} + u_{0j} + u_{1j} * x_{ij} + \varepsilon_{ij} \tag{4.5}$$

Essa representação é chamada de modelo de efeitos mistos porque é composta de duas partes: a parte fixa contendo os coeficientes fixos que informa a tendência média (dois primeiros termos), e a parte aleatória contendo os termos de efeitos aleatórios que indica como as observações variam ao redor da média de acordo com os níveis (demais termos).

Para a estimação dos parâmetros, geralmente utiliza-se os métodos de máxima verossimilhança ou máxima verossimilhança restrita e a escolha de um método de otimização. Essas estimativas são utilizadas para estimar as variâncias e covariâncias no primeiro e segundo nível da hierarquia.

Uma vez calculado os componentes de variância, podemos mensurar a dependência entre as observações através do coeficiente de correlação intraclasse (CCI). Este indica a proporção da variância explicada pela estrutura de agrupamento, ou seja, a parcela da variância de um nível comparado com a variância total. Assim, coeficientes muito baixos indicam que não há necessidade de realizar uma análise em mais de um nível. Por outro lado, quanto maior este coeficiente, maior é a variabilidade das notas associado à variável daquele nível. Tais informações possibilitam uma adequada compreensão dos fenômenos ocorridos no nível 1, uma vez que é possível verificar como as relações condicionais estabelecidas entre as variáveis explicativas de diferentes níveis é exercida sobre ele (NAKAGAWA; JOHNSON; SCHIELZETH, 2017).

Dado que na análise hierárquica vários modelos podem ser ajustados aos dados, estes devem ser escolhidos de forma cautelosa. Isso significa analisar previamente quais níveis serão considerados, quais informações individuais e contextuais serão incluídas, que efeitos de interação são esperados entre variáveis de níveis diferentes, com o intuito de obter um modelo capaz de explicar a maior quantidade de variância a partir do menor número de variáveis independentes (PUENTE-PALACIOS; LAROS, 2009).

4.6 Árvores de Decisão

Do inglês, *Decicion Tree*, é um algoritmo que divide iterativamente o conjunto de dados, criando uma estrutura semelhante a uma árvore em que cada nó representa uma tomada de decisão. Esse processo de divisão que se inicia no nó raiz até o nó terminal (folha), é realizado buscando a melhor variável independente e o melhor ponto de corte até que algum critério de parada seja alcançado (SONG; YING, 2015).

Em problemas de regressão, é comum minimizar o MSE (Mean Squared Error), ou seja, a variabilidade de cada subconjunto resultante da divisão do nó. A média no nó terminal será a previsão final.

É um método rápido na previsão de novos registros, de fácil interpretação e robusto na presença de pontos extremos, porém, é um método instável e sensível a pequenas mudanças no conjunto de dados, podendo levar a má precisão das estimativas.

4.7 Florestas Aleatórias

Do inglês, *Random Forest*, é um algoritmo que combina a predição de várias árvores de decisão no processo de tomada de decisão (floresta). Cada árvore é treinada a partir de uma amostra *bootstrap* [2] de registros e uma subamostra aleatória de variáveis independentes que capturam diferentes tendências no conjunto de dados. Esse processo, evita que as árvores fiquem correlacionadas aumentando a capacidade de generalização, porém não são facilmente interpretáveis (CUTLER; CUTLER; STEVENS, 2011).

Este algoritmo é considerado um método de aprendizado do tipo "*ensemble*", onde um grupo de algoritmos fracos são combinados para formar um algoritmo mais forte.

[2] Amostra aleatória gerada a partir de um método de amostragem com reposição da amostra original.

4.8 Avaliação e Validação

De maneira geral, métodos de previsão eficientes requerem boa capacidade de explicação dos dados conhecidos e de generalização para novas observações. Para isso, podemos utilizar algumas métricas para avaliar a qualidade do ajuste de cada método.

É comum dividir o conjunto de dados em dados de treino (informações que serão utilizadas no desenvolvimento/treinamento do modelo/algoritmo) e dados de teste (informações que serão utilizadas na validação do modelo/algoritmo) de forma aleatória, para a obtenção dessas métricas. A proximidade dos valores obtidos no cálculo dessas métricas em ambas amostras é um indicativo da capacidade de generalização do modelo ajustado para informações não vistas anteriormente [3].

Para comparação entre metodologias distintas, será calculado o RMSE (Root Mean Squared Error) que reflete a dispersão dos erros/resíduos, obtida pela média dos erros quadráticos entre as previsões estimadas e os valores reais. Quanto menor esse indicador, menor é a dispersão e portanto melhor é o ajuste do método aos dados.

Também será calculado o R^2 (Coeficiente de Determinação) que mensura a proporção da variabilidade da variável resposta explicada pelas variáveis independentes selecionadas pelo método aplicado.

Para comparação de dois modelos de regressão aninhados, medidas como AIC (Akaike Information Criterion) e BIC (Bayesian Information Criterion) podem ser complementadas, onde baixos valores dessas medidas também indicam melhor ajuste do modelo proposto (BATES *et al.*, 2015).

Quanto a validação das suposições dos modelos de regressão hierárquicos, será realizada uma análise de resíduos e uma análise dos efeitos aleatórios (NOBRE; SINGER, 2007).

[3] Há várias outras formas de detectar o problema do overfitting, que é justamente quando o modelo/algoritmo se ajusta perfeitamente aos dados mas perde a capacidade de generalização para novas informações como: a redução da sua complexidade com a remoção de variáveis independentes irrelevantes ou altamente correlacionadas, análise de uma amostra externa, ou seja, de validação, aplicação de técnicas de regularização como a L1 (Lasso) e L2 (Ridge), para penalizar os seus coeficientes.

5 METODOLOGIA

5.1 Casuística

Este estudo será conduzido utilizando os microdados do ENEM de 2018 a 2022 disponibilizados anualmente pelo INEP (INEP, 2022). Tais informações contemplam os dados do participante, da escola, do local de aplicação da prova, da prova objetiva, da redação e do questionário socioeconômico. Para a adequação à LGPD (Lei Geral de Proteção de Dados), adotou um modelo de questionário simplificado com o objetivo de eliminar da base pública, variáveis que facilitassem a identificação do participante.

O ambiente Google Colaboratory (Google COLAB), foi utilizado para armazenamento e pré-processamento de todas as bases tendo como interpretador a linguagem de programação Python. Já para o ajuste/treinamento dos modelos/algoritmos propostos foi utilizado a linguagem de programação *R* (4.3.1) através do ambiente *RStudio*.

5.2 Variáveis e Volumetria

Em cada ano de realização do ENEM aqui estudado, havia mais de 70 colunas e aproximadamente 5.5MM, 5.1MM, 5.8MM, 3.4MM e 3.5MM de registros respectivamente para os anos de 2018 a 2022. Destas foram selecionadas aproximadamente 50 colunas [1], utilizadas como variável dependente, independente ou como critério de inclusão/exclusão. É importante ressaltar que o campo "NU_INSCRICAO" é uma máscara gerada sequencialmente que substitui o número real de inscrição do participante.

5.2.1 Critérios de Inclusão/Exclusão

- Não ter feito a opção de treineiro;
- Ter frequentado a modalidade de ensino regular do Ensino Médio;
- Ter finalizado todos os testes objetivos e a redação;
- Não ter sido eliminado em quaisquer teste objetivo;
- Ter respondido ao questionário socioeconômico, em sua maioria, principalmente às questões referentes à escola.

[1] Foram desconsideradas infomações sobre o local de aplicação da prova, código/cor da prova, vetor com respostas e gabaritos e especificidades da correção da prova de redação.

5.2.2 Variáveis Dependentes/Target

- Notas obtidas em cada competência: Linguagens, Códigos e suas Tecnologias; Ciências Humanas e suas Tecnologias; Ciências da Natureza e suas Tecnologias; e Matemática e suas Tecnologias;
- Nota obtida na Redação.

Tais informações foram utilizadas para cálculo da média aritmética do participante, aqui denominado por nota geral.

5.2.3 Variáveis Independentes/Explicativas

- Características Individuais:

  Que correspondem às informações de sexo, cor/raça, faixa etária, estado civil e tipo de língua estrangeira;

- Características Familiares;

  Que correspondem aos itens relacionados à bens de consumo, renda familiar, estrutura física da residência, escolaridade e ocupação dos pais/responsáveis;

- Características Escolares;

  Que correspondem aos itens relacionados à escola, como tipo de localização, tipo de dependência administrativa e região geográfica.

Uma vez que as informações referentes às características familiares deste questionário expressam de forma unilateral o contexto social do participante, a elaboração de um indicador que sintetize o nível socioeconômico do mesmo se faz importante para a compreensão da desigualdade educacional e sua relação com o desempenho escolar. Assim, esse constructo não observável diretamente passa a ser mensurado através da união dessas informações do questionário.

O Critério de Classificação Econômica Brasil (CCEB) ou simplesmente Critério Brasil, (BRASIL, 2022), é um estudo socioeconômico desenvolvido pela Associação Brasileira de Empresas de Pesquisa (ABEP) para estimar o poder de compra das classes econômicas brasileiras. Aspectos como estrutura física da residência, bens de consumo e escolaridade do chefe da família são levados em consideração para o cálculo de uma pontuação que situa o indivíduo nas classes A, B1, B2, C1, C2, D e E, ou seja, do maior ao menor poder aquisitivo.

Tais resultados são considerados muito fidedignos quanto a capacidade de consumo das famílias com rendimento mensal de até R$30.000,00, o que compreende a maioria absoluta da população.

A partir desse critério, quatro indicadores foram criados:

1. Indicador de Nível socioeconômico (INSE):

   Consolida os itens associados a bens de consumo e estrutura física da residência;

2. Indicador de Nível socioeconômico Ampliado (INSE_A):

   Consolida todos os itens associados a bens de consumo e estrutura física da residência;

3. Indicador de Tecnologia e Conectividade (ITC):

   Consolida os itens associados a comunicação;

4. Indicador de Capital Cultural (ICC):

   Consolida os itens associados a escolaridade e ocupação dos pais/responsáveis.

Entretanto, dado que nem todos os itens do questionário socioeconômico respondido pelo participante do ENEM são contemplados no Critério Brasil, a criação desses indicadores exigiu decisões metodológicas para operacionalizar tais constructos. Os detalhes do sistema de pontuação final adotado pode ser visto no Anexo B. O cálculo final de cada indicador é dado pela soma dos pontos de cada grupo de itens.

