Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise do impacto da pandemia de COVID-19 no desempenho dos candidatos do Enem de 2020 em comparação ao Enem de 2019 utilizando técnicas de mineração de dados.

Christian Braga de Almeida Pires

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Jan Mendonça Correa

Brasília
2023

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise do impacto da pandemia de COVID-19 no desempenho dos candidatos do Enem de 2020 em comparação ao Enem de 2019 utilizando técnicas de mineração de dados.

Christian Braga de Almeida Pires

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Jan Mendonça Correa (Orientador)
CIC/UnB

Prof.a Dr.a Aletéia Patrícia Favacho de Araújo
CIC/UnB

Prof. Dr. Marcelo Grandi Mandelli
CIC/UnB

Prof. Dr. Marcelo Grandi Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 20 de Fevereiro de 2023

Dedicatória

Dedico este trabalho a minha família, em especial aos meus pais, e a todos os meus amigos que me acompanharam e fizeram parte desta jornada.

Agradecimentos

Agradeço primeiramente à minha família, especialmente aos meus pais, que sempre estiveram presentes na minha vida, me apoiando sempre que preciso. Gostaria de agradecer também a todos os meus amigos, que acompanharam e participaram de toda a trajetória até aqui, ajudando a suportar os momentos difíceis e tornar mais agradável esta jornada. Sou grato também ao Prof. Dr. Jan Mendonça Correa pela orientação.

Resumo

Neste trabalho foram utilizados os microdados do Enem dos anos de 2019 e 2020, disponibilizados pelo Inep, para conduzir uma análise a respeito do impacto da pandemia de COVID-19 nos candidatos do exame. O Enem de 2020 teve o maior número de abstenções da sua história, e considerando as desigualdades sociais presentes no país e a maneira como o governo gerenciou a pandemia, surgiu a motivação de investigar e obter informações a respeito do perfil dos inscritos, tanto dos presentes quanto dos ausentes, no ano de 2020 e realizar uma comparação com os dados do ano de 2019, onde as provas foram realizadas em situações regulares. Para isso, foi utilizada a linguagem Python, em específico a biblioteca Pandas, para analisar os microdados. Foram feitas análises para obter distribuições dos inscritos em relação a algumas variáveis, como cor/raça, renda, escolaridade dos pais, acesso a computador, celular e internet. Também foram investigadas as médias de notas do participantes com relação a estas mesmas variáveis somadas aos estados de realização das provas. Com relação aos estados, os resultados foram comparadas com os dados do Índice de Desenvolvimento Humano Municipal (IDHM) disponibilizados pelo Atlas do Desenvolvimento Humano no Brasil. Por fim, foi utilizado o algoritmo de classificação JRip, disponível no software Weka, para tentar entender quais variáveis podem influenciar na presença em um dia de prova. Foram obtidos indícios de que a pandemia teve um impacto maior sobre as populações menos favorecidas, que vivem sob níveis de fatores socioeconômicos piores, como menores valores de renda, escolaridade dos pais, estados com menor IDHM e maior dificuldade de acesso à internet. Observou-se também que estes fatores socioeconômicos podem ter influência sobre o desempenho dos candidatos nas provas.

Palavras-chave: Microdados do Enem, Mineração de dados, pandemia de COVID-19

Abstract

In this work, microdata from Enem from the years 2019 and 2020, made available by Inep, were used to conduct an analysis regarding the impact of the COVID-19 pandemic on exam candidates. The 2020 Enem had the highest number of abstentions in its history, and considering the social inequalities present in the country and the way in which the government managed the pandemic, the motivation arose to investigate and obtain information about the profile of those enrolled, both present and absent, in 2020 and make a comparison with the data for 2019, where the tests occurred in regular situations. Python language was used, specifically the Pandas library, to analyze the microdata. Analyzes were carried out to obtain distributions of those enrolled in relation to some variables, such as color/race, income, parental education, access to computer, cell phone and internet. The participant's grade averages were also investigated in relation to these same variables added to the states in which the tests were done. Regarding the states, the results were compared with the data from Índice de Desenvolvimento Humano Municipal (IDHM) provided by the Atlas of Human Development in Brazil. Finally, the JRip classification algorithm, available in the Weka software, was used to try to understand which variables can influence the presence on a test day. Evidence was obtained that the pandemic had a greater impact on disadvantaged populations, who live under worse socioeconomic levels, such as lower income levels, parental education, states with lower IDHM and greater difficulty in accessing healthcare. Internet. It was also observed that these socioeconomic factors may have an influence on the performance of candidates in the tests.

Keywords: Enem microdata, Data mining, COVID-19

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ANPD Autoridade Nacional de Proteção de Dados.

Cetic.br Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação.

CGU Controladoria-Geral da União.

CNE Conselho Nacional de Educação.

COVID-19 Coronavirus Disease 2019.

CRISP-DM Cross-Industry Standard Process for Data Mining.

DAEB Diretoria de Avaliação da Educação Básica.

Enem Exame Nacional do Ensino Médio.

ESPII Emergência de Saúde Pública de Importância Internacional.

Fiocruz Fundação Oswaldo Cruz.

IDHM Índice de Desenvolvimento Humano Municipal.

IES Instituição de Ensino Superior.

Inep Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

Ipea Instituto de Pesquisa Econômica Aplicada.

KDD Knowledge Discovery in Databases.

MEC Ministério da Educação.

OMS Organização Mundial da Saúde.

OPAS Organização Pan-Americana da Saúde.

PPL Pessoas Privadas de Liberdade e Jovens sob Medida Socioeducativa que inclua privação de liberdade.

Projur Procuradoria Federal especializada junto ao Inep.

ProUni Programa Universidade para Todos.

Saeb Sistema Nacional de Avaliação da Educação Básica.

Sisu Sistema de Seleção Unificada.

TED Termo de Execução Descentralizada.

TIC Tecnologias da informação e comunicação.

TRI Teoria de Resposta ao Item.

UF Unidade da Federação.

UFMG Universidade Federal de Minas Gerais.

Weka Waikato Environment for Knowledge Analysis.

Capítulo 1

Introdução

A pandemia de COVID-19 estabeleceu um cenário atípico no mundo. Suas consequências para o futuro, em diversas áreas, ainda são desconhecidas. Segundo o Boletim Observatório COVID-19 [4], produzido pela Fiocruz, a pandemia vem produzindo repercussões e impactos sociais, econômicos, políticos, culturais e históricos como há muito tempo não era visto. As injustiças estruturais ficaram mais explícitas, relembrando a existência das grandes desigualdades sociais do país [4]. Dessa maneira, ainda segundo o Boletim, tendo como referência a Constituição de 1988 e a complexidade do enfrentamento à COVID-19 no Brasil, é de grande importância ampliar e fortalecer políticas públicas e ações do Estado com o objetivo de proporcionar justiça social no enfrentamento desta crise sanitária, econômica e social.

Dentre os objetivos fundamentais da República Federativa do Brasil, segundo a Constituição de 1988 [5], estão: "reduzir as desigualdades sociais e regionais"; e "promover o bem de todos, sem preconceitos de origem, raça, sexo, cor, idade e quaisquer outras formas de discriminação". Além disso, a educação é categorizada como um direito social, sendo direito de todos e dever do Estado e da família. No Artigo 214, é dito que "a lei estabelecerá o plano nacional de educação, de duração plurianual, visando à articulação e ao desenvolvimento do ensino em seus diversos níveis e à integração das ações do Poder Público que conduzam à erradicação do analfabetismo; universalização do atendimento escolar; melhoria da qualidade do ensino; formação para o trabalho; e promoção humanística, científica e tecnológica do País". Atualmente, o Plano Nacional de Educação [6], aprovado pela lei nº 13.005/2014 com vigência de dez anos, é composto por vinte metas em cumprimento do disposto na Constituição de 1988.

Entretanto, a pandemia representa uma ameaça para o avanço da educação segundo um estudo do Grupo Banco Mundial [7]. Se não houverem grandes esforços para enfrentar os impactos desse período, o fechamento das escolas e universidades pode interromper o processo de aprendizagem e aumentar a evasão escolar e a desigualdade social. Por estes

motivos, são necessários estudos que visem investigar estes impactos da pandemia de COVID-19 na educação. De maneira que seja possível identificar populações e grupos mais afetados por esse período.

O Exame Nacional do Ensino Médio (Enem) é um exame de grande importância para a avaliação da educação no Brasil, além do seu resultado ser muito importante para o futuro dos estudantes. Anualmente, são produzidos conjuntos de dados sobre as realizações das provas que podem revelar muitas informações úteis sobre a situação da educação no país [8]. Neste sentido, a mineração de dados pode ter grandes contribuições na investigação de padrões e informações ocultas em grandes conjuntos de dados.

1.1 Definição do Problema

A edição de 2020 do Exame Nacional do Ensino Médio (Enem) foi marcada por diversos problemas em sua aplicação. O surgimento da pandemia de COVID-19 causou discussões sobre possível cancelamento ou remarcação das provas [9], além de provocar medo e outras preocupações nos candidatos que pretendiam realizar o exame [10] [11] [12]. Os impactos da pandemia e consequentes decisões do governo [13] [14] [15] fizeram com que esta edição tenha tido o maior número de abstenções da história do exame [16] [17] [18].

Como visto, as desigualdades sociais do país ficaram ainda mais evidentes durante essa crise. Então, tem-se a hipótese de que parcelas da população tenham sidos mais prejudicadas na realização do exame em 2020. Isso pode atrasar o ingresso destas pessoas na educação superior, interrompendo sua aprendizagem e aumentando ainda mais a diferença de oportunidades das populações mais favorecidas em relação às menos favorecidas. Assim, é útil realizar comparações com os anos anteriores para analisar quais serão as consequências no futuro dos candidatos. Diante do exposto, surgiu a motivação para realizar as investigações do presente trabalho.

1.2 Justificativa

Além de ser o principal meio de acesso à educação superior hoje, o Enem também é uma das ferramentas de avaliação sobre o fim do ciclo de educação básica no Brasil. Investigar os dados produzidos a cada ano em sua série histórica é de grande importância para avaliar a qualidade do ensino no país, quais os impactos de ações realizadas neste campo e entender quais necessidades ainda existem para propor e adequar políticas públicas e outras ações relacionadas.

Estes conjuntos de dados proporcionam a oportunidade de aplicar técnicas de mineração de dados para tentar encontrar informações sobre populações mais afetadas pela

pandemia. Descobrir características sobre o perfil destas pessoas pode permitir o direcionamento de novos estudos e ações que visem responder ao impacto da pandemia no futuro.

1.3 Objetivo

O objetivo geral deste trabalho é realizar uma análise comparativa nos microdados do Enem, dos anos de 2019 e 2020, com o intuito de tentar encontrar indicativos do impacto da pandemia de COVID-19 na realização da edição de 2020 do exame.

Para atingi-lo, foram definidos os seguintes objetivos específicos:

- Identificar e comparar informações básicas da realização das provas, como número total de inscritos, ausentes e distribuição geral das notas;
- Recuperar informações sobre a distribuição de características dos indivíduos que estiveram presentes nas duas provas, comparando as duas edições do exame;
- Informações a respeito da distribuição de ausentes em 2020 por características dos indivíduos, da escola e fatores do questionário socioeconômico;
- Distribuição dos inscritos em 2020 que responderam ao questionário socioeconômico com acesso a computador, celular e Internet;
- Obter informações sobre a distribuição das notas nas provas em 2019 e 2020 por características dos indivíduos, das escolas e fatores do questionário socioeconômico;
- Comparar as notas em 2020 por Unidade da Federação (UF) com os últimos dados do Índice de Desenvolvimento Humano Municipal (IDHM);
- Utilizar o algoritmo JRip, do software Weka, para tentar identificar variáveis que influenciam na ausência em uma prova.

1.4 Estrutura deste Trabalho

Este trabalho está dividido em seis partes. Além deste capítulo introdutório, a estrutura do trabalho é formada também pelos seguintes capítulos:

- Capítulo 2: Contextualização do Problema;
- Capítulo 3: Referencial Teórico;
- Capítulo 4: Entendimento e Preparação dos Dados;

- Capítulo 5: Análise dos Dados e Resultados;
- Capítulo 6: Conclusão.

No Capítulo 2 é apresentado o contexto em que se insere o problema abordado no trabalho. No Capítulo 3 é realizada uma fundamentação teórica, abordando conceitos importantes para a compreensão do trabalho. Também são apresentadas as ferramentas utilizadas nas análises. O Capítulo 4 apresenta a metodologia utilizada, os arquivos dos microdados e as preparações necessárias para as análises. O Capítulo 5 apresenta as análises feitas e seus resultados. No Capítulo 6 são apresentadas conclusões a respeito deste estudo e sugestões de trabalhos futuros.

Capítulo 2

Contextualização do Problema

Neste capítulo, serão abordados alguns pontos importantes para se compreender melhor o contexto do problema tratado neste trabalho. Na primeira seção, é possível conhecer um pouco mais sobre o Inep e suas responsabilidades. Em seguida, são expostas informações sobre o Enem, a estrutura das provas e detalhes da realização do exame dos anos de 2019 e 2020. Logo depois, serão apresentados momentos importantes na linha do tempo da pandemia de COVID-19, cuja relevância é grande para este estudo. Também serão apresentados resultados de relatórios e pesquisas que tratam das dificuldades enfrentadas na realização do ensino remoto, devido às desigualdades sociais existentes no Brasil. Por fim, será tratado das contribuições da utilização de mineração de dados na área da educação.

2.1 O Inep e o Enem

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) foi criado pela Lei n.º 378, de 13 de Janeiro de 1937 [19], com o objetivo de realizar estudos para identificar problemas do ensino nacional e propor políticas públicas. Atualmente, é uma autarquia federal vinculada ao Ministério da Educação (MEC), sendo responsável pelas evidências educacionais, atuando em três esferas: avaliações e exames educacionais, pesquisas estatísticas e indicadores educacionais, e gestão do conhecimento e estudos [20] [21]. Como exemplos da atuação do Inep é possível citar o Censo escolar e o Enem. O Enem é o foco de análise deste trabalho.

O Exame Nacional do Ensino Médio (Enem) [22] [23] foi criado em 1998, com o objetivo de avaliar o desempenho escolar dos estudantes ao final da etapa de educação básica. Sua primeira edição foi realizada no dia 20 de agosto de 1998, registrou 157.221 inscrições e 115.575 participantes presentes. A nota do Enem também é utilizada para o acesso à educação superior, com mais universidades aceitando os resultados a cada ano. A

criação do Programa Universidade para Todos (ProUni), em 2004, aumentou ainda mais sua popularidade nos anos seguintes. Em 2008, o Inep e o MEC anunciaram que o Enem se tornaria o processo nacional de seleção para ingresso na educação superior e certificação do ensino médio. Em 2009, foi criado o Sistema de Seleção Unificada (Sisu). A prova foi reformulada, e o exame passou a ser realizado em dois dias. Em 2014, o exame também passou a ser aceito em algumas universidades de Portugal.

Atualmente, o Enem é realizado em dois dias. A prova possui 180 questões objetivas e abrange quatro áreas do conhecimento: linguagens, códigos e suas tecnologias; ciências humanas e suas tecnologias; ciências da natureza e suas tecnologias; e matemática e suas tecnologias. Além disso, uma redação, do tipo textual dissertativo-argumentativo, também faz parte do exame [23]. Durante os dias de realização do Enem, são aplicados diferentes tipos de provas, contendo as mesmas questões, mas organizadas em ordens distintas.

Um dos requisitos para a realização da inscrição é responder a um questionário socioeconômico. Este questionário, junto das provas, gabaritos, informações sobre os itens e as notas, formam os microdados do Enem. Eles são divulgados alguns meses após sua realização, com seu conteúdo anonimizado, não sendo possível identificar os participantes por meio dos dados [24].

De acordo com o documento "Leia-me" dos Microdados do Enem de 2019 [25], a partir de 2017, as provas passaram a ser realizadas em dois domingos consecutivos. No ano de 2019, as provas foram realizadas nos dias 3 e 10 de novembro de 2019, sendo que no primeiro dia os participantes realizaram as provas de Linguagens, Códigos e suas tecnologias e Redação e de Ciências Humanas e suas tecnologias e , no segundo, as provas de Ciências da Natureza e suas tecnologias e Matemática e suas tecnologias. A segunda aplicação do Enem 2019, aconteceu nos dias 10 e 11 de dezembro de 2019, para Pessoas Privadas de Liberdade e Jovens sob Medida Socioeducativa que inclua privação de liberdade (PPL), e para os participantes com direito à reaplicação.

Segundo o documento "Leia-me" dos Microdados do Enem de 2020 [26], devido à pandemia de COVID-19, as provas de 2020 foram realizadas nos dias 17 e 24 de janeiro de 2021, sendo que no primeiro dia os participantes realizaram as provas de Linguagens, Códigos e suas tecnologias, Ciências Humanas e suas tecnologias e Redação. No segundo, as provas de Ciências da Natureza e suas tecnologias e Matemática e suas tecnologias. A segunda aplicação do Enem 2020, por sua vez, ocorreu nos dias 24 e 25 de fevereiro de 2021, para Pessoas Privadas de Liberdade e Jovens sob Medida Socioeducativa que inclua privação de liberdade (PPL), além dos participantes com direito à reaplicação.

Também na edição de 2020, o Enem foi aplicado pela primeira vez em formato digital, nos dias 31 de janeiro e 07 de fevereiro de 2021, com 100 mil vagas para participação [26].

Espera-se que este novo formato ganhe espaço e seja ampliado de forma progressiva.

2.2 Pandemia de COVID-19 e seu Impacto na Educação

Segundo a Organização Pan-Americana da Saúde (OPAS) [27], em 7 de janeiro de 2020, foi confirmado pelas autoridades chinesas a identificação de um novo tipo de coronavírus, após vários casos de pneumonia na cidade de Wuhan na semana anterior. Nas últimas décadas, esse tipo de vírus raramente causava doenças mais graves em humanos além de resfriados comuns. Porém desta vez seria diferente. O novo coronavírus foi nomeado SARS-CoV-2 e é responsável pela doença COVID-19. Em 30 de janeiro de 2020, a Organização Mundial da Saúde (OMS) declarou uma Emergência de Saúde Pública de Importância Internacional (ESPII), seu nível mais alto de alerta segundo o Regulamento Sanitário Internacional. Este alerta indica que o evento pode ser um risco de saúde pública para outros países devido a disseminação internacional de doenças, precisando de uma resposta internacional coordenada e imediata. Com o crescente número de casos de infecção pela doença ao redor do mundo, a COVID-19 foi caracterizada pela OMS como uma pandemia no dia 11 de março de 2020.

No Brasil, o primeiro caso confirmado foi em 26 de fevereiro de 2020, no estado de São Paulo. Em 12 de março de 2020, foi confirmado o primeiro óbito pela COVID-19 [28]. Na data de 20 de março de 2020, o Ministério da Saúde declarou estado de transmissão comunitária em todo o território nacional, segundo a portaria nº 454 [29]. No dia 22 de março de 2020, todas as unidades federativas do país já haviam notificado casos da doença [28].

Diante do agravamento do cenário mundial de infecção pela doença, fez-se necessária a adoção de medidas de enfrentamento, conforme previstas pela Lei 13.979 [30], de 6 de fevereiro de 2020. Dentre estas medidas estavam a possibilidade de realizar o isolamento e a quarentena, que previam a separação de pessoas doentes ou contaminadas, ou com suspeita de contaminação, de outras que não estariam doentes ou contaminadas, além da restrição de atividades, incluindo fechamento de locais e interrupção de serviços. Essas medidas foram regulamentadas posteriormente por outros decretos, como o nº 10.282 [31], de 20 de março de 2020, que definiu quais serviços seriam considerados essenciais e excluídos da possibilidade de interrupção. As escolas e universidades não estavam incluídas nesta lista, e suas atividades continuariam a distância assim que ajustadas para a nova realidade. A portaria nº 343 [32] do Ministério da Educação, de 17 de março de 2020, resolve autorizar, em caráter excepcional, a substituição de disciplinas presenciais por aulas em meios digitais, ou a suspensão das atividades acadêmicas, com eventual reposição,

enquanto durar a pandemia de COVID-19. Para a educação básica, técnica e superior, o Conselho Nacional de Educação (CNE), divulgou o parecer CNE/CP nº 5/2020 [33], homologado parcialmente, contendo sugestões para a reorganização do calendário escolar e realização de atividades não presenciais, apontando também desafios a serem superados. Em um de seus trechos, o documento expressa a importância de se considerar as fragilidades e desigualdades estruturais da sociedade brasileira, que se agravam diante da pandemia, principalmente em relação à educação. Existem grandes diferenças de proficiência, alfabetização, taxa líquida de matrícula e acesso ao mundo digital relacionadas a fatores socioeconômicos e étnico-raciais. Além disso, também devem ser consideradas as consequências socioeconômicas que podem surgir do impacto da doença na economia, como aumento da taxa de desemprego e redução da renda familiar.

2.2.1 Discrepâncias no Acesso ao Ensino Remoto

Segundo a pesquisa TIC Educação 2019 [34], edição de uma pesquisa realizada anualmente pelo Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br) a respeito do uso de Tecnologias da informação e comunicação (TIC) nas escolas brasileiras, a utilização de tecnologias digitais se tornou uma das principais estratégias para a continuidade de diversas atividades impactadas pela COVID-19, como a interação social, o desenvolvimento de atividades profissionais, as operações de comércio e as atividades educacionais. Entretanto, as diferenças sociais no acesso de recursos digitais ficaram mais evidentes, da mesma forma que outras questões como acesso à alimentação, moradia, ao saneamento, a medidas de prevenção de contágio e tratamentos de saúde. Dessa maneira, um dos grandes problemas para a realização das atividades a distância, era o acesso dos estudantes às Tecnologias da informação e comunicação. Segundo outra pesquisa, a TIC Domicílios 2019 [35], 61% dos domicílios brasileiros não contavam com computador e 28% não possuíam acesso à Internet, sendo 86% e 50%, respectivamente, para as classes DE.

De acordo com a TIC Educação 2019 [34], 99% das escolas públicas e particulares localizadas em áreas urbanas possuíam ao menos um computador com acesso à Internet e, em 92% delas, havia também a presença de rede WiFi. Porém, em apenas 34% das escolas públicas, o acesso à rede WiFi estava disponível para os alunos, percentual que era de 49% entre as escolas particulares. Para as escolas em áreas rurais, 40% possuíam ao menos um computador (de mesa, notebook ou tablet) com acesso à Internet, e em 9% das instituições não havia computadores mas a escola acessava a Internet por outros dispositivos, como o celular. A proporção de escolas rurais sem infraestrutura de conexão é de 51%.

Sem uma perspectiva para o fim da pandemia, o governo passou a considerar o ensino totalmente remoto ou híbrido, e as escolas precisariam se adaptar. Segundo a TIC Educação 2019 [34], apenas 14% das escolas públicas e 10% das municipais contavam com uma plataforma ou ambiente virtual de aprendizagem que permitisse a disponibilização de atividades para os alunos de forma remota. Nas escolas particulares, o percentual era de 64%. Plataformas como o Facebook, Instagram e WhatsApp se tornaram ferramentas muito utilizadas para transmissão de aulas, compartilhamento de conteúdos e materiais e comunicação em geral.

Ainda segundo a mesma pesquisa [34], identificou-se uma desigualdade de acesso à Internet entre as regiões do país. Dentre os alunos da Educação Básica, aqueles que haviam utilizado a rede nos três meses anteriores à realização da pesquisa, era de 83%. Por regiões, Sudeste (88%), Sul (87%) e Centro-Oeste (86%) apresentaram os maiores percentuais, enquanto Nordeste e Norte registraram 78% e 73%, respectivamente. Este acesso foi, em geral, pelo telefone celular, dispositivo mais utilizado para este motivo desde 2015. Além disso, foi o único dispositivo utilizado para acessar a rede por 18% dos alunos, dentre eles 21% de escolas da rede pública e 3% de escolas rede privada. 39% dos alunos de escolas públicas não possuíam computador em casa. 62% acessavam a rede em lugares com acesso livre ou gratuito, e 37% em centros públicos de acesso. Estes dados apontam que com o fechamento destes locais durante a pandemia, muitos devem ter ficado sem condições de acessar a rede.

Nas áreas rurais, os percentuais de acesso à rede foram bem menores, e o impacto pode ter sido ainda maior. Segundo dados da TIC Domicílios 2019 [35], 82% dos domicílios localizados em áreas rurais não possuíam computadores e 48% não contavam com acesso à Internet. E de acordo com a TIC Educação 2019 [34], apenas 49% das escolas possuíam computador de mesa, 30% computador portátil e 4% tablet. Os dados de acesso à Internet mostraram que as instituições das regiões Norte (21%), Nordeste (38%) e Sudeste (51%) apresentaram proporções de acesso à rede menores quando comparados com o observado nas regiões Centro-Oeste (74%) e Sul (83%). Dentre os maiores obstáculos citados para a ampliação do acesso à rede estavam a oferta de conexão e o custo. Muitos responsáveis por escolas nessas áreas afirmaram também que possuíam outras prioridades, como melhorar a infraestrutura básica da escola, garantir a manutenção dos equipamentos, a ampliação do espaço físico e o investimento em segurança geral. Dessa forma, assim como nas áreas urbanas, o telefone celular e as redes sociais foram os mais utilizados para acesso a Internet e interação entre escolas e famílias.

Outro ponto abordado pela TIC Educação 2019 [34], foi que a evasão escolar se tornou uma das grandes preocupações causadas pela interrupção das aulas. Além dos problemas citados até aqui, ainda existem casos onde os alunos precisam abandonar os estudos para

trabalhar. As dificuldades e falta de motivação para seguir com os estudos podem levar à evasão e consequentemente causar um impacto no acesso ao Ensino Superior, limitando ainda mais as oportunidades desta parcela da população.

Em setembro de 2021, foi divulgada a nota técnica nº 88 [36] do Instituto de Pesquisa Econômica Aplicada (Ipea), a respeito do acesso domiciliar à Internet e ensino remoto durante a pandemia. Segundo o documento, os dados apontam que grande parte dos estudantes brasileiros de instituições públicas de ensino não possui condições necessárias para acompanhar as atividades de ensino remoto durante a pandemia. Muitos deles não tem acesso a dispositivos para a transmissão de dados ou mecanismos de transmissão, como internet ou sinal de televisão digital. Além disso, a nota técnica encerra afirmando que a dificuldade em estudar durante o período de isolamento na pandemia pode ser uma fonte de ampliação da desigualdade no futuro, deixando estes em desvantagem em relação aos que puderam ter acesso ao ensino remoto. Dessa forma, as desigualdades seriam ampliadas, uma vez que os estudantes mais afetados são os que já se encontram em desvantagens de oportunidades devido às suas condições socioeconômicas.

Todos os aspectos citados contribuem para que a edição do Enem de 2020 tenha tido, em hipótese, maiores número de abstenções e um pior desempenho dos participantes em relação aos anos anteriores, principalmente entre os candidatos cercados por piores condições socioeconômicas, dado o impacto da pandemia e as dificuldades enfrentadas por todos, mas que afetou com maior ênfase essas parcelas da população.

2.3 Análise de Dados Educacionais

Segundo Albernaz, Ferreira e Franco [8], a disponibilização de microdados como os do Sistema Nacional de Avaliação da Educação Básica (Saeb), do Enem, e de outros sistemas de avaliação da educação realizados pelo Inep, tornaram possíveis a investigação dos determinantes de uma medida de desempenho escolar com base em rendimentos de alunos em testes padronizados de conhecimento. De acordo com Soares [37], também permitiu perceber que, observando-se análises já publicadas, a escola básica brasileira tem determinantes de qualidade parecidos com outros países, de forma que a literatura internacional já existente sobre a área pode ter grande contribuição.

Segundo Baker, Isotani e Carvalho [38], com o aumento no número da utilização de tecnologias digitais e plataformas educacionais na educação, surgiu também o interesse de pesquisadores em realizar investigações na área utilizando mineração de dados. Esta área tem sido conhecida como Mineração de Dados Educacionais, e tem como objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Dessa maneira, espera-se compreender melhor os alunos, como e em que

contexto eles aprendem, abordagens mais adequadas e outros fatores que influenciam a aprendizagem. Dentre as contribuições dessa área recente, pode-se citar também a redução no tempo gasto pelos alunos para desenvolver habilidades acadêmicas e uma expansão do conhecimento científico relacionado aos estados emocionais dos alunos. De forma que seja possível investigar a relação de fatores que influenciam seus comportamentos no processo de aprendizagem.

Por fim, é importante que sejam feitas análises sobre dados educacionais com o objetivo de produzir informações e conhecimento que possam auxiliar os responsáveis por tomar decisões a alcançar seus objetivos, de forma que se tenha um acesso mais igualitário à educação, com mais eficácia e eficiência, trazendo consequentemente mais oportunidades para todas as pessoas.

Capítulo 3

Referencial Teórico

Neste capítulo será feita uma fundamentação teórica para compreender melhor os resultados deste trabalho. São discutidas definições de conceitos importantes na visão de diferentes autores. Também são apresentadas as principais ferramentas utilizadas nas análises do capítulo seguinte.

3.1 Dados, Informação e Conhecimento

Para começar, é importante visitar a literatura para tentar esclarecer as definições de alguns termos que serão utilizados ao longo deste trabalho. Dados, informação e conhecimento são termos presentes em muitas áreas, e possuem definições que divergem em alguns aspectos de acordo com diferentes autores.

Para Ralph Stair e George Reynolds [39], dados são fatos brutos, que podem ser alfanuméricos, áudio, imagem ou vídeo. Esta definição concorda com a de Laudon e Laudon [40], que diz que dados são sequências de fatos brutos representando eventos que ocorrem nas organizações ou no ambiente físico, antes de serem organizados e arranjados de forma que as pessoas possam entendê-los e usá-los. Ainda para Laudon e Laudon [40] informação é um dado moldado em formato significativo e útil para seres humanos. Já para Stair e Reynolds [39], informação seria uma coleção de fatos organizados e processados de maneira que tenham um valor adicional, além do valor dos fatos individuais. Conhecimento seria a consciência e compreensão de um conjunto de informações e maneiras como essas informações podem ser úteis para apoiar uma tarefa específica ou para chegar a uma decisão.

Segundo Ackoff [41], a diferença entre dado e informação seria funcional e não estrutural. Dados são símbolos que representam propriedades de objetos, eventos e seus ambientes, produtos de observação e que não possuem valor até serem transformados em uma forma utilizável. Já informação seria inferida dos dados e estariam contidas em des-

crições, respostas para perguntas que comecem com palavras como quem, o que, quando e quantos. O conhecimento seria como fazer, o que torna possível a transformação de informação para instruções, sendo obtido ou por transmissão de outro que já o possui, ou extraído pela experiência.

Valdemar W. Setzer [42] define dado como uma sequência de símbolos quantificados ou quantificáveis, uma entidade matemática e, portanto, puramente sintática. Informação seria caracterizada como uma abstração informal que está na mente de alguém, com uma representação significativa para essa pessoa. Por fim, conhecimento seria uma abstração interior, pessoal, de algo que foi experimentado, vivenciado, por alguém. Para ele, conhecimento não pode ser descrito, o que se descreve é a informação (se entendida pelo receptor), ou o dado, e requer uma vivência do objeto do conhecimento.

3.2 Informação como Apoio à Tomada de Decisão

De acordo com Han, Kamber e Pei [43], dados são gerados e armazenados em quantidades e velocidades cada vez maiores. O contínuo desenvolvimento tecnológico diversifica e aprimora maneiras de gerar, armazenar e transmitir dados. Organizações possuem quantidades gigantescas de dados científicos, comerciais, governamentais, educacionais, dentre outros. Analisar essas grandes quantidades de dados é uma necessidade e uma tarefa que não é possível sem o apoio de ferramentas computacionais.

Segundo Ralph Stair [39], transformar os dados em informação é um processo e definir as relações entre os dados para criar informações requer conhecimento. Ainda segundo o autor, o valor da informação está diretamente relacionado a como ela apoia os responsáveis por tomar decisões a alcançar os objetivos da organização.

Para Goldschmidt e Passos [44], dentro desse contexto, surgiu a área de Descoberta de Conhecimento em Bases de Dados, do inglês KDD (*Knowledge Discovery in Databases*), muitas vezes referida como mineração de dados. Uma evolução natural da tecnologia da informação, segundo Han, Kamber e Pei [43], com o objetivo de transformar as enormes quantidades de dados em informações, e utilizá-las como apoio na tomada de decisões.

3.3 Mineração de Dados

É interessante ver como alguns autores definem mineração de dados e fazem sua distinção de Descoberta de Conhecimento em Bases de Dados, que frequentemente será referenciado neste capítulo como KDD.

Laudon e Laudon [40] define mineração de dados como a análise de grandes quantidades de dados a fim de encontrar padrões e regras que possam ser usados para orientar

a tomada de decisão e prever o comportamento futuro. Já Aggarwal [45] define o mesmo termo como o estudo de coleta, limpeza, processamento, análise e obtenção de informações úteis a partir dos dados.

Fayyad, Piatetsky-Shapiro e Smyth [46] reforçam a necessidade de distinção entre KDD e mineração de dados. KDD seria um processo de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados. Enquanto mineração de dados seria uma etapa no processo de KDD em que são aplicados algoritmos de análise e descoberta de dados que, sob limitações de eficiência computacional aceitáveis, produzem uma enumeração particular de padrões sobre os dados. Goldschmidt e Passos [44] concordam com estas definições.

Para Fayyad, Piatetsky-Shapiro e Smyth [46], o processo de KDD é composto de várias etapas que envolvem atividades de preparação dos dados, busca por padrões, avaliação dos resultados, e a consolidação da descoberta de conhecimento. Além disso, este processo pode conter várias iterações entre as etapas até se chegar em resultados aceitáveis. De acordo com Goldschmidt e Passos [44], a etapa de pré-processamento compreende atividades relacionadas à captação, à organização e ao tratamento dos dados. E segundo Han, Kamber e Pei [43], a etapa de preparação dos dados pode conter diversas atividades como:

1. Limpeza de dados: Para remover ruído e inconsistências nos dados.
2. Integração dos dados: Caso haja necessidade de combinar diferentes fontes de dados.
3. Seleção de dados: Onde dados relevantes para a análise são recuperados da base de dados.
4. Transformação de dados: Atividade em que os dados são transformados em formatos apropriados para as atividades de mineração de dados.

Já na etapa de Mineração de Dados, segundo Goldschmidt e Passos [44], é realizada a busca por padrões e conhecimentos úteis no contexto da aplicação de KDD. Na fase de avaliação, de acordo com Fayyad, Piatetsky-Shapiro e Smyth [46], é feita a visualização e interpretação dos padrões obtidos na etapa de mineração de dados, avaliando a necessidade de novas iterações. Por fim, em uma última fase, os novos conhecimentos são organizados e documentados para apresentar as partes interessadas.

3.4 Técnicas de Mineração de Dados

Segundo Goldschmidt e Passos [44], todo o processo de KDD deve ser orientado pelos objetivos estabelecidos para o projeto. Dessa forma, as aplicações de KDD podem ser

orientadas para a verificação de hipóteses ou descobertas de conhecimentos. No caso da segunda, de acordo com Fayyad, Piatetsky-Shapiro e Smyth [46], ainda pode ser subdividida em predição, onde o sistema encontra padrões com a finalidade de prever o comportamento futuro de algumas entidades; e descrição, na qual o sistema encontra padrões com a finalidade de apresentá-los a um usuário de forma compreensível para humanos. A seguir será brevemente descrito o objetivo de algumas técnicas de mineração de dados. Para cada uma delas, existem várias implementações de algoritmos.

Han, Kamber e Pei [43], definem classificação como o processo de encontrar um modelo ou função que descreva e diferencie classes de dados ou conceitos. As funções ou modelos são derivados por meio da análise de um conjunto de dados de treinamento. Segundo Goldschmidt e Passos [44], de forma que seja possível mapear um conjunto de dados em categorias predefinidas, chamadas de classes. Assim, é possível aplicar a função em novos dados, com classes desconhecidas, para prever qual a classe em que eles se encaixam.

Segundo Goldschmidt e Passos [44], a técnica de regressão é parecida com a classificação, mas o mapeamento é feito para valores numéricos ao invés de classes. A intenção é encontrar funções que possam prever valores futuros, mapeando registros de dados em valores reais. O caso mais simples é uma função linear, mas também existem técnicas de regressão não linear.

De acordo com Fayyad, Piatetsky-Shapiro e Smyth [46], a sumarização consiste em encontrar descrições compactas para subconjuntos de dados. Para Goldschmidt e Passos [44], tem como objetivo identificar e apresentar as principais características dos dados em um conjunto, de maneira breve e compreensível. Sua intenção é caracterizar de maneira resumida os dados, podendo ser utilizada para descobrir características e criar perfis de identificação.

Para Larose e Larose [47], clusterização consiste em agrupar dados em classes de objetos similares. Um *cluster* é uma coleção de registros similares entre si e diferentes de registros em outros *clusters*. O objetivo é segmentar os dados em subgrupos relativamente homogêneos, em que a similaridade com registros dentro do *cluster* é maximizada e a com registros de outros *clusters* sejam minimizadas. O que diferencia a clusterização da classificação é que a primeira não atribui rótulos para os grupos como é feito com as classes na segunda, e também não tenta classificar ou prever o valor da variável alvo. Han, Kamber e Pei [43] afirmam que em muitos casos os rótulos de classes não existem inicialmente, e a clusterização pode ser usada para gerá-los. Segundo Goldschmidt e Passos [44], esse processo geralmente necessita que seja determinado qual o número de grupos a serem considerados para a segmentação dos dados.

Para Goldschmidt e Passos [44], a detecção de desvios busca identificar mudanças em padrões que já foram anteriormente identificados, padrões de pouca incidência com valores

suficientemente diferentes dos padrões normalmente identificados. A maioria das técnicas de mineração de dados descarta esses valores atípicos como ruído ou exceções, mas em algumas aplicações como análise de fraudes, estes valores podem ser de grande interesse. Segundo Han, Kamber e Pei [43], tais valores podem ser identificados usando modelos estatísticos que assumem uma determinada distribuição ou modelo de probabilidade para os dados, ou analisando pontos remotos, longes de qualquer outro *cluster*, usando medidas de distância.

Segundo Larose e Larose [47], o objetivo da técnica de associação é encontrar regras para quantificar o relacionamento entre dois ou mais atributos, usando parâmetros de suporte e confiança. De acordo com Goldschmidt e Passos [44], a associação consiste em identificar conjuntos de itens que ocorram de forma simultânea e frequente nos dados. O parâmetro de suporte mínimo tem como objetivo identificar se a associação é frequente, enquanto o parâmetro de confiança é utilizado para validá-la.

3.5 Ferramentas de Mineração de Dados

A linguagem Python [48] foi criada no início dos anos 1990 por Guido van Rossum. É uma linguagem de programação interpretada, orientada a objetos e com suporte a vários outros paradigmas de programação. Possui uma sintaxe simples com ênfase na legibilidade e uma lista de bibliotecas e pacotes muito extensa, permitindo o desenvolvimento de aplicações em diversas áreas e classes de problemas. As bibliotecas Pandas e Matplotlib são muito utilizados para atividades de mineração de dados, conforme descrição a seguir:

1. **Pandas** [49] é um pacote Python de código aberto, rápido, poderoso, flexível e fácil de usar para análise de dados. Possui estruturas de dados eficientes e permite fazer operações complexas de maneira simples;

2. **Matplotlib** [50] é uma biblioteca para criar visualizações estáticas, animadas e interativas em Python de maneira simples. Muito útil para criar diversos tipos de gráficos de forma a visualizar os resultados de análises.

Jupyter [51] é uma aplicação *web* que fornece um ambiente interativo para desenvolvimento em diversas linguagens. É um *notebook* computacional que pode ser utilizado por meio de um navegador *web*. O projeto é de código aberto e sem fins lucrativos.

Anaconda [52] é uma plataforma que contem diversas ferramentas utilizadas em ciência de dados. A sua instalação contém o conda, que é um gerenciador de pacotes e ambientes, o Python, e diversos outros pacotes científicos. Também existe a opção de utilizar uma interface gráfica para facilmente iniciar aplicações, como o Jupyter Notebook, e gerenciar os pacotes. O repositório do Anaconda possui centenas de pacotes muito utilizados em

ciências de dados disponíveis. Essas facilidades tornam a plataforma muito interessante para as atividades de mineração de dados.

3.5.1 Weka e o Algoritmo JRip

Waikato Environment for Knowledge Analysis (Weka) [53] é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Fornece implementações de vários algoritmos para serem aplicados em conjuntos de dados de maneira simples. Também inclui diversas ferramentas para realizar atividades de pré-processamento, classificação, regressão, clusterização, regras de associação e visualização dos dados. Além disso, fornece uma interface gráfica para auxiliar nas tarefas e utilização do programa.

Um algoritmo importante do Weka para este trabalho será o JRip [54]. Ele é um algoritmo de classificação baseado em regras. Segundo Han, Kamber e Pei [43], um algoritmo deste tipo tem como produto um modelo representado na forma de um conjunto de regras do tipo SE-ENTÃO. Este tipo de regra é uma expressão no formato: SE condição ENTÃO conclusão. A parte "SE" da regra, chamada de antecedente, é formada por um ou mais testes de atributos. Já a parte "ENTÃO", chamada de consequente, contém uma predição de classe.

O JRip implementa uma versão semelhante ao algoritmo RIPPER, proposto por Cohen [55]. O algoritmo RIPPER é uma versão com desempenho otimizado do algoritmo IREP, proposto por Furnkranz e Widmer em 1994 [56]. Dentre as melhorias estão o melhor desempenho e precisão em conjuntos de dados com muito ruído. Uma das vantagem de algoritmos de classificação baseados em regras é a facilidade de entendimento das regras geradas.

Do ponto de vista prático, o algoritmo JRip fornece um conjunto de regras do tipo SE-ENTÃO utilizando uma classe de predição previamente escolhida. Estas regras são acompanhadas de informações a respeito da taxa de acertos do algoritmo, para que seja possível realizar uma avaliação do modelo gerado.

3.6 Conceitos de Estatística

Nesta seção serão brevemente descritos alguns conceitos de estatística que podem ser necessários para compreender os resultados das análises deste trabalho.

3.6.1 Medidas de Posição Central e de Dispersão

Para Moretin e Bussab [57], a média aritmética é a soma das observações divididas pelo número total de observações. Pode ser representado pela Equação 3.1:

$$X = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n} \tag{3.1}$$

onde X é o valor da média aritmética, n é o número total de observações e a_i é a i-ésima observação.

Ainda segundo Moretin e Bussab [57], a mediana é o valor que ocupa a posição central da série de valores quando ordenados em ordem crescente. Caso o número de observações seja par, a mediana é a média aritmética das duas observações centrais.

A moda, segundo Moretin e Bussab [57], é definida como a observação mais frequente no conjunto de valores. Estas três medidas – média, mediana e moda - são utilizadas para resumir um conjunto de dados, utilizando valores que sejam representativos para este conjunto. Dessa forma, são chamadas de medidas de posição central.

De acordo com Moretin e Bussab [57], desvio Padrão e variância são medidas de dispersão dos dados em relação a média. Ambas fornecem uma medida que indica, em média, qual erá o valor do erro ao tentar substituir os valores pela média do conjunto.

3.6.2 População e Amostra

Segundo Moretin e Bussab [57], população é o conjunto de todos os elementos ou resultados sob investigação, enquanto amostra é qualquer subconjunto da população. Quanto mais conhecimento explícito ou implícito de uma população, mais informativas são as observações dentro de uma amostra da mesma população. É preciso também ter cuidado ao selecionar amostras para que o resultado não contenha um viés de seleção, de forma que prejudique as análises ao mostrar resultados tendenciosos que não representam a realidade com a precisão adequada.

Desta maneira, encerra-se a apresentação do referencial teórico necessário para compreender as próximas etapas deste estudo. Foram vistos neste capítulo: os conceitos de dados, informação e conhecimento; as definições de KDD e mineração de dados; objetivos de algumas técnicas de mineração de dados; apresentação das ferramentas utilizadas neste estudo; e alguns conceitos de estatística. A seguir, inicia-se o Capítulo 4, que trata do entendimento e preparação dos dados utilizados neste trabalho.

Capítulo 4

Entendimento e Preparação dos dados

Neste capítulo, primeiro será apresentada a metodologia utilizada no trabalho. Posteriormente, serão feitas considerações iniciais sobre o formato de apresentação dos microdados do Enem e seu conteúdo. Em seguida, serão expostas as atividades de preparação dos dados para as análises.

4.1 CRISP-DM

O processo de produção deste trabalho foi orientado pelo modelo CRISP-DM [58], acrônimo de *Cross-Industry Standard Process for Data Mining*, foi criado com a intenção de ser um modelo padrão de processos não proprietário e disponível gratuitamente, para guiar as atividades de um projeto de mineração de dados.

Sua metodologia é descrita como um modelo de processos hierárquicos consistindo de conjuntos de tarefas em diferentes níveis de abstração, de forma que seja possível cobrir todo o processo de mineração de dados e aplicações possíveis. Além disso, o modelo é flexível e permite ser customizado para contextos específicos de maneira individual [58].

No modelo de referência [58], é apresentado uma visão geral do ciclo de vida de um projeto de mineração de dados, que consiste em seis fases, as quais serão brevemente introduzidas a seguir.

1. **Entendimento do Negócio**: consiste em entender os objetivos e requisitos do projeto da perspectiva do negócio, convertendo posteriormente esse conhecimento para a definição de um projeto de mineração de dados com um plano preliminar para atingir os objetivos;

2. **Entendimento dos Dados**: esta fase começa com uma coleta de dados iniciais e segue com atividades para se familiarizar com os dados, identificar problemas de qualidade e detectar possíveis subconjuntos de interesse para formar hipóteses a respeito de informações ocultas;

3. **Preparação dos Dados**: a preparação dos dados abrange todas as atividades com objetivo de elaborar o conjunto final de dados, que serão utilizados para realizar as análises. Podem ser realizadas várias vezes e sem ordem prescrita. Como exemplo tem-se a seleção de atributos, transformação e limpeza de dados para ferramentas de mineração de dados;

4. **Modelagem**: nessa parte, ferramentas e técnicas de mineração de dados são selecionadas e aplicadas, seguida da calibração de seus parâmetros. Muitas vezes é necessário voltar a fase de preparação de dados para ajustar os dados de forma a satisfazer os requisitos de técnicas e ferramentas específicas;

5. **Avaliação**: nesse ponto, um ou mais modelos de aparente alta qualidade já devem ter sido construídos. É o momento para revisar todos os passos que levaram a essa construção e avaliar se o modelo atingiu os objetivos de negócio, ou se ainda existe algo que não foi suficientemente considerado;

6. **Implantação**: por fim, é preciso organizar e apresentar o resultado de uma maneira que o cliente entenda e possa utilizar.

Seguindo as etapas dessa metodologia, foi realizado o entendimento do negócio, descrito nos Capítulos 1 a 2. A seguir tem-se a etapa de entendimento dos dados, composta pelas próximas duas seções. Logo em seguida, se inicia a etapa de preparação dos dados.

4.2 Considerações Iniciais sobre a Reestruturação da Apresentação dos Dados Utilizados

Esta análise utilizará os microdados do Enem dos anos de 2019 e 2020, disponíveis no site do Inep [59] [3]. Os microdados reúnem um conjunto de informações detalhadas sobre pesquisas, avaliações e exames realizados pelo Inep. Estes dados permitem que gestores, pesquisadores, instituições e outros interessados na área da educação possam realizar análises para subsidiar diagnósticos, estudos, pesquisas e acompanhamento de estatísticas e informações educacionais. Recentemente, no ano de 2022, os formatos de apresentação dos dados estão sendo reestruturados para eliminar a possibilidade de identificação de pessoas. Estes novos formatos também estão sendo avaliados para verificar possíveis melhorias. Segundo o Portal Gov.br [60], as mudanças ocorrem baseadas em estudos técnicos

e análise jurídica da Procuradoria Federal especializada junto ao Inep (Projur), além de terem sido, posteriormente, objeto de análises pela Autoridade Nacional de Proteção de Dados (ANPD) e pela Controladoria-Geral da União (CGU).

Segundo o Inep [25], os microdados se constituem no menor nível de desagregação de dados recolhidos por pesquisas, avaliações e exames realizados. Nos microdados do Enem, eles são listados por participante, mas não constam nestes variáveis que permitam a identificação direta do participante. No entanto, em um estudo técnico formalizado no Termo de Execução Descentralizada (TED) 8750 [61], firmado entre o Inep e a Universidade Federal de Minas Gerais (UFMG), foi constatado a possibilidade de reidentificação dos candidatos.

De acordo com a TED 8750 [61], o Inep utiliza apenas técnicas de desidentificação, em que se removem possíveis identificadores individuais óbvios dos registros, e de pseudonimização, em que tais identificadores individuais óbvios são substituídos por um código único de identificação artificialmente criado. Entretanto, mesmo quando se removem ou se criptografam identificadores explícitos dos microdados, outros dados distintos, chamados de quaseidentificadores, podem se combinar de maneira inadequada e ser vinculados a informações publicamente disponíveis para reidentificar os indivíduos.

Ainda segundo o estudo técnico [61], foi concluído que para o Censo da Educação Básica de 2018, sem informação auxiliar, nenhum indivíduo pode ser reidentificado com absoluta certeza na base, mas é possível reidentificar até 14.54% dos indivíduos com uma combinação de 3 quaseidentificadores, 33.12% com a combinação de 4 quaseidentificadores e com o uso de todos os 10 quaseidentificadores o risco chega a 60.90%. Estas taxas são de sucesso determinístico, é medido como a fração dos indivíduos da base que podem ser reidentificados com absoluta certeza. Dentre os 10 quaseidentificadores estavam mês e ano do nascimento, gênero, cor/raça, código do município de nascimento, nacionalidade, código do país de origem, código do município de residência, código da entidade, dependência administrativa. Para o Censo da Educação Superior de 2018, também não é possível identificar indivíduos sem informação auxiliar. Para combinações de quaseidentificadores, os resultados foram de 38.87% dos indivíduos na base com 3 quaseidentificadores, 79.20% com 4 quaseidentificadores e 97.22% com todos os 11 quaseidentificadores. Dentre estes quaseidentificadores estavam dados como dia, mês e ano do nascimento, código do curso, gênero, cor/raça, código do município de nascimento, nacionalidade, código do país de origem, código da IES e escola de conclusão do ensino médio.

Um dos argumentos que forçou a adaptação dos microdados foi que, com a possibilidade de reidentificação dos indivíduos, a base não estaria anonimizada. Dessa forma, como propostas para suprimir a possibilidade de reidentificação, a Diretoria de Avaliação da Educação Básica (DAEB) [62] sugeriu as seguintes alterações: Exclusão do código da

escola de conclusão do ensino médio; Exclusão das informações referentes aos pedidos de atendimento especializado e específico, recursos de atendimento especializado e específico para a realização da prova; Substituição da Idade por Faixa Etária; Exclusão de informações referentes aos municípios de nascimento e residência do participante. Por isso, os microdados utilizados para as análises deste trabalho possuem consideravelmente menos variáveis disponíveis que suas versões anteriores, e podem conter um número diferente dos disponibilizados no futuro em caso de necessidade de novas alterações.

Segundo Posicionamento Público da sociedade civil [63] divulgado por 33 entidades, esta adequação realizada pelo Inep dos dados representa um retrocesso em termos da transparência da administração pública e causará um grande impacto em termos de avaliação educacional, prejudicando a elaboração de políticas públicas que respondam às necessidades da população.

4.3 Descrição dos Dados

Os arquivos dos microdados são obtidos em um arquivo compactado, ao extraí-los observa-se que existe uma estrutura de diretórios. A seguir será apresentada, de maneira breve, essa estrutura e o seu conteúdo. O formato segue um padrão e é comum para os microdados das edições de 2019 e 2020, tanto para nomes quanto para estrutura e arquivos, eventuais diferenças serão pontuadas. O conjunto é composto por cinco pastas:

- DADOS: contém os arquivos *.csv* com as bases de dados sobre os itens, notas e o questionário socioeconômico;
- DICIONÁRIO: contém informações sobre as variáveis presentes em cada base;
- INPUTS: arquivos de entrada para a leitura das bases de dados em outros softwares como SAS, SPSS e R;
- LEIA-ME E DOCUMENTOS TÉCNICOS: possui documentos a respeito dos dados e sobre o Enem;
- PROVAS E GABARITOS: contém os arquivos de todas as provas e gabaritos aplicadas na edição.

Os dados do participante, da escola, do local de aplicação de prova, de respostas e presença na prova objetiva, redação e do questionário socioeconômico, estão no arquivo MICRODADOS_ENEM_2020.csv (o número no final do nome do arquivo corresponde à edição dos microdados). O arquivo ITENS_PROVA_2020.csv contém informações sobre as provas, como a posição do item na prova, área de conhecimento, se é de língua estrangeira, cor da prova, gabarito, parâmetros dos itens do modelo de Teoria de

Resposta ao Item (TRI), entre outros. O dicionário dos dados está contido no arquivo Dicionário_Microdados_Enem_2020, nos formatos *.xlsx* e *.ods*. Por meio deste arquivo é possível entender como estão estruturados os arquivos das bases de dados, o que significa cada variável, seu tamanho, tipo de dados e possibilidades de valores armazenados. Como exemplo, uma captura de tela contendo parte do dicionário dos microdados de 2020 pode ser vista na Figura 4.1. O arquivo Leia_Me_Enem_2020.pdf contém uma apresentação dos microdados, algumas considerações a respeito da sua elaboração e uma breve descrição sobre o conteúdo de cada item da pasta principal do conjunto de dados.

DICIONÁRIO DE VARIÁVEIS - ENEM 2020					
NOME DA VARIÁVEL	Descrição	Variáveis Categóricas		Tamanho	Tipo
		Categoria	Descrição		
DADOS DO PARTICIPANTE					
NU_INSCRICAO	Número de inscrição[1]			12	Numérica
NU_ANO	Ano do Enem			4	Numérica
TP_FAIXA_ETARIA	Faixa etária[2]	1	Menor de 17 anos	2	Numérica
		2	17 anos		
		3	18 anos		
		4	19 anos		
		5	20 anos		
		6	21 anos		
		7	22 anos		
		8	23 anos		
		9	24 anos		
		10	25 anos		
		11	Entre 26 e 30 anos		
		12	Entre 31 e 35 anos		
		13	Entre 36 e 40 anos		
		14	Entre 41 e 45 anos		
		15	Entre 46 e 50 anos		
		16	Entre 51 e 55 anos		
		17	Entre 56 e 60 anos		
		18	Entre 61 e 65 anos		
		19	Entre 66 e 70 anos		
		20	Maior de 70 anos		
TP_SEXO	Sexo	M	Masculino	1	Alfanumérica
		F	Feminino		
TP_ESTADO_CIVIL	Estado Civil	0	Não informado	1	Numérica
		1	Solteiro(a)		
		2	Casado(a)/Mora com companheiro(a)		
		3	Divorciado(a)/Desquitado(a)/Separado(a)		
		4	Viúvo(a)		
TP_COR_RACA	Cor/raça	0	Não declarado	1	Numérica
		1	Branca		
		2	Preta		
		3	Parda		
		4	Amarela		
		5	Indígena		
TP_NACIONALIDADE	Nacionalidade	0	Não informado	1	Numérica
		1	Brasileiro(a)		
		2	Brasileiro(a) Naturalizado(a)		
		3	Estrangeiro(a)		
		4	Brasileiro(a) Nato(a), nascido(a) no exterior		

MICRODADOS_ENEM_2020 | ITENS_PROVA_2020

Figura 4.1: Parte do arquivo de Dicionário dos Microdados do Enem 2020.

Com relação à reestruturação da apresentação dos microdados citada na Seção 4.2, foram realizadas algumas mudanças, segundo os arquivos Leia-me dos dados [25] [26]. Para o Enem 2019 e 2020, foram realizadas as seguintes alterações nas tabelas MICRODADOS_ENEM_2019 e MICRODADOS_ENEM_2020:

- Excluir a variável CO_ESCOLA;

- Excluir dos microdados informações referentes aos pedidos de atendimento especializado e específico, recursos de atendimento especializado e específico para a realização da prova;
- Substituir a variável NU_IDADE por TP_FAIXA_ETARIA, mostrando uma faixa de idade ao invés da idade exata do participante;
- Excluir informações referentes aos municípios de nascimento e residência do participante.

Ainda segundo os documentos [25] [26], também foram excluídos do conjunto de dados os registros de indivíduos que realizaram tipos de provas com um número total muito pequeno, o que permitiria sua identificação. Para o Enem 2019, foram excluídos da base do microdados os registros dos participantes que realizaram as provas: 543, 544, 545 e 546 de Ciências da Natureza; 547, 548, 549, 550 e 564 de Ciências Humanas; 551, 552, 553, 554 e 565 de Linguagens e Códigos; e 555, 556, 557 e 558 de Matemática. Para o Enem 2020, foram excluídos os registros dos participantes que realizaram as provas: 601, 602 e 684 de Ciências da Natureza; 571, 572 e 654 de Ciências Humanas; 581, 582 e 664 de Linguagens e Códigos; e 591, 592 e 674 de Matemática.

4.4 Preparação dos Dados

O próximo passo, seguindo a metodologia do CRISP-DM, é a preparação dos dados. Para a realização da análise, os dados precisam ser manipulados para que fiquem em formatos aceitáveis pelas ferramentas utilizadas e para eliminar ruído ou informações irrelevantes levando em conta os objetivos e natureza de determinadas tarefas.

Para compreender melhor as tarefas de preparação e análise dos dados, precisa-se conhecer também as ferramentas utilizadas, que foram as seguintes:

- Distribuição Anaconda 2022.05;
- Gerenciador de pacotes e ambientes Conda 4.12.0;
- Python 3.9.12;
- Jupyter Notebook 6.4.8;
- Biblioteca Pandas 1.4.2;
- Biblioteca Matplotlib 3.5.1 ;
- Weka 3.8.6;

- Microsoft Windows 10 Home 21H2.

A instalação da distribuição Anaconda permite também a instalação e configuração de algumas das outras ferramentas acima de maneira simples, isso foi um ponto positivo que levou a escolha da ferramenta. A linguagem de programação Python foi escolhida pela facilidade e simplicidade em executar as tarefas necessárias, graças a grande quantidade de bibliotecas e recursos que facilitam as atividades na área de ciência de dados, como as bibliotecas Pandas e Matplotlib. Como interface e ambiente de desenvolvimento foi usado o Jupyter Notebook, devido a sua flexibilidade para trabalhar com a linguagem ao mesmo tempo que possui recursos muito interessantes para a visualização de resultados.

Dado os objetivos específicos apresentados no Capítulo 1, os arquivos que serão utilizados são apenas MICRODADOS_ENEM_2019.csv e MICRODADOS_ENEM_2020.csv, que contém todas as variáveis necessárias para obter as respostas de interesse. Para abrir corretamente estes dados, foi necessário descobrir qual era o formato de codificação dos arquivos, caso contrário os caracteres poderiam ser interpretados de maneira incorreta. Para isso, foi utilizada a biblioteca *chardet*, em Python. Dessa forma, descobriu-se que a codificação dos arquivos esta no formato *ISO-8859-1*. Além disso, também foi preciso saber qual é o caractere utilizado como separador na base de dados. No caso, o caractere ponto e vírgula (“ ; ”), informação disponível nos arquivos leia-me (Leia_Me_Enem_2019.pdf [25] e Leia_Me_Enem_2020.pdf [26]). Com essas informações, é possível utilizar o método *read_csv()* do Pandas para ler os arquivos dos microdados e armazená-los em um objeto da classe *Dataframe*.

Na biblioteca Pandas são utilizadas duas estruturas de dados fundamentais, as Séries e os *Dataframes*. Série é um vetor unidimensional, com rótulo e índice, que pode armazenar qualquer tipo de dados. *Dataframe* é uma estrutura de dados bidimensional com colunas que podem ter tipos de dados diferentes, semelhante a uma planilha. Também pode-se dizer que cada coluna de um *Dataframe*, somadas ao seu índice e rótulo, configuram uma Série.

Após a abertura, foi utilizado o método *info()* da biblioteca Pandas para obter informações básicas sobre os arquivos. Para os microdados do Enem 2019, existem 5.095.171 entradas (linhas), com 76 colunas. Para o Enem 2020, foi observado um total de 5.783.109 entradas e 76 colunas. Devido a essa grande quantidade de dados, foi preciso manipular as estruturas que armazenam os dados para reduzir a quantidade de linhas e colunas com que se trabalhava simultaneamente, de forma a otimizar o desempenho e reduzir o consumo de tempo e recursos computacionais para executar as tarefas. Estas alterações foram planejadas levando em conta o objetivo de cada análise, de maneiras que a omissão de algumas variáveis não prejudicassem os resultados.

A maioria das manipulações nessas estruturas envolveram selecionar e filtrar colunas de *Dataframes* e Séries, criando novas estruturas de um desses tipos e aplicar métodos da linguagem de programação Python e da biblioteca Pandas nos resultados. De forma a diminuir o uso de tempo e recursos computacionais, além de facilitar as análises. Devido a grande quantidade dessas alterações, elas serão detalhadas apenas quando necessário na análise para justificar escolhas da metodologia que possam ter impacto nos resultados finais.

Como muitas das colunas não serão utilizadas nas análises, devido a pouca relevância do conteúdo para os objetivos, foi decidido carregar o arquivo de dados filtrando apenas as colunas que poderiam ser necessárias. Assim, o *dataframe* dos dados utilizado tem o mesmo número de linhas que o arquivo original, mas foi reduzido para apenas 32 colunas. Isso foi feito passando uma lista com as colunas desejadas em um parâmetro adicional, chamado de *usecols*, para o método *read_csv()*. A Figura 4.2 mostra um exemplo de como foi feita essa seleção na abertura dos dados. As variáveis das colunas restantes possuem os dados referentes ao número de inscrição, cor/raça, características da escola, UF de realização da prova, presença e notas em cada uma das provas, vetor das respostas do candidato e gabaritos de cada uma das provas, opção de língua estrangeira e respostas das perguntas do questionário socioeconômico com relação a escolaridade dos pais, renda e acesso a celular, computador e internet.

Utilizando o método *info()* é possível obter informações sobre um *dataframe*. A Figura 4.3 mostra as informações dos *dataframes* com os microdados dos anos de 2019 e 2020. Posteriormente, colunas com muitos dados ausentes ou alta frequência de valores que indiquem que a questão não foi respondida, foram retiradas da análise para não produzir resultados tendenciosos. Como por exemplo, atributos com informações sobre as escolas dos inscritos.

A Figura 4.4 mostra a visualização das cinco primeiras linhas dos microdados de 2019 utilizando o método *head()*. Ele é útil para verificar se os arquivos foram abertos corretamente e como estão organizados. Além disso pode ser passado um parâmetro para alterar o número de linhas apresentadas na saída.