5.3 Pré Processamento

Verificação e validação dos microdados disponibilizados, no período analisado. Pontos importantes observados:

- Alteração do número de categorias e da descrição de algumas questões de um ano para outro. Ação: padronização de todas as categorias das variáveis;
- Presença de questões com alto percentual de missings. Ação: criação de uma categoria específica para esse grupo e análise da sua relevância;
- Diferentes pontos de corte adotados para a definição da faixa de renda mensal da família. Apesar dessa diferença, entende-se que esta reflete um ajuste anual da faixa, que equivale a mesma proporção de renda ano a ano.
- Verificação de duplicidade e pontos discrepantes;
- Normalização e padronização das variáveis quantitativas;

- Recategorização de informações com baixa prevalência, mas priorizando sua relevância.
- Análise exploratória de dados (descritivas e gráficas).

5.4 Processamento

Principais pontos observados nessa etapa:

- Amostragem e divisão dos dados em treino, para o ajuste dos modelos (75%), e teste, para verificação do seu desempenho (25%);
- Seleção de atributos mais relevantes e criação de novos atributos visando a melhoria no desempenho do algoritmo;
- Análise da Correlação entre as variáveis quantitativas através do cálculo dos coeficientes de correlação de Pearson e Spearman;
- Codificação das variáveis;
- Ajuste dos modelos de regressão hierárquicos em cada ano de realização do ENEM estudado através da biblioteca *lme4* (BATES *et al.*, 2023) e verificação do grau de importância das variáveis independentes através do módulo das estimativas dos efeitos fixos ajustados;
- Cálculo e interpretação do CCI através da biblioteca *performance* (LÜDECKE *et al.*, 2021);
- Treinamento do algoritmo Random Forest através da biblioteca *tidymodels* da linguagem de programação R, tendo como *engine* a estrutura *ranger* (WRIGHT; WAGER; PROBST, 2020) e verificação do grau de importância das variáveis independentes através da biblioteca *vip* (GREENWELL, 2023);
- Análise do perfil associado com o desempenho baseado nas variáveis mais relevantes presentes em cada método;

5.5 Avaliação e Validação

Análise da performance de cada método através do:

- Cálculo das medidas: RMSE, R^2, AIC e BIC na amostra de treinamento;
- Cálculo e análise do RMSE na amostra de teste;
- Análise de Resíduos e construção dos gráficos:
    1. Histograma dos Resíduos: verificação do comportamento da sua distribuição;
    2. Normalidade dos Resíduos: verificação da adequação dessa suposição;
    3. Resíduos vs Ordem: verificação de pontos discrepantes, tendências e variação residual.
- Análise de Efeitos Aleatórios: idem à análise de resíduos.

6 RESULTADOS

6.1 Análise Descritiva

A Figura 5 descreve a volumetria informada na seção 5.2 de acordo com a situação de conclusão do participante por ano de realização do ENEM[1].

No primeiro painel, observa-se a alteração da distribuição percentual de inscritos pós-pandemia em cada situação de conclusão, com o aumento da participação de Cursistas/Concluintes e a diminuição da participação dos Egressos. Já no segundo painel, exceto o ano de 2020, ano de início da pandemia, o percentual de inscritos que compareceram nos dois dias de realização da prova foi aproximadamente constante ao longo dos anos, de acordo com a situação de conclusão, sendo os Cursistas, os que mais compareceram (acima de 81%) e os Não concluintes e Não Egressos, os que menos compareceram (em torno de 50%). O terceiro painel destaca apenas os participantes Concluintes, cujo público apresentava características escolares, o que correpondia a 65%, chegando a 89% da base total de Concluintes presentes no periodo estudado.

A Figura 6 descreve a volumetria dos municípios referente à escola do participante Concluinte. De maneira geral, observa-se aproximadamente a mesma quantidade de municípios ao longo dos anos, o que indica a representatividade das escolas de todo o Brasil nesse exame. Seguindo a tendência do país de concentração de pessoas nos grandes centros urbanos, cerca de 5% dos municípios com mais inscritos correspondiam a mais de 58% do total de alunos Concluintes presentes (painel 1).

Com o intuito de ter representatividade de alunos por município, optou-se por seguir as análises apenas com os municípios que apresentavam mais de cem Concluintes, o que correspondia a 15% em 2020, chegando a 29% em 2018 do número total de municipios (painel 2). Já em número de alunos Concluintes, expressou uma perda de 12% a 17% do total, o que significou um público final menor que 1MM (milhão) por ano de realização de ENEM (painel 3).

---

[1]

- Cursista: Participante que cursava o Ensino Médio e não concluiria no mesmo ano de realização do ENEM.
- Concluinte: Participante que cursava o último ano do Ensino Médio e o concluiria no mesmo ano de realização do ENEM.
- Egresso: Participante que já concluiu seus estudos no Ensino Médio.
- Não concluinte e Não egresso: Participante que saiu do sistema escolar e que não efetivou nova matrícula em anos seguintes.

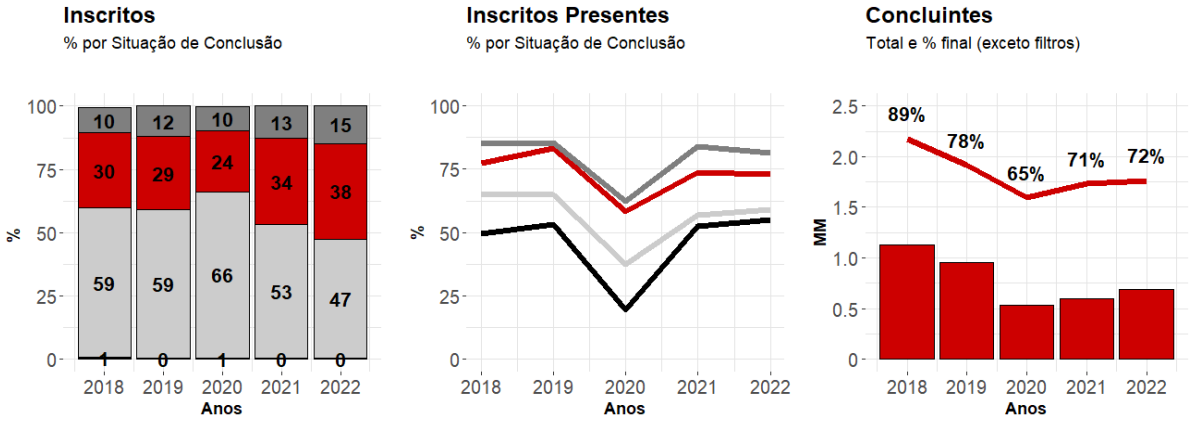


Figura 5 – Quantidade de Inscritos (painel 1), Inscritos Presentes por Situação de Conclusão (**Cursista**, **Concluinte**, Egresso e **Não Concluinte_Não Egresso**) (painel 2) e quantidade de Concluintes exceto filtros (painel 3) por ano de realização do ENEM.

A Figura 16 (Anexo C) apresenta a distribuição percentual dos Concluintes de acordo com as características individuais, familiares e escolares por ano de realização do ENEM. De maneira geral, houve uma alteração na distribuição percentual de várias informações relacionadas com o perfil dos estudantes nos anos pós-pandemia, principalmente de acordo com o "Tipo de Dependência Administrativa", com o aumento da participação do ensino "Federal" e "Privado" e diminuição do ensino "Estadual".

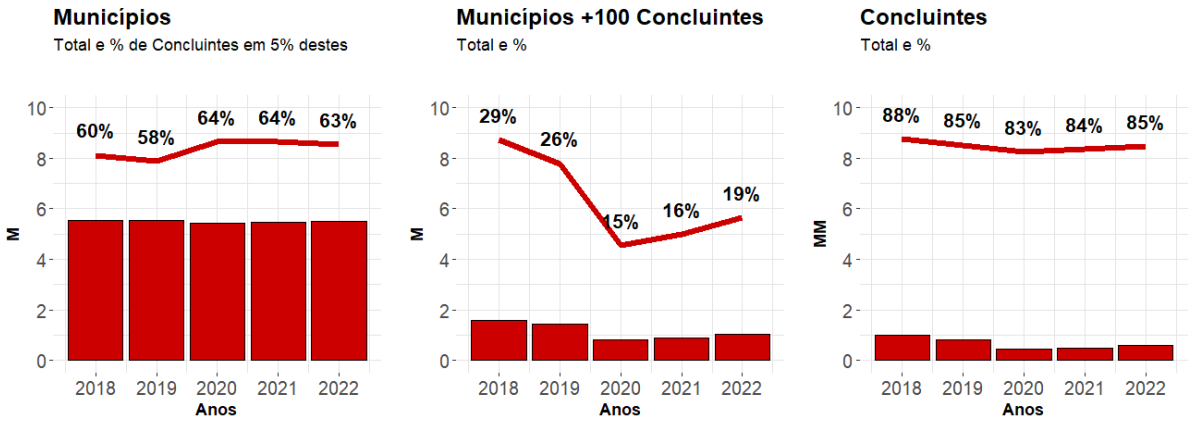


Figura 6 – Quantidade de Municípios (painel 1), Municípios com mais de cem alunos Concluintes (painel 2) e quantidade de alunos Concluintes final (painel 3) por ano de realização do ENEM.

Também nota-se um aumento da participação dos estudantes com maior "background familiar". Já entre as características individuais, destacam-se as variáveis "Cor_Raça", "Tipo de Língua Estrangeira" e "Faixa Etária" com as maiores alterações.

A Figura 17 (Anexo C) apresenta a média da nota geral para cada categoria de acordo com as características individuais, familiares e escolares, por ano de realização do ENEM. De maneira geral, todos os subgrupos de características observados apresentam diferenças significativas entre as categorias com relação ao desempenho médio dos estudantes, que perduram ao longo do tempo de maneira constante. Destaque para a variável "Renda¨ que apresentou maior variabilidade no desempenho médio entre as categorias.

6.2 Análise da Correlação

A Tabela 1 apresenta o cálculo da correlação de Spearman entre as variáveis independentes quantitativas e estas com a variável resposta [2].

	INSE_A	ICC	ITC	Renda Familiar	Nota Geral
INSE	0,948 a 0,963	0,524 a 0,617	0,705 a 0,738	0,669 a 0,738	0,389 a 0,452
INSE_A		0,543 a 0,633	0,727 a 0,758	0,683 a 0,756	0,411 a 0,468
ICC			0,553 a 0,608	0,637 a 0,716	0,476, a 0,498
ITC				0,651 a 0,721	0,410 a 0,455
Renda					0,473 a 0,511

Tabela 1 – Correlação entre as variáveis independentes quantitativas e estas com a variável resposta. Valor mínimo e máximo obtido entre os anos estudados de realização do ENEM.