Abrindos os arquivos de dados somente com as colunas utilizadas

```
In [8]: microdados2019 = pd.read_csv(caminho_arquivo_2019, sep=";", encoding = "ISO-8859-1",
                                     usecols=['NU_INSCRICAO', 'TP_COR_RACA', 'TP_ESCOLA', 'TP_ENSINO', 'TP_DEPENDENCIA_ADM_ESC',
                                              'TP_LOCALIZACAO_ESC', 'TP_SIT_FUNC_ESC', 'SG_UF_PROVA',
                                              'TP_PRESENCA_CN', 'TP_PRESENCA_CH', 'TP_PRESENCA_LC',
                                              'TP_PRESENCA_MT', 'NU_NOTA_CN', 'NU_NOTA_CH', 'NU_NOTA_LC', 'NU_NOTA_MT',
                                              'NU_NOTA_REDACAO', 'TX_RESPOSTAS_CN', 'TX_RESPOSTAS_CH', 'TX_RESPOSTAS_LC',
                                             'TX_RESPOSTAS_MT', 'TP_LINGUA', 'TX_GABARITO_CN', 'TX_GABARITO_CH', 'TX_GABARITO_LC',
                                             'TX_GABARITO_MT' ,'Q001', 'Q002', 'Q006','Q022','Q024', 'Q025'])

In [13]: microdados2020 = pd.read_csv(caminho_arquivo_2020, sep=";", encoding = "ISO-8859-1",
                                     usecols=['NU_INSCRICAO', 'TP_COR_RACA', 'TP_ESCOLA', 'TP_ENSINO', 'TP_DEPENDENCIA_ADM_ESC',
                                              'TP_LOCALIZACAO_ESC', 'TP_SIT_FUNC_ESC', 'SG_UF_PROVA',
                                              'TP_PRESENCA_CN', 'TP_PRESENCA_CH', 'TP_PRESENCA_LC',
                                              'TP_PRESENCA_MT', 'NU_NOTA_CN', 'NU_NOTA_CH', 'NU_NOTA_LC', 'NU_NOTA_MT',
                                              'NU_NOTA_REDACAO', 'TX_RESPOSTAS_CN', 'TX_RESPOSTAS_CH', 'TX_RESPOSTAS_LC',
                                             'TX_RESPOSTAS_MT', 'TP_LINGUA', 'TX_GABARITO_CN', 'TX_GABARITO_CH', 'TX_GABARITO_LC',
                                             'TX_GABARITO_MT' , 'Q001', 'Q002', 'Q006','Q022','Q024', 'Q025'])
```

Figura 4.2: Exemplo da abertura dos microdados utilizando o parâmetro *usecols*.

```
In [10]: microdados2019.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5095171 entries, 0 to 5095170
Data columns (total 32 columns):
 #   Column                  Dtype
---  ------                  -----
 0   NU_INSCRICAO            int64
 1   TP_COR_RACA             int64
 2   TP_ESCOLA               int64
 3   TP_ENSINO               float64
 4   TP_DEPENDENCIA_ADM_ESC  float64
 5   TP_LOCALIZACAO_ESC      float64
 6   TP_SIT_FUNC_ESC         float64
 7   SG_UF_PROVA             object
 8   TP_PRESENCA_CN          int64
 9   TP_PRESENCA_CH          int64
 10  TP_PRESENCA_LC          int64
 11  TP_PRESENCA_MT          int64
 12  NU_NOTA_CN              float64
 13  NU_NOTA_CH              float64
 14  NU_NOTA_LC              float64
 15  NU_NOTA_MT              float64
 16  TX_RESPOSTAS_CN         object
 17  TX_RESPOSTAS_CH         object
 18  TX_RESPOSTAS_LC         object
 19  TX_RESPOSTAS_MT         object
 20  TP_LINGUA               int64
 21  TX_GABARITO_CN          object
 22  TX_GABARITO_CH          object
 23  TX_GABARITO_LC          object
 24  TX_GABARITO_MT          object
 25  NU_NOTA_REDACAO         float64
 26  Q001                    object
 27  Q002                    object
 28  Q006                    object
 29  Q022                    object
 30  Q024                    object
 31  Q025                    object
dtypes: float64(9), int64(8), object(15)
memory usage: 1.2+ GB
```

```
In [15]: microdados2020.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5783109 entries, 0 to 5783108
Data columns (total 32 columns):
 #   Column                  Dtype
---  ------                  -----
 0   NU_INSCRICAO            int64
 1   TP_COR_RACA             int64
 2   TP_ESCOLA               int64
 3   TP_ENSINO               float64
 4   TP_DEPENDENCIA_ADM_ESC  float64
 5   TP_LOCALIZACAO_ESC      float64
 6   TP_SIT_FUNC_ESC         float64
 7   SG_UF_PROVA             object
 8   TP_PRESENCA_CN          int64
 9   TP_PRESENCA_CH          int64
 10  TP_PRESENCA_LC          int64
 11  TP_PRESENCA_MT          int64
 12  NU_NOTA_CN              float64
 13  NU_NOTA_CH              float64
 14  NU_NOTA_LC              float64
 15  NU_NOTA_MT              float64
 16  TX_RESPOSTAS_CN         object
 17  TX_RESPOSTAS_CH         object
 18  TX_RESPOSTAS_LC         object
 19  TX_RESPOSTAS_MT         object
 20  TP_LINGUA               int64
 21  TX_GABARITO_CN          object
 22  TX_GABARITO_CH          object
 23  TX_GABARITO_LC          object
 24  TX_GABARITO_MT          object
 25  NU_NOTA_REDACAO         float64
 26  Q001                    object
 27  Q002                    object
 28  Q006                    object
 29  Q022                    object
 30  Q024                    object
 31  Q025                    object
dtypes: float64(9), int64(8), object(15)
memory usage: 1.4+ GB
```

Figura 4.3: Método *info()* utilizado para obter informações sobre um dataframe.

```
In [9]: microdados2019.head()
```

Out[9]:

	NU_INSCRICAO	TP_COR_RACA	TP_ESCOLA	TP_ENSINO	TP_DEPENDENCIA_ADM_ESC	TP_LOCALIZACAO_ESC	TP_SIT_FUNC_ESC	SG_UF_PROVA	TP_PRESENC
0	190001595656	3	1	NaN	NaN	NaN	NaN	SP	
1	190001421546	1	1	1.0	NaN	NaN	NaN	BA	
2	190001133210	3	1	1.0	NaN	NaN	NaN	CE	
3	190001199383	1	1	NaN	NaN	NaN	NaN	TO	
4	190001237802	1	1	1.0	NaN	NaN	NaN	MG	

5 rows × 32 columns

Figura 4.4: Visualização das cinco primeiras linhas dos microdados de 2019 utilizando o método *head()*.

Com o objetivo de reduzir o tempo de processamento, muitas vezes foi necessário criar novos *dataframes* com um subconjunto dos dados originais, contendo apenas as colunas de relevância para obter uma determinada informação. Como exemplo, tem-se a Figura 4.5. Pode-se ver a aplicação do método *filter()* no *dataframe* original dos microdados de 2019. Esse método recebe como parâmetro uma lista com as colunas desejadas e cria uma nova visualização apenas desse subconjunto. O resultado dessa operação pode ser armazenado em um novo *dataframe*, como feito nesta figura. Dessa forma, podem ser feitas operações em um conjunto de dados menor, otimizando a utilização de recursos e tempo.

```
In [10]: colunas_presenca = ['TP_PRESENCA_CN', 'TP_PRESENCA_CH', 'TP_PRESENCA_LC',
                            'TP_PRESENCA_MT']

In [17]: presenca2019 = microdados2019.filter(items = colunas_presenca)

In [18]: presenca2019.head()
```

Out[18]:

	TP_PRESENCA_CN	TP_PRESENCA_CH	TP_PRESENCA_LC	TP_PRESENCA_MT
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	1	1	1	1
4	1	1	1	1

Figura 4.5: Criando novo *Dataframe* utilizando o método *filter()*.

Um ponto importante para as análises envolvendo inscritos presentes nos dois dias, é filtrar as linhas correspondentes às colunas em que o valor do atributo de presença, nos dois dias de provas, fosse igual a 1, que significa que a pessoa estava presente na

prova. Como são aplicadas mais de uma prova no mesmo dia, e estas fazem parte de um mesmo caderno, se uma pessoa esteve presente em uma das provas, isso significa que ela obrigatoriamente esteve presente em todas as outras provas do mesmo dia. Não é possível entregar apenas uma parte do caderno de provas. Então, primeiro foi verificado se os dados podiam ter alguma inconsistência neste sentido, com linhas em que valores de presenças fossem diferentes para provas de um mesmo dia. Para isso, foi utilizado o método *equals()*. A Figura 4.6 mostra essa operação. Um detalhe a ser notado, é que quando se especifica uma coluna em um *dataframe* (adicionando o nome dela entre os colchetes na frente do nome do *dataframe*), o Pandas entende que está trabalhando com uma Série. Normalmente, existem métodos parecidos, com mesmo objetivo, para os dois tipos de estruturas. Dessa forma, o método *equals()* retorna *True* caso os valores sejam iguais, ou *False* caso contrário. Como pode ser visto na figura, os valores de presença para provas feitas em um mesmo dia eram iguais, mostrando que os dados não apresentam inconsistência neste caso. Assim, para as análises sobre inscritos presentes nos dois dias de provas, foi verificado a presença em apenas uma das provas de cada dia. O mesmo processo foi feito para as análises envolvendo os inscritos ausentes, fazendo as alterações de acordo com o valor desejado.

```
In [19]: presenca2019['TP_PRESENCA_CH'].equals(presenca2019['TP_PRESENCA_LC'])
Out[19]: True

In [20]: presenca2019['TP_PRESENCA_CN'].equals(presenca2019['TP_PRESENCA_MT'])
Out[20]: True
```

Figura 4.6: Utilização do método *equals()* para comparar os valores de duas Séries.

Em seguida, para finalmente filtrar as linhas correspondentes às colunas em que o valor do atributo de presença fosse o desejado, foi utilizada a propriedade *loc* do *Dataframe*, como pode ser visto na Figura 4.7. Uma maneira de utilizá-la é passar uma condição. Neste caso, na primeira linha é retornado e armazenado na variável "renda2019" todas as linhas em que as presenças nas provas de Linguagens e Códigos e Matemática forem iguais a 1, simultaneamente. Em seguida, é filtrado desse resultado, apenas a coluna relativa a variável que possui as respostas da questão sobre a classe de renda. Logo depois, pode ser visto o resultado do método *value_counts()*, que neste caso está retornando o número de ocorrências para cada valor único possível de classe de renda. Também foi utilizado o método *sort_index()*, que sendo aplicado após o *value_counts()*, ordena o *dataframe* resultante pelo índice, de acordo com o parâmetro *ascending*. Neste caso, em ordem

crescente para a letra correspondente à classe de renda. Caso seja especificado dentro do método *value_counts()* o parâmetro *normalize* com valor *True*, é retornada a frequência relativa ao total, e não o número total de ocorrências. Isso pode ser visto na Figura 4.8.

```
In [11]: renda2019 = microdados2019.loc[(microdados2019.TP_PRESENCA_LC == 1) & (microdados2019.TP_PRESENCA_MT == 1)]

In [12]: renda2019 = renda2019.filter(items=['Q006'])
         renda2019.info()

         <class 'pandas.core.frame.DataFrame'>
         Int64Index: 3701910 entries, 3 to 5095165
         Data columns (total 1 columns):
          #   Column  Dtype
         ---  ------  -----
          0   Q006    object
         dtypes: object(1)
         memory usage: 56.5+ MB

In [13]: renda2019.value_counts().sort_index(ascending=True)

Out[13]: Q006
         A       158088
         B       892317
         C       904985
         D       356864
         E       345980
         F       172193
         G       234290
         H       154765
         I       117955
         J        66755
         K        48519
         L        36796
         M        36098
         N        51645
         O        41915
         P        36086
         Q        46659
         dtype: int64
```

Figura 4.7: Utilização da propriedade *loc* para filtrar linhas correspondentes às colunas com valores desejados.

```
In [26]: escolaridade2019_pai = microdados2019.loc[(microdados2019.TP_PRESENCA_LC == 1) & (microdados2019.TP_PRESENCA_MT == 1)]

In [27]: escolaridade2019_pai = escolaridade2019_pai.filter(items=['Q001'])

In [28]: escolaridade2019_pai.value_counts(normalize=True)

Out[28]: Q001
         E       0.271815
         B       0.203286
         C       0.139326
         D       0.117016
         F       0.083283
         H       0.082704
         G       0.052277
         A       0.050294
         dtype: float64
```

Figura 4.8: Utilização do método *value_counts()* com o parâmetro *normalize* = *True*.

Outro método muito solicitado foi o *groupby()*. Uma maneira de utilizá-lo é escolhendo um critério de agrupamento, as colunas a serem agrupadas e a função que será aplicada

nestas colunas. A Figura 4.9 apresenta um exemplo. Pode-se observar que o critério de agrupamento foi a cor/raça. Neste caso não foram especificadas colunas a serem agrupadas, então todas as colunas presentes no *Dataframe* "notas2020_raca" foram utilizadas. A função aplicada calculou a média dos valores dessas colunas, que armazenam as notas em cada prova. O método *round* é utilizado para arredondar os valores de acordo com um número de casas decimais passado como parâmetro.

```
In [161]: notas2020_raca.groupby(['TP_COR_RACA']).mean().round(1)
```

Out[161]:

TP_COR_RACA	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT	NU_NOTA_REDACAO
1.0	513.4	541.6	547.6	557.9	633.1
2.0	470.8	493.8	512.0	486.6	557.2
3.0	476.8	498.0	512.4	499.0	573.2
4.0	491.2	509.5	522.3	522.7	592.5
5.0	453.6	469.4	483.8	463.4	522.3

Figura 4.9: Exemplo de utilização do método *groupby*.

Algumas análises envolveram o escore bruto dos participantes. O escore bruto é o total de questões que um candidato acertou em uma prova. Os microdados do Enem não disponibilizam o escore bruto dos candidatos diretamente. Para obtê-los foi necessário comparar as respostas de cada prova de um candidato com as respostas do gabarito da respectiva prova. Apesar das provas serem de tipos diferentes, os dados da coluna do gabarito nos microdados correspondem ao gabarito da prova que o candidato realizou. A Figura 4.10 apresenta como exemplo a função utilizada para realizar este cálculo para a prova de Matemática de 2019. A variável "respostas_gabaritos2019" é um *dataframe* que possui as colunas com os vetores de respostas dos candidatos e com os vetores de respostas do gabarito de cada prova. Então, para cada inscrito, foi comparada a resposta de cada questão com o gabarito da prova que ele realizou. No final, o valor do seu escore bruto foi armazenado em uma nova coluna chamada de "ESCORE_MT". Esse cálculo foi feito para todas as provas objetivas. Dessa forma, com esses valores no *Dataframe*, puderam ser feitos agrupamentos utilizando como critério a renda e a nota da respectiva prova. Neste último caso, para descobrir valores médios de nota para cada valor possível de escore bruto (0 a 45). A Figura 4.11 mostra o exemplo para este caso.

```
In [25]: x=0
while x < len(respostas_gabaritos2019):
    respostas = respostas_gabaritos2019.at[x,'TX_RESPOSTAS_MT']
    gabarito = respostas_gabaritos2019.at[x, 'TX_GABARITO_MT']
    i=0
    escore=0
    while i < (len(respostas)):
        if (respostas[i] == gabarito[i]):
            escore=escore+1
        i=i+1
    respostas_gabaritos2019.at[x,'ESCORE_MT'] = escore
    x=x+1
```

Figura 4.10: Função para calcular o escore bruto dos candidatos na prova de Matemática de 2019.

```
In [75]: respostas_gabaritos2020.groupby(['ESCORE_CN'])['NU_NOTA_CN'].mean()

Out[75]: ESCORE_CN
         0.0      39.969099
         1.0     339.263889
         2.0     346.633333
         3.0     353.787981
         4.0     362.130447
         5.0     370.884537
         6.0     381.051290
         7.0     392.001670
         8.0     403.889094
         9.0     416.891343
         10.0    431.427787
         11.0    446.914170
         12.0    463.500997
         13.0    480.730834
         14.0    498.029850
         15.0    515.250456
         16.0    531.154346
         17.0    545.810066
         18.0    559.093191
         19.0    571.129149
         20.0    581.922111
         21.0    591.946722
         22.0    601.046486
         23.0    609.829794
         24.0    618.273185
         25.0    626.405463
         26.0    634.567180
         27.0    642.460763
         28.0    650.899842
         29.0    659.357860
         30.0    668.057478
         31.0    677.063493
         32.0    686.249096
         33.0    695.235873
         34.0    705.288959
         35.0    715.525731
         36.0    726.123789
         37.0    736.096189
         38.0    748.143413
         39.0    760.036352
         40.0    774.583541
         41.0    790.397436
         42.0    805.025397
         43.0    828.443478
         Name: NU_NOTA_CN, dtype: float64
```

Figura 4.11: Exemplo de agrupamento com o escore bruto como critério.

4.4.1 Utilização do Matplotlib

A construção dos gráficos utilizando a biblioteca Matplotlib para apresentar os resultados, foi uma das etapas mais trabalhosas deste estudo. Cada gráfico precisa ser gerado individualmente, por meio de seu respectivo código. Um dos principais elementos do Matplotlib é a Figura. Este objeto contém todos os elementos do gráfico. A Figura 4.12 mostra os componentes de uma Figura do Matplotlib. Nota-se que existem muitos elementos personalizáveis para adaptar os gráficos de acordo com a necessidade. É possível alterar o tamanho da figura, o título, os rótulos e divisões dos eixos, cores, estilo do gráfico, dentre outros.

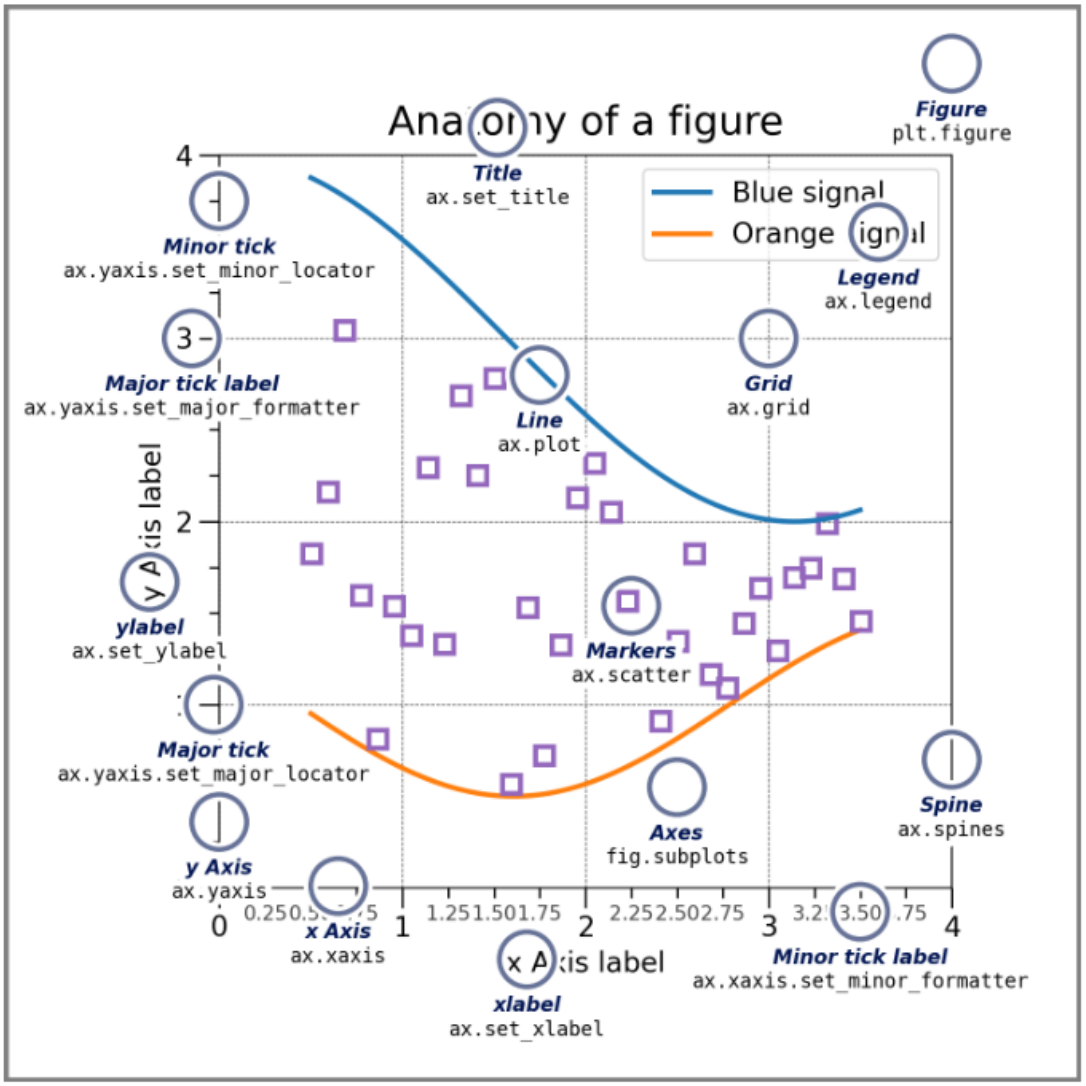

Figura 4.12: Componentes de uma Figura no Matplotlib. Fonte: Matplotlib Quick start guide [1].

A Figura 4.13 mostra um exemplo de código utilizado para gerar um gráfico com a biblioteca Matplotlib. Cada visualização de um conjunto de dados é desenhada no gráfico a partir de um objeto da classe *Axes* do Matplotlib. No código, isso é representado pelas variáveis "ax" e "ax2". Em muitos casos, os dados dos gráficos eram resultados de agrupamentos utilizando o método *groupby()*. O tipo de gráfico que será gerado, depende do método utilizado. Neste exemplo, o método *plot()* produz um gráfico em linhas. Mas também foram utilizados neste trabalho gráficos nos formatos de barra e pizza, gerados pelos métodos *bar()* e *pie()*, respectivamente. Os métodos da classe *Axes* fornecem a interface para alterar os elementos do desenho gerado. Como alguns exemplos na Figura 4.13 tem-se: *set_title()* para formatar o título do gráfico; *set_xlabel()* e *set_ylabel()* para formatar os eixos x e y, respectivamente; e *text()* para inserir e formatar textos no gráfico.

Com a intenção de proporcionar um melhor entendimento dos gráficos, foram geradas legendas com o significado dos valores de algumas variáveis analisadas, como cor/raça, escolaridade dos pais, renda, dentre outras. Para isso, foram construídos alguns dicionários traduzindo estes valores. A Figura 4.14 mostra estas estruturas.

```
In [123]: labels = ['2019', '2020']

ax = notas2019_renda_plot.groupby(['Q006'])['NU_NOTA_MT'].mean().round(1).\
plot(figsize=(15, 8), title = 'Média das notas na prova de Matemática por renda em 2019 e 2020', xlabel='Categorias de Renda', ylabe
     fontsize=12, color='blue')

ax2 = notas2020_renda_plot.groupby(['Q006'])['NU_NOTA_MT'].mean().round(1).\
plot(figsize=(15, 8), title = 'Média das notas na prova de Matemática por renda em 2019 e 2020', xlabel='Categorias de Renda', ylabe
     fontsize=12, color='red')


texto_legenda = ""

for key,value in dicionario_Q006.items():
    texto_legenda = texto_legenda + "{k} : {v} \n".format(k=key, v=value)

ax.text(17.2, 570, texto_legenda, fontsize=10)

ax.set_title('Média das notas na prova de Matemática por renda em 2019 e 2020',fontdict={'fontsize': 18}, pad=22)
ax.set_xlabel('Categorias de Renda', fontdict={'fontsize': 15}, labelpad=10)
ax.set_ylabel('Média da nota', fontdict={'fontsize': 15}, labelpad=10)

ax.legend(labels, fontsize=12)
ax.grid()

plt.savefig('notamt_renda2019_2020.png', facecolor='w', bbox_inches="tight")
plt.show()
```

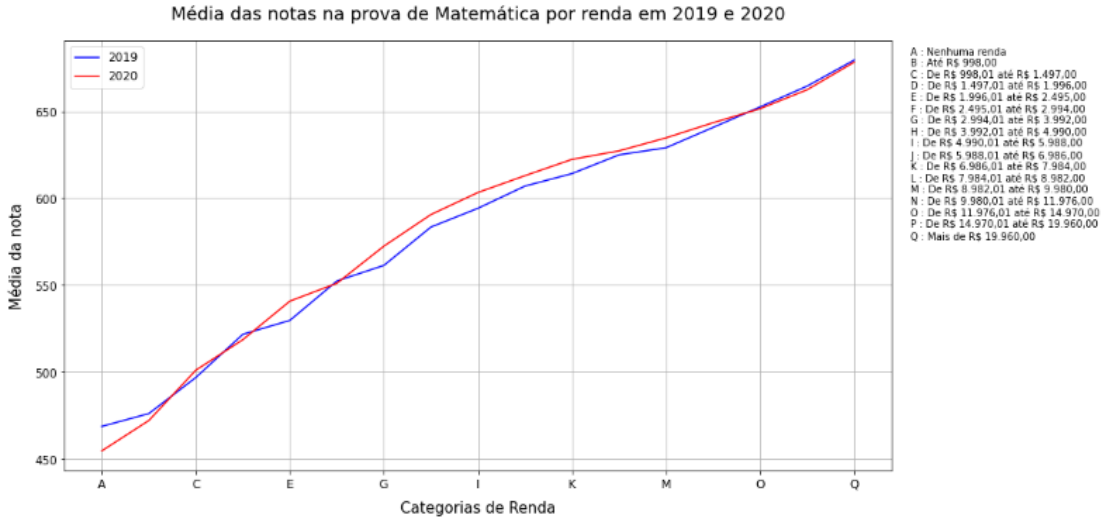

Figura 4.13: Código utilizado para gerar o gráfico de média das notas na prova de Matemática por renda em 2019 e 2020.

```
In [8]: dicionario_Q001eQ002 = {'A':'Nunca estudou',
                                'B':'Não completou a 4ª série/5º ano do Ensino Fundamental',
                                'C':'Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental',
                                'D':'Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio',
                                'E':'Completou o Ensino Médio, mas não completou a Faculdade',
                                'F':'Completou a Faculdade, mas não completou a Pós-graduação',
                                'G':'Completou a Pós-graduação',
                                'H':'Não sei'}

In [9]: dicionario_Q022eQ024 = {'A':'Não',
                                'B':'Sim, um',
                                'C':'Sim, dois',
                                'D':'Sim, três',
                                'E':'Sim, quatro ou mais'}

In [10]: dicionario_Q025 = {'A':'Não',
                            'B':'Sim'}

In [11]: dicionario_presenca = {0 : 'Faltou à prova',
                                1 : 'Presente na prova',
                                2 : 'Eliminado na prova'}

In [12]: dicionario_Q006 = {'A':'Nenhuma renda',
                            'B': 'Até R\$ 998,00',
                            'C': 'De R\$ 998,01 até R\$ 1.497,00',
                            'D': 'De R\$ 1.497,01 até R\$ 1.996,00',
                            'E': 'De R\$ 1.996,01 até R\$ 2.495,00',
                            'F': 'De R\$ 2.495,01 até R\$ 2.994,00',
                            'G': 'De R\$ 2.994,01 até R\$ 3.992,00',
                            'H': 'De R\$ 3.992,01 até R\$ 4.990,00',
                            'I': 'De R\$ 4.990,01 até R\$ 5.988,00',
                            'J': 'De R\$ 5.988,01 até R\$ 6.986,00',
                            'K': 'De R\$ 6.986,01 até R\$ 7.984,00',
                            'L': 'De R\$ 7.984,01 até R\$ 8.982,00',
                            'M': 'De R\$ 8.982,01 até R\$ 9.980,00',
                            'N': 'De R\$ 9.980,01 até R\$ 11.976,00',
                            'O': 'De R\$ 11.976,01 até R\$ 14.970,00',
                            'P': 'De R\$ 14.970,01 até R\$ 19.960,00',
                            'Q': 'Mais de R\$ 19.960,00'}
```

Figura 4.14: Código dos dicionários utilizados para criar as legendas dos gráficos.

4.4.2 Preparação dos Dados para o Weka e JRip

A execução do algoritmo no Weka demanda uma grande quantidade de memória do computador, por isso, fez-se necessário reduzir o conjunto de dados. Assim, foi escolhido um subconjunto da população de 5.783.109 registros, contendo o total de 10.005 linhas. Segundo Conroy [64], tal valor satisfaz o tamanho de amostra mínima para que os resultados tenham 95% de nível de confiança, com uma margem de erro de 1%. Esta amostra foi gerada utilizando o Pandas, como mostra a Figura 4.15. Foram exportados 10.005 registros, selecionados de maneira aleatória dos microdados de 2020, em um arquivo com extensão *.csv*. O método *sample()* permite fazer essa seleção de uma fração dos dados de maneira aleatória. Depois de sua aplicação, foi necessário resetar o índice com o método *reset_Index()*. Isso acontece porque as linhas retornadas ficam com o número da posição da qual foram retiradas no *dataframe* original.

```
amostra_microdados2020 = amostra_microdados2020.sample(frac=0.00173)
amostra_microdados2020.reset_index(drop=True, inplace=True)
amostra_microdados2020.head()
```

Out[5]:

	NU_INSCRICAO	NU_ANO	TP_FAIXA_ETARIA	TP_SEXO	TP_ESTADO_CIVIL	TP_COR_RACA	TP_NACIONALIDADE	TP_ST_CONCLUSAO	TP_ANO_CONCLUIU	TP_ES
0	200002677098	2020	10	M	1	1	1	1	6	
1	200002735645	2020	4	F	1	3	1	1	2	
2	200003176041	2020	2	F	1	1	1	2	0	
3	200004579309	2020	13	F	2	3	1	1	0	
4	200005044021	2020	4	F	1	2	1	1	2	

5 rows × 76 columns

Figura 4.15: Geração da amostra de 10.005 registros no Pandas, para utilização do Weka.

Algumas colunas que não teriam relevância na análise foram retiradas nessa exportação dos dados pois estavam causando problemas ao importar o arquivo resultante no Weka. As colunas com nome do município da escola e nome de município de realização da prova, por exemplo, tinham dados que ainda poderiam ser identificados pelas coluna com código relativo a estes municípios. Então, como estavam causando problemas na importação, foram retiradas. O número de inscrição também foi retirado pois não teria relevância nas análises.

O segundo passo no pré-processamento dos dados foi aplicar um filtro do Weka, de forma que os dados fossem adequados para utilizar o algoritmo JRip. O filtro transforma atributos com valores numéricos em atributos com valores nominais, como o algoritmo exige. Estes atributos são as colunas dos microdados. Durante a primeira execução do filtro, notou-se que seria necessário retornar ao Pandas e remover também as colunas relacionados aos vetores de respostas das provas dos participantes. Durante a transformação para valores nominais, o filtro compara todas as linhas de uma coluna, buscando valores

únicos para utilizar como categorias e atribuindo o número de ocorrências deste valor à categoria. Como é pouco provável que existam muitos vetores de respostas iguais, ou seja, vários participantes responderem à todas as questões da mesma prova de maneira igual, quase todas as linhas dariam origem a uma categoria diferente, e o filtro levaria muito tempo realizando estas comparações. A Figura 4.16 mostra uma captura de tela do Weka após a aplicação do filtro *NumericToNominal*.

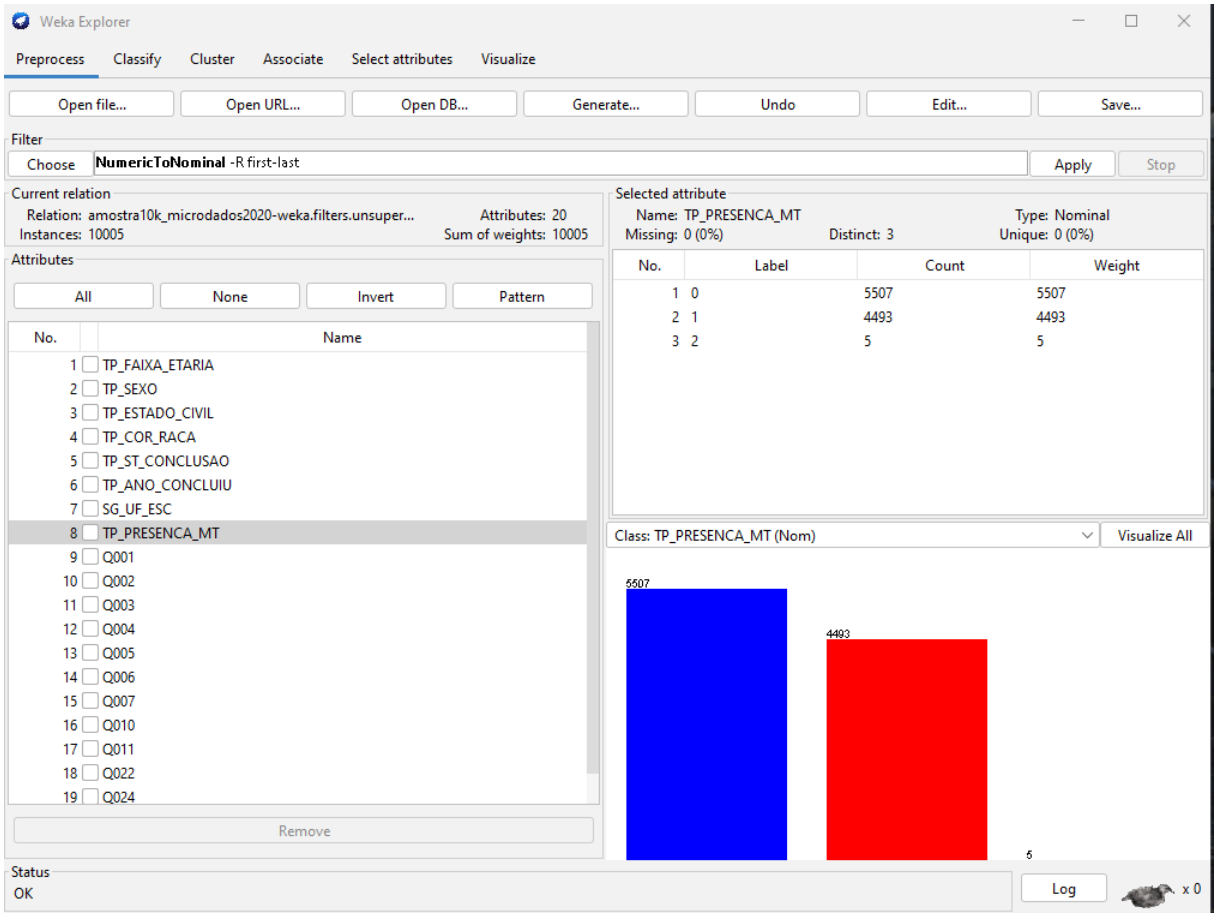

Figura 4.16: Captura de tela do Weka após aplicação do filtro *NumericToNominal*.

Antes de executar o JRip é preciso escolher qual será a classe de predição para a qual serão geradas as regras. Então, o terceiro passo na etapa de pré-processamento, foi retirar da lista de atributos aqueles que possam gerar regras óbvias. Por exemplo, é um evento comum faltar ao segundo dia de provas já tendo faltado ao primeiro. Então, caso existam dois ou mais atributos relativos à presença do inscrito em um dia de provas, serão produzidas regras do tipo: SE TP_PRESENCA_CH = 0, ENTÃO TP_PRESENCA_MT = 0. Ou seja, se estava ausente na prova de Ciências Humanas, então estava ausente na prova de Matemática. O mesmo se aplica para atributos relacionados às provas, como notas, códigos da provas e gabaritos. Também foram retirados do conjunto, atributos

que contêm muitos valores correspondentes a opções do tipo não respondeu ou "*NaN*", como aqueles que fazem referência a características da escola, já citados anteriormente. Outra decisão foi retirar atributos referentes às perguntas do questionário socioeconômico de pouca relevância, com baixo poder de predição em comparação às demais. Dentre elas, estavam perguntas relacionados à existência e quantidade de alguns cômodos (Q008 e Q009) e eletrodomésticos na residência (Q012 a Q021, e Q023). Dessa forma, os 20 atributos restantes foram os seguintes:

- TP_FAIXA_ETARIA - Faixa Etária;
- TP_SEXO - Sexo;
- TP_ESTADO_CIVIL - Estado Civil;
- TP_COR_RACA - Cor/raça;
- TP_ST_CONCLUSAO - Situação de conclusão do Ensino Médio;
- TP_ANO_CONCLUIU - Ano de Conclusão do Ensino Médio;
- SG_UF_PROVA - Sigla da Unidade da Federação da aplicação da prova;
- TP_PRESENCA_MT - Presença na prova objetiva de Matemática;
- Q001 - Até que série seu pai, ou o homem responsável por você, estudou?
- Q002 - Até que série sua mãe, ou a mulher responsável por você, estudou?
- Q003 - A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você (se ele não estiver trabalhando, escolha uma ocupação pensando no último trabalho dele);
- Q004 - A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da sua mãe ou da mulher responsável por você (se ela não estiver trabalhando, escolha uma ocupação pensando no último trabalho dela);
- Q005 - Incluindo você, quantas pessoas moram atualmente em sua residência?
- Q006 - Qual é a renda mensal de sua família? (some a sua renda com a dos seus familiares);
- Q007 - Em sua residência trabalha empregado(a) doméstico(a)?
- Q010 - Na sua residência tem carro?

- Q011 - Na sua residência tem motocicleta?
- Q022 - Na sua residência tem telefone celular?
- Q024 - Na sua residência tem computador?
- Q025 - Na sua residência tem acesso à Internet?

Com isso, concluem-se as preparações de dados necessárias para as análises do próximo capítulo. O código completo está disponível em: `https://github.com/christianpires/Monografia_Enem`.

Capítulo 5

Análise dos Dados e Resultados

Neste capítulo, serão apresentados os resultados das pesquisas e análises deste trabalho. Os resultados foram divididos em seções. A primeira apresenta algumas estatísticas obtidas dos conjuntos dos microdados de 2019 e 2020. Em seguida, são feitas análises das notas levando em conta as classes de renda dos participantes. A terceira seção tem as notas e o escore bruto dos participantes como foco das análises. A quarta seção inclui análises a respeito das notas por cor/raça e classe de renda. A quinta seção apresenta o efeito do Índice de Desenvolvimento Humano Municipal (IDHM) nas notas. Por último, é utilizado o algoritmo JRip para encontrar regras de classificação nos conjuntos de dados.

5.1 Estatísticas

Considerando que para estas bases de dados, cada linha do arquivo corresponde aos dados de um participante inscrito na edição, o número de entradas corresponde ao total de inscritos. Já para o número de presentes e ausentes nas provas, foram considerados os números relativos a cada dia de prova, desconsiderando o número de candidatos eliminados. Dessa forma, tem-se que o número de inscritos no ano de 2019, foi de 5.095.171, com 3.923.046 presentes e 1.168.053 ausentes no primeiro dia. No segundo dia de provas foram 3.710.335 presentes e 1.382.924 ausentes. Para o ano de 2020, se inscreveram 5.783.109 candidatos. Foram 2.754.140 de presentes e 3.024.590 ausentes no primeiro dia, No segundo dia, os números de ausentes chegaram a 3.184.243, com apenas 2.597.440 presentes. Os valores percentuais podem ser vistos nas Figura 5.1 e Figura 5.2, para os anos de 2019 e 2020 respectivamente. Percebe-se que mais de 50% dos candidatos se ausentaram em cada dia de provas na edição de 2020, o maior número de abstenções que o Enem já teve.

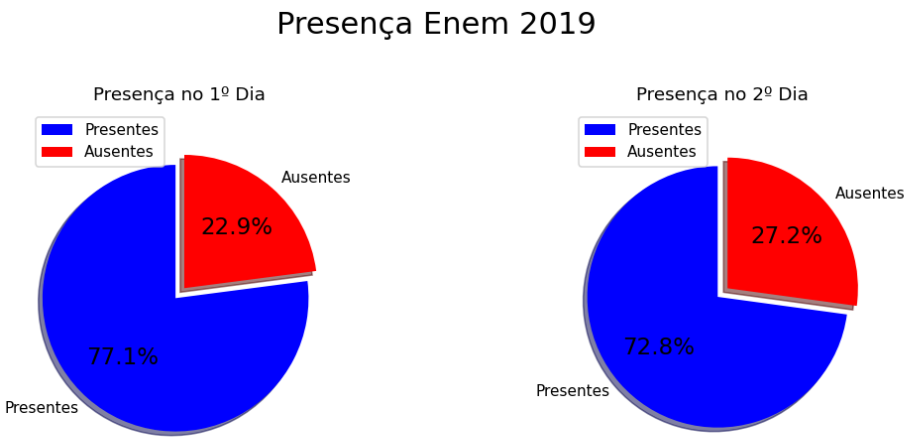

Figura 5.1: Presença no Enem de 2019.

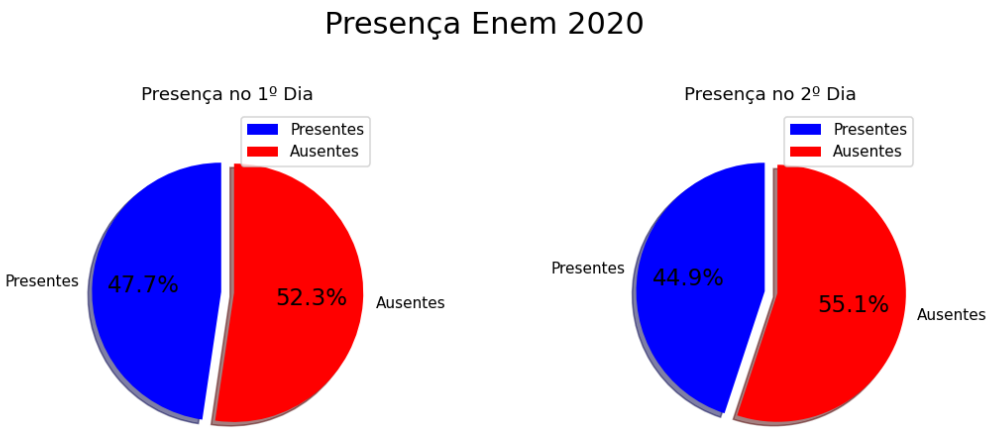

Figura 5.2: Presença no Enem de 2020.

Em seguida, procurou-se conhecer a distribuição geral das notas, sem olhar para outras variáveis ainda. Para isso, foram considerados apenas os candidatos que realizaram as provas, excluindo-se entradas sem valor válido nos campos de nota (valores "*NaN*" ou "*null*" por exemplo). Assim, foram selecionadas também apenas as notas maiores que zero, de forma a obter valores que representem melhor as médias das notas dos alunos que realizaram as provas de maneira regular. Outra observação é que os valores obtidos consideram as notas de cada prova individualmente, ou seja, para cada prova, independente do aluno estar presente nos dois dias ou em apenas um deles, foi calculada uma

Tabela 5.1: Distribuição geral das notas de 2019.

	Ciências Humanas	**Linguagens e Códigos**	**Ciências da Natureza**	**Matemática**	**Redação**
Média	508.0	520.9	477.9	523.2	592.9
Mínima	315.9	322.0	327.9	359.0	40.0
Máxima	835.1	801.7	860.9	985.5	1000.0
1º quartil	448.2	483.6	417.8	435.2	500.0
2º quartil	510.8	526.2	470.3	501.1	580.0
3º quartil	566.7	565.3	533.2	597.8	680.0
Desvio padrão	80.1	62.5	75.9	108.8	155.3

Tabela 5.2: Distribuição geral das notas de 2020.

	Ciências Humanas	**Linguagens e Códigos**	**Ciências da Natureza**	**Matemática**	**Redação**
Média	512.1	524.2	490.5	520.8	590.4
Mínima	313.7	288.7	323.9	327.1	40.0
Máxima	862.6	801.1	854.8	975.0	1000.0
1º quartil	435.7	478.2	427.1	425.8	480.0
2º quartil	512.7	530.0	483.7	505.2	580.0
3º quartil	580.7	576.4	548.7	602.3	700.0
Desvio padrão	93.8	73.0	79.6	116.9	176.3

distribuição considerando todas as notas das provas finalizadas regularmente e maiores que zero. A Tabela 5.1 mostra o resultado para o ano de 2019, e a Tabela 5.2 para o ano de 2020.

Observa-se que houve um pequeno aumento na média das notas das provas de Ciências Humanas, Linguagens e Códigos e Ciências da Natureza, enquanto as de Matemática e Redação tiveram uma pequena redução. A mediana (segundo quartil) de 2020 também teve um aumento em quatro das cinco variáveis analisadas. Essas informações não necessariamente representam um melhor desempenho dos candidatos no ano de 2020. O número bem menor de pessoas presentes nas provas e os pontos levantados no Capítulo 2, a respeito das populações mais afetadas pela pandemia, também deve ser considerado como causa para um aumento das médias. Lembrando ainda que o exame usa a TRI, um método que atribui pesos diferentes para as questões de acordo com sua dificuldade, calculada por pré-testes e pelo desempenho dos candidatos. A seguir serão apresentadas informações sobre a análise de fatores socioeconômicos relacionados aos candidatos presentes nas duas

provas, para conhecer um pouco das diferenças entre as populações que realizaram as provas nos dois anos.

A Figura 5.3 mostra um gráfico com a distribuição dos inscritos presentes nos dois dias de provas por cada categoria de cor/raça. No ano de 2019, existem 3.701.910 registros de inscritos que estiveram presentes nos dois dias para a variável de cor/raça. Para a edição de 2020, o número é de 2.588.681. A Figura 5.4 apresenta um gráfico com a distribuição percentual dos inscritos presentes nos dois dias de provas por cor/raça. Os percentuais são em relação ao total do respectivo ano. É interessante também calcular a redução percentual para cada grupo no ano de 2020 em relação a 2019. Houve uma diminuição de 28,38% para os candidatos que se declararam de cor/raça Branca, 30,87% para os de cor/raça Preta, 30,98% para Parda, 32,21% para Amarela, e 35,12% para Indígena. O gráfico com a redução percentual para o ano de 2020 pode ser visto na Figura 5.5.

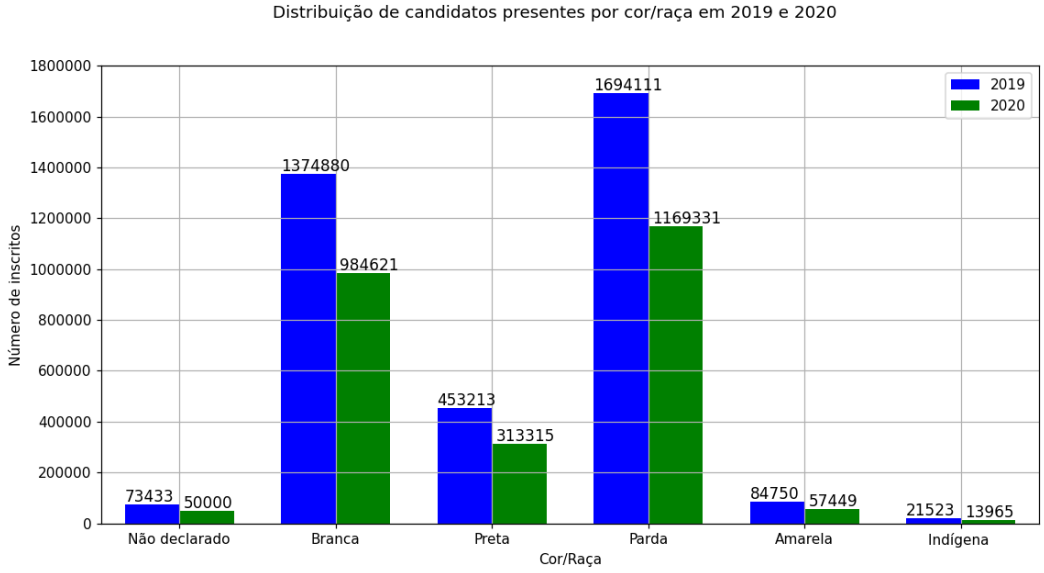

Figura 5.3: Distribuição de indivíduos presentes nos dois dias de provas por cor/raça em 2019 e 2020.

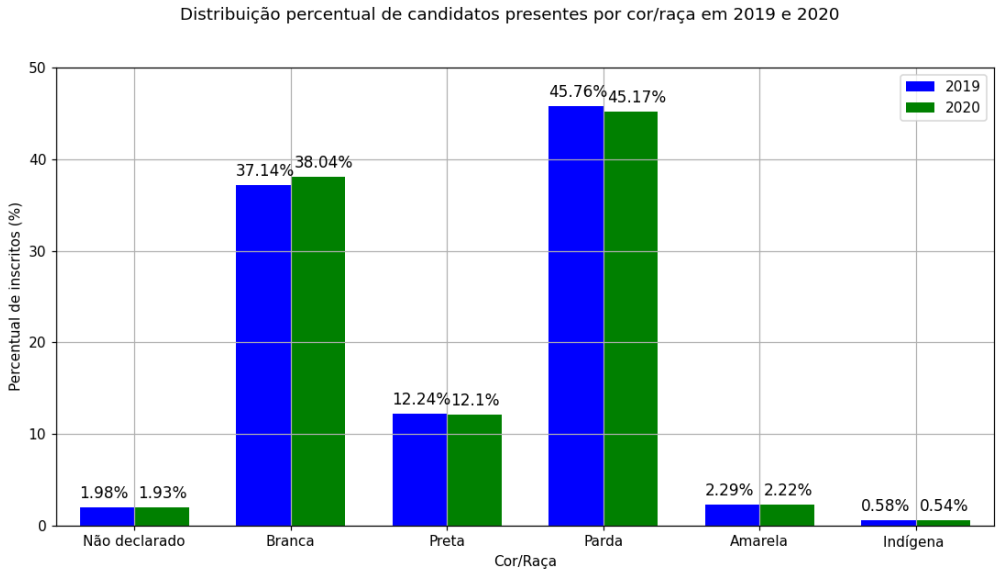

Figura 5.4: Distribuição percentual de indivíduos presentes nos dois dias de prova por cor/raça em 2019 e 2020.

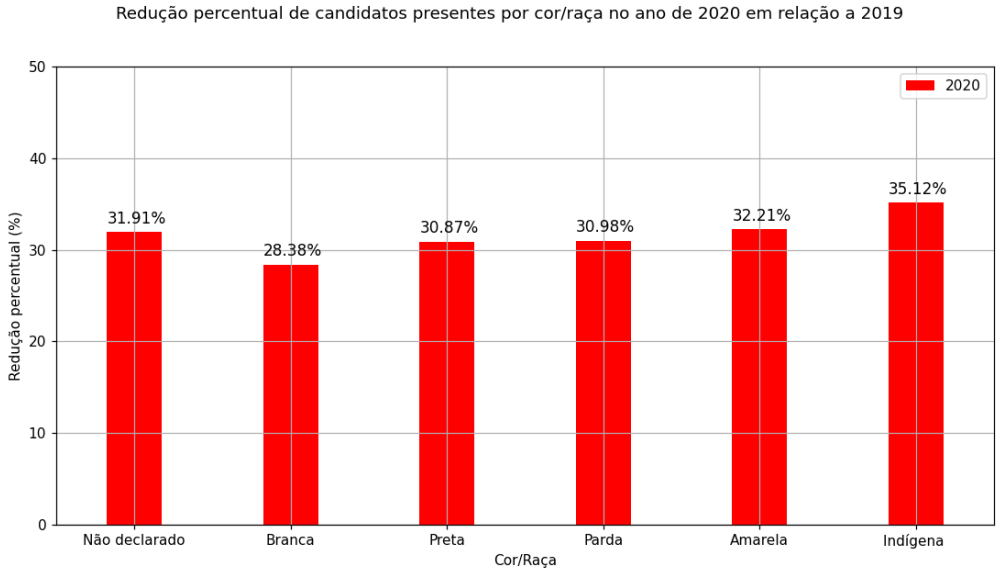

Figura 5.5: Redução percentual de indivíduos presentes nos dois dias de provas por cor/raça no ano de 2020 em relação a 2019.

A respeito da distribuição de candidatos presentes por categorias de renda nas duas edições do exame, observando a Figura 5.6 é possível perceber uma grande redução, em números absolutos, no total de participantes em todas as categorias de renda, com destaque para a classe C, de R$998,01 até R$1497,00, que teve uma queda de 53,37%

no número de candidatos presentes nas provas em 2020. Observa-se que a maioria dos participantes em 2019 era das classes de renda B e C (somadas são 48.55% do total), seguidos das classes D e E (somadas são 18,99% do total). Entretanto, em 2020, apesar das classes B e C ainda possuírem mais inscritos que as demais (46% do total), vê-se que a proporção de inscritos da classe C em relação às classes D e E é bem menor que no ano anterior, onde a diferença chegava a mais de 500 mil em relação a cada uma. Somando-se os número das classes B, C e D no ano de 2020 tem-se 58,36% do total de inscritos. A classe B passou a ser maioria no ano de 2020. Interessante observar também, um aumento de 12.691 (8%) participantes presentes que declararam não ter nenhuma renda em 2020. A Figura 5.7 apresenta a distribuição percentual de indivíduos presentes nos dois dias de provas por renda em 2019 e 2020, com percentuais relativos ao total do respectivo ano. Para o ano de 2020, os dados dos inscritos presentes nos dois dias de provas possuíam 27.377 valores "NaN" no campo de renda, que foram removidos das análises, totalizando então 2.561.304 inscritos presentes nos dois dias para o ano de 2020.

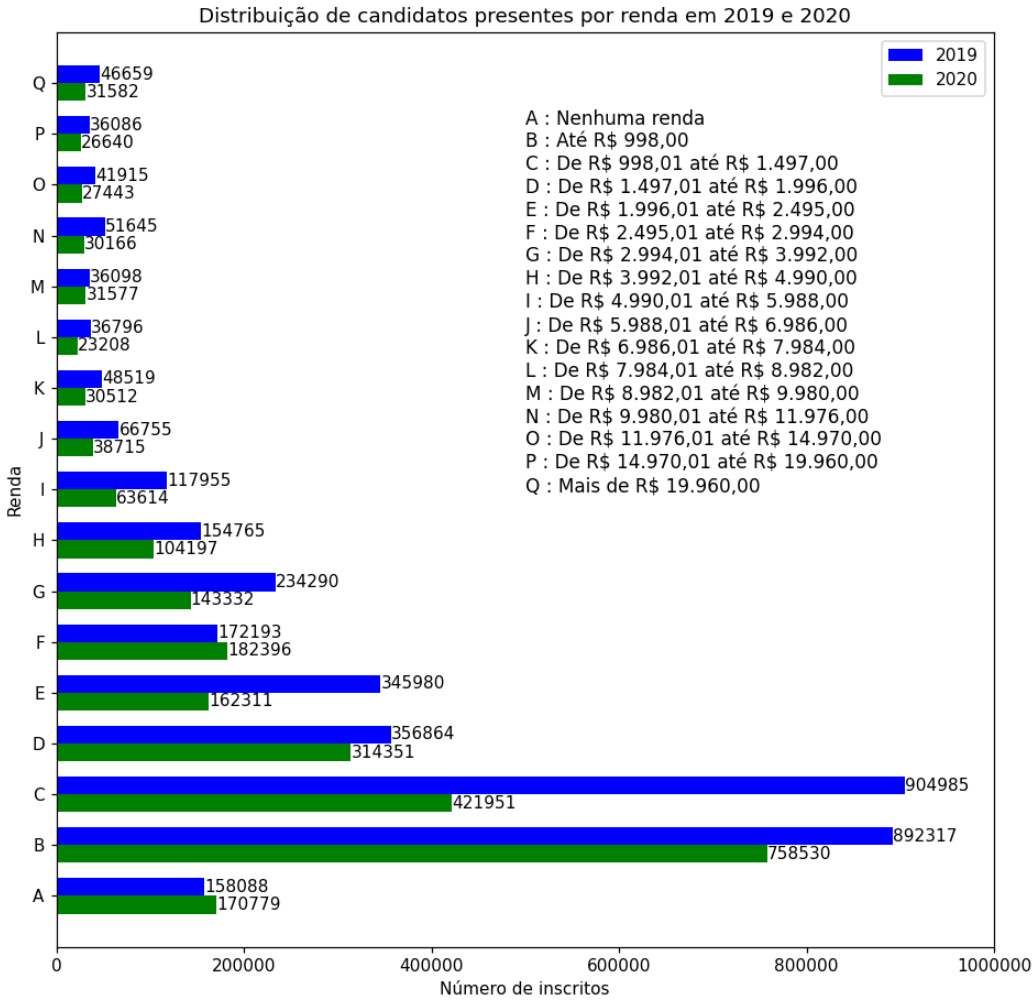

Figura 5.6: Distribuição de indivíduos presentes nos dois dias de provas por renda em 2019 e 2020.

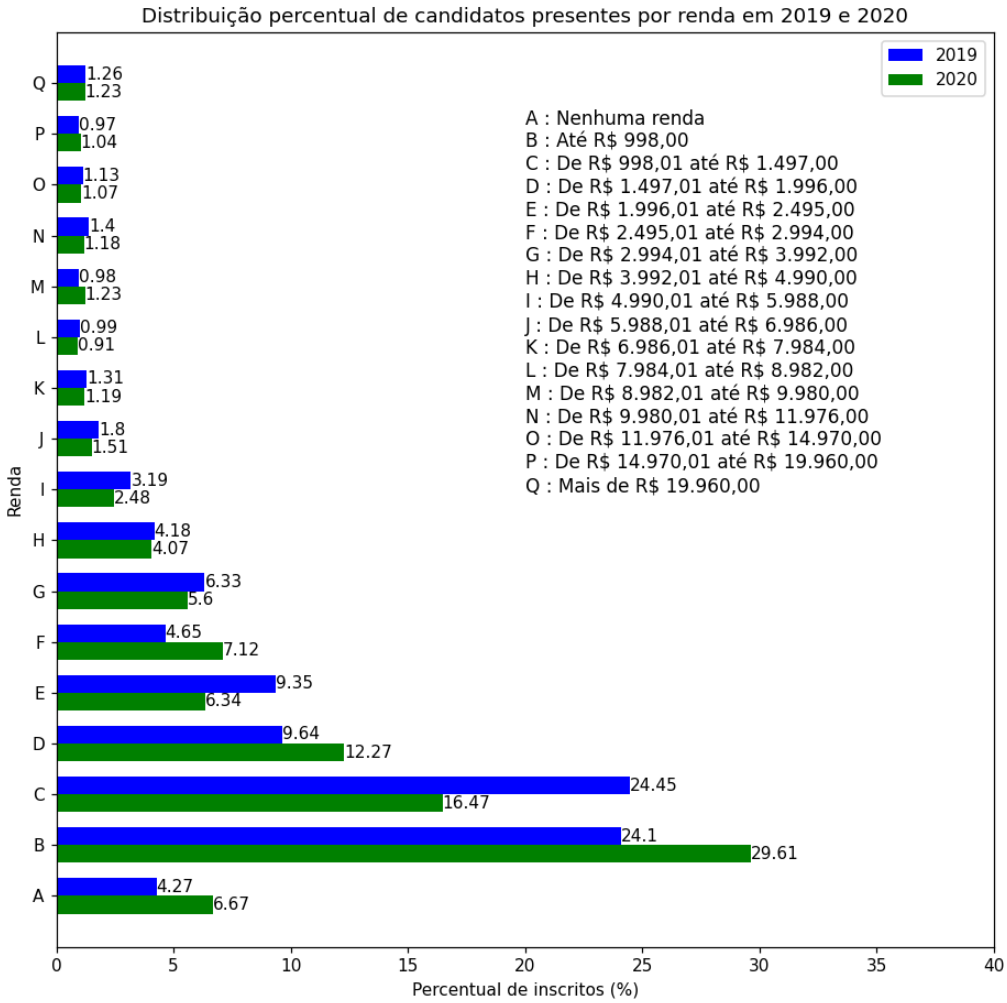

Figura 5.7: Distribuição percentual de indivíduos presentes nos dois dias de provas por renda em 2019 e 2020.

Em relação à escolaridade dos pais, os dados são divididos entre escolaridade do pai e da mãe. Na Figura 5.8 e na Figura 5.9 tem-se a distribuição obtida para os inscritos presentes nos dois dias de provas em 2019 e 2020 com relação à escolaridade do pai e da mãe, respectivamente. Já na Figura 5.10 e na Figura 5.11 tem-se a mesma distribuição utilizando os valores percentuais.

Para a escolaridade do pai, em ambas as edições, a maioria dos números é referente a candidatos com pais que completaram o ensino médio, mas não completaram a faculdade, sendo 27,18% do total em 2019 e 28,75% em 2020. Em seguida foram os inscritos com pais que não completaram a 4ª série/5º ano do ensino fundamental, com 20,33% em 2019 e 18,87% em 2020. Houveram grandes reduções de inscritos presentes em todas as categorias no ano de 2020, com destaque para a dos inscritos com pais que nunca estudaram (38,88%), pais que não completaram a 4ª série/5º ano do ensino fundamental (35,78%), pais que completaram a 8ª série/9º ano do Ensino Fundamental mas não completaram o Ensino Médio (34,4%) e dos que não sabiam a escolaridade do pai (34,76%). A Figura 5.12 mostra o gráfico com estes resultados. Todos os valores de redução percentual em 2020 foram calculados em relação a 2019.

Com relação a escolaridade da mãe, assim como na dos pais, os maiores números dizem respeito a candidatos com mães que completaram o ensino médio, mas não completaram a faculdade, com 33,35% em 2019 e 34,42% para o ano seguinte. As demais categorias tiveram valores mais próximos entre si, com exceção dos inscritos com mães que nunca estudaram e dos que não sabiam a escolaridade da mãe. Esta última categoria apresentou a maior redução percentual no ano de 2020 com relação a 2019, 42,37% dos inscritos. Em seguida, outras quatro categorias tiveram valores de redução muito próximos. Inscritos com mães que completaram a 8ª série/9º ano do Ensino Fundamental, mas não completaram o Ensino Médio (38,01%), mães que nunca estudaram (36,92%), que completaram a 4ª série/5º ano, mas não completaram a 8ª série/9º ano do Ensino Fundamental (35,05%) e mães que não completaram a 4ª série/5º ano do Ensino Fundamental (35,04%). Os resultados da redução percentual podem ser vistos na Figura 5.13. Outra informação observada é que dentre os inscritos que responderam "Não sei" sobre a escolaridade dos pais, a proporção dos que não sabem a escolaridade do pai é mais que três vezes maior que a dos que não sabem a escolaridade da mãe.

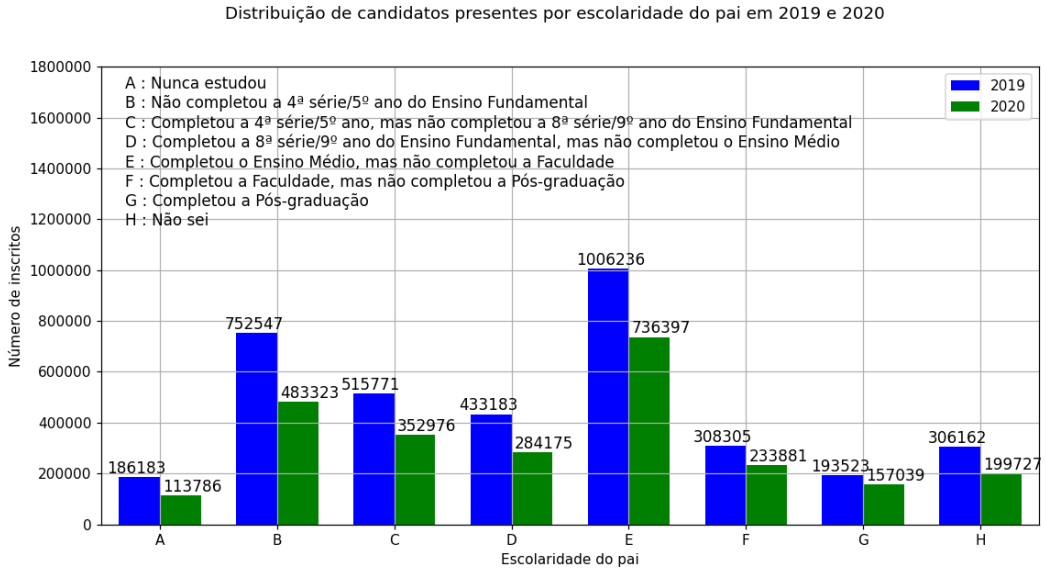

Figura 5.8: Distribuição de indivíduos presentes nos dois dias de provas por escolaridade do pai em 2019 e 2020.

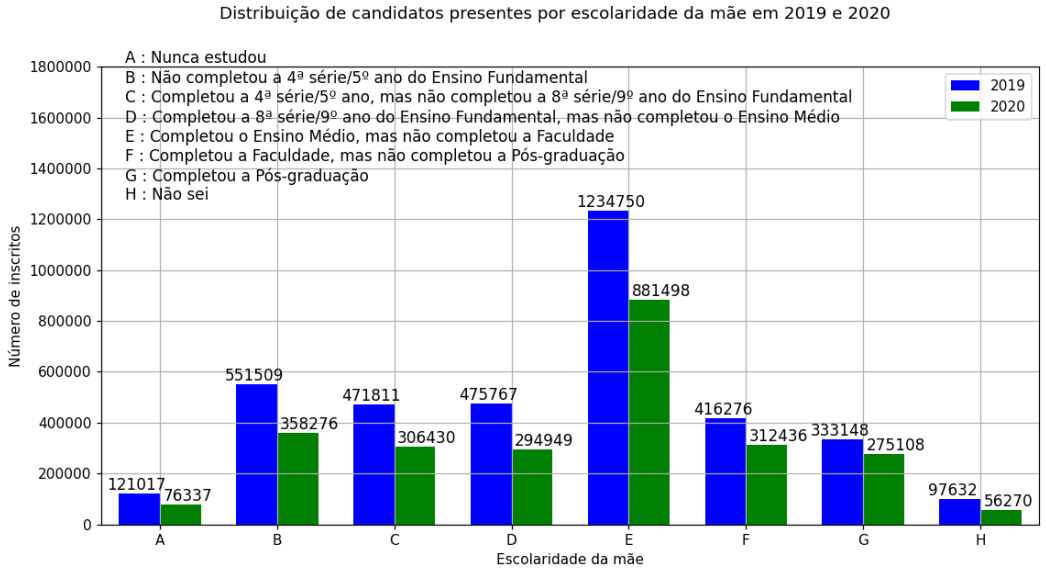

Figura 5.9: Distribuição de indivíduos presentes nos dois dias de provas por escolaridade da mãe em 2019 e 2020.