Como era esperado, os indicadores "INSE" e "INSE_A" apresentaram uma forte correlação, seguido do indicador "ITC" e da variável "Renda Familiar"[3]. Também observa-se uma correlação moderada entre a variável resposta e todas as variáveis independentes. Desse modo, para evitar a multicolinearidade no ajuste dos modelos de regressão, foram excluídos os indicadores "INSE" e "INSE_A".

[2] Também foi calculado a correlação de Pearson, sendo seus resultados e conclusões obtidas similares aos resultados da correlação de Spearman aqui apresentados.

[3] Originalmente era dividida em dezessete faixas. Para ter uma idéia da correlação desta com as demais variáveis, optou-se por adotar o ponto médio de cada faixa. Para a primeira faixa, cuja descrição era 'Nenhuma Renda', convencionou-se adotar 0. Já a última faixa, não havendo limite de faixa, convencionou-se adotar o valor inicial da faixa.

6.3 Ajuste de Modelos Hierárquicos

6.3.1 Capacidade Preditiva

Com o intuito de verificar a relação entre o perfil socioeconômico do aluno Concluinte e o seu desempenho geral no ENEM através dos modelos de regressão hierárquicos, foram elaborados diferentes cenários considerando os dados de 2022.

Como variáveis independentes, foram utilizadas cada subgrupo de características e respectivas codificações descritos no Anexo C, exceto para os indicadores ICC e ITC, que foram utilizados na sua forma original, quantitativa. Também foi excluída a variável 'Estado Civil' dado sua intuitiva relação com a variável 'Faixa Etária'. Como níveis de hierarquia, adotou-se o aluno como nível 1 e municipio/unidade federativa como nível 2.

Uma vez que as estimativas dos efeitos fixos eram muito similiares, estas foram agrupadas em uma só categoria (agrupamento das categorias 'Preta_Parda' com 'NA' da variável 'Cor_Raça'). Também optou-e por agrupar as categorias 'Municipal' e 'Estadual' da variável 'Tipo de Dependência Administrativa' dada sua baixa representatividade por unidade federativa.

Um resumo das principais informações e métricas de validação obtidas em cada um dos cenários pode ser visto na Tabela 2 seguindo a sintaxe sugerida pela biblioteca *lme4* (Ver Anexo D) [4].

No cenário 1, foi ajustado o modelo nulo, ou seja, sem nenhum tipo de característica. Isso significa dizer que as diferenças de desempenho entre os alunos Concluintes será dado somente pela diferença entre municípios e unidades federativas. Nos cenários 2 ao 4, são observados a contribuição de algum tipo de característica no intercepto do modelo além das diferenças regionais. Nos demais cenários, são observados a contribuição de cada tipo de característica não só no intercepto, mas também no parâmetro de inclinação do modelo. Isso significa estimar um efeito para cada um dos municípios/unidades federativas (cenário 5) ou somente para a unidade federativa (cenários 6 a 8) daquela característica. Isso demonstra a flexibilidade da metodologia em estimar um efeito específico para cada localidade geográfica e sua relação com determinada característica de interesse.

As variáveis, tipo de dependência administrativa e o indicador ITC são em particular, informações de grande influência no desempenho escolar e que apresentam diferenças regionais significativas, principalmente neste período pandêmico. Estas foram consideradas tanto como efeito fixo (efeito geral) quanto efeito aleatório (efeito específico) nos cenários analisados.

[4] Valores de CCI multiplicados por cem.

CENÁRIO	EFEITO ALEATÓRIO		MODELO	TREINO				TESTE	CCI
	Intercepto	Inclinação		RMSE	R2	AIC	BIC	RMSE	
1	S		Nota_Geral ~ (1|UF/Município)	84,5	0,119	5.133.111	5.133.155	84,8	11,9
2	S		Nota_Geral ~ CI + (1|UF/Município)	78,9	0,212	5.073.478	5.073.588	79,2	8,1
3	S		Nota_Geral ~ CI + CF + (1|UF/Município)	72,1	0,367	4.994.587	4.994.785	72,4	6,9
4	S		Nota_Geral ~ CI + CF + CE + (1|UF/Município)	69,1	0,413	4.957.373	4.957.604	69,5	5,8
5	S	S	Nota_Geral ~ CI + CF + CE + (1 + CF1|UF/Município)	69,0	0,407	4.956.461	4.956.736	69,4	6,8
6	S	S	Nota_Geral ~ CI + CF + CE + (1|Município) + (1 + CF1|UF)	69,1	0,413	4.956.883	4.957.136	69,5	6,1
7	S	S	Nota_Geral ~ CI + CF + CE + (1|Município) + (1 + CE1|UF)	68,9	0,416	4.954.807	4.955.092	69,3	6,1
8	S	S	Nota_Geral ~ CI + CF + CE + (1|Município) + (1 + CE1 + CF1|UF)	68,9	0,416	4.954.648	4.954.978	69,3	6,3

Tabela 2 – Resumo das métricas de validação referente aos ajustes dos modelos de regressão hierárquicos para o ano de 2022 considerando diferentes parametrizações e os tipos de características: individuais (CI), familiares (CF) ou escolares (CE) tanto na amostra de treinamento quanto na amostra de teste. A notação 1 ao lado da sigla indica apenas uma variável escolhida.

De maneira geral, a medida que incluimos algum tipo de característica associada ao aluno Concluinte, as métricas de validação indicam uma melhoria no ajuste tanto na amostra de treinamento, quanto na amostra de teste. A partir do cenário 4, não houve grandes alterações nas métricas de validação analisadas. Para todos esses cenários foram realizados a análise dos resíduos e de efeitos aleatórios. Tais análises indicaramm grande heterogeneidade entre os municípios e entre as unidades federativas. Mais detalhes dessa análise para o cenário 8 podem ser vistos no Anexo E.

Uma vez ajustado o modelo completo, é importante entender o quão homogêneo os municipios e unidades federativas são e o quão essa variação depende de uma determinada característica de interesse, através do coeficiente de correlação intraclasse (CCI). A medida que incluímos algum tipo de característica no modelo, menor é a variabilidade do desempenho do aluno atribuída às diferenças entre os municípios e/ou unidades federativas e maior é a atribuição dada ao perfil socioeconômico. No cenário 8, por exemplo, aproximadamente 6% da variação das notas pode ser atribuída às diferenças regionais.

Na Figura 7 se encontra um resumo dos efeitos aleatórios obtidos no cenário 8 entre as unidades federativas de acordo com a sua magnitude por região geográfica [5]. De maneira geral, a região sudeste se destaca por possuir as unidades federativas com os maiores efeitos positivos tanto no intercepto, quanto atrelado ao tipo de dependência administrativa federal, seguido pelas regiões sul e nordeste. Quanto ao indicador ITC, destaca-se a região nordeste com o maior percentual de unidades federativas com maior impacto positivo do uso de tecnologia no desempenho da nota geral [6].

[5] Notação: Para os efeitos aleatórios referentes ao Intercepto, Tipos de Dependência Administrativa (Federal e Privada): Baixo (<-5), Nulo (-5 a 5) e Alto (>5). Para o efeito aleatório relacionado ao ITC: Baixo (<-0,5), Nulo (-0,5 a 0,5) e Alto (> 0,5).

[6] Aqui, optou-se por apresentar o percentual de unidades federativas por região geográfica. Outra visão seria apresentar o percentual de Concluintes correspondentes à cada unidade federativa e esta por região geográfica.

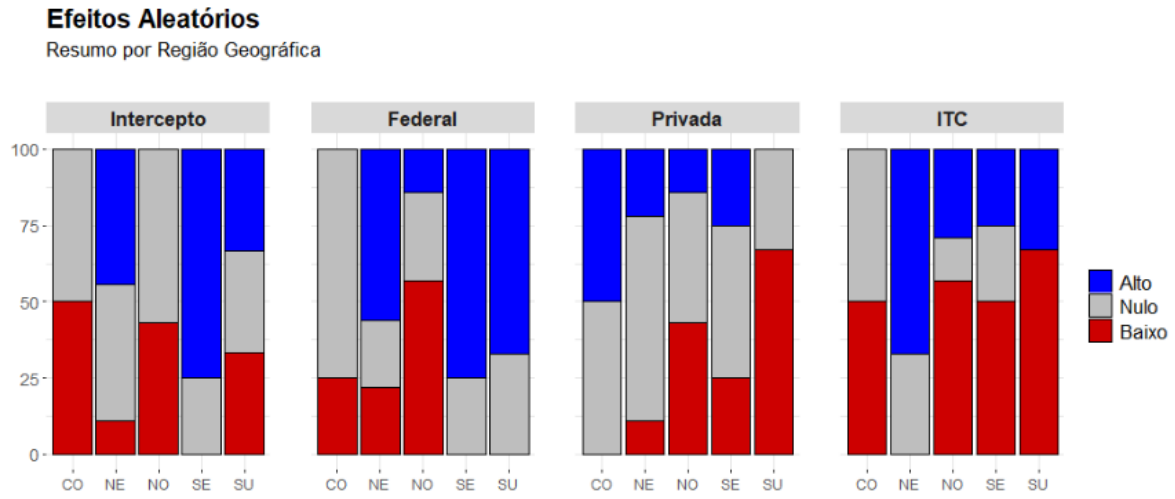


Figura 7 – Resumo dos Efeitos Aleatórios estimados entre as unidades federativas por região geográfica de acordo com a magnitude do efeito considerando o ajuste dos dados de 2022.

De fato, a região sudeste se destaca individualmente em cada competência, tendo os maiores percentuais de alunos Concluintes com as maiores notas. Em Matemática, por exemplo, mais de 51% dos alunos Concluintes com nota superior a 750, era dessa região, haja vista que a representatividade do sudeste no total de alunos é em torno de 39%. Há de se destacar também o crescente desempenho das regiões nordeste e sul nos últimos anos.

Apesar do ajuste ser considerado satisfatório do ponto de vista geral, nota-se uma baixa capacidade preditiva na faixa de baixíssimo desempenho com a presença de altos valores discrepantes.

Outro modo de verificar a capacidade preditiva do modelo de forma sintética, é resumir a nota geral em um critério baixo/alto desempenho no exame (Figura 8).

Dado a mediana da nota geral como ponto de corte, para o cenário 8, observamos uma acurácia de aproximadamente 74%, ou seja, a quantidade de alunos cujo desempenho foi corretamente classificado. Outros 15% corresponderam àqueles que tiveram alto desempenho no exame, porém a predição foi baixa e outros 11% tiveram baixo desempenho no exame, mas a predição foi alta [7].