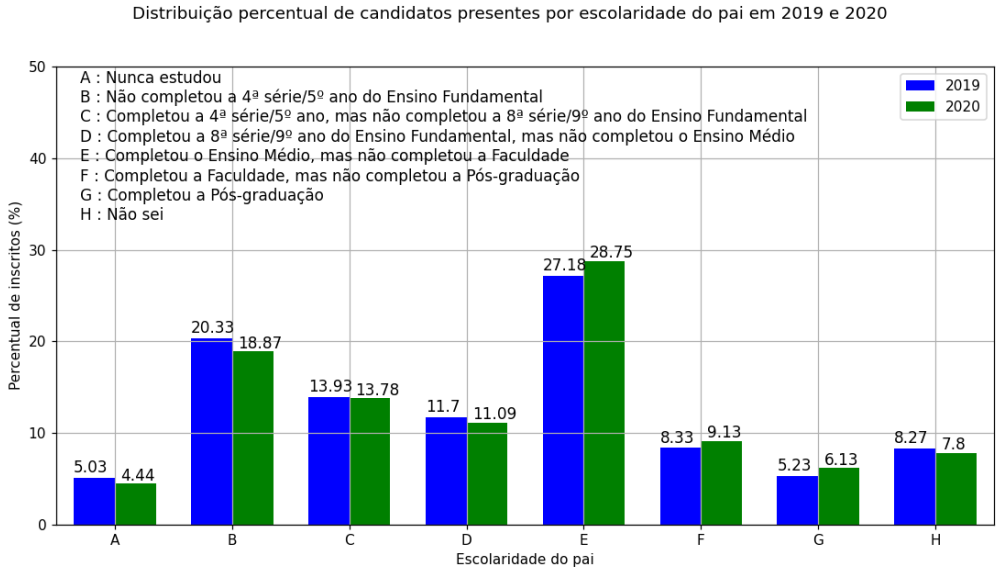

Figura 5.10: Distribuição percentual de indivíduos presentes nos dois dias de provas por escolaridade do pai em 2019 e 2020.

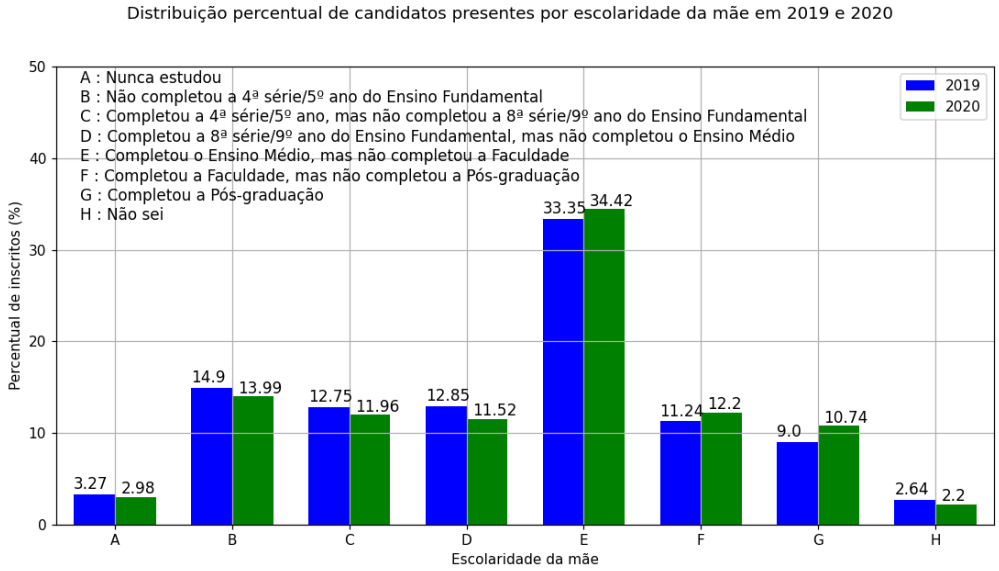

Figura 5.11: Distribuição percentual de indivíduos presentes nos dois dias de provas por escolaridade da mãe em 2019 e 2020.

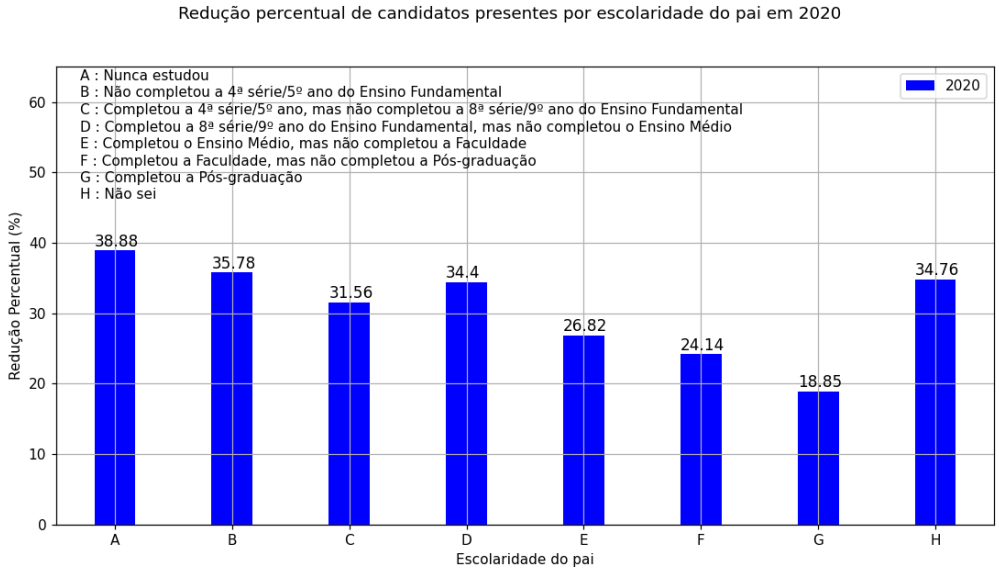

Figura 5.12: Redução percentual de indivíduos presentes nos dois dias de provas por escolaridade do pai em 2020.

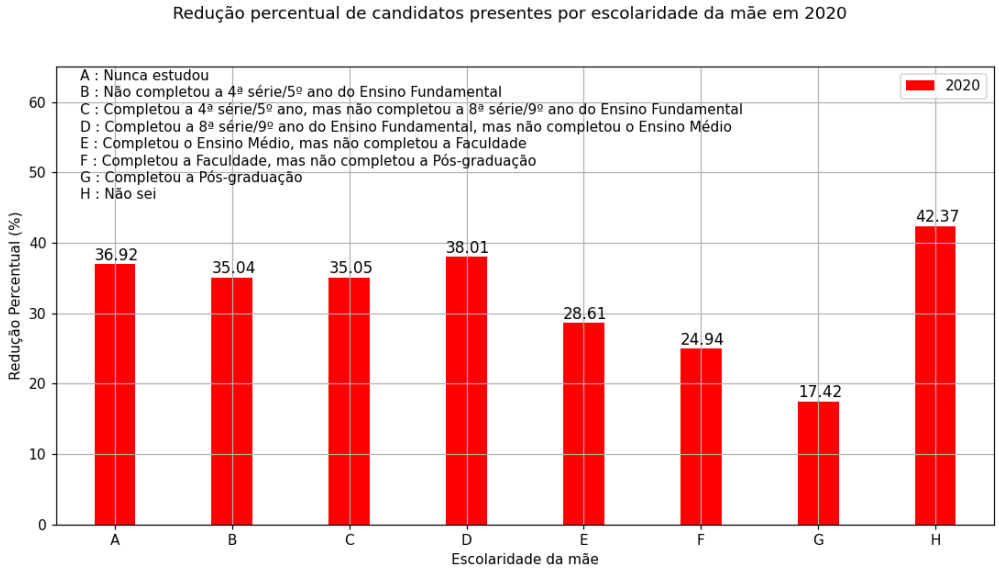

Figura 5.13: Redução percentual de indivíduos presentes nos dois dias de provas por escolaridade da mãe em 2020.

Ao analisar os dados a respeito de características das escolas, como o tipo (público, privada, exterior) ou a localização (urbana ou rural), foram encontrados muitos valores ausentes ("*NaN*") ou com o código que significa "Não Respondeu". Por este motivo, foi tomada a decisão de não considerar estes dados na análise. A quantidade de dados com respostas válidas era menor do que 50% do total de inscritos, e poderia gerar informações tendenciosas e com pouca precisão sobre a população.

A próxima etapa dos resultados trata de informações a respeito dos candidatos ausentes nos dois dias de provas de 2020. Conhecer as características desses indivíduos é tão importante quanto a dos presentes. Assim sendo, foram analisados os dados de inscritos ausentes por cor/raça, renda, escolaridade dos pais e acesso a celular, computador e Internet. Começando pelos resultados de cor/raça, existem 3.016.082 registros para ausentes nos dois dias. Desse total, tem-se que 48,64% são de cor/raça Parda, 31,9% Branca, 14,39% Preta, 2,22% Amarela, e 0,75% Indígena. Destaca-se que 60% do total de inscritos de cor/raça Indígena se ausentaram das duas provas (22.706 ausentes de 37.846 inscritos), sendo este o maior valor percentual entre as categorias, seguido da cor/raça Preta com 56,25% (434.069 ausentes de 771.740 inscritos). Os de cor/raça branca tiveram o menor valor de candidatos ausentes. A Figura 5.14 mostra a distribuição em números absolutos, e a Figura 5.15 mostra a distribuição percentual.

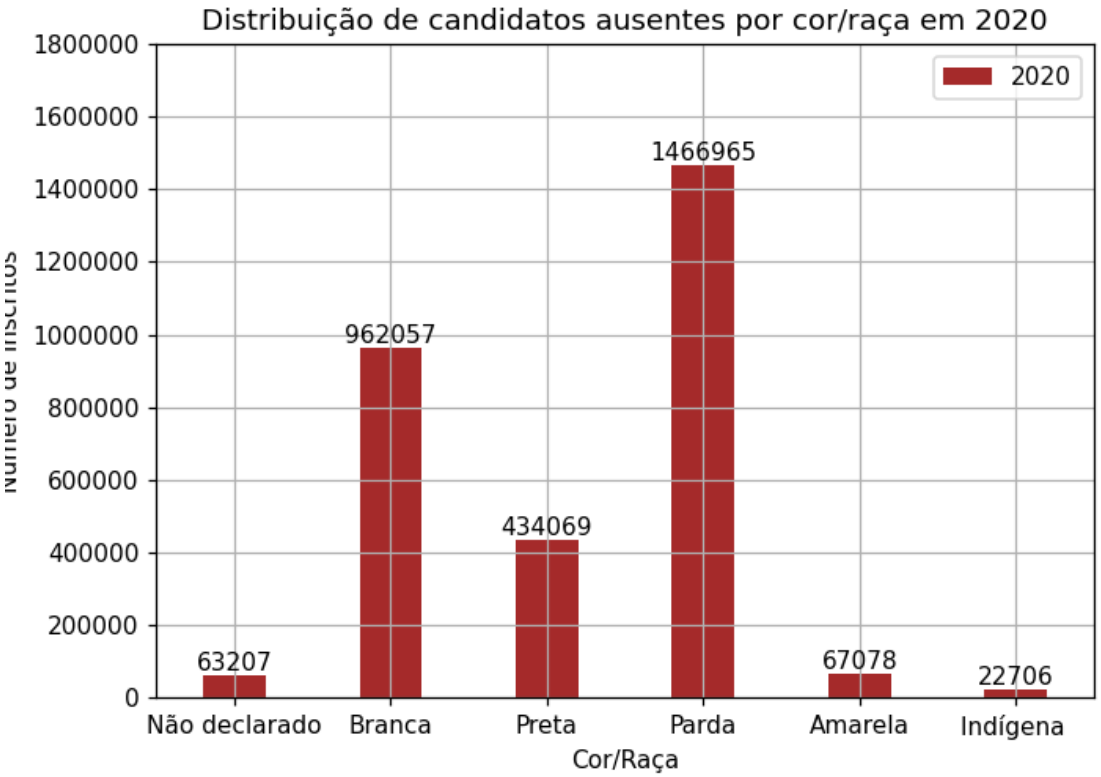

Figura 5.14: Distribuição de indivíduos ausentes nas duas provas por cor/raça em 2020.

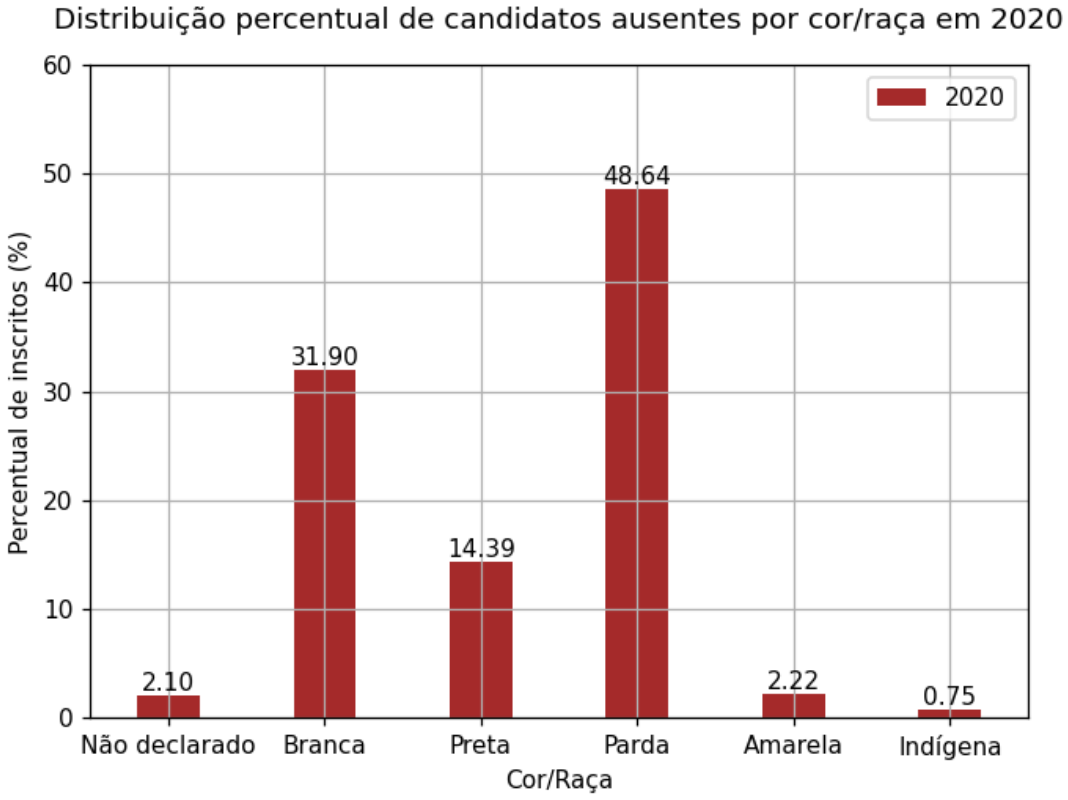

Figura 5.15: Distribuição percentual de indivíduos ausentes nas duas provas por cor/raça em 2020.

Sobre a renda, percebe-se que os inscritos da classe B, com renda até R$ 998,00, foram a maioria e com grande diferença para as demais classes. Eles representam 36,10% dos ausentes, seguidos por 19,85% da classe C, e 13,12% da classe D. Os gráficos da Figura 5.16 e da Figura 5.17 apresentam os números.

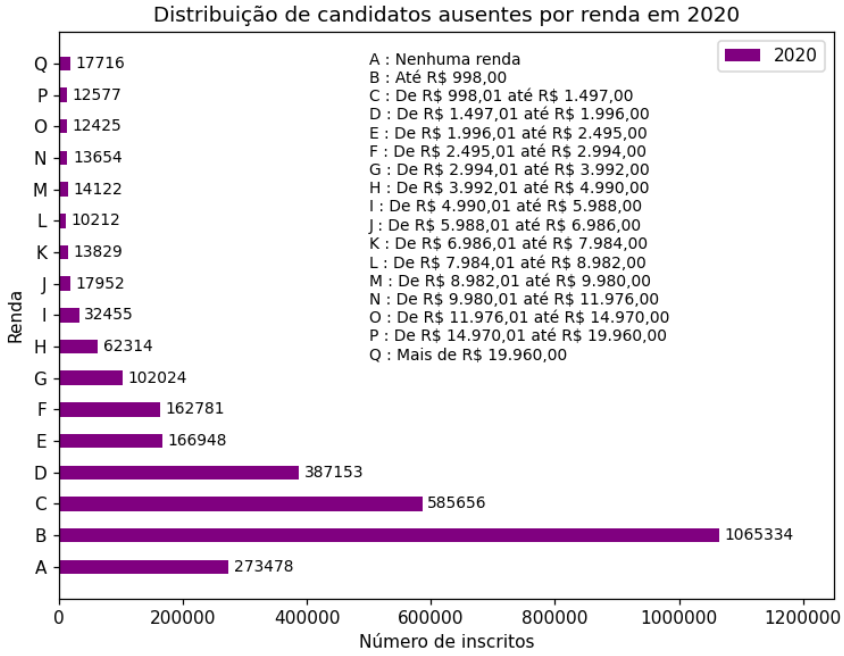

Figura 5.16: Distribuição de indivíduos ausentes nas duas provas por renda em 2020.

Para a escolaridade do pai, tem-se que os inscritos com pais que não completaram a 4ª série/5º ano do ensino fundamental foram a maioria, e representam 27,17% dos ausentes. Interessante observar que o maior número de ausentes foi de uma categoria que não possui o maior número de inscritos. A categoria "E", com pais que completaram o Ensino Médio, mas não completaram a Faculdade, somaram o maior número de inscrições em 2020. Para a escolaridade da mãe, candidatos com mães que completaram o ensino médio, mas não completaram a faculdade, foram o maior número com 29,08%. Os números podem ser vistos na Figura 5.18, para escolaridade do pai, e na Figura 5.20 para escolaridade da mãe. As distribuições percentuais são apresentadas na Figura 5.19 e na Figura 5.21, para escolaridade do pai e da mãe, respectivamente.

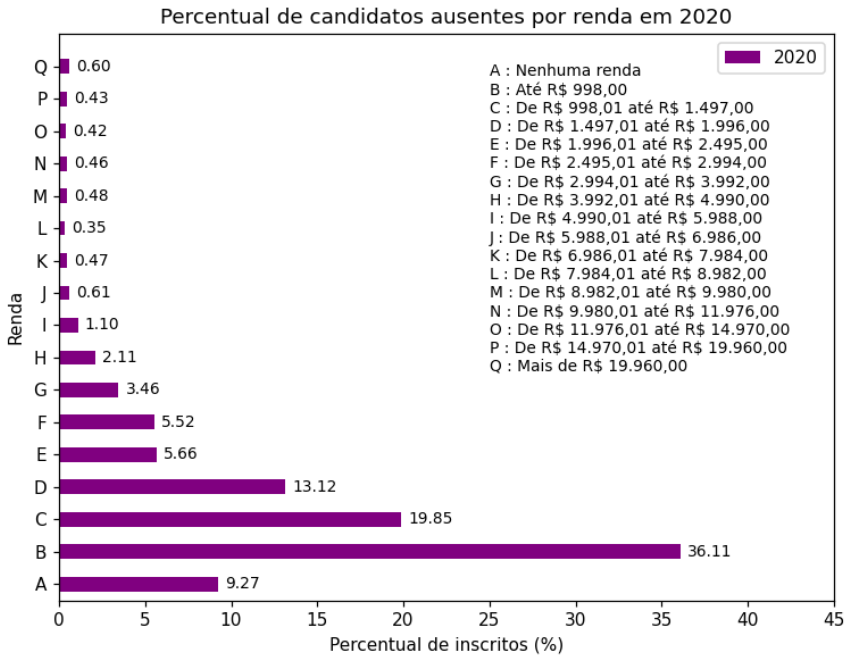

Figura 5.17: Distribuição percentual de indivíduos ausentes nas duas provas por renda em 2020.

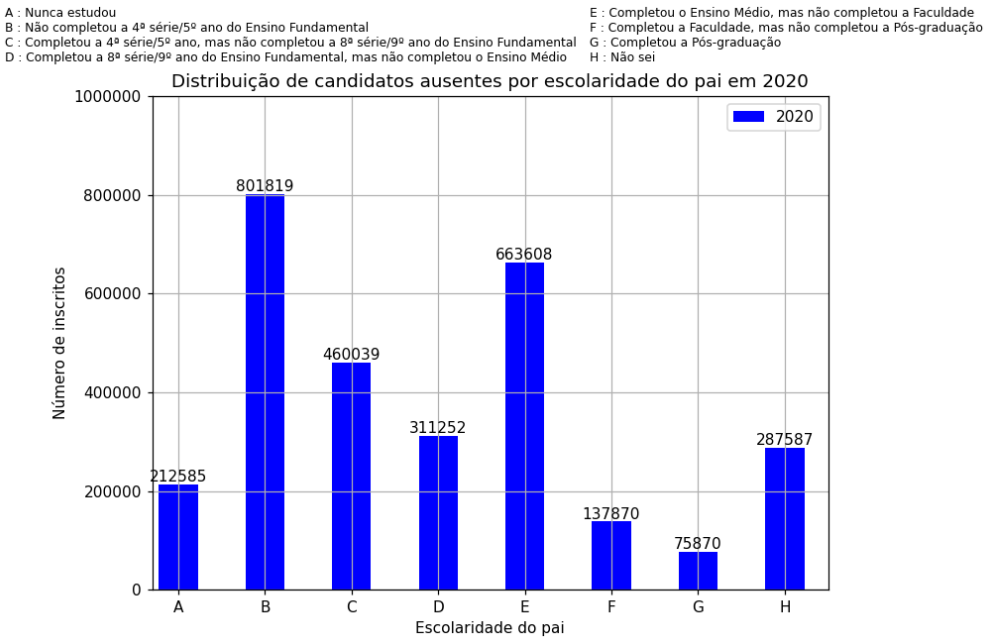

Figura 5.18: Distribuição de indivíduos ausentes nas duas provas por escolaridade do pai em 2020.

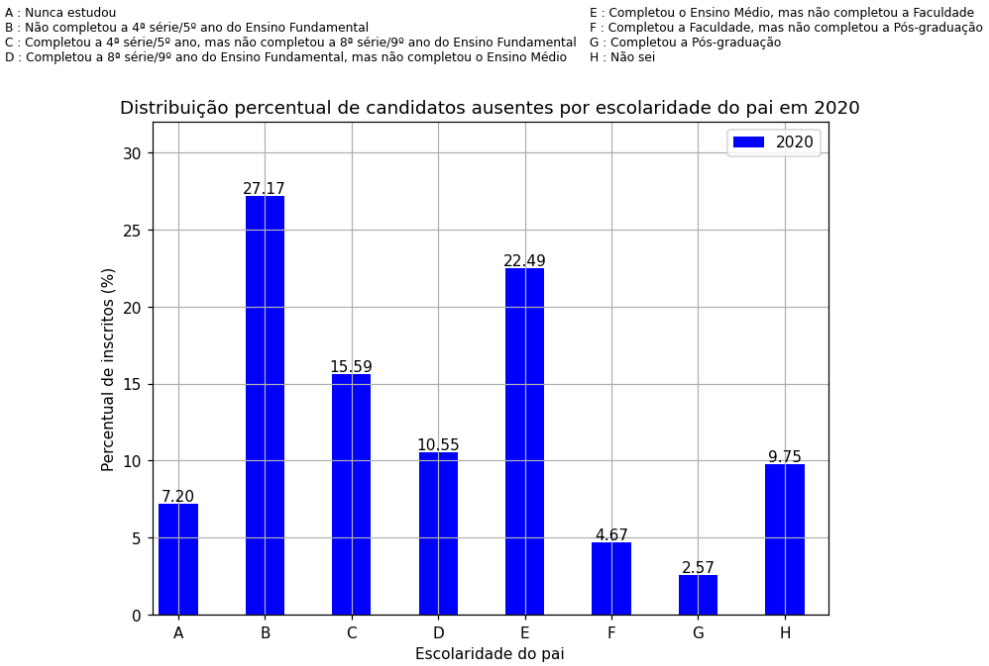

Figura 5.19: Distribuição percentual de indivíduos ausentes nas duas provas por escolaridade do pai em 2020.

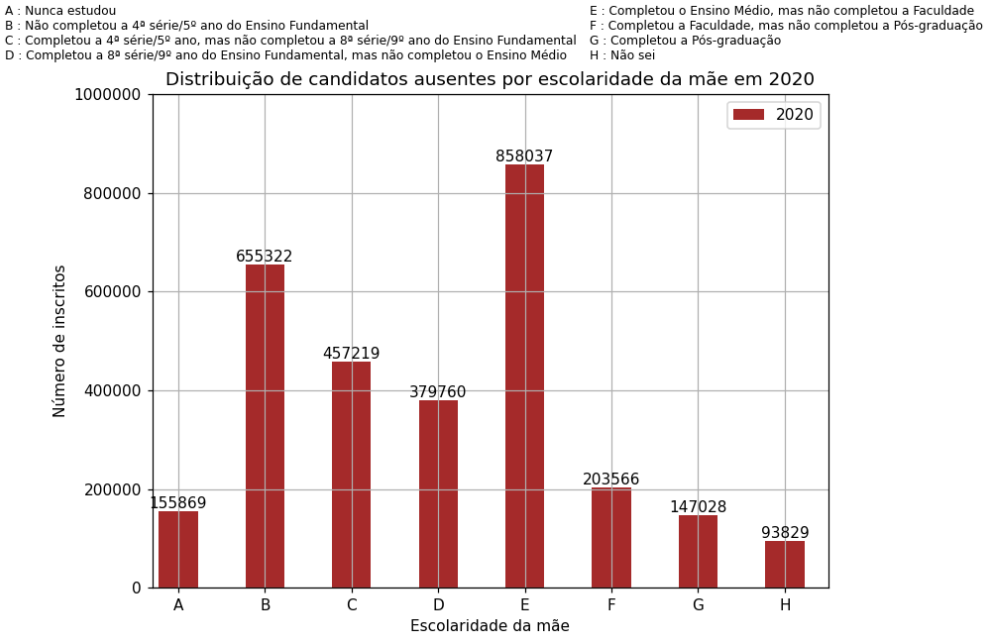

Figura 5.20: Distribuição de indivíduos ausentes nas duas provas por escolaridade da mãe em 2020.

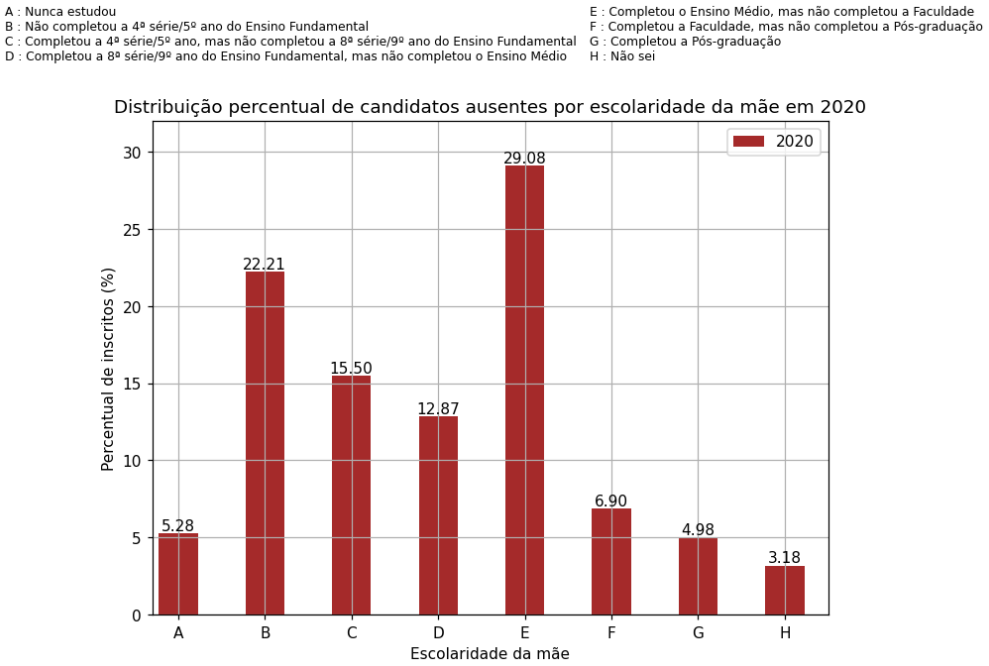

Figura 5.21: Distribuição percentual de indivíduos ausentes nas duas provas por escolaridade da mãe em 2020.

Considerando a contextualização feita no Capítulo 2, decidiu-se incluir também na análise um levantamento sobre o acesso dos inscritos às TIC, em específico o acesso a Internet, celular e computador. É possível observar na Figura 5.22 que 54,85% dos candidatos que faltaram nas duas provas não possuíam computador, e 20,58% não possuía acesso a Internet. Como esperado, de acordo com a pesquisa TIC Educação 2019 [34], a grande maioria possuía celular. A Figura 5.23 apresenta outra visualização dessa distribuição.

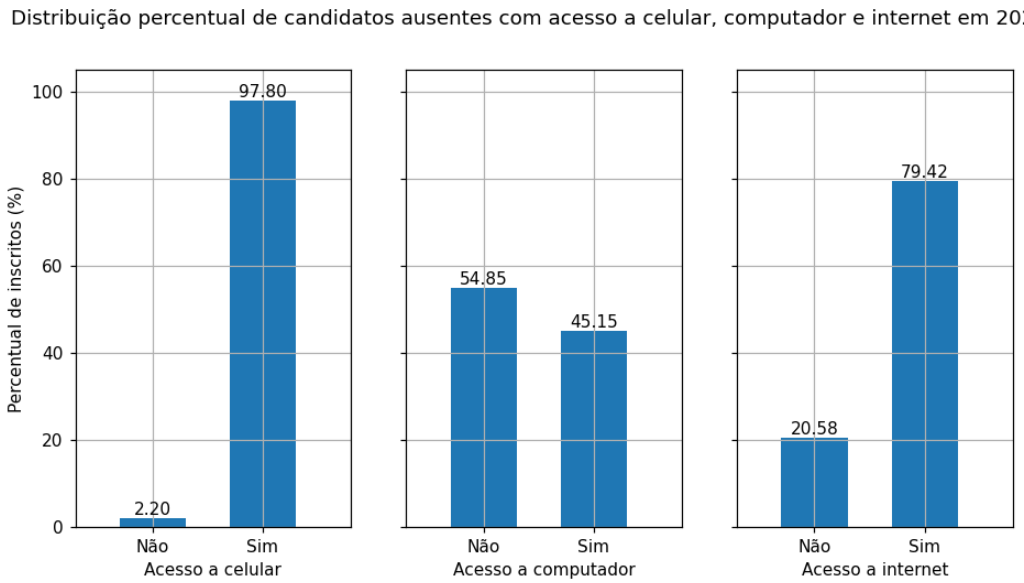

Figura 5.22: Distribuição percentual de indivíduos ausentes nas duas provas por acesso a celular, computador e internet em 2020.

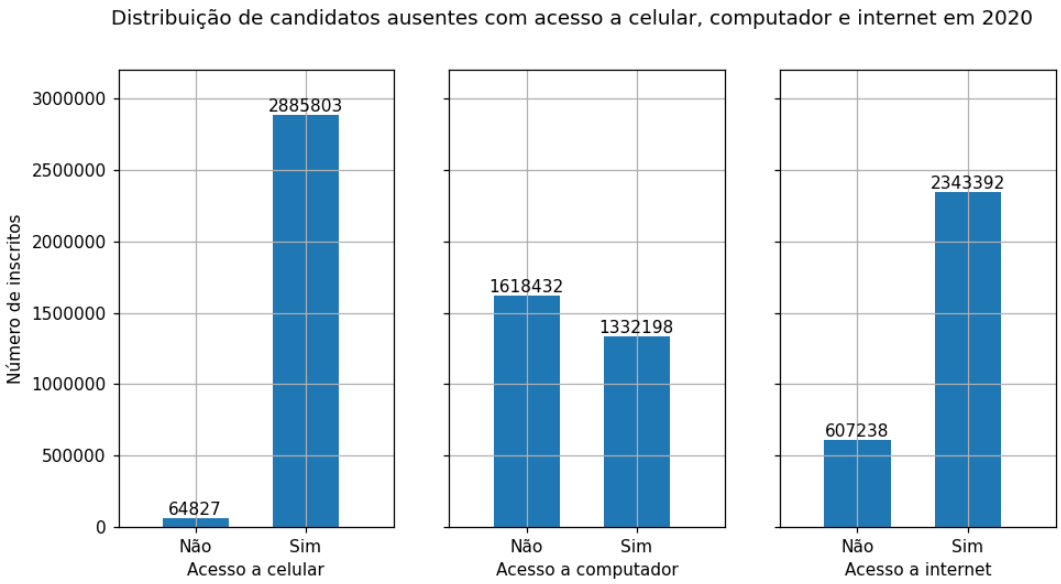

Figura 5.23: Distribuição de indivíduos ausentes nas duas provas por acesso a celular, computador e internet em 2020.

Também foram feitas análises combinando algumas variáveis. Para isso, consideraram-se os participantes que estiveram presentes nos dois dias de provas para os anos mencionados em cada análise. A Figura 5.24 apresenta os resultados de uma comparação das médias das notas entre as categorias de cor/raça, levando em conta o acesso à Internet. É possível observar que independente da cor/raça, se o indivíduo possuí acesso à Internet, a média de suas notas aumenta consideravelmente. Além disso, nota-se que indivíduos de cor/raça Indígena possuem as menores médias, seguido dos de cor/raça Preta. Por outro lado, os indivíduos de cor/raça Branca têm as maiores médias.

Cor/raça	Acesso à Internet	Ciências da Natureza	Ciências Humanas	Linguagens e Códigos	Matemática	Redação
Branca	Não	465.3	486.1	502.1	480.2	543.6
Branca	Sim	517.7	546.5	551.6	564.9	641.2
Preta	Não	450.8	468.4	487.7	455.6	519.0
Preta	Sim	476.1	500.4	518.5	494.9	567.3
Parda	Não	451.1	467.0	483.4	458.5	526.4
Parda	Sim	483.6	506.1	520.1	509.8	585.6
Amarela	Não	453.6	467.7	488.5	460.9	531.6
Amarela	Sim	498.3	517.3	528.7	534.6	604.2
Indígena	Não	437.6	451.6	461.0	438.8	493.5
Indígena	Sim	461.4	477.8	494.8	475.0	535.3

Figura 5.24: Média das notas em 2020 agrupadas por cor/raça e acesso à internet.

5.2 Análises das Notas por Classe de Renda

A renda também mostrou ter influência sobre as médias das notas dos participantes. Na Figura 5.25 e na Figura 5.26 pode-se perceber que as médias das notas das provas têm a tendência de aumentarem conforme maior a faixa de renda do indivíduo. Nota-se que as médias das notas não aumentam de maneira linear, mas que existe de fato influência da renda no desempenho dos candidatos.

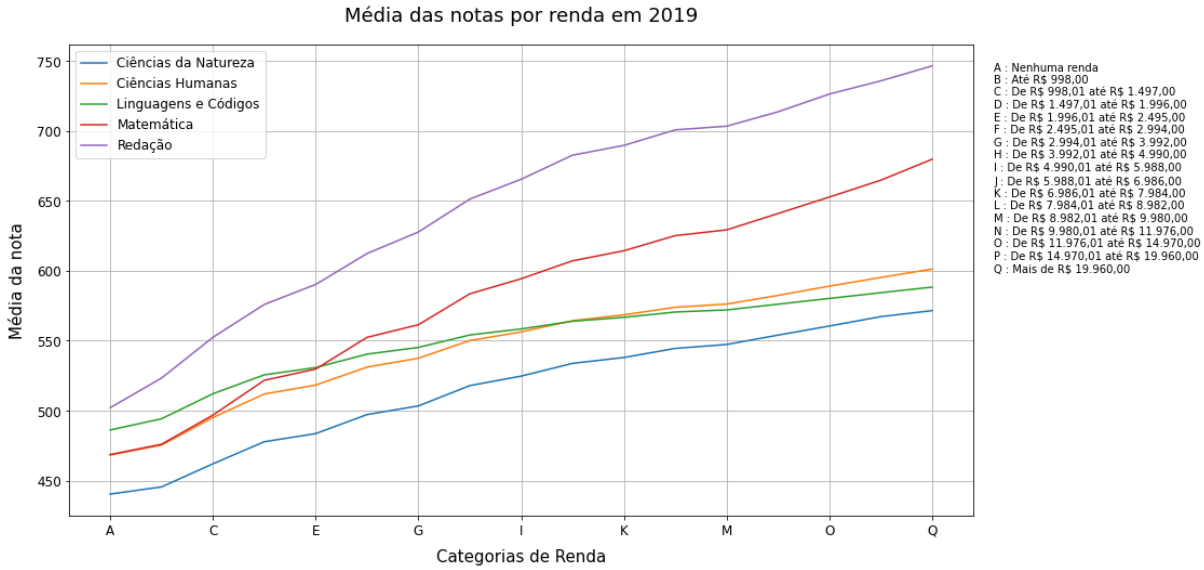

Figura 5.25: Média das notas por renda em 2019.

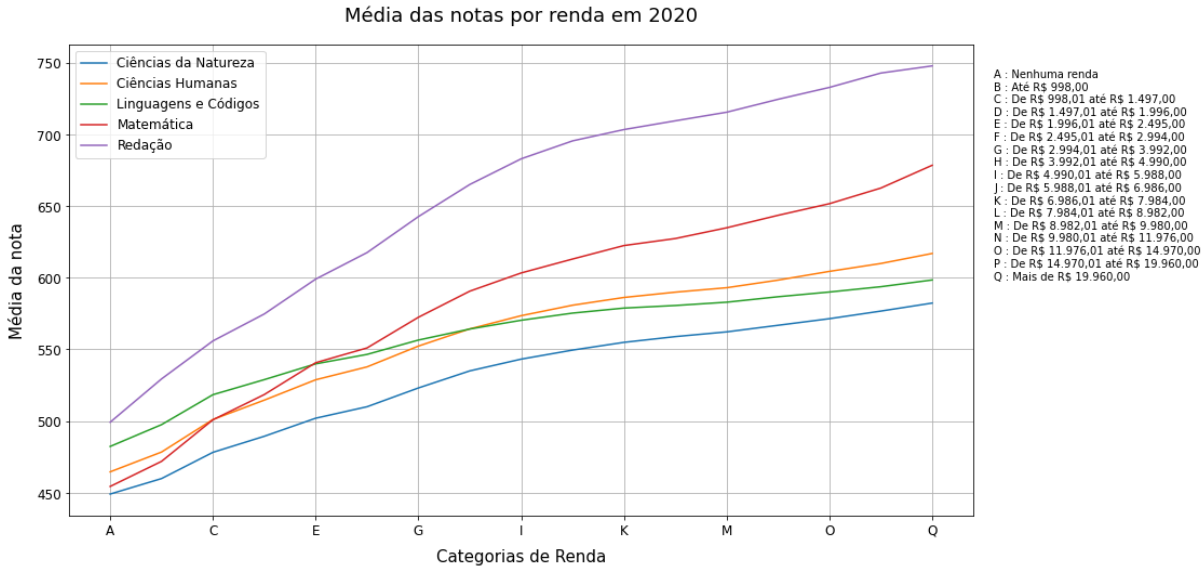

Figura 5.26: Média das notas por renda em 2020.

Nas Figuras 5.27 a 5.31 estão os gráficos com as médias de cada prova por renda nos anos de 2019 e 2020.

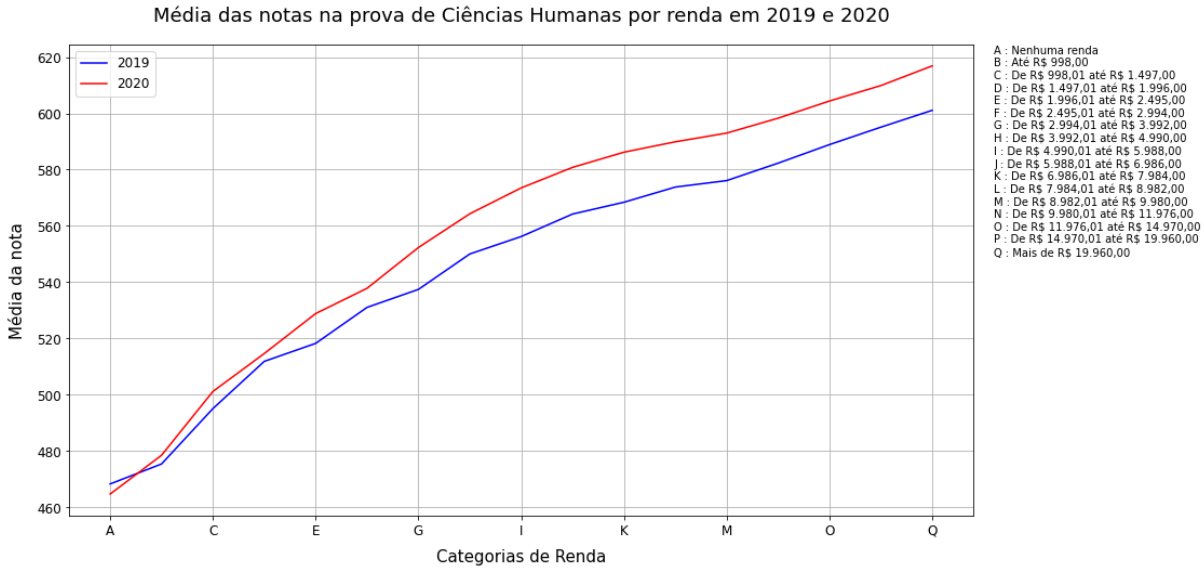

Figura 5.27: Média das notas de Ciências Humanas por renda em 2019 e 2020.

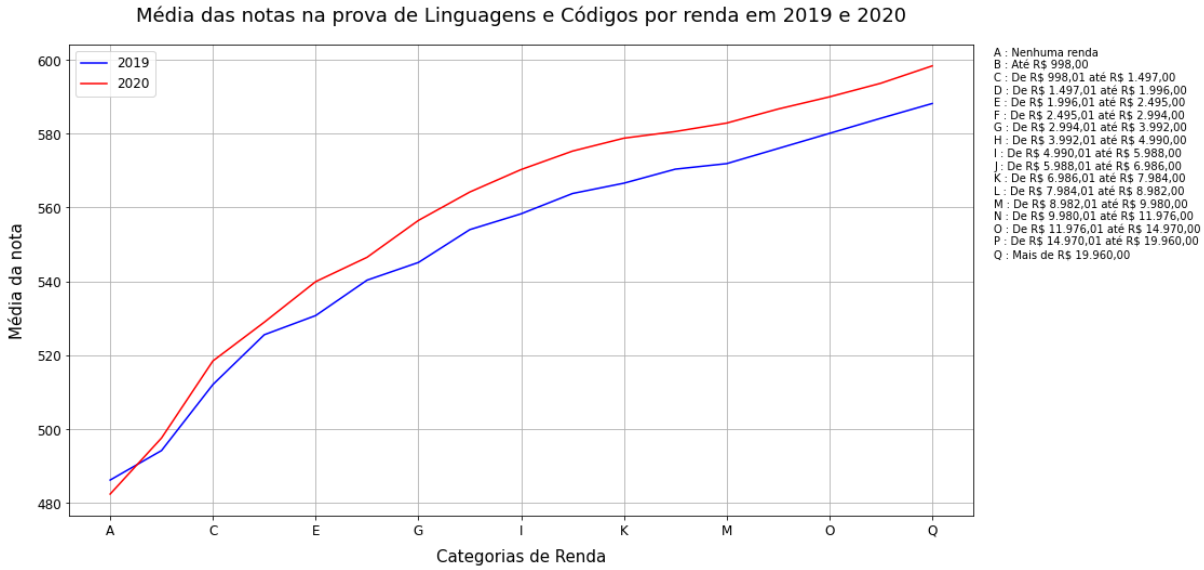

Figura 5.28: Média das notas de Linguagens e Códigos por renda em 2019 e 2020.

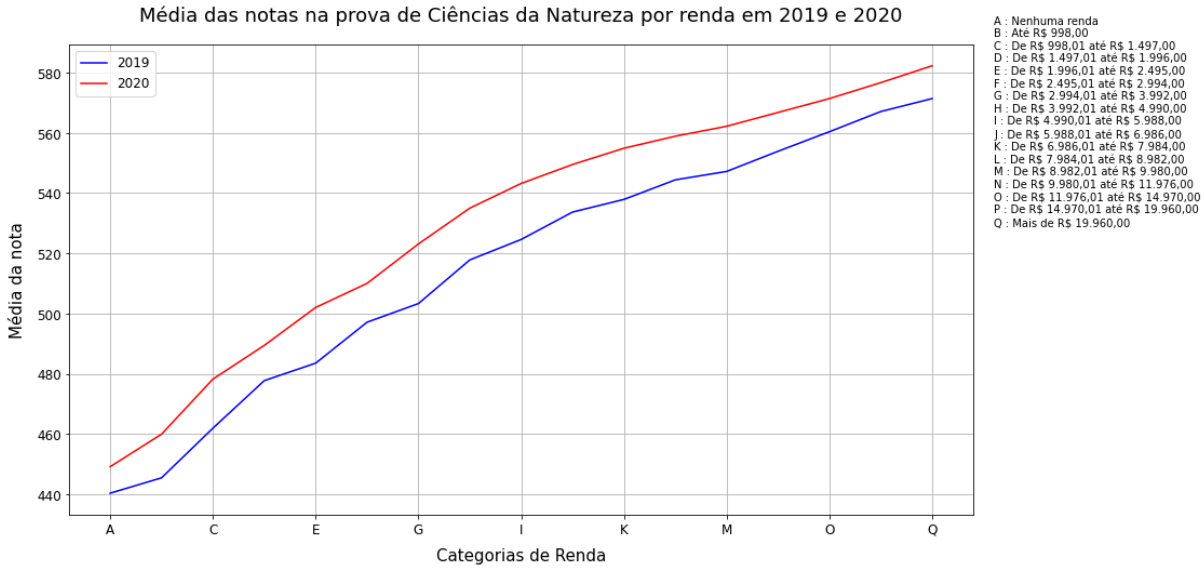

Figura 5.29: Média das notas de Ciências da Natureza por renda em 2019 e 2020.

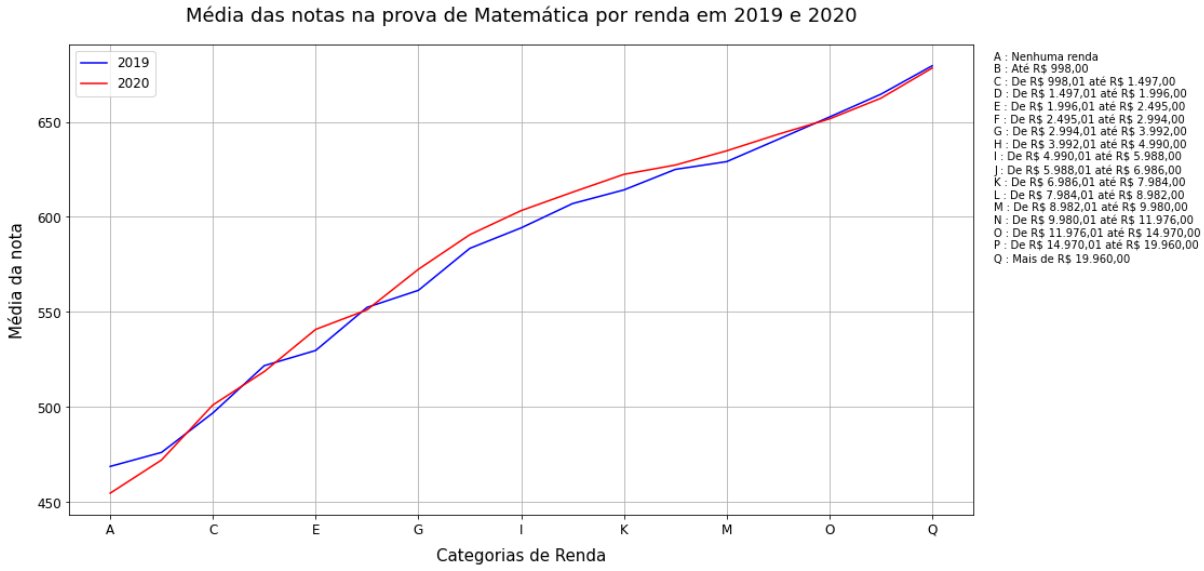

Figura 5.30: Média das notas de Matemática por renda em 2019 e 2020.

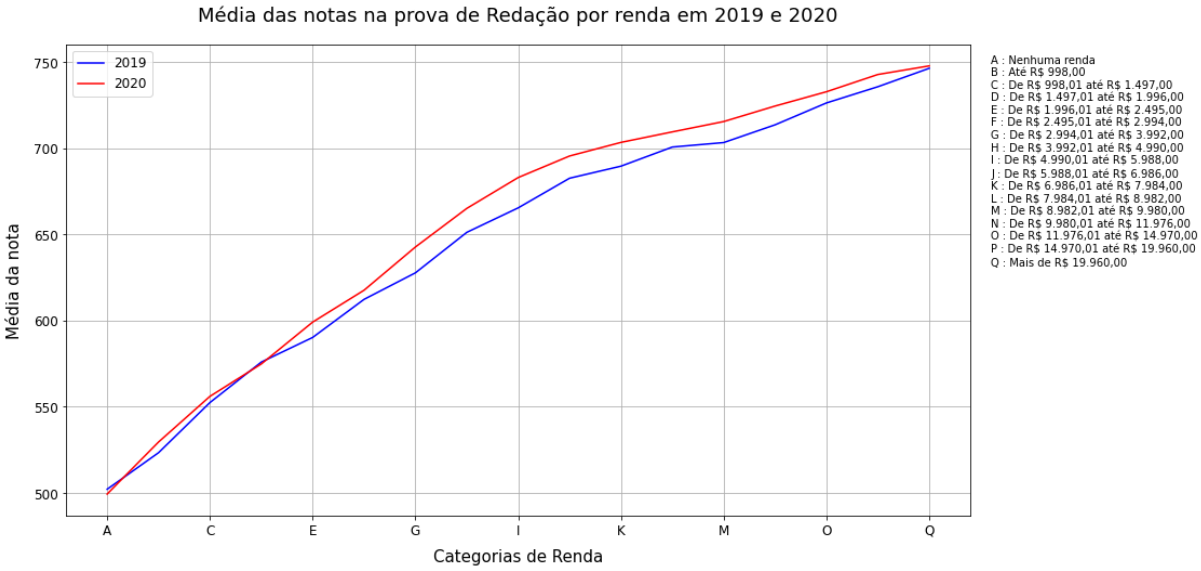

Figura 5.31: Média das notas de Redação por renda em 2019 e 2020.

Nas Figuras 5.32 a 5.35 podem ser vistas tabelas com as diferenças de média das notas entre cada categoria de renda em 2019 e 2020 para as provas de Ciências Humanas, Linguagens e Códigos, Ciências da Natureza e Matemática, respectivamente. Os valores abaixo da diagonal com zeros representam a diferença de média das notas entre uma categoria de renda mais alta para uma mais baixa, e seus valores tendem a ser positivos. Isso significa que as categorias de renda mais altas têm a tendência de possuírem maiores médias de notas, como visto nos gráficos das Figuras 5.27 a 5.30. Os valores acima da diagonal com zeros representam o oposto, a diferença de média das notas entre uma categorias de renda mais baixa para uma mais alta, e possuem em sua maioria valores negativos.

Diferença da média das notas entre as categorias de renda na prova de Ciências Humanas em 2019 e 2020

2019

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-7.11	-26.85	-43.6	-50.02	-62.78	-69.22	-81.83	-88.05	-96.05	-100.19	-105.63	-107.88	-114.11	-120.68	-126.89	-132.9
B	7.11	0.0	-19.74	-36.49	-42.91	-55.67	-62.11	-74.72	-80.94	-88.94	-93.08	-98.52	-100.77	-107.0	-113.57	-119.78	-125.79
C	26.85	19.74	0.0	-16.75	-23.17	-35.93	-42.37	-54.98	-61.2	-69.2	-73.34	-78.78	-81.03	-87.26	-93.83	-100.04	-106.05
D	43.6	36.49	16.75	0.0	-6.42	-19.18	-25.62	-38.23	-44.45	-52.45	-56.59	-62.03	-64.28	-70.51	-77.08	-83.29	-89.3
E	50.02	42.91	23.17	6.42	0.0	-12.76	-19.2	-31.81	-38.03	-46.03	-50.17	-55.61	-57.86	-64.09	-70.66	-76.87	-82.88
F	62.78	55.67	35.93	19.18	12.76	0.0	-6.44	-19.05	-25.27	-33.27	-37.41	-42.85	-45.1	-51.33	-57.9	-64.11	-70.12
G	69.22	62.11	42.37	25.62	19.2	6.44	0.0	-12.61	-18.83	-26.83	-30.97	-36.41	-38.66	-44.89	-51.46	-57.67	-63.68
H	81.83	74.72	54.98	38.23	31.81	19.05	12.61	0.0	-6.22	-14.22	-18.36	-23.8	-26.05	-32.28	-38.85	-45.06	-51.07
I	88.05	80.94	61.2	44.45	38.03	25.27	18.83	6.22	0.0	-8.0	-12.14	-17.58	-19.83	-26.06	-32.63	-38.84	-44.85
J	96.05	88.94	69.2	52.45	46.03	33.27	26.83	14.22	8.0	0.0	-4.14	-9.58	-11.83	-18.06	-24.63	-30.84	-36.85
K	100.19	93.08	73.34	56.59	50.17	37.41	30.97	18.36	12.14	4.14	0.0	-5.44	-7.69	-13.92	-20.49	-26.7	-32.71
L	105.63	98.52	78.78	62.03	55.61	42.85	36.41	23.8	17.58	9.58	5.44	0.0	-2.25	-8.48	-15.05	-21.26	-27.27
M	107.88	100.77	81.03	64.28	57.86	45.1	38.66	26.05	19.83	11.83	7.69	2.25	0.0	-6.23	-12.8	-19.01	-25.02
N	114.11	107.0	87.26	70.51	64.09	51.33	44.89	32.28	26.06	18.06	13.92	8.48	6.23	0.0	-6.57	-12.78	-18.79
O	120.68	113.57	93.83	77.08	70.66	57.9	51.46	38.85	32.63	24.63	20.49	15.05	12.8	6.57	0.0	-6.21	-12.22
P	126.89	119.78	100.04	83.29	76.87	64.11	57.67	45.06	38.84	30.84	26.7	21.26	19.01	12.78	6.21	0.0	-6.01
Q	132.9	125.79	106.05	89.3	82.88	70.12	63.68	51.07	44.85	36.85	32.71	27.27	25.02	18.79	12.22	6.01	0.0

2020

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-13.79	-36.46	-49.96	-64.15	-73.2	-87.69	-99.67	-108.85	-116.23	-121.55	-125.27	-128.35	-133.68	-139.76	-145.29	-152.29
B	13.79	0.0	-22.67	-36.17	-50.36	-59.41	-73.9	-85.88	-95.06	-102.44	-107.76	-111.48	-114.56	-119.89	-125.97	-131.5	-138.5
C	36.46	22.67	0.0	-13.5	-27.69	-36.74	-51.23	-63.21	-72.39	-79.77	-85.09	-88.81	-91.89	-97.22	-103.3	-108.83	-115.83
D	49.96	36.17	13.5	0.0	-14.19	-23.24	-37.73	-49.71	-58.89	-66.27	-71.59	-75.31	-78.39	-83.72	-89.8	-95.33	-102.33
E	64.15	50.36	27.69	14.19	0.0	-9.05	-23.54	-35.52	-44.7	-52.08	-57.4	-61.12	-64.2	-69.53	-75.61	-81.14	-88.14
F	73.2	59.41	36.74	23.24	9.05	0.0	-14.49	-26.47	-35.65	-43.03	-48.35	-52.07	-55.15	-60.48	-66.56	-72.09	-79.09
G	87.69	73.9	51.23	37.73	23.54	14.49	0.0	-11.98	-21.16	-28.54	-33.86	-37.58	-40.66	-45.99	-52.07	-57.6	-64.6
H	99.67	85.88	63.21	49.71	35.52	26.47	11.98	0.0	-9.18	-16.56	-21.88	-25.6	-28.68	-34.01	-40.09	-45.62	-52.62
I	108.85	95.06	72.39	58.89	44.7	35.65	21.16	9.18	0.0	-7.38	-12.7	-16.42	-19.5	-24.83	-30.91	-36.44	-43.44
J	116.23	102.44	79.77	66.27	52.08	43.03	28.54	16.56	7.38	0.0	-5.32	-9.04	-12.12	-17.45	-23.53	-29.06	-36.06
K	121.55	107.76	85.09	71.59	57.4	48.35	33.86	21.88	12.7	5.32	0.0	-3.72	-6.8	-12.13	-18.21	-23.74	-30.74
L	125.27	111.48	88.81	75.31	61.12	52.07	37.58	25.6	16.42	9.04	3.72	0.0	-3.08	-8.41	-14.49	-20.02	-27.02
M	128.35	114.56	91.89	78.39	64.2	55.15	40.66	28.68	19.5	12.12	6.8	3.08	0.0	-5.33	-11.41	-16.94	-23.94
N	133.68	119.89	97.22	83.72	69.53	60.48	45.99	34.01	24.83	17.45	12.13	8.41	5.33	0.0	-6.08	-11.61	-18.61
O	139.76	125.97	103.3	89.8	75.61	66.56	52.07	40.09	30.91	23.53	18.21	14.49	11.41	6.08	0.0	-5.53	-12.53
P	145.29	131.5	108.83	95.33	81.14	72.09	57.6	45.62	36.44	29.06	23.74	20.02	16.94	11.61	5.53	0.0	-7.0
Q	152.29	138.5	115.83	102.33	88.14	79.09	64.6	52.62	43.44	36.06	30.74	27.02	23.94	18.61	12.53	7.0	0.0

A: Nenhuma Renda C: De R$ 1.045,01 até R$ 1.567,50 E: De R$ 2.090,01 até R$ 2.612,50 G: De R$ 3.135,01 até R$ 4.180,00 I: De R$ 5.225,01 até R$ 6.270,00 K: De R$ 7.315,01 até R$ 8.360,00 M: De R$ 9.405,01 até R$ 10.450,00 O: De R$ 12.540,01 até R$ 15.675,00 Q: Acima de R$ 20.900,00
B: Até R$ 1.045,00 D: De R$ 1.567,51 até R$ 2.090,00 F: De R$ 2.612,51 até R$ 3.135,00 H: De R$ 4.180,01 até R$ 5.225,00 J: De R$ 6.270,01 até R$ 7.315,00 L: De R$ 8.360,01 até R$ 9.405,00 N: De R$ 10.450,01 até R$ 12.540,00 P: De R$ 15.675,01 até R$ 20.900,00

Figura 5.32: Diferença de média das notas de Ciências Humanas por renda em 2019 e 2020.

Diferença da média das notas entre as categorias de renda na prova de Linguagens e Códigos em 2019 e 2020

2019

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-8.0	-25.92	-39.48	-44.64	-54.2	-59.01	-67.92	-72.28	-77.78	-80.55	-84.34	-85.85	-89.91	-94.04	-98.14	-102.1
B	8.0	0.0	-17.92	-31.48	-36.64	-46.2	-51.01	-59.92	-64.28	-69.78	-72.55	-76.34	-77.85	-81.91	-86.04	-90.14	-94.1
C	25.92	17.92	0.0	-13.56	-18.72	-28.28	-33.09	-42.0	-46.36	-51.86	-54.63	-58.42	-59.93	-63.99	-68.12	-72.22	-76.18
D	39.48	31.48	13.56	0.0	-5.16	-14.72	-19.53	-28.44	-32.8	-38.3	-41.07	-44.86	-46.37	-50.43	-54.56	-58.66	-62.62
E	44.64	36.64	18.72	5.16	0.0	-9.56	-14.37	-23.28	-27.64	-33.14	-35.91	-39.7	-41.21	-45.27	-49.4	-53.5	-57.46
F	54.2	46.2	28.28	14.72	9.56	0.0	-4.81	-13.72	-18.08	-23.58	-26.35	-30.14	-31.65	-35.71	-39.84	-43.94	-47.9
G	59.01	51.01	33.09	19.53	14.37	4.81	0.0	-8.91	-13.27	-18.77	-21.54	-25.33	-26.84	-30.9	-35.03	-39.13	-43.09
H	67.92	59.92	42.0	28.44	23.28	13.72	8.91	0.0	-4.36	-9.86	-12.63	-16.42	-17.93	-21.99	-26.12	-30.22	-34.18
I	72.28	64.28	46.36	32.8	27.64	18.08	13.27	4.36	0.0	-5.5	-8.27	-12.06	-13.57	-17.63	-21.76	-25.86	-29.82
J	77.78	69.78	51.86	38.3	33.14	23.58	18.77	9.86	5.5	0.0	-2.77	-6.56	-8.07	-12.13	-16.26	-20.36	-24.32
K	80.55	72.55	54.63	41.07	35.91	26.35	21.54	12.63	8.27	2.77	0.0	-3.79	-5.3	-9.36	-13.49	-17.59	-21.55
L	84.34	76.34	58.42	44.86	39.7	30.14	25.33	16.42	12.06	6.56	3.79	0.0	-1.51	-5.57	-9.7	-13.8	-17.76
M	85.85	77.85	59.93	46.37	41.21	31.65	26.84	17.93	13.57	8.07	5.3	1.51	0.0	-4.06	-8.19	-12.29	-16.25
N	89.91	81.91	63.99	50.43	45.27	35.71	30.9	21.99	17.63	12.13	9.36	5.57	4.06	0.0	-4.13	-8.23	-12.19
O	94.04	86.04	68.12	54.56	49.4	39.84	35.03	26.12	21.76	16.26	13.49	9.7	8.19	4.13	0.0	-4.1	-8.06
P	98.14	90.14	72.22	58.66	53.5	43.94	39.13	30.22	25.86	20.36	17.59	13.8	12.29	8.23	4.1	0.0	-3.96
Q	102.1	94.1	76.18	62.62	57.46	47.9	43.09	34.18	29.82	24.32	21.55	17.76	16.25	12.19	8.06	3.96	0.0

2020

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-15.21	-36.14	-46.62	-57.66	-64.22	-74.18	-81.92	-88.02	-93.03	-96.51	-98.33	-100.61	-104.38	-107.73	-111.37	-116.08
B	15.21	0.0	-20.93	-31.41	-42.45	-49.01	-58.97	-66.71	-72.81	-77.82	-81.3	-83.12	-85.4	-89.17	-92.52	-96.16	-100.87
C	36.14	20.93	0.0	-10.48	-21.52	-28.08	-38.04	-45.78	-51.88	-56.89	-60.37	-62.19	-64.47	-68.24	-71.59	-75.23	-79.94
D	46.62	31.41	10.48	0.0	-11.04	-17.6	-27.56	-35.3	-41.4	-46.41	-49.89	-51.71	-53.99	-57.76	-61.11	-64.75	-69.46
E	57.66	42.45	21.52	11.04	0.0	-6.56	-16.52	-24.26	-30.36	-35.37	-38.85	-40.67	-42.95	-46.72	-50.07	-53.71	-58.42
F	64.22	49.01	28.08	17.6	6.56	0.0	-9.96	-17.7	-23.8	-28.81	-32.29	-34.11	-36.39	-40.16	-43.51	-47.15	-51.86
G	74.18	58.97	38.04	27.56	16.52	9.96	0.0	-7.74	-13.84	-18.85	-22.33	-24.15	-26.43	-30.2	-33.55	-37.19	-41.9
H	81.92	66.71	45.78	35.3	24.26	17.7	7.74	0.0	-6.1	-11.11	-14.59	-16.41	-18.69	-22.46	-25.81	-29.45	-34.16
I	88.02	72.81	51.88	41.4	30.36	23.8	13.84	6.1	0.0	-5.01	-8.49	-10.31	-12.59	-16.36	-19.71	-23.35	-28.06
J	93.03	77.82	56.89	46.41	35.37	28.81	18.85	11.11	5.01	0.0	-3.48	-5.3	-7.58	-11.35	-14.7	-18.34	-23.05
K	96.51	81.3	60.37	49.89	38.85	32.29	22.33	14.59	8.49	3.48	0.0	-1.82	-4.1	-7.87	-11.22	-14.86	-19.57
L	98.33	83.12	62.19	51.71	40.67	34.11	24.15	16.41	10.31	5.3	1.82	0.0	-2.28	-6.05	-9.4	-13.04	-17.75
M	100.61	85.4	64.47	53.99	42.95	36.39	26.43	18.69	12.59	7.58	4.1	2.28	0.0	-3.77	-7.12	-10.76	-15.47
N	104.38	89.17	68.24	57.76	46.72	40.16	30.2	22.46	16.36	11.35	7.87	6.05	3.77	0.0	-3.35	-6.99	-11.7
O	107.73	92.52	71.59	61.11	50.07	43.51	33.55	25.81	19.71	14.7	11.22	9.4	7.12	3.35	0.0	-3.64	-8.35
P	111.37	96.16	75.23	64.75	53.71	47.15	37.19	29.45	23.35	18.34	14.86	13.04	10.76	6.99	3.64	0.0	-4.71
Q	116.08	100.87	79.94	69.46	58.42	51.86	41.9	34.16	28.06	23.05	19.57	17.75	15.47	11.7	8.35	4.71	0.0

A: Nenhuma Renda C: De R$ 1.045,01 até R$ 1.567,50 E: De R$ 2.090,01 até R$ 2.612,50 G: De R$ 3.135,01 até R$ 4.180,00 I: De R$ 5.225,01 até R$ 6.270,00 K: De R$ 7.315,01 até R$ 8.360,00 M: De R$ 9.405,01 até R$ 10.450,00 O: De R$ 12.540,01 até R$ 15.675,00 Q: Acima de R$ 20.900,00
B: Até R$ 1.045,00 D: De R$ 1.567,51 até R$ 2.090,00 F: De R$ 2.612,51 até R$ 3.135,00 H: De R$ 4.180,01 até R$ 5.225,00 J: De R$ 6.270,01 até R$ 7.315,00 L: De R$ 8.360,01 até R$ 9.405,00 N: De R$ 10.450,01 até R$ 12.540,00 P: De R$ 15.675,01 até R$ 20.900,00

Figura 5.33: Diferença de média das notas de Linguagens e Códigos por renda em 2019 e 2020.