[7] Tais medidas são muito usadas num contexto de classificação. Aqui, optou-se por calculá-las apenas para ter uma visão mais intuitiva da performance geral do modelo proposto. A medida *Precision* reflete o percentual de classificações corretas entre as classificações positivas (ou alto desempenho) e a medida *Recall* reflete as classificações corretas entre os que de fato tiveram alto desempenho. Também foi obtido o percentual dos falsos negativos e falsos positivos ao invés da taxa, dando ênfase no total de alunos incorretamente classificados pelo método proposto.

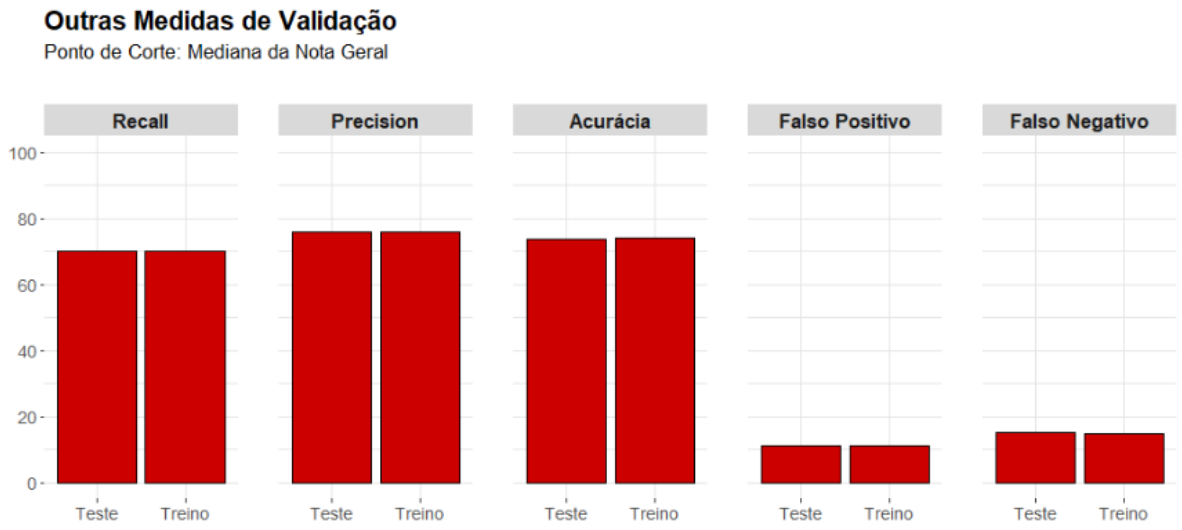


Figura 8 – Medidas de Validação obtidas para os dados de 2022, nas amostras de treino e teste, considerando a mediana da nota geral como ponto de corte para baixo/alto desempenho no exame.

6.3.2 Predição Ano a Ano

Com o intuito de verificar a alteração do perfil socioeconômico ao longo dos anos de realização do ENEM e sua relação com o desempenho geral do aluno Concluinte, foi ajustado um modelo de regressão hierárquico considerando as premissas adotadas no cenário 8 visto na subseção anterior. Na Figura 9 se encontram o grau de importância relacionada as estimativas de efeitos fixos dos parâmetros e respectivos erros-padrão para cada modelo ajustado anualmente, com destaque para os anos pós-pandemia.

De maneira geral, as variáveis "Tipo de Dependência Administrativa", a "Renda Familiar"e a "Faixa Etária", foram as informações mais relevantes na predição da nota geral do aluno Concluinte. Desse modo, alunos provenientes de escolas federais ou privadas, alta renda e idade até 18 anos, tendem a apresentar maior desempenho neste exame em quaisquer ano. Destacam-se também as variáveis 'Cor_Raça' e 'Tipo de Língua Estrangeira' pelo aumento das estimativas nos anos pós-pandemia.

Além disso, as estimativas dos erros-padrão associado a cada efeito fixo nos anos pós-pandemia, tenderam a ser maiores do que as mesmas estimativas nos anos antes da pandemia, indicando maior variabilidade das notas atreladas à essas características. De fato, a de ressaltar que as amostras referentes aos anos pós-pandemia são menores do que as amostras referentes aos anos antes da pandemia e isso impacta o cálculo dessa estimativa.

Na Tabela 3 está uma comparação entre as predições obtidas anualmente considerando dois distintos perfis:

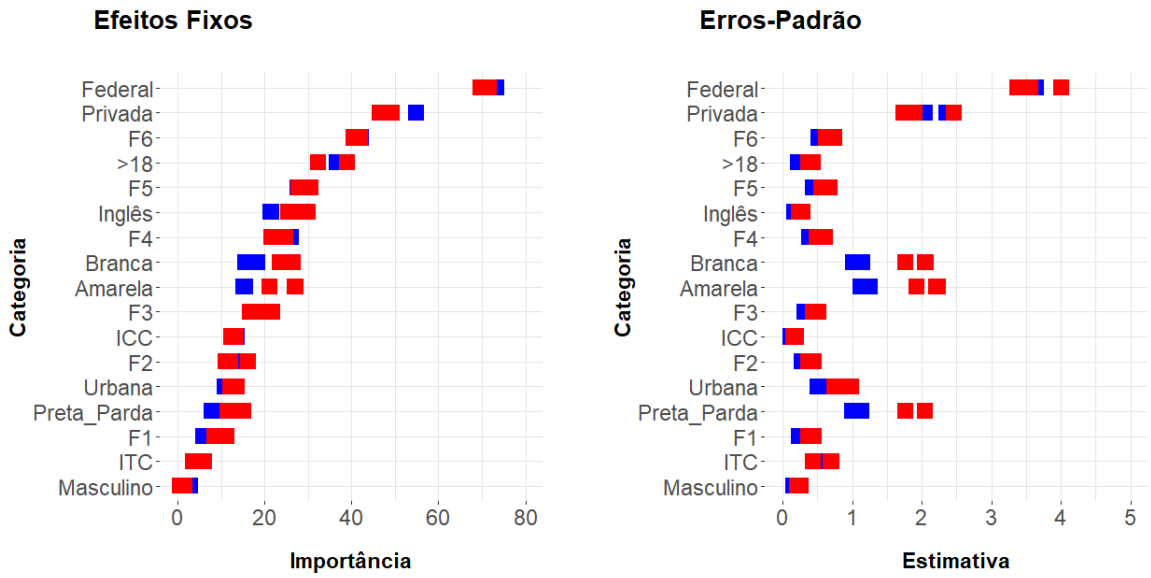


Figura 9 – Grau de importância associado às estimativas de efeitos fixos e respectivos erros-padrão estimados anualmente através dos modelos de regressão hierárquicos (**2018 e 2019** e **2020, 2021 e 2022**).

- Perfil(A), aluna, indígena, maior que dezoito anos, baixa renda e nível educacional dos pais, sem acesso tecnológico, cuja escola era de dependência administrativa estadual/municipal localizada na zona rural e optou pelo espanhol como língua estrangeira;

- Perfil (B), aluno, branco, com até dezoito anos, alta renda e nível educacional dos pais, que teve acesso à vários meios tecnológicos, cuja escola era de dependência administrativa federal localizada na zona urbana e optou pelo inglês como língua estrangeira.

Uma vez que as questões do ENEM são corrigidas pela TRI, elas são comparáveis ano a ano. Para estes casos, nota-se uma diferença na previsão da nota geral aproximadamente constante ao longo dos anos, exceto para o ano de início da pandemia. Isso traz um indicativo que, independente da alteração ocorrida no perfil socioeconômico do aluno Concluinte entre os anos, a diferença esperada entre a nota geral obtida pelos alunos mais propensos e menos propensos a ter um bom desempenho é aproximadamente a mesma.

É importante ressaltar que em 2020, ano de início da pandemia, a nota geral esperada para o Perfil A diminuiu e a do Perfil B aumentou. Tal resultado sugere que os alunos em maior situação de vulnerabilidade sofreram mais os efeitos negativos da pandemia.

Perfil	2018	2.019	2020	2021	2022
A	411	404	393	400	411
B	692	684	699	686	691
Diferença	281	280	306	286	280

Tabela 3 – Comparação de previsões da nota geral estimadas pelos modelos de regressão hierárquicos ano a ano fixando dois perfis de aluno, considerando apenas os efeitos fixos.

Para verificar a capacidade preditiva de um modelo ajustado em determinado ano com relação aos anos seguintes, foi calculado a medida RMSE na amostra de teste entre os mesmos municípios de todos os anos (Tabela 4). De maneira geral, apesar do ajuste anual ser mais preciso e portanto menor valor do RMSE, não houve grandes alterações na predição da nota geral nos anos posteriores. De fato, os valores do RMSE nos anos de realização do ENEM pós-pandemia são maiores do que os valores do RMSE antes da pandemia, indicando maior variabilidade entre as notas nesses anos.

ANOS	Previsão				
	2018	2.019	2020	2021	2022
2018	65,3	65,4	72,3	71,2	70,0
2019		65,0	72,4	70,9	70,2
2020			71,8	70,9	69,8
2021				70,3	70,2
2022					69,2

Tabela 4 – Métrica de validação RMSE obtida na amostra teste em cada modelo de regressão hierárquico ajustado anualmente e sua previsão nos anos posteriores.

6.3.3 Comparação com outro Método

Com o intuito de comparar a performance dos modelos de regressão hierárquicos com outro método preditivo, foi realizado o treinamento do algoritmo *Random Forest* (RF) para os dados de 2022. Para a parametrização deste algoritmo, foi realizado previamente o *tuning* dos principais hiperparâmetros em uma estrutura *cross validation*, com 5 *folds* para obter o melhor *grid* a ser utilizado no algoritmo final, tendo como métrica de validação o RMSE e o R^2 [8].

[8] Os principais hiperparâmetros observados foram: número de árvores geradas no treinamento, quantidade de variáveis independentes selecionadas por árvore, número mínimo de observações em um nó para divisão.

As variáveis independentes utilizadas bem como as amostra de treinamento e teste foram as mesmas utilizadas nos modelos de regressão hierárquicos, com a inclusão das variáveis 'INSE_A' e 'Estado Civil'. Os níveis de hierarquia, município e unidade federativa, também foram incluídos como variáveis independentes qualitativas a serem testadas.

Na Tabela 5 se encontra um resumo dos principais hiperparâmetros observados e respectivas medidas de validação para os cinco melhores ajustes. De modo geral, a predição da nota geral pelo algoritmo *Random Forest* se aproximou da predição obtida pelo modelo de regressão hierárquico tanto na amostra de treinamento quanto na amostra teste.