Diferença da média das notas entre as categorias de renda na prova de Ciências da Natureza em 2019 e 2020

2019

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-5.07	-21.59	-37.37	-43.16	-56.81	-63.01	-77.51	-84.26	-93.36	-97.58	-104.07	-106.86	-113.58	-120.11	-126.74	-131.1
B	5.07	0.0	-16.52	-32.3	-38.09	-51.74	-57.94	-72.44	-79.19	-88.29	-92.51	-99.0	-101.79	-108.51	-115.04	-121.67	-126.03
C	21.59	16.52	0.0	-15.78	-21.57	-35.22	-41.42	-55.92	-62.67	-71.77	-75.99	-82.48	-85.27	-91.99	-98.52	-105.15	-109.51
D	37.37	32.3	15.78	0.0	-5.79	-19.44	-25.64	-40.14	-46.89	-55.99	-60.21	-66.7	-69.49	-76.21	-82.74	-89.37	-93.73
E	43.16	38.09	21.57	5.79	0.0	-13.65	-19.85	-34.35	-41.1	-50.2	-54.42	-60.91	-63.7	-70.42	-76.95	-83.58	-87.94
F	56.81	51.74	35.22	19.44	13.65	0.0	-6.2	-20.7	-27.45	-36.55	-40.77	-47.26	-50.05	-56.77	-63.3	-69.93	-74.29
G	63.01	57.94	41.42	25.64	19.85	6.2	0.0	-14.5	-21.25	-30.35	-34.57	-41.06	-43.85	-50.57	-57.1	-63.73	-68.09
H	77.51	72.44	55.92	40.14	34.35	20.7	14.5	0.0	-6.75	-15.85	-20.07	-26.56	-29.35	-36.07	-42.6	-49.23	-53.59
I	84.26	79.19	62.67	46.89	41.1	27.45	21.25	6.75	0.0	-9.1	-13.32	-19.81	-22.6	-29.32	-35.85	-42.48	-46.84
J	93.36	88.29	71.77	55.99	50.2	36.55	30.35	15.85	9.1	0.0	-4.22	-10.71	-13.5	-20.22	-26.75	-33.38	-37.74
K	97.58	92.51	75.99	60.21	54.42	40.77	34.57	20.07	13.32	4.22	0.0	-6.49	-9.28	-16.0	-22.53	-29.16	-33.52
L	104.07	99.0	82.48	66.7	60.91	47.26	41.06	26.56	19.81	10.71	6.49	0.0	-2.79	-9.51	-16.04	-22.67	-27.03
M	106.86	101.79	85.27	69.49	63.7	50.05	43.85	29.35	22.6	13.5	9.28	2.79	0.0	-6.72	-13.25	-19.88	-24.24
N	113.58	108.51	91.99	76.21	70.42	56.77	50.57	36.07	29.32	20.22	16.0	9.51	6.72	0.0	-6.53	-13.16	-17.52
O	120.11	115.04	98.52	82.74	76.95	63.3	57.1	42.6	35.85	26.75	22.53	16.04	13.25	6.53	0.0	-6.63	-10.99
P	126.74	121.67	105.15	89.37	83.58	69.93	63.73	49.23	42.48	33.38	29.16	22.67	19.88	13.16	6.63	0.0	-4.36
Q	131.1	126.03	109.51	93.73	87.94	74.29	68.09	53.59	46.84	37.74	33.52	27.03	24.24	17.52	10.99	4.36	0.0

2020

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-10.79	-29.16	-40.33	-52.91	-60.94	-74.06	-85.94	-94.13	-100.42	-105.86	-109.8	-113.15	-117.69	-122.33	-127.63	-133.2
B	10.79	0.0	-18.37	-29.54	-42.12	-50.15	-63.27	-75.15	-83.34	-89.63	-95.07	-99.01	-102.36	-106.9	-111.54	-116.84	-122.41
C	29.16	18.37	0.0	-11.17	-23.75	-31.78	-44.9	-56.78	-64.97	-71.26	-76.7	-80.64	-83.99	-88.53	-93.17	-98.47	-104.04
D	40.33	29.54	11.17	0.0	-12.58	-20.61	-33.73	-45.61	-53.8	-60.09	-65.53	-69.47	-72.82	-77.36	-82.0	-87.3	-92.87
E	52.91	42.12	23.75	12.58	0.0	-8.03	-21.15	-33.03	-41.22	-47.51	-52.95	-56.89	-60.24	-64.78	-69.42	-74.72	-80.29
F	60.94	50.15	31.78	20.61	8.03	0.0	-13.12	-25.0	-33.19	-39.48	-44.92	-48.86	-52.21	-56.75	-61.39	-66.69	-72.26
G	74.06	63.27	44.9	33.73	21.15	13.12	0.0	-11.88	-20.07	-26.36	-31.8	-35.74	-39.09	-43.63	-48.27	-53.57	-59.14
H	85.94	75.15	56.78	45.61	33.03	25.0	11.88	0.0	-8.19	-14.48	-19.92	-23.86	-27.21	-31.75	-36.39	-41.69	-47.26
I	94.13	83.34	64.97	53.8	41.22	33.19	20.07	8.19	0.0	-6.29	-11.73	-15.67	-19.02	-23.56	-28.2	-33.5	-39.07
J	100.42	89.63	71.26	60.09	47.51	39.48	26.36	14.48	6.29	0.0	-5.44	-9.38	-12.73	-17.27	-21.91	-27.21	-32.78
K	105.86	95.07	76.7	65.53	52.95	44.92	31.8	19.92	11.73	5.44	0.0	-3.94	-7.29	-11.83	-16.47	-21.77	-27.34
L	109.8	99.01	80.64	69.47	56.89	48.86	35.74	23.86	15.67	9.38	3.94	0.0	-3.35	-7.89	-12.53	-17.83	-23.4
M	113.15	102.36	83.99	72.82	60.24	52.21	39.09	27.21	19.02	12.73	7.29	3.35	0.0	-4.54	-9.18	-14.48	-20.05
N	117.69	106.9	88.53	77.36	64.78	56.75	43.63	31.75	23.56	17.27	11.83	7.89	4.54	0.0	-4.64	-9.94	-15.51
O	122.33	111.54	93.17	82.0	69.42	61.39	48.27	36.39	28.2	21.91	16.47	12.53	9.18	4.64	0.0	-5.3	-10.87
P	127.63	116.84	98.47	87.3	74.72	66.69	53.57	41.69	33.5	27.21	21.77	17.83	14.48	9.94	5.3	0.0	-5.57
Q	133.2	122.41	104.04	92.87	80.29	72.26	59.14	47.26	39.07	32.78	27.34	23.4	20.05	15.51	10.87	5.57	0.0

A: Nenhuma Renda C: De R$ 1.045,01 até R$ 1.567,50 E: De R$ 2.090,01 até R$ 2.612,50 G: De R$ 3.135,01 até R$ 4.180,00 I: De R$ 5.225,01 até R$ 6.270,00 K: De R$ 7.315,01 até R$ 8.360,00 M: De R$ 9.405,01 até R$ 10.450,00 O: De R$ 12.540,01 até R$ 15.675,00 Q: Acima de R$ 20.900,00
B: Até R$ 1.045,00 D: De R$ 1.567,51 até R$ 2.090,00 F: De R$ 2.612,51 até R$ 3.135,00 H: De R$ 4.180,01 até R$ 5.225,00 J: De R$ 6.270,01 até R$ 7.315,00 L: De R$ 8.360,01 até R$ 9.405,00 N: De R$ 10.450,01 até R$ 12.540,00 P: De R$ 15.675,01 até R$ 20.900,00

Figura 5.34: Diferença de média das notas de Ciências da Natureza por renda em 2019 e 2020.

Diferença da média das notas entre as categorias de renda na prova de Matemática em 2019 e 2020

2019

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-7.38	-28.22	-53.1	-61.13	-83.8	-92.79	-114.85	-125.7	-138.53	-145.65	-156.47	-160.59	-172.33	-184.1	-196.11	-211.09
B	7.38	0.0	-20.84	-45.72	-53.75	-76.42	-85.41	-107.47	-118.32	-131.15	-138.27	-149.09	-153.21	-164.95	-176.72	-188.73	-203.71
C	28.22	20.84	0.0	-24.88	-32.91	-55.58	-64.57	-86.63	-97.48	-110.31	-117.43	-128.25	-132.37	-144.11	-155.88	-167.89	-182.87
D	53.1	45.72	24.88	0.0	-8.03	-30.7	-39.69	-61.75	-72.6	-85.43	-92.55	-103.37	-107.49	-119.23	-131.0	-143.01	-157.99
E	61.13	53.75	32.91	8.03	0.0	-22.67	-31.66	-53.72	-64.57	-77.4	-84.52	-95.34	-99.46	-111.2	-122.97	-134.98	-149.96
F	83.8	76.42	55.58	30.7	22.67	0.0	-8.99	-31.05	-41.9	-54.73	-61.85	-72.67	-76.79	-88.53	-100.3	-112.31	-127.29
G	92.79	85.41	64.57	39.69	31.66	8.99	0.0	-22.06	-32.91	-45.74	-52.86	-63.68	-67.8	-79.54	-91.31	-103.32	-118.3
H	114.85	107.47	86.63	61.75	53.72	31.05	22.06	0.0	-10.85	-23.68	-30.8	-41.62	-45.74	-57.48	-69.25	-81.26	-96.24
I	125.7	118.32	97.48	72.6	64.57	41.9	32.91	10.85	0.0	-12.83	-19.95	-30.77	-34.89	-46.63	-58.4	-70.41	-85.39
J	138.53	131.15	110.31	85.43	77.4	54.73	45.74	23.68	12.83	0.0	-7.12	-17.94	-22.06	-33.8	-45.57	-57.58	-72.56
K	145.65	138.27	117.43	92.55	84.52	61.85	52.86	30.8	19.95	7.12	0.0	-10.82	-14.94	-26.68	-38.45	-50.46	-65.44
L	156.47	149.09	128.25	103.37	95.34	72.67	63.68	41.62	30.77	17.94	10.82	0.0	-4.12	-15.86	-27.63	-39.64	-54.62
M	160.59	153.21	132.37	107.49	99.46	76.79	67.8	45.74	34.89	22.06	14.94	4.12	0.0	-11.74	-23.51	-35.52	-50.5
N	172.33	164.95	144.11	119.23	111.2	88.53	79.54	57.48	46.63	33.8	26.68	15.86	11.74	0.0	-11.77	-23.78	-38.76
O	184.1	176.72	155.88	131.0	122.97	100.3	91.31	69.25	58.4	45.57	38.45	27.63	23.51	11.77	0.0	-12.01	-26.99
P	196.11	188.73	167.89	143.01	134.98	112.31	103.32	81.26	70.41	57.58	50.46	39.64	35.52	23.78	12.01	0.0	-14.98
Q	211.09	203.71	182.87	157.99	149.96	127.29	118.3	96.24	85.39	72.56	65.44	54.62	50.5	38.76	26.99	14.98	0.0

2020

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-17.5	-46.56	-64.19	-86.34	-96.62	-118.07	-136.22	-148.92	-158.65	-168.03	-172.94	-180.44	-189.19	-197.27	-208.17	-224.08
B	17.5	0.0	-29.06	-46.69	-68.84	-79.12	-100.57	-118.72	-131.42	-141.15	-150.53	-155.44	-162.94	-171.69	-179.77	-190.67	-206.58
C	46.56	29.06	0.0	-17.63	-39.78	-50.06	-71.51	-89.66	-102.36	-112.09	-121.47	-126.38	-133.88	-142.63	-150.71	-161.61	-177.52
D	64.19	46.69	17.63	0.0	-22.15	-32.43	-53.88	-72.03	-84.73	-94.46	-103.84	-108.75	-116.25	-125.0	-133.08	-143.98	-159.89
E	86.34	68.84	39.78	22.15	0.0	-10.28	-31.73	-49.88	-62.58	-72.31	-81.69	-86.6	-94.1	-102.85	-110.93	-121.83	-137.74
F	96.62	79.12	50.06	32.43	10.28	0.0	-21.45	-39.6	-52.3	-62.03	-71.41	-76.32	-83.82	-92.57	-100.65	-111.55	-127.46
G	118.07	100.57	71.51	53.88	31.73	21.45	0.0	-18.15	-30.85	-40.58	-49.96	-54.87	-62.37	-71.12	-79.2	-90.1	-106.01
H	136.22	118.72	89.66	72.03	49.88	39.6	18.15	0.0	-12.7	-22.43	-31.81	-36.72	-44.22	-52.97	-61.05	-71.95	-87.86
I	148.92	131.42	102.36	84.73	62.58	52.3	30.85	12.7	0.0	-9.73	-19.11	-24.02	-31.52	-40.27	-48.35	-59.25	-75.16
J	158.65	141.15	112.09	94.46	72.31	62.03	40.58	22.43	9.73	0.0	-9.38	-14.29	-21.79	-30.54	-38.62	-49.52	-65.43
K	168.03	150.53	121.47	103.84	81.69	71.41	49.96	31.81	19.11	9.38	0.0	-4.91	-12.41	-21.16	-29.24	-40.14	-56.05
L	172.94	155.44	126.38	108.75	86.6	76.32	54.87	36.72	24.02	14.29	4.91	0.0	-7.5	-16.25	-24.33	-35.23	-51.14
M	180.44	162.94	133.88	116.25	94.1	83.82	62.37	44.22	31.52	21.79	12.41	7.5	0.0	-8.75	-16.83	-27.73	-43.64
N	189.19	171.69	142.63	125.0	102.85	92.57	71.12	52.97	40.27	30.54	21.16	16.25	8.75	0.0	-8.08	-18.98	-34.89
O	197.27	179.77	150.71	133.08	110.93	100.65	79.2	61.05	48.35	38.62	29.24	24.33	16.83	8.08	0.0	-10.9	-26.81
P	208.17	190.67	161.61	143.98	121.83	111.55	90.1	71.95	59.25	49.52	40.14	35.23	27.73	18.98	10.9	0.0	-15.91
Q	224.08	206.58	177.52	159.89	137.74	127.46	106.01	87.86	75.16	65.43	56.05	51.14	43.64	34.89	26.81	15.91	0.0

A: Nenhuma Renda C: De R$ 1.045,01 até R$ 1.567,50 E: De R$ 2.090,01 até R$ 2.612,50 G: De R$ 3.135,01 até R$ 4.180,00 I: De R$ 5.225,01 até R$ 6.270,00 K: De R$ 7.315,01 até R$ 8.360,00 M: De R$ 9.405,01 até R$ 10.450,00 O: De R$ 12.540,01 até R$ 15.675,00 Q: Acima de R$ 20.900,00
B: Até R$ 1.045,00 D: De R$ 1.567,51 até R$ 2.090,00 F: De R$ 2.612,51 até R$ 3.135,00 H: De R$ 4.180,01 até R$ 5.225,00 J: De R$ 6.270,01 até R$ 7.315,00 L: De R$ 8.360,01 até R$ 9.405,00 N: De R$ 10.450,01 até R$ 12.540,00 P: De R$ 15.675,01 até R$ 20.900,00

Figura 5.35: Diferença de média das notas de Matemática por renda em 2019 e 2020.

É possível observar também nas Figuras 5.32 a 5.35 que, em geral, houve um aumento nas diferenças entre as notas das classes de renda mais altas e mais baixas. Para ilustrar isto, foi feita a Figura 5.36. Nela pode-se observar que a diferença nas notas da prova de Matemática entre a classe que não possui nenhuma renda (classe "A") e a classe com maior renda (classe "Q") aumentou 12,99 pontos em 2020, com relação a 2019. Também percebe-se que a variação das notas para as classes A e B em relação às classes mais altas, são na sua grande maioria negativas. Isso pode ser um indicativo de que a pandemia de COVID-19 prejudicou mais o desempenho das classes de renda mais baixas que o das mais altas.

Diferença entre as variações das notas de Matemática por categoria de renda em 2020 com relação a 2019

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-10.12	-18.34	-11.09	-25.21	-12.82	-25.28	-21.37	-23.22	-20.12	-22.38	-16.47	-19.85	-16.86	-13.17	-12.06	-12.99
B	10.12	0.0	-8.22	-0.97	-15.09	-2.7	-15.16	-11.25	-13.1	-10.0	-12.26	-6.35	-9.73	-6.74	-3.05	-1.94	-2.87
C	18.34	8.22	0.0	7.25	-6.87	5.52	-6.94	-3.03	-4.88	-1.78	-4.04	1.87	-1.51	1.48	5.17	6.28	5.35
D	11.09	0.97	-7.25	0.0	-14.12	-1.73	-14.19	-10.28	-12.13	-9.03	-11.29	-5.38	-8.76	-5.77	-2.08	-0.97	-1.9
E	25.21	15.09	6.87	14.12	0.0	12.39	-0.07	3.84	1.99	5.09	2.83	8.74	5.36	8.35	12.04	13.15	12.22
F	12.82	2.7	-5.52	1.73	-12.39	0.0	-12.46	-8.55	-10.4	-7.3	-9.56	-3.65	-7.03	-4.04	-0.35	0.76	-0.17
G	25.28	15.16	6.94	14.19	0.07	12.46	0.0	3.91	2.06	5.16	2.9	8.81	5.43	8.42	12.11	13.22	12.29
H	21.37	11.25	3.03	10.28	-3.84	8.55	-3.91	0.0	-1.85	1.25	-1.01	4.9	1.52	4.51	8.2	9.31	8.38
I	23.22	13.1	4.88	12.13	-1.99	10.4	-2.06	1.85	0.0	3.1	0.84	6.75	3.37	6.36	10.05	11.16	10.23
J	20.12	10.0	1.78	9.03	-5.09	7.3	-5.16	-1.25	-3.1	0.0	-2.26	3.65	0.27	3.26	6.95	8.06	7.13
K	22.38	12.26	4.04	11.29	-2.83	9.56	-2.9	1.01	-0.84	2.26	0.0	5.91	2.53	5.52	9.21	10.32	9.39
L	16.47	6.35	-1.87	5.38	-8.74	3.65	-8.81	-4.9	-6.75	-3.65	-5.91	0.0	-3.38	-0.39	3.3	4.41	3.48
M	19.85	9.73	1.51	8.76	-5.36	7.03	-5.43	-1.52	-3.37	-0.27	-2.53	3.38	0.0	2.99	6.68	7.79	6.86
N	16.86	6.74	-1.48	5.77	-8.35	4.04	-8.42	-4.51	-6.36	-3.26	-5.52	0.39	-2.99	0.0	3.69	4.8	3.87
O	13.17	3.05	-5.17	2.08	-12.04	0.35	-12.11	-8.2	-10.05	-6.95	-9.21	-3.3	-6.68	-3.69	0.0	1.11	0.18
P	12.06	1.94	-6.28	0.97	-13.15	-0.76	-13.22	-9.31	-11.16	-8.06	-10.32	-4.41	-7.79	-4.8	-1.11	0.0	-0.93
Q	12.99	2.87	-5.35	1.9	-12.22	0.17	-12.29	-8.38	-10.23	-7.13	-9.39	-3.48	-6.86	-3.87	-0.18	0.93	0.0

A: Nenhuma Renda C: De R$ 1.045,01 até R$ 1.567,50 E: De R$ 2.090,01 até R$ 2.612,50 G: De R$ 3.135,01 até R$ 4.180,00 I: De R$ 5.225,01 até R$ 6.270,00
B: Até R$ 1.045,00 D: De R$ 1.567,51 até R$ 2.090,00 F: De R$ 2.612,51 até R$ 3.135,00 H: De R$ 4.180,01 até R$ 5.225,00 J: De R$ 6.270,01 até R$ 7.315,00
K: De R$ 7.315,01 até R$ 8.360,00 M: De R$ 9.405,01 até R$ 10.450,00 O: De R$ 12.540,01 até R$ 15.675,00 Q: Acima de R$ 20.900,00
L: De R$ 8.360,01 até R$ 9.405,00 N: De R$ 10.450,01 até R$ 12.540,00 P: De R$ 15.675,01 até R$ 20.900,00

Figura 5.36: Diferença entre as variações das notas de Matemática por categoria de renda em 2020 com relação a 2019.

Entretanto, as médias das notas podem não refletir variações nas taxas de acerto dos candidatos devido à Teoria de Resposta ao Item (TRI). Por esta razão, foi feita uma análise utilizando as médias dos escores brutos dos candidatos. Foram realizadas comparações entre as médias dos escores brutos em cada prova, por categoria de renda, dos candidatos presentes nos dois dias de provas de 2019 e 2020. As Figuras 5.37 a 5.40 mostram os resultados para as prova de Ciências Humanas, Linguagens e Códigos, Ciências da Natureza e Matemática, respectivamente. Na prova de Ciências Humanas, observa-se que a média dos escores brutos foram bem próximas dentro de cada categoria de renda. Na prova de Linguagens e Códigos, a média dos escores brutos de 2020 foi maior até a categoria de renda "H" (de R$ 3.992,01 até R$ 4.990,00), logo em seguida a média de 2019 começou a ser superior. Já nas provas de Ciências da Natureza e Matemática, as médias de 2020 foram superiores em todas as categorias de renda. Em todos os casos, as médias dos escores brutos aumentaram conforme aumentava a faixa de renda. Observou-se também que apenas em Ciências da Natureza, as médias das notas de 2020 foram

maiores que 2019, assim como as médias do escores brutos. Em Matemática, apesar das médias do escores brutos serem maiores em 2020, as médias das notas de 2019 foram maiores, tanto para a maioria das categorias de renda, quanto no desempenho geral do exame.

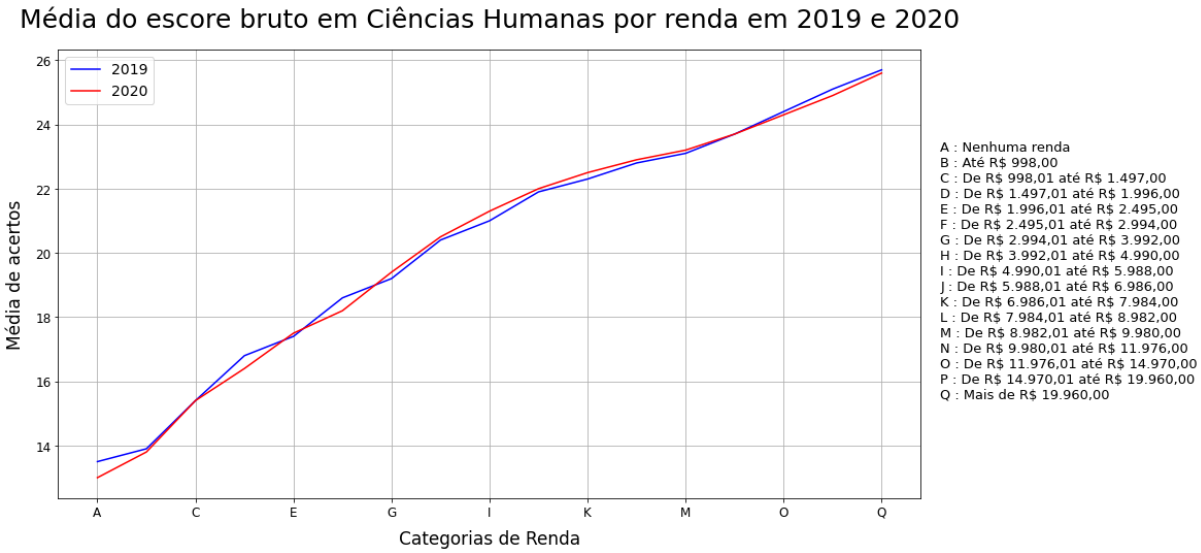

Figura 5.37: Média do escore bruto na prova de Ciências Humanas por renda em 2019 e 2020.

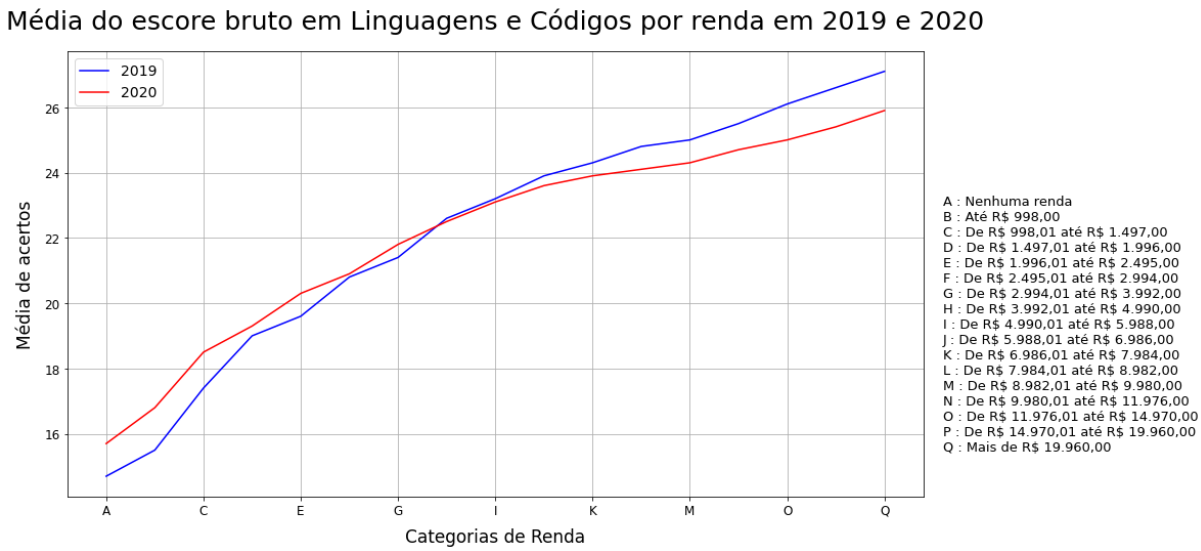

Figura 5.38: Média do escore bruto na prova de Linguagens e Códigos por renda em 2019 e 2020.

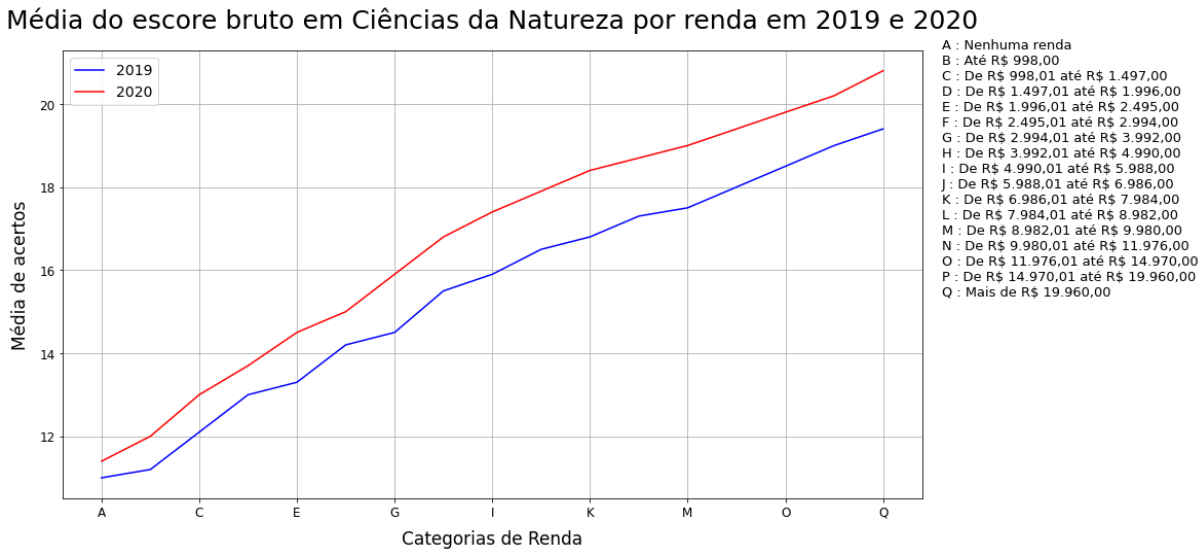

Figura 5.39: Média do escore bruto na prova de Ciências da Natureza por renda em 2019 e 2020.

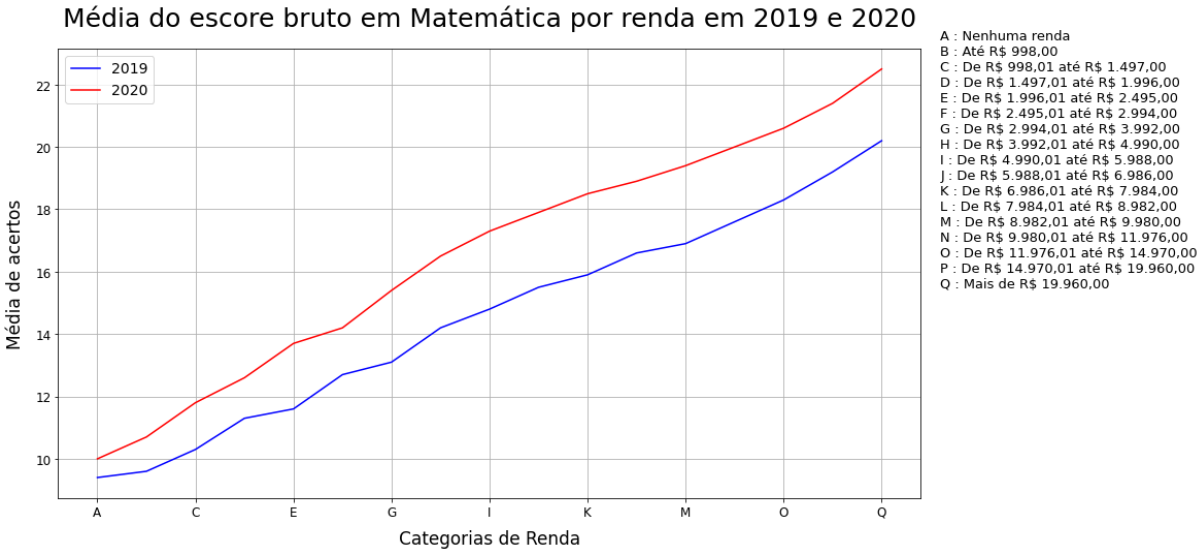

Figura 5.40: Média do escore bruto na prova de Matemática por renda em 2019 e 2020.