CENÁRIO	GRID			TREINO		TESTE
	Árvores	Variáveis	Observações	RMSE	R2	RMSE
1	50	8	500	69,2	0,410	69,3
2	50	5	500	69,2	0,409	69,4
3	20	8	500	69,2	0,408	69,4
4	20	5	500	69,2	0,408	69,4
5	50	8	1000	69,3	0,407	69,5

Tabela 5 – Resumo dos melhores hiperparâmetros obtidos no treinamento do algoritmo Random Forest e respectivas métricas de validação na amostra de treinamento e teste considerando os dados de 2022.

Na Figura 10 se encontra as variáveis independentes com maior grau de importância obtidas no cenário 1 do algoritmo *Random Forest* e do cenário 8 do modelo de regressão hierárquico considerando os dados de 2022. Em ambos métodos preditivos, as variáveis mais associadas com o desempenho geral na prova coincidem em sua maioria, com destaque para as variáveis 'Tipo de Dependência Administrativa' e 'Renda Familiar'. De fato, uma metodologia demonstra uma visão independente do município/unidade federativa e outra, uma visão dependente destes [9].

[9] Para melhor visualização e comparação de ambos os métodos, optou-se por dividir o grau de importância por 10^7, no caso do algoritmo *Random Forest* e por representar pela categoria de maior coeficiente absoluto no caso do Modelo de Regressão Hierárquico.

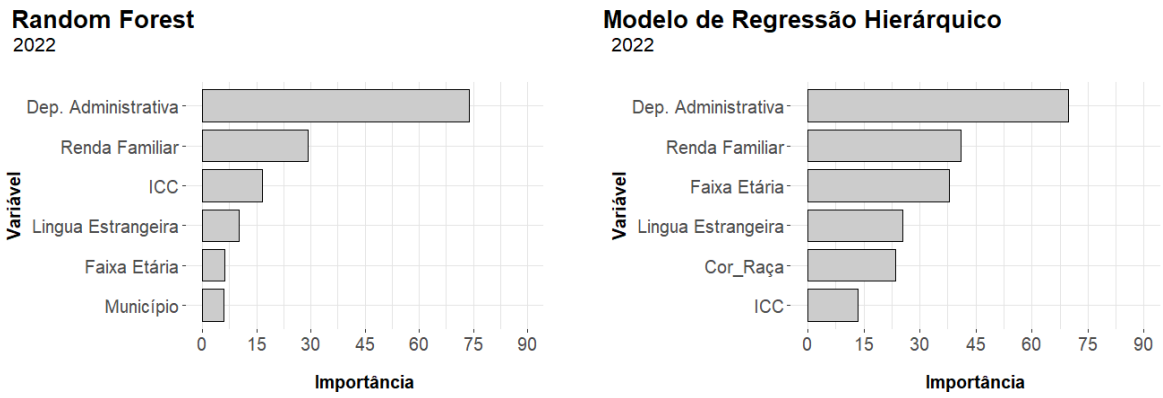


Figura 10 – Grau de importância associado às principais variáveis obtidas pelo algoritmo Random Forest e pelo modelo de regressão hierárquico para os dados de 2022.

7 DISCUSSÃO

Tendo em vista a relevância que o ENEM tem no cenário educacional brasileiro, é importante investigar os mecanismos que afetam o desempenho dos estudantes e sua associação com diversos fatores, de tal forma que identifique subgrupos mais e menos propensos a ir bem no exame.

Diante disso, muitos estudos já foram realizados, sendo evidente a identificação de dois grupos: um caracterizado pela alta renda, acesso a bens duráveis e incentivo à escolarização, o que lhe confere maior chance de ingressar em qualquer curso, principalmente àqueles de maior prestígio, e outro, caracterizado pela precariedade em relação à renda, à escolaridade/profissão parental e pela menor chance de ingressar no ensino superior.

Nesse estudo, os resultados observados reforçam essa desigualdade em dados recentes e contribui para o debate sobre os desafios educacionais a serem enfrentados e a necessidade de investir em ações efetivas e abrangentes que estimulem a equidade do acesso dos alunos à universidade. Isso pode ser observado não só pela nota média esperada para um determinado aluno ao longo dos últimos anos, mas também pela alteração do perfil interessado em prestar esse exame, principalmente após o início da pandemia.

A utilização de modelos de regressão hierárquicos permitiu verificar com maior especificidade as diferenças regionais, sejam essas atribuídas aos municípios ou unidades federativas, a medida que favoreceu a organização da estrutura dos dados em níveis sendo estas observadas de acordo com alguma característica de interesse. Pode-se observar a distribuição e dispersão dos efeitos aleatórios, quais localidades apresentaram os maiores e menores efeitos, se houve alguma tendência positiva/negativa regional, entre outros.

Em termos de efeitos fixos, uma vez que a variável "Tipo de Dependência Administrativa" se sobrepôs às variáveis relativas à condição socioeconômica do estudante e suas características individuais, traz à tona o quão o fortalecimento da educação é capaz de minimizar o impacto do contexto social e sua influência no desempenho do aluno.

Independente disso, é importante ressaltar que as condições sociais, as características individuais e escolares favorecem ao menos 74% dos alunos Concluintes, conforme aponta a previsão geral do modelo. Porém, ela não é determinante para outros 26% dos alunos, dos quais, 15% tiveram alto desempenho, mesmo diante de um contexto desfavorável e outros 11% tiveram baixo desempenho, apesar de uma situação mais promissora.

Um levantamento feito pela plataforma QEdu [1] através das informações do INEP, constatou a grande diferença entre o percentual de alunos com aprendizado adequado no Ensino Médio considerando as disciplinas de Português e Matemática dependendo do Tipo de Dependência Administrativa (Tabela 6).

Tipo de Dependência Administrativa	Português			Matemática		
	2017	2019	2021	2017	2019	2021
Estadual	23	32	30	4	6	4
Federal	71	75	75	42	41	35
Privada		74	70		41	34

Tabela 6 – Percentual de estudantes com aprendizado adequado no Ensino Médio em Português e Matemática no período de 2017 a 2021 por Tipo de Dependência Administrativa. Fonte INEP.

Isso reforça a importância dessa informação como preditora do desempenho do aluno nos últimos anos e o impacto negativo que a pandemia trouxe no aprendizado dos alunos em todos os segmentos.

Outros indicadores como taxa de rendimento (aprovação, reprovação, abandono), distorção idade-série e IDEB (Índice de Desenvolvimento da Educação Básica) também podem ser observados nessa plataforma que apresenta um panorama geral da educação brasileira por etapa escolar e estado brasileiro nos últimos anos, tendo como fonte de dados o INEP.

Sendo a educação um investimento de longo prazo, com custos presentes e benefícios futuros, quais as necessidades educacionais que surgem em um ambiente pós pandemia? São novos problemas frente a desafios antigos, para os quais ainda estavam por resolver?

Moraes, Peres e Pedreira (2021) sugerem que é necessário superar as barreiras tecnológicas com políticas de inclusão digital e apoio a carreira docente, qualificando-o para o ensino online. Além disso, os gestores devem desenvolver políticas que estimulem o acesso e a permanência escolar desde os primeiros anos de estudo até o final do ensino médio, com a melhoria da qualidade pedagógica e humana no contexto social que cada um vive.

[1] Site: https://qedu.org.br/brasil. Acesso em: 10 nov. 2023

O monitoramento contínuo dos indicadores educacionais bem como a análise de desempenho dos alunos periodicamente também se faz necessário uma vez que diversas ações/propostas vão sendo colocadas em prática, como é o caso do ENEM Seriado[2] ou mesmo sobre a discussão do Novo Ensino Médio[3] que afetam, direta ou indiretamente as avaliações em larga escala.

Paralelo a isso, está o público que costuma não comparecer nos dois dias de prova, mas que também carece de políticas específicas que estimulem ao retorno no sistema educacional.

A educação sempre será um dos principais pontos de atenção na elaboração de estratégias para a redução das desigualdades sociais.

[2] Nova forma de ingressar no ensino superior no país, através de provas anuais, iniciando no primeiro ano do ensino médio. Ao final, a pontuação das três avaliações poderá ser utilizada em programas como o PROUNI, FIES, ... Essa proposta tem o intuito de medir os conhecimentos dos estudantes de forma gradual e progressiva, permitindo-o acompanhar sua evolução em cada ano de ensino. Fonte: <http://portal.mec.gov.br/component/content/article?id=89391.>

[3] Nova organização curricular, mais flexível com oferta de itinerários formativos, com foco nas áreas de conhecimento e na formação técnica e profissional. Fonte: <http://portal.mec.gov.br/component/content/article?id=40361>

8 CONCLUSÃO

Avaliar o desempenho cognitivo de estudantes, cada qual em sua realidade é algo muito complexo. Diversos fatores inerentes ao processo de ensino/aprendizagem sejam estes sociais, familiares, escolares, até o próprio conhecimento adquirido por meio da sua vida acadêmica e experiências extracurriculares podem implicar em resultados individuais na prova.

No que diz respeito à análise, a proposta do estudo se limitou a estudar apenas as informações socioeconômicas relacionadas ao estudante. Portanto, há a necessidade de integração entre os dados do ENEM e outras fontes de informação a fim de investigar interferências mais específicas para uma melhor compreensão dessa relação, sempre em conformidade com a LGPD.

Informações presentes no Censo Escolar consistem em indicadores relevantes relativos a alguma etapa do processo educacional, sendo consolidados na visão escola, municipio ou unidade federativa. Estes refletem a complexidade da gestão escolar, a formação e regularidade do corpo docente, o nível socioeconômico escolar, entre outros e são disponibilizados pelo INEP, uns anualmente, outros a cada dois anos.

O IBGE é outra fonte de informação que também pode ser utilizada para agregar especificidades relativas aos municípios ou unidades federativas como: densidade ou estimativa populacional, PIB (Produto Interno Bruto), IDHM (Índice de Desenvolvimento Humano Municipal) entre outras informações contextuais.

Além disso, é necessário estudos locais para compreendermos melhor a pluralidade regional em relação à abrangência do país, bem como estudos específicos para determinados segmentos da sociedade como estudantes ausentes, deficientes, análises por situação da conclusão do ensino médio, entre outros, que contribuam para o fomento de políticas públicas no país.

Outra estratégia a ser utilizada para aprimorar a criação de indicadores que sintetizam um constructo latente específico, é a utilização da TRI. Nesse sentido, Barros, Senkevics e Oliveira (2019) utiliza o questionário socioeconômico respondido pelos candidatos do ENEM (2011 a 2017) para construir uma escala de nível socioeconômico específica elaborada por meio do modelo de respostas graduais da TRI para em seguida correlacioná-la com a nota média das quatro provas objetivas do exame. A correlação de Pearson para estes dados foi de 0,49. Aqui, este cálculo foi de acima 0,41 para o indicador INSE e acima de 0,43 para o indicador INSE_A quando utilizado o Critério Brasil além da inclusão da nota da redação no cálculo da média anualmente.

Em termos metodológicos, optou-se por ajustar o modelo de regressão hierárquico completo dada a pequena quantidade de váriaveis independentes a serem testadas, favorecendo assim a comparação dos modelos ano a ano. Porém, é importante ressaltar que é possível a realização da seleção automática de variáveis (KUZNETSOVA; BROCKHOFF; CHRISTENSEN, 2017) ou mesmo a aplicação de algum método de regularização (GROLL; GROLL, 2017) em modelos de regressão hierárquicos. No entanto, essa implementação pode ser complexa e exige um bom entendimento de como isso afetará a estrutura hierárquica dos dados.

Já a aplicação do algoritmo Random Forest nesse conjunto de dados apresentou resultados muito semelhantes aos dos Modelos de Regressão Hierárquicos, porém não fica claro o efeito das evidências por localidade geográfica. De fato, é necessário destacar que este algoritmo é muito eficaz quando temos muitas informações e muitas variáveis independentes a serem testadas.

REFERÊNCIAS

AGARWAL, S. Data mining: Data mining concepts and techniques. *In*: **2013 International Conference on Machine Intelligence and Research Advancement**. [*S.l.: s.n.*], 2013. p. 203–207.

ANDRADE, D. F. de; TAVARES, H. R.; VALLE, R. da C. Teoria da resposta ao item: conceitos e aplicações. 2000.

ANTONAKIS, J.; BASTARDOZ, N.; RÖNKKÖ, M. On ignoring the random effects assumption in multilevel models: Review, critique, and recommendations. **Organizational Research Methods**, Sage Publications Sage CA: Los Angeles, CA, v. 24, n. 2, p. 443–483, 2021.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. **Revista Brasileira de informática na educação**, v. 19, n. 02, p. 03, 2011.

BARROS, G. T. de F.; SENKEVICS, A. S.; OLIVEIRA, A. S. de. Indicador de nível socioeconômico dos inscritos no enem. **Textos para discussão. Serie Documental**, INEP. MEC. Diretoria de Estudos Educacionais, n. 47, p. 01–72, 2019.

BATES, D. *et al.* Fitting linear mixed-effects models using lme4. **Journal of Statistical Software**, v. 67, n. 1, p. 1–48, 2015.

BATES, D. *et al.* Package 'lme4'. 2023. Available at: <https://cran.r-project.org/web/packages/lme4/lme4.pdf>.

BRASIL, C. Critério de classificação econômica brasil. **Associação Brasileira de Empresas de Pesquisa (ABEP)**, 2022. Available at: <https://www.abep.org/criterio-brasil>.

BRITO, W. H. de; PEDROSO, F. P. Impactos de variáveis socioeconômicas no desempenho no enem no primeiro biênio da pandemia de covid-19. **Revista de Gestão e Avaliação Educacional**, p. e84069–e84069, 2023.

BROWN, V. A. An introduction to linear mixed-effects modeling in r. **Advances in Methods and Practices in Psychological Science**, SAGE Publications Sage CA: Los Angeles, CA, v. 4, n. 1, p. 1–19, 2021.

COSTA, E. *et al.* Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. **Jornada de Atualização em Informática na Educação**, v. 1, n. 1, p. 1–29, 2013.

CRISTO, H. S. de. A quem serve o exame nacional do ensino médio em tempos de pandemia da covid-19 no brasil? **Revista Espaço Acadêmico**, v. 20, n. 224, p. 262–273, 2020.

CRUZ, R. C. d. *et al.* Uma avaliação empírica do exame nacional do ensino médio–enem: impacto da pandemia do covid-19 no desempenho dos participantes do enem 2020. Universidade Católica de Brasília, 2022.

CUTLER, A.; CUTLER, D.; STEVENS, J. Random forests. *In*: **Machine Learning - ML**. [*S.l.: s.n.*], 2011. v. 45, p. 157–176.

DUARTE, H. F. F. L. Estudo sobre o desempenho dos estudantes com deficiência no enem 2019. Universidade Federal de Viçosa, 2020.

DUTRA, J. F.; JÚNIOR, J. B. F.; FERNANDES, D. Y. de S. Fatores que podem interferir no desempenho de estudantes no enem: uma revisão sistemática da literatura. **Revista Brasileira de Informática na Educação**, v. 31, p. 323–351, 2023.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996.

FRITOLI, S. T. de L.; POLATO, A. D. M. Enem 2020 em tempos de pandemia: A análise de uma charge em perspectiva dialógica. **Humanidades & Inovação**, v. 8, n. 38, p. 334–348, 2021.

GREENWELL, B. M. Package 'vip' - variable importance plots. 2023. Available at: <https://cran.r-project.org/web/packages/vip/vip.pdf>.

GROLL, A.; GROLL, M. A. **Package 'glmmLasso' - Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation**. 2017.

INEP. **Sinopse Estatísticas do Exame Nacional de Ensino Médio 2019**. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Dados Abertos., 2019. Available at: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/sinopses-estatisticas/enem>. Access at: 27 feb. 2023.

INEP. **Enem 2020 - Resultados edição impressa, digital e PPL**. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Notícias. ENEM., 2020. Available at: <https://download.inep.gov.br/enem/resultados/2020/apresentacao_resultados_finais.pdf>. Access at: 04 mar. 2023.

INEP. **Histórico**. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Avaliações e Exames Educacionais. ENEM., 2020. Available at: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/historico>. Access at: 25 feb. 2023.

INEP. **Pesquisas Estatísticas e Indicadores Educacionais. Censo Escolar. Resultados**. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Censo Escolar., 2020. Available at: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar/resultados>. Access at: 25 feb. 2023.

INEP. **Microdados. ENEM.** Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Acesso á Informação. Dados Abertos., 2022. Available at: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados>. Access at: 04 mar. 2023.

INEP. **Painel da Pesquisa ENEM 2022 - Hábitos de Estudo na Pandemia**. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Censo Escolar., 2023. Available at: <https://www.gov.br/inep/pt-br/assuntos/noticias/enem/painel-apresenta-pesquisa-sobre-estudo-na-pandemia>. Access at: 09 sep. 2023.

JALOTO, A.; PRIMI, R. Fatores socioeconômicos associados ao desempenho no enem. **Em Aberto**, v. 34, n. 112, 2021.

JUNIOR, V. F. de S. Uma breve história do exame nacional do ensino médio-enem: Avanços e ranços até a era digital a brief history of the exame nacional do ensino médio-enem (national high school exam): Advances and weaknesses until the digital age. **Brazilian Journal of Development**, v. 7, n. 12, p. 120314–120325, 2021.

KUZNETSOVA, A.; BROCKHOFF, P. B.; CHRISTENSEN, R. H. lmertest package: Tests in linear mixed effects models. **Journal of Statistical Software**, v. 82, p. 1–26, 2017.

LÜDECKE, D. *et al.* performance: An r package for assessment, comparison and testing of statistical models. **Journal of Open Source Software**, v. 6, n. 60, 2021.

MORAES, C. P. de; PERES, R. T.; PEDREIRA, C. E. Eficácia escolar e variáveis familiares em tempos de pandemia: um estudo a partir de dados do enem. **Interfaces da educação**, v. 12, n. 35, p. 635–658, 2021.

NAKAGAWA, S.; JOHNSON, P. C.; SCHIELZETH, H. The coefficient of determination r 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. **Journal of the Royal Society Interface**, The Royal Society, v. 14, n. 134, p. 1–11, 2017.

NETO, N. W. Minerando dados para entender o impacto da pandemia da covid-19 no exame nacional do ensino médio. Universidade Federal do Maranhão, 2023.

NOBRE, J. S.; SINGER, J. da M. Residual analysis for linear mixed models. **Biometrical Journal: Journal of Mathematical Methods in Biosciences**, Wiley Online Library, v. 49, n. 6, p. 863–875, 2007.

NOGUERA, V. E. R.; AGUIAR, C. D. d. Análise de dados do enem baseada em data warehousing, mineração de dados, estatística inferencial e processamento paralelo e distribuído. 2023.

PUENTE-PALACIOS, K. E.; LAROS, J. A. Análise multinível: contribuições para estudos sobre efeito do contexto social no comportamento individual. **Estudos de Psicologia (Campinas)**, SciELO Brasil, v. 26, p. 349–361, 2009.

SCHEUER, B. M. M. O. **Educational Data Mining**. Encyclopedia of the Sciences of Learning, Springer, 2011. Available at: <https://www.cs.cmu.edu/~bmclaren/pubs/ScheuerMcLaren-EducationalDataMining-EncyOfLearningScience2011.pdf>. Access at: 28 may. 2023.

SONG, Y.-Y.; YING, L. Decision tree methods: applications for classification and prediction. **Shanghai archives of psychiatry**, Shanghai Mental Health Center, v. 27, n. 2, p. 130, 2015.

TODOS-PELA-EDUCAÇÃO. **Balanço 2020: Impacto da pandemia na educação vai além do fechamento de escolas.** 2021. Available at: <https://todospelaeducacao.org.br/noticias/relatorio-do-todos-impacto-da-pandemia-na-educacao-basica-tem-ido-alem-do-fechamento-\de-escolas/>. Access at: 03 mar. 2023.

UNESCO. **COVID-19 educational disruption and response**. United Nations Educational, Scientific, and Cultural Organization, 2020. Last Update: 21 apr. 2022. Available at: <https://www.unesco.org/en/articles/covid-19-educational-disruption-and-response>. Access at: 23 feb. 2023.

WRIGHT, M. N.; WAGER, S.; PROBST, P. Ranger: A fast implementation of random forests. **R package version 0.12**, v. 1, 2020.

WU, F. *et al.* A new coronavirus associated with human respiratory disease in china. **Nature**, v. 579, n. 7798, p. 265–269, 2020.

ANEXOS

ANEXO A – NOTÍCIAS RELACIONADAS

- BETIM, FELIPE. Governo adia Enem após pressão que trouxe à tona o fosso entre ensino público e privado. EL PAÍS, São Paulo, 20 mai. 2021. Sociedade, Disponível em: https://brasil.elpais.com/sociedade/2020-05-20/governo-adia-enem-apos-pressao-que-trouxe-a-tona-o-fosso-entre-ensino-publico-e-privado.html. Acesso em: 21 fev. 2023.

- FALCÃO, Lucas. Desigualdade e empobrecimento: os efeitos da pandemia nas notas do ENEM. UOL, São Paulo, 28 nov. 2021. Diálogos Públicos, Disponível em: https://noticias.uol.com.br/colunas/dialogos-publicos/2021/11/28/desigualdade-e-empobrecimento-os-efeitos-da-pandemia-nas-notas-do-enem.htm. Acesso em: 21 fev. 2023.

- IDOETA, Paula Adamo. Enem 2020: 5 pontos cruciais sobre o exame 'mais desafiador' de todos. BBC News Brasil, São Paulo, 6 jan. 2021. Disponível em: https://www.bbc.com/portuguese/brasil-55549706. Acesso em: 21 fev. 2023.

- IDOETA, Paula Adamo. Enem vai expor nova camada de exclusão entre alunos mais pobres, diz estudioso de desigualdade na educação. BBC News Brasil, São Paulo, 10 jan. 2021. Disponível em: https://www.bbc.com/portuguese/brasil-55596205. Acesso em: 21 fev. 2023.

- NOGUEIRA, Fernanda. Para ser inclusivo, Enem pós-pandemia precisa ir além da mudança de data. PorVir, 6 jul. 2020. Inovações em Educação, Disponível em: https://porvir.org/para-ser-inclusivo-enem-pos-pandemia-precisa-ir-alem-da-mudanca-de-data/. Acesso em: 21 fev. 2023.

- OLIVEIRA, Elida. Pandemia, trabalho e 'desorganização' do Enem estão por trás das histórias de quem desistiu da prova este ano. EL PAÍS, São Paulo, 16 nov. 2021. Brasil. Educação, Disponível em: https://brasil.elpais.com/brasil/2021-11-16/pandemia-trabalho-e-desorganizacao-do-enem-estao-por-tras-das-historias-de-quem-desistiu-da-prova-este-ano.html. Acesso em: 21 fev. 2023.

- SALDAÑA, Paulo. Manutenção do Enem durante a pandemia coloca Brasil na contramão da tendência mundial. Folha de São Paulo, 8 mai. 2020. Educação, Disponível em: https://www1.folha.uol.com.br/educacao/2020/05/manutencao-do-enem-durante-a-pandemia-coloca-brasil-na-contramao-da-tendencia-mundial.shtml. Acesso em: 21 fev. 2023.

ANEXO B – INDICADORES

TABELA C.1: Itens de Conforto e contratação de trabalhador doméstico.

Variáveis	Quantidade				
	0	1	2	3	4 ou +
Banheiros	0	3	7	10	14
Empregados Domésticos	0	3	7	10	13
Automóveis	0	3	5	8	11
Microcomputador	0	3	6	8	11
Lava louça	0	3	6	6	6
Geladeira	0	2	3	5	5
Freezer	0	2	4	6	6
Lava roupa	0	2	4	6	6
DVD	0	1	3	4	6
Micro-ondas	0	2	4	4	4
Motocicleta	0	1	3	3	3
Secadora roupa	0	2	2	2	2

TABELA C.2: Grau de instrução do chefe da família (pessoa que contribui com a maior parte da renda do domicílio)

Grau de instrução	Quantidade
Analfabeto / Fundamental I Incompleto	0
Fundamental I completo / Fundamental II incompleto	1
Fundamental II completo / Médio incompleto	2
Médio completo / Superior incompleto	4
Superior completo	7

TABELA C.3: Serviços Públicos.

Serviços	Quantidade	
	Não	Sim
Água encanada	0	4
Rua pavimentada	0	2

Figura 11 – Sistema de Pontos sugerida pela metodologia do Critério de Classificação Econômica Brasil (CCEB).

TABELA C.4: Itens de Conforto e contratação de trabalhador doméstico.

Variáveis	Quantidade				
	Não	Sim, 1	Sim, 2	Sim, 3	Sim, 4 ou +
Banheiros	0	3	7	10	14
Empregados Domésticos	0		5	10	13
Automóveis	0	3	5	8	11
Lava louça	0	3	6	6	6
Geladeira	0	2	3	5	5
Freezer	0	2	4	6	6
Lava roupa	0	2	4	6	6
DVD	0	4			
Micro-ondas	0	2	4	4	4
Motocicleta	0	1	3	3	3
Secadora roupa	0	2	2	2	2

Figura 12 – Critério adotado para cálculo do Indicador de Nível Socioeconômico (INSE) baseado no sistema de pontos CCEB. Em azul, os itens cuja classificação diferia em ambos questionários, sendo adotado uma pontuação arbitrária.

TABELA C.5: Itens de Conforto adicionais.

Variáveis	Quantidade				
	Não	Sim, 1	Sim, 2	Sim, 3	Sim, 4 ou +
Quartos	0	3	7	10	14
TV	0	2	3	5	5
Aspirador de pó	0	5			

TABELA C.6: Quantidade de pessoas que moram atualmente na residência.

Variável	Quantidade				
	10 ou +	8 a 9	6 a 8	5	1 a 4
Quantidade de pessoas	0	1	2	6	8

Figura 13 – Critério adicionado ao INSE para cálculo do Indicador de Nível Socioeconômico Ampliado (INSE_A). Em vermelho a pontuação adotada de forma arbitrária nos itens presentes no questionário sócioeconomico respondido pelos participantes do ENEM, mas ausentes no CCEB (item "Quartos", idêntica ao item "Banheiros"e item "TV", idêntico ao item "Geladeira").

TABELA C.7: Grau de instrução do responsável.

Grau de instrução	Quantidade
Nunca estudou / Não completou a 4ª série/5º ano do Ensino Fundamental / Não sei.	0
Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.	1
Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.	2
Completou o Ensino Médio, mas não completou a Faculdade.	4
Completou a Faculdade, mas não completou a Pós-graduação.	7
Completou a Pós-graduação.	10

TABELA C.8: Ocupação mais próxima do responsável.

Ocupação	Quantidade
Grupo 1: Lavrador, agricultor sem empregados, bóia fria, criador de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultor, pescador, lenhador, seringueiro, extrativista / Não sei.	0
Grupo 2: Diarista, empregado doméstico, cuidador de idosos, babá, cozinheiro (em casas particulares), motorista particular, jardineiro, faxineiro de empresas e prédios, vigilante, porteiro, carteiro, office-boy, vendedor, caixa, atendente de loja, auxiliar administrativo, recepcionista, servente de pedreiro, repositor de mercadoria.	1
Grupo 3: Padeiro, cozinheiro industrial ou em restaurantes, sapateiro, costureiro, joalheiro, torneiro mecânico, operador de máquinas, soldador, operário de fábrica, trabalhador da mineração, pedreiro, pintor, eletricista, encanador, motorista, caminhoneiro, taxista.	2
Grupo 4: Professor (de ensino fundamental ou médio, idioma, música, artes etc.), técnico (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretor de imóveis, supervisor, gerente, mestre de obras, pastor, microempresário (proprietário de empresa com menos de 10 empregados), pequeno comerciante, pequeno proprietário de terras, trabalhador autônomo ou por conta própria.	4
Grupo 5: Médico, engenheiro, dentista, psicólogo, economista, advogado, juiz, promotor, defensor, delegado, tenente, capitão, coronel, professor universitário, diretor em empresas públicas ou privadas, político, proprietário de empresas com mais de 10 empregados.	7

Figura 14 – Critério adotado para cálculo do Indicador de Capital Cultural (ICC). Em vermelho a pontuação adotada de forma arbitrária nos itens presentes no questionário sócioeconomico respondido pelos participantes do ENEM, mas ausentes no CCEB, tanto relacionados com o pai/responsável como da mãe/responsável (item "Ocupação", idêntica ao item "Grau de Instrução").

TABELA C.9: Itens de comunicação..

Variáveis	Quantidade				
	Não	Sim, 1	Sim, 2	Sim, 3	Sim, 4 ou +
TV por assinatura	0	5			
Telefone Celular	0	3	6	8	11
Telefone Fixo	0	4			
Microcomputador	0	3	6	8	11
Acesso à Internet	0	4			

Figura 15 – Critério adotado para cálculo do Indicador de Tecnologia e Conectividade (ITC). Em vermelho a pontuação adotada de forma arbitrária nos itens presentes no questionário sócioeconomico respondido pelos participantes do ENEM, mas ausentes no CCEB, (item "Telefone Celular", idêntica ao item "Microcomputador").

ANEXO C – ANÁLISES DESCRITIVAS

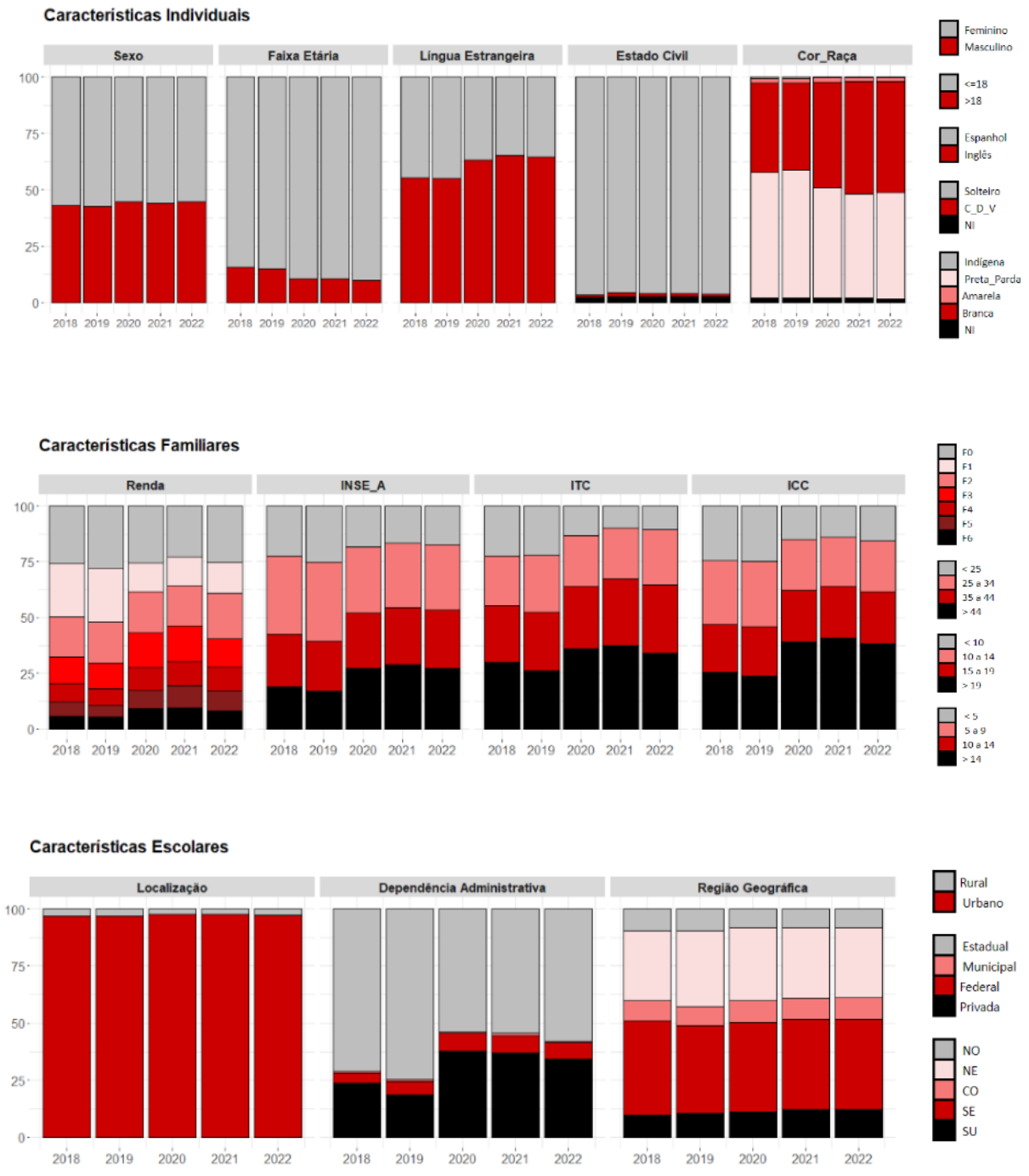


Figura 16 – Distribuição Percentual das características individuais, familiares e escolares por ano de realização do ENEM

.

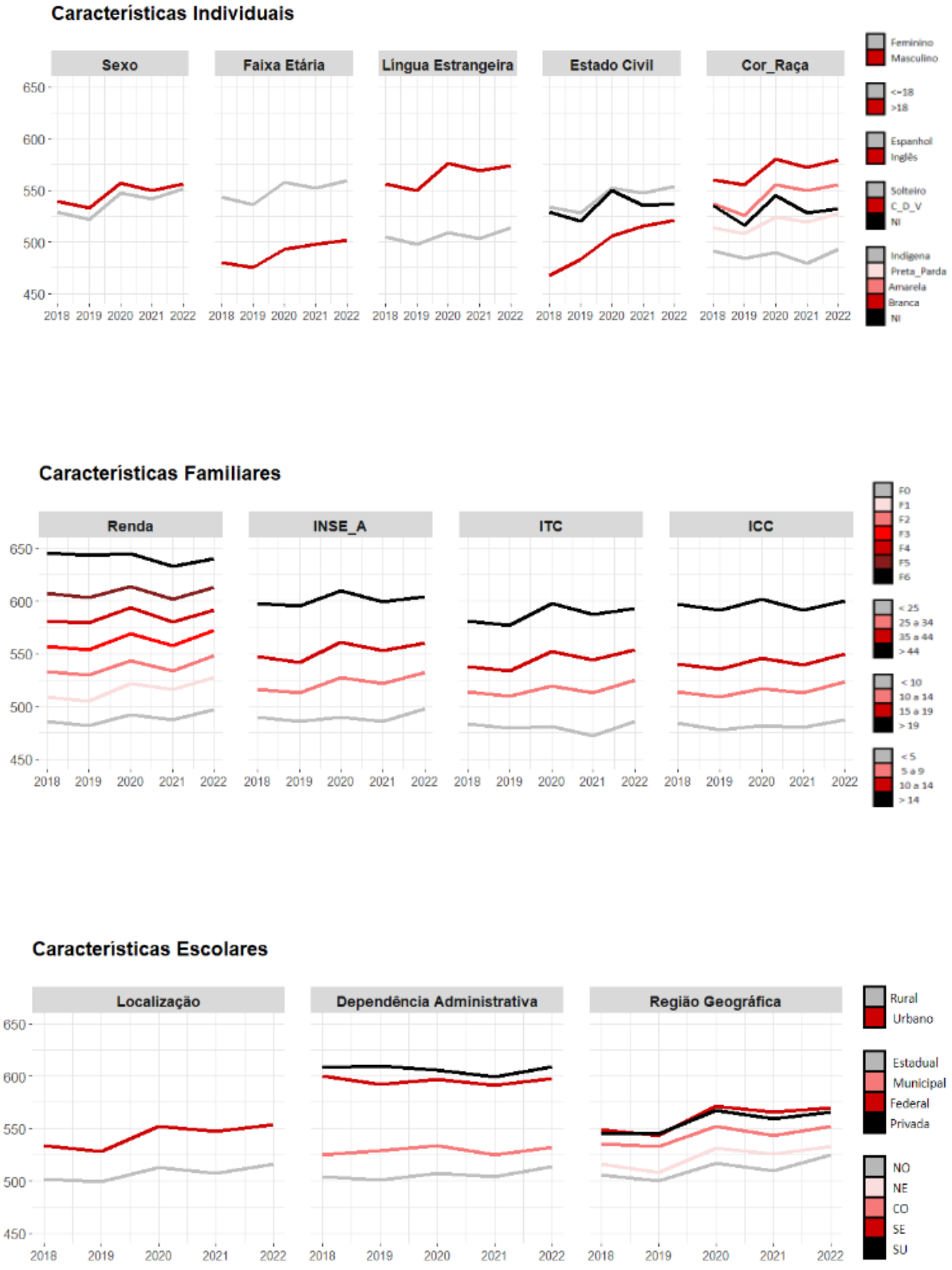


Figura 17 – Médias da Nota Geral por categoria de acordo com as características individuais, familiares e escolares por ano de realização do ENEM.

.

ANEXO D – SINTAXE DE MODELOS HIERÁRQUICOS NO AMBIENTE R

D.1 Exemplo:

lme4::

lmer(variavel resposta $\sim$ efeitos fixos + (. | efeitos aleatórios), data = Dataframe)

Onde: (.) refere-se a escolha da estrutura (intercepto, coeficiente angular)

D.2 Outras variações:

- (1 | g): Intercepto aleatório;
- (1 | g1/g2): Intercepto variando entre g1 e entre g1 dentro de g2;
- (1 | g1) + (1 | g2): Intercepto variando entre g1 e g2
- x + (x | g): Intercepto e coeficiente de inclinação correlacionados

ANEXO E – ANÁLISES ADICIONAIS

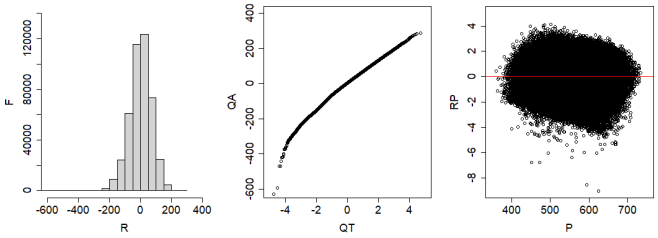
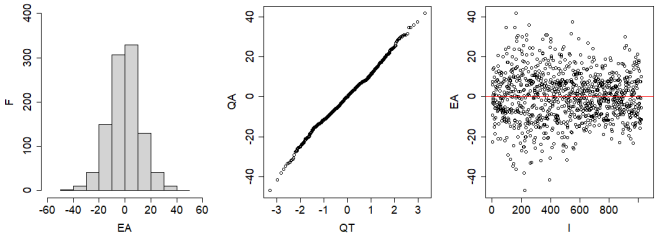
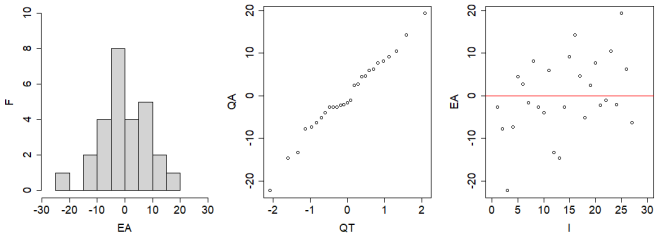
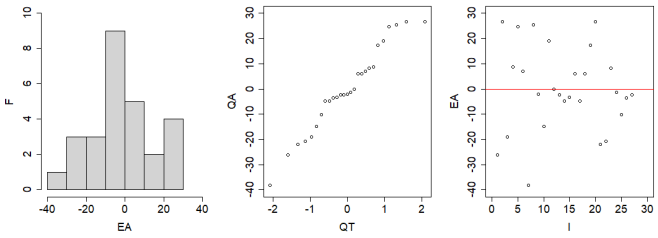
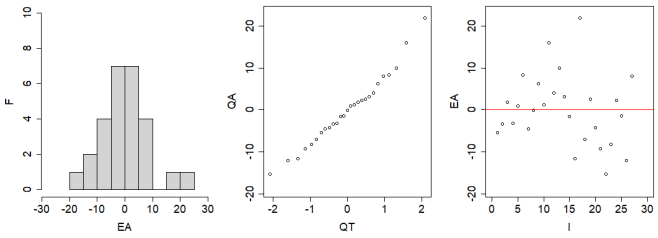
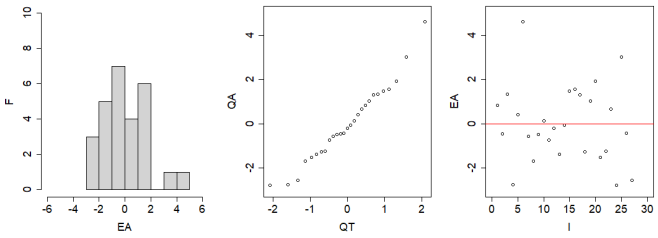
Resíduos	
Municípios	
Unidade Federativa (UF)	
UF - Dep. Administrativa Federal	
UF - Dep. Administrativa Privada	
UF - ITC	

Tabela 7 – Análise de Resíduos e de Efeitos Aleatórios referente ao ajuste do modelo hierárquico para o cenário 8 considerando os dados do ENEM 2022. Notação: 'EA': efeito aleatório, 'F': frequência, 'I': índice, 'P': predito, 'QA': quantil amostral, 'QT': quantil teórico, 'R': resíduo, 'RP': resíduo padronizado,