Nas Figuras 5.41 a 5.44 estão as tabelas com as diferenças de acerto entre cada categoria de renda para as provas de Ciências Humanas, Linguagens e Códigos, Ciências da Natureza e Matemática, respectivamente. Os valores abaixo da diagonal com zeros representam a diferença de acertos entre uma categoria de renda mais alta para uma mais baixa, e seus valores tendem a ser positivos. Isso significa que as categorias de renda mais altas tem a tendência de possuírem mais acertos, como visto nos gráficos das Figuras 5.37 a 5.40. Os valores acima da diagonal com zeros representam o oposto, a diferença de

acertos entre uma categorias de renda mais baixa para uma mais alta, e possuem em sua maioria valores negativos.

Diferença de acertos entre as categorias de renda na prova de Ciências Humanas em 2019 e 2020

2019

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-0.46	-1.97	-3.39	-3.96	-5.13	-5.73	-6.98	-7.59	-8.41	-8.84	-9.39	-9.62	-10.25	-10.95	-11.62	-12.28
B	0.46	0.0	-1.51	-2.93	-3.5	-4.67	-5.27	-6.52	-7.13	-7.95	-8.38	-8.93	-9.16	-9.79	-10.49	-11.16	-11.82
C	1.97	1.51	0.0	-1.42	-1.99	-3.16	-3.76	-5.01	-5.62	-6.44	-6.87	-7.42	-7.65	-8.28	-8.98	-9.65	-10.31
D	3.39	2.93	1.42	0.0	-0.57	-1.74	-2.34	-3.59	-4.2	-5.02	-5.45	-6.0	-6.23	-6.86	-7.56	-8.23	-8.89
E	3.96	3.5	1.99	0.57	0.0	-1.17	-1.77	-3.02	-3.63	-4.45	-4.88	-5.43	-5.66	-6.29	-6.99	-7.66	-8.32
F	5.13	4.67	3.16	1.74	1.17	0.0	-0.6	-1.85	-2.46	-3.28	-3.71	-4.26	-4.49	-5.12	-5.82	-6.49	-7.15
G	5.73	5.27	3.76	2.34	1.77	0.6	0.0	-1.25	-1.86	-2.68	-3.11	-3.66	-3.89	-4.52	-5.22	-5.89	-6.55
H	6.98	6.52	5.01	3.59	3.02	1.85	1.25	0.0	-0.61	-1.43	-1.86	-2.41	-2.64	-3.27	-3.97	-4.64	-5.3
I	7.59	7.13	5.62	4.2	3.63	2.46	1.86	0.61	0.0	-0.82	-1.25	-1.8	-2.03	-2.66	-3.36	-4.03	-4.69
J	8.41	7.95	6.44	5.02	4.45	3.28	2.68	1.43	0.82	0.0	-0.43	-0.98	-1.21	-1.84	-2.54	-3.21	-3.87
K	8.84	8.38	6.87	5.45	4.88	3.71	3.11	1.86	1.25	0.43	0.0	-0.55	-0.78	-1.41	-2.11	-2.78	-3.44
L	9.39	8.93	7.42	6.0	5.43	4.26	3.66	2.41	1.8	0.98	0.55	0.0	-0.23	-0.86	-1.56	-2.23	-2.89
M	9.62	9.16	7.65	6.23	5.66	4.49	3.89	2.64	2.03	1.21	0.78	0.23	0.0	-0.63	-1.33	-2.0	-2.66
N	10.25	9.79	8.28	6.86	6.29	5.12	4.52	3.27	2.66	1.84	1.41	0.86	0.63	0.0	-0.7	-1.37	-2.03
O	10.95	10.49	8.98	7.56	6.99	5.82	5.22	3.97	3.36	2.54	2.11	1.56	1.33	0.7	0.0	-0.67	-1.33
P	11.62	11.16	9.65	8.23	7.66	6.49	5.89	4.64	4.03	3.21	2.78	2.23	2.0	1.37	0.67	0.0	-0.66
Q	12.28	11.82	10.31	8.89	8.32	7.15	6.55	5.3	4.69	3.87	3.44	2.89	2.66	2.03	1.33	0.66	0.0

2020

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-0.83	-2.35	-3.35	-4.45	-5.19	-6.42	-7.5	-8.33	-9.0	-9.53	-9.88	-10.17	-10.72	-11.31	-11.91	-12.63
B	0.83	0.0	-1.52	-2.52	-3.62	-4.36	-5.59	-6.67	-7.5	-8.17	-8.7	-9.05	-9.34	-9.89	-10.48	-11.08	-11.8
C	2.35	1.52	0.0	-1.0	-2.1	-2.84	-4.07	-5.15	-5.98	-6.65	-7.18	-7.53	-7.82	-8.37	-8.96	-9.56	-10.28
D	3.35	2.52	1.0	0.0	-1.1	-1.84	-3.07	-4.15	-4.98	-5.65	-6.18	-6.53	-6.82	-7.37	-7.96	-8.56	-9.28
E	4.45	3.62	2.1	1.1	0.0	-0.74	-1.97	-3.05	-3.88	-4.55	-5.08	-5.43	-5.72	-6.27	-6.86	-7.46	-8.18
F	5.19	4.36	2.84	1.84	0.74	0.0	-1.23	-2.31	-3.14	-3.81	-4.34	-4.69	-4.98	-5.53	-6.12	-6.72	-7.44
G	6.42	5.59	4.07	3.07	1.97	1.23	0.0	-1.08	-1.91	-2.58	-3.11	-3.46	-3.75	-4.3	-4.89	-5.49	-6.21
H	7.5	6.67	5.15	4.15	3.05	2.31	1.08	0.0	-0.83	-1.5	-2.03	-2.38	-2.67	-3.22	-3.81	-4.41	-5.13
I	8.33	7.5	5.98	4.98	3.88	3.14	1.91	0.83	0.0	-0.67	-1.2	-1.55	-1.84	-2.39	-2.98	-3.58	-4.3
J	9.0	8.17	6.65	5.65	4.55	3.81	2.58	1.5	0.67	0.0	-0.53	-0.88	-1.17	-1.72	-2.31	-2.91	-3.63
K	9.53	8.7	7.18	6.18	5.08	4.34	3.11	2.03	1.2	0.53	0.0	-0.35	-0.64	-1.19	-1.78	-2.38	-3.1
L	9.88	9.05	7.53	6.53	5.43	4.69	3.46	2.38	1.55	0.88	0.35	0.0	-0.29	-0.84	-1.43	-2.03	-2.75
M	10.17	9.34	7.82	6.82	5.72	4.98	3.75	2.67	1.84	1.17	0.64	0.29	0.0	-0.55	-1.14	-1.74	-2.46
N	10.72	9.89	8.37	7.37	6.27	5.53	4.3	3.22	2.39	1.72	1.19	0.84	0.55	0.0	-0.59	-1.19	-1.91
O	11.31	10.48	8.96	7.96	6.86	6.12	4.89	3.81	2.98	2.31	1.78	1.43	1.14	0.59	0.0	-0.6	-1.32
P	11.91	11.08	9.56	8.56	7.46	6.72	5.49	4.41	3.58	2.91	2.38	2.03	1.74	1.19	0.6	0.0	-0.72
Q	12.63	11.8	10.28	9.28	8.18	7.44	6.21	5.13	4.3	3.63	3.1	2.75	2.46	1.91	1.32	0.72	0.0

A: Nenhuma Renda B: Até R$ 1.045,00 C: De R$ 1.045,01 até R$ 1.567,50 D: De R$ 1.567,51 até R$ 2.090,00 E: De R$ 2.090,01 até R$ 2.612,50 F: De R$ 2.612,51 até R$ 3.135,00 G: De R$ 3.135,01 até R$ 4.180,00 H: De R$ 4.180,01 até R$ 5.225,00 I: De R$ 5.225,01 até R$ 6.270,00 J: De R$ 6.270,01 até R$ 7.315,00 K: De R$ 7.315,01 até R$ 8.360,00 L: De R$ 8.360,01 até R$ 9.405,00 M: De R$ 9.405,01 até R$ 10.450,00 N: De R$ 10.450,01 até R$ 12.540,00 O: De R$ 12.540,01 até R$ 15.675,00 P: De R$ 15.675,01 até R$ 20.900,00 Q: Acima de R$ 20.900,00

Figura 5.41: Diferença de acertos das categorias de renda para a prova de Ciências Humanas em 2019 e 2020.

Diferença de acertos entre as categorias de renda na prova de Linguagens e Códigos em 2019 e 2020

2019

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-0.73	-2.64	-4.22	-4.85	-6.06	-6.67	-7.85	-8.43	-9.16	-9.54	-10.03	-10.25	-10.78	-11.34	-11.88	-12.42
B	0.73	0.0	-1.91	-3.49	-4.12	-5.33	-5.94	-7.12	-7.7	-8.43	-8.81	-9.3	-9.52	-10.05	-10.61	-11.15	-11.69
C	2.64	1.91	0.0	-1.58	-2.21	-3.42	-4.03	-5.21	-5.79	-6.52	-6.9	-7.39	-7.61	-8.14	-8.7	-9.24	-9.78
D	4.22	3.49	1.58	0.0	-0.63	-1.84	-2.45	-3.63	-4.21	-4.94	-5.32	-5.81	-6.03	-6.56	-7.12	-7.66	-8.2
E	4.85	4.12	2.21	0.63	0.0	-1.21	-1.82	-3.0	-3.58	-4.31	-4.69	-5.18	-5.4	-5.93	-6.49	-7.03	-7.57
F	6.06	5.33	3.42	1.84	1.21	0.0	-0.61	-1.79	-2.37	-3.1	-3.48	-3.97	-4.19	-4.72	-5.28	-5.82	-6.36
G	6.67	5.94	4.03	2.45	1.82	0.61	0.0	-1.18	-1.76	-2.49	-2.87	-3.36	-3.58	-4.11	-4.67	-5.21	-5.75
H	7.85	7.12	5.21	3.63	3.0	1.79	1.18	0.0	-0.58	-1.31	-1.69	-2.18	-2.4	-2.93	-3.49	-4.03	-4.57
I	8.43	7.7	5.79	4.21	3.58	2.37	1.76	0.58	0.0	-0.73	-1.11	-1.6	-1.82	-2.35	-2.91	-3.45	-3.99
J	9.16	8.43	6.52	4.94	4.31	3.1	2.49	1.31	0.73	0.0	-0.38	-0.87	-1.09	-1.62	-2.18	-2.72	-3.26
K	9.54	8.81	6.9	5.32	4.69	3.48	2.87	1.69	1.11	0.38	0.0	-0.49	-0.71	-1.24	-1.8	-2.34	-2.88
L	10.03	9.3	7.39	5.81	5.18	3.97	3.36	2.18	1.6	0.87	0.49	0.0	-0.22	-0.75	-1.31	-1.85	-2.39
M	10.25	9.52	7.61	6.03	5.4	4.19	3.58	2.4	1.82	1.09	0.71	0.22	0.0	-0.53	-1.09	-1.63	-2.17
N	10.78	10.05	8.14	6.56	5.93	4.72	4.11	2.93	2.35	1.62	1.24	0.75	0.53	0.0	-0.56	-1.1	-1.64
O	11.34	10.61	8.7	7.12	6.49	5.28	4.67	3.49	2.91	2.18	1.8	1.31	1.09	0.56	0.0	-0.54	-1.08
P	11.88	11.15	9.24	7.66	7.03	5.82	5.21	4.03	3.45	2.72	2.34	1.85	1.63	1.1	0.54	0.0	-0.54
Q	12.42	11.69	9.78	8.2	7.57	6.36	5.75	4.57	3.99	3.26	2.88	2.39	2.17	1.64	1.08	0.54	0.0

2020

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-1.1	-2.73	-3.6	-4.57	-5.15	-6.07	-6.79	-7.37	-7.84	-8.2	-8.38	-8.61	-8.96	-9.3	-9.68	-10.16
B	1.1	0.0	-1.63	-2.5	-3.47	-4.05	-4.97	-5.69	-6.27	-6.74	-7.1	-7.28	-7.51	-7.86	-8.2	-8.58	-9.06
C	2.73	1.63	0.0	-0.87	-1.84	-2.42	-3.34	-4.06	-4.64	-5.11	-5.47	-5.65	-5.88	-6.23	-6.57	-6.95	-7.43
D	3.6	2.5	0.87	0.0	-0.97	-1.55	-2.47	-3.19	-3.77	-4.24	-4.6	-4.78	-5.01	-5.36	-5.7	-6.08	-6.56
E	4.57	3.47	1.84	0.97	0.0	-0.58	-1.5	-2.22	-2.8	-3.27	-3.63	-3.81	-4.04	-4.39	-4.73	-5.11	-5.59
F	5.15	4.05	2.42	1.55	0.58	0.0	-0.92	-1.64	-2.22	-2.69	-3.05	-3.23	-3.46	-3.81	-4.15	-4.53	-5.01
G	6.07	4.97	3.34	2.47	1.5	0.92	0.0	-0.72	-1.3	-1.77	-2.13	-2.31	-2.54	-2.89	-3.23	-3.61	-4.09
H	6.79	5.69	4.06	3.19	2.22	1.64	0.72	0.0	-0.58	-1.05	-1.41	-1.59	-1.82	-2.17	-2.51	-2.89	-3.37
I	7.37	6.27	4.64	3.77	2.8	2.22	1.3	0.58	0.0	-0.47	-0.83	-1.01	-1.24	-1.59	-1.93	-2.31	-2.79
J	7.84	6.74	5.11	4.24	3.27	2.69	1.77	1.05	0.47	0.0	-0.36	-0.54	-0.77	-1.12	-1.46	-1.84	-2.32
K	8.2	7.1	5.47	4.6	3.63	3.05	2.13	1.41	0.83	0.36	0.0	-0.18	-0.41	-0.76	-1.1	-1.48	-1.96
L	8.38	7.28	5.65	4.78	3.81	3.23	2.31	1.59	1.01	0.54	0.18	0.0	-0.23	-0.58	-0.92	-1.3	-1.78
M	8.61	7.51	5.88	5.01	4.04	3.46	2.54	1.82	1.24	0.77	0.41	0.23	0.0	-0.35	-0.69	-1.07	-1.55
N	8.96	7.86	6.23	5.36	4.39	3.81	2.89	2.17	1.59	1.12	0.76	0.58	0.35	0.0	-0.34	-0.72	-1.2
O	9.3	8.2	6.57	5.7	4.73	4.15	3.23	2.51	1.93	1.46	1.1	0.92	0.69	0.34	0.0	-0.38	-0.86
P	9.68	8.58	6.95	6.08	5.11	4.53	3.61	2.89	2.31	1.84	1.48	1.3	1.07	0.72	0.38	0.0	-0.48
Q	10.16	9.06	7.43	6.56	5.59	5.01	4.09	3.37	2.79	2.32	1.96	1.78	1.55	1.2	0.86	0.48	0.0

A: Nenhuma Renda B: Até R$ 1.045,00 C: De R$ 1.045,01 até R$ 1.567,50 D: De R$ 1.567,51 até R$ 2.090,00 E: De R$ 2.090,01 até R$ 2.612,50 F: De R$ 2.612,51 até R$ 3.135,00 G: De R$ 3.135,01 até R$ 4.180,00 H: De R$ 4.180,01 até R$ 5.225,00 I: De R$ 5.225,01 até R$ 6.270,00 J: De R$ 6.270,01 até R$ 7.315,00 K: De R$ 7.315,01 até R$ 8.360,00 L: De R$ 8.360,01 até R$ 9.405,00 M: De R$ 9.405,01 até R$ 10.450,00 N: De R$ 10.450,01 até R$ 12.540,00 O: De R$ 12.540,01 até R$ 15.675,00 P: De R$ 15.675,01 até R$ 20.900,00 Q: Acima de R$ 20.900,00

Figura 5.42: Diferença de acertos das categorias de renda para a prova de Linguagens e Códigos em 2019 e 2020.

Diferença de acertos entre as categorias de renda na prova de Ciências da Natureza em 2019 e 2020

2019

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-0.25	-1.14	-2.04	-2.36	-3.19	-3.55	-4.49	-4.96	-5.56	-5.87	-6.31	-6.53	-7.03	-7.54	-8.06	-8.45
B	0.25	0.0	-0.89	-1.79	-2.11	-2.94	-3.3	-4.24	-4.71	-5.31	-5.62	-6.06	-6.28	-6.78	-7.29	-7.81	-8.2
C	1.14	0.89	0.0	-0.9	-1.22	-2.05	-2.41	-3.35	-3.82	-4.42	-4.73	-5.17	-5.39	-5.89	-6.4	-6.92	-7.31
D	2.04	1.79	0.9	0.0	-0.32	-1.15	-1.51	-2.45	-2.92	-3.52	-3.83	-4.27	-4.49	-4.99	-5.5	-6.02	-6.41
E	2.36	2.11	1.22	0.32	0.0	-0.83	-1.19	-2.13	-2.6	-3.2	-3.51	-3.95	-4.17	-4.67	-5.18	-5.7	-6.09
F	3.19	2.94	2.05	1.15	0.83	0.0	-0.36	-1.3	-1.77	-2.37	-2.68	-3.12	-3.34	-3.84	-4.35	-4.87	-5.26
G	3.55	3.3	2.41	1.51	1.19	0.36	0.0	-0.94	-1.41	-2.01	-2.32	-2.76	-2.98	-3.48	-3.99	-4.51	-4.9
H	4.49	4.24	3.35	2.45	2.13	1.3	0.94	0.0	-0.47	-1.07	-1.38	-1.82	-2.04	-2.54	-3.05	-3.57	-3.96
I	4.96	4.71	3.82	2.92	2.6	1.77	1.41	0.47	0.0	-0.6	-0.91	-1.35	-1.57	-2.07	-2.58	-3.1	-3.49
J	5.56	5.31	4.42	3.52	3.2	2.37	2.01	1.07	0.6	0.0	-0.31	-0.75	-0.97	-1.47	-1.98	-2.5	-2.89
K	5.87	5.62	4.73	3.83	3.51	2.68	2.32	1.38	0.91	0.31	0.0	-0.44	-0.66	-1.16	-1.67	-2.19	-2.58
L	6.31	6.06	5.17	4.27	3.95	3.12	2.76	1.82	1.35	0.75	0.44	0.0	-0.22	-0.72	-1.23	-1.75	-2.14
M	6.53	6.28	5.39	4.49	4.17	3.34	2.98	2.04	1.57	0.97	0.66	0.22	0.0	-0.5	-1.01	-1.53	-1.92
N	7.03	6.78	5.89	4.99	4.67	3.84	3.48	2.54	2.07	1.47	1.16	0.72	0.5	0.0	-0.51	-1.03	-1.42
O	7.54	7.29	6.4	5.5	5.18	4.35	3.99	3.05	2.58	1.98	1.67	1.23	1.01	0.51	0.0	-0.52	-0.91
P	8.06	7.81	6.92	6.02	5.7	4.87	4.51	3.57	3.1	2.5	2.19	1.75	1.53	1.03	0.52	0.0	-0.39
Q	8.45	8.2	7.31	6.41	6.09	5.26	4.9	3.96	3.49	2.89	2.58	2.14	1.92	1.42	0.91	0.39	0.0

2020

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-0.58	-1.63	-2.31	-3.13	-3.65	-4.57	-5.43	-6.06	-6.57	-7.02	-7.32	-7.59	-7.98	-8.38	-8.81	-9.37
B	0.58	0.0	-1.05	-1.73	-2.55	-3.07	-3.99	-4.85	-5.48	-5.99	-6.44	-6.74	-7.01	-7.4	-7.8	-8.23	-8.79
C	1.63	1.05	0.0	-0.68	-1.5	-2.02	-2.94	-3.8	-4.43	-4.94	-5.39	-5.69	-5.96	-6.35	-6.75	-7.18	-7.74
D	2.31	1.73	0.68	0.0	-0.82	-1.34	-2.26	-3.12	-3.75	-4.26	-4.71	-5.01	-5.28	-5.67	-6.07	-6.5	-7.06
E	3.13	2.55	1.5	0.82	0.0	-0.52	-1.44	-2.3	-2.93	-3.44	-3.89	-4.19	-4.46	-4.85	-5.25	-5.68	-6.24
F	3.65	3.07	2.02	1.34	0.52	0.0	-0.92	-1.78	-2.41	-2.92	-3.37	-3.67	-3.94	-4.33	-4.73	-5.16	-5.72
G	4.57	3.99	2.94	2.26	1.44	0.92	0.0	-0.86	-1.49	-2.0	-2.45	-2.75	-3.02	-3.41	-3.81	-4.24	-4.8
H	5.43	4.85	3.8	3.12	2.3	1.78	0.86	0.0	-0.63	-1.14	-1.59	-1.89	-2.16	-2.55	-2.95	-3.38	-3.94
I	6.06	5.48	4.43	3.75	2.93	2.41	1.49	0.63	0.0	-0.51	-0.96	-1.26	-1.53	-1.92	-2.32	-2.75	-3.31
J	6.57	5.99	4.94	4.26	3.44	2.92	2.0	1.14	0.51	0.0	-0.45	-0.75	-1.02	-1.41	-1.81	-2.24	-2.8
K	7.02	6.44	5.39	4.71	3.89	3.37	2.45	1.59	0.96	0.45	0.0	-0.3	-0.57	-0.96	-1.36	-1.79	-2.35
L	7.32	6.74	5.69	5.01	4.19	3.67	2.75	1.89	1.26	0.75	0.3	0.0	-0.27	-0.66	-1.06	-1.49	-2.05
M	7.59	7.01	5.96	5.28	4.46	3.94	3.02	2.16	1.53	1.02	0.57	0.27	0.0	-0.39	-0.79	-1.22	-1.78
N	7.98	7.4	6.35	5.67	4.85	4.33	3.41	2.55	1.92	1.41	0.96	0.66	0.39	0.0	-0.4	-0.83	-1.39
O	8.38	7.8	6.75	6.07	5.25	4.73	3.81	2.95	2.32	1.81	1.36	1.06	0.79	0.4	0.0	-0.43	-0.99
P	8.81	8.23	7.18	6.5	5.68	5.16	4.24	3.38	2.75	2.24	1.79	1.49	1.22	0.83	0.43	0.0	-0.56
Q	9.37	8.79	7.74	7.06	6.24	5.72	4.8	3.94	3.31	2.8	2.35	2.05	1.78	1.39	0.99	0.56	0.0

A: Nenhuma Renda B: Até R$ 1.045,00 C: De R$ 1.045,01 até R$ 1.567,50 D: De R$ 1.567,51 até R$ 2.090,00 E: De R$ 2.090,01 até R$ 2.612,50 F: De R$ 2.612,51 até R$ 3.135,00 G: De R$ 3.135,01 até R$ 4.180,00 H: De R$ 4.180,01 até R$ 5.225,00 I: De R$ 5.225,01 até R$ 6.270,00 J: De R$ 6.270,01 até R$ 7.315,00 K: De R$ 7.315,01 até R$ 8.360,00 L: De R$ 8.360,01 até R$ 9.405,00 M: De R$ 9.405,01 até R$ 10.450,00 N: De R$ 10.450,01 até R$ 12.540,00 O: De R$ 12.540,01 até R$ 15.675,00 P: De R$ 15.675,01 até R$ 20.900,00 Q: Acima de R$ 20.900,00

Figura 5.43: Diferença de acertos das categorias de renda para a prova de Ciências da Natureza em 2019 e 2020.

Diferença de acertos entre as categorias de renda na prova de Matemática em 2019 e 2020

2019

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-0.23	-0.97	-1.94	-2.26	-3.29	-3.7	-4.83	-5.44	-6.15	-6.57	-7.2	-7.47	-8.17	-8.97	-9.78	-10.84
B	0.23	0.0	-0.74	-1.71	-2.03	-3.06	-3.47	-4.6	-5.21	-5.92	-6.34	-6.97	-7.24	-7.94	-8.74	-9.55	-10.61
C	0.97	0.74	0.0	-0.97	-1.29	-2.32	-2.73	-3.86	-4.47	-5.18	-5.6	-6.23	-6.5	-7.2	-8.0	-8.81	-9.87
D	1.94	1.71	0.97	0.0	-0.32	-1.35	-1.76	-2.89	-3.5	-4.21	-4.63	-5.26	-5.53	-6.23	-7.03	-7.84	-8.9
E	2.26	2.03	1.29	0.32	0.0	-1.03	-1.44	-2.57	-3.18	-3.89	-4.31	-4.94	-5.21	-5.91	-6.71	-7.52	-8.58
F	3.29	3.06	2.32	1.35	1.03	0.0	-0.41	-1.54	-2.15	-2.86	-3.28	-3.91	-4.18	-4.88	-5.68	-6.49	-7.55
G	3.7	3.47	2.73	1.76	1.44	0.41	0.0	-1.13	-1.74	-2.45	-2.87	-3.5	-3.77	-4.47	-5.27	-6.08	-7.14
H	4.83	4.6	3.86	2.89	2.57	1.54	1.13	0.0	-0.61	-1.32	-1.74	-2.37	-2.64	-3.34	-4.14	-4.95	-6.01
I	5.44	5.21	4.47	3.5	3.18	2.15	1.74	0.61	0.0	-0.71	-1.13	-1.76	-2.03	-2.73	-3.53	-4.34	-5.4
J	6.15	5.92	5.18	4.21	3.89	2.86	2.45	1.32	0.71	0.0	-0.42	-1.05	-1.32	-2.02	-2.82	-3.63	-4.69
K	6.57	6.34	5.6	4.63	4.31	3.28	2.87	1.74	1.13	0.42	0.0	-0.63	-0.9	-1.6	-2.4	-3.21	-4.27
L	7.2	6.97	6.23	5.26	4.94	3.91	3.5	2.37	1.76	1.05	0.63	0.0	-0.27	-0.97	-1.77	-2.58	-3.64
M	7.47	7.24	6.5	5.53	5.21	4.18	3.77	2.64	2.03	1.32	0.9	0.27	0.0	-0.7	-1.5	-2.31	-3.37
N	8.17	7.94	7.2	6.23	5.91	4.88	4.47	3.34	2.73	2.02	1.6	0.97	0.7	0.0	-0.8	-1.61	-2.67
O	8.97	8.74	8.0	7.03	6.71	5.68	5.27	4.14	3.53	2.82	2.4	1.77	1.5	0.8	0.0	-0.81	-1.87
P	9.78	9.55	8.81	7.84	7.52	6.49	6.08	4.95	4.34	3.63	3.21	2.58	2.31	1.61	0.81	0.0	-1.06
Q	10.84	10.61	9.87	8.9	8.58	7.55	7.14	6.01	5.4	4.69	4.27	3.64	3.37	2.67	1.87	1.06	0.0

2020

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
A	0.0	-0.66	-1.82	-2.6	-3.69	-4.22	-5.41	-6.48	-7.3	-7.92	-8.53	-8.87	-9.38	-9.96	-10.53	-11.34	-12.53
B	0.66	0.0	-1.16	-1.94	-3.03	-3.56	-4.75	-5.82	-6.64	-7.26	-7.87	-8.21	-8.72	-9.3	-9.87	-10.68	-11.87
C	1.82	1.16	0.0	-0.78	-1.87	-2.4	-3.59	-4.66	-5.48	-6.1	-6.71	-7.05	-7.56	-8.14	-8.71	-9.52	-10.71
D	2.6	1.94	0.78	0.0	-1.09	-1.62	-2.81	-3.88	-4.7	-5.32	-5.93	-6.27	-6.78	-7.36	-7.93	-8.74	-9.93
E	3.69	3.03	1.87	1.09	0.0	-0.53	-1.72	-2.79	-3.61	-4.23	-4.84	-5.18	-5.69	-6.27	-6.84	-7.65	-8.84
F	4.22	3.56	2.4	1.62	0.53	0.0	-1.19	-2.26	-3.08	-3.7	-4.31	-4.65	-5.16	-5.74	-6.31	-7.12	-8.31
G	5.41	4.75	3.59	2.81	1.72	1.19	0.0	-1.07	-1.89	-2.51	-3.12	-3.46	-3.97	-4.55	-5.12	-5.93	-7.12
H	6.48	5.82	4.66	3.88	2.79	2.26	1.07	0.0	-0.82	-1.44	-2.05	-2.39	-2.9	-3.48	-4.05	-4.86	-6.05
I	7.3	6.64	5.48	4.7	3.61	3.08	1.89	0.82	0.0	-0.62	-1.23	-1.57	-2.08	-2.66	-3.23	-4.04	-5.23
J	7.92	7.26	6.1	5.32	4.23	3.7	2.51	1.44	0.62	0.0	-0.61	-0.95	-1.46	-2.04	-2.61	-3.42	-4.61
K	8.53	7.87	6.71	5.93	4.84	4.31	3.12	2.05	1.23	0.61	0.0	-0.34	-0.85	-1.43	-2.0	-2.81	-4.0
L	8.87	8.21	7.05	6.27	5.18	4.65	3.46	2.39	1.57	0.95	0.34	0.0	-0.51	-1.09	-1.66	-2.47	-3.66
M	9.38	8.72	7.56	6.78	5.69	5.16	3.97	2.9	2.08	1.46	0.85	0.51	0.0	-0.58	-1.15	-1.96	-3.15
N	9.96	9.3	8.14	7.36	6.27	5.74	4.55	3.48	2.66	2.04	1.43	1.09	0.58	0.0	-0.57	-1.38	-2.57
O	10.53	9.87	8.71	7.93	6.84	6.31	5.12	4.05	3.23	2.61	2.0	1.66	1.15	0.57	0.0	-0.81	-2.0
P	11.34	10.68	9.52	8.74	7.65	7.12	5.93	4.86	4.04	3.42	2.81	2.47	1.96	1.38	0.81	0.0	-1.19
Q	12.53	11.87	10.71	9.93	8.84	8.31	7.12	6.05	5.23	4.61	4.0	3.66	3.15	2.57	2.0	1.19	0.0

A: Nenhuma Renda B: Até R$ 1.045,00 C: De R$ 1.045,01 até R$ 1.567,50 D: De R$ 1.567,51 até R$ 2.090,00 E: De R$ 2.090,01 até R$ 2.612,50 F: De R$ 2.612,51 até R$ 3.135,00 G: De R$ 3.135,01 até R$ 4.180,00 H: De R$ 4.180,01 até R$ 5.225,00 I: De R$ 5.225,01 até R$ 6.270,00 J: De R$ 6.270,01 até R$ 7.315,00 K: De R$ 7.315,01 até R$ 8.360,00 L: De R$ 8.360,01 até R$ 9.405,00 M: De R$ 9.405,01 até R$ 10.450,00 N: De R$ 10.450,01 até R$ 12.540,00 O: De R$ 12.540,01 até R$ 15.675,00 P: De R$ 15.675,01 até R$ 20.900,00 Q: Acima de R$ 20.900,00

Figura 5.44: Diferença de acertos das categorias de renda para a prova de Matemática em 2019 e 2020.

5.3 Análise das Notas por Escore Bruto

Como complemento para as análises a respeito do escore bruto, procurou-se descobrir qual o valor médio da nota obtida pelos participantes com base nos possíveis valores de escore bruto (0 a 45 acertos para cada prova). Os resultados podem ser vistos nas Figuras 5.45 a 5.48. Observa-se que na prova de Ciências da Natureza foram obtidos os valores mais próximos para as médias das provas nos anos de 2019 e 2020. Para Ciências Humanas, a partir de aproximadamente 15 acertos a média notas de 2020 foram maiores. Em Linguagens e Códigos, as médias de 2019 foram maiores até aproximadamente 20 acertos, quando as médias de 2020 passaram à frente. Para a prova de Matemática, as

médias foram maiores em 2019 para todos os valores de escore bruto.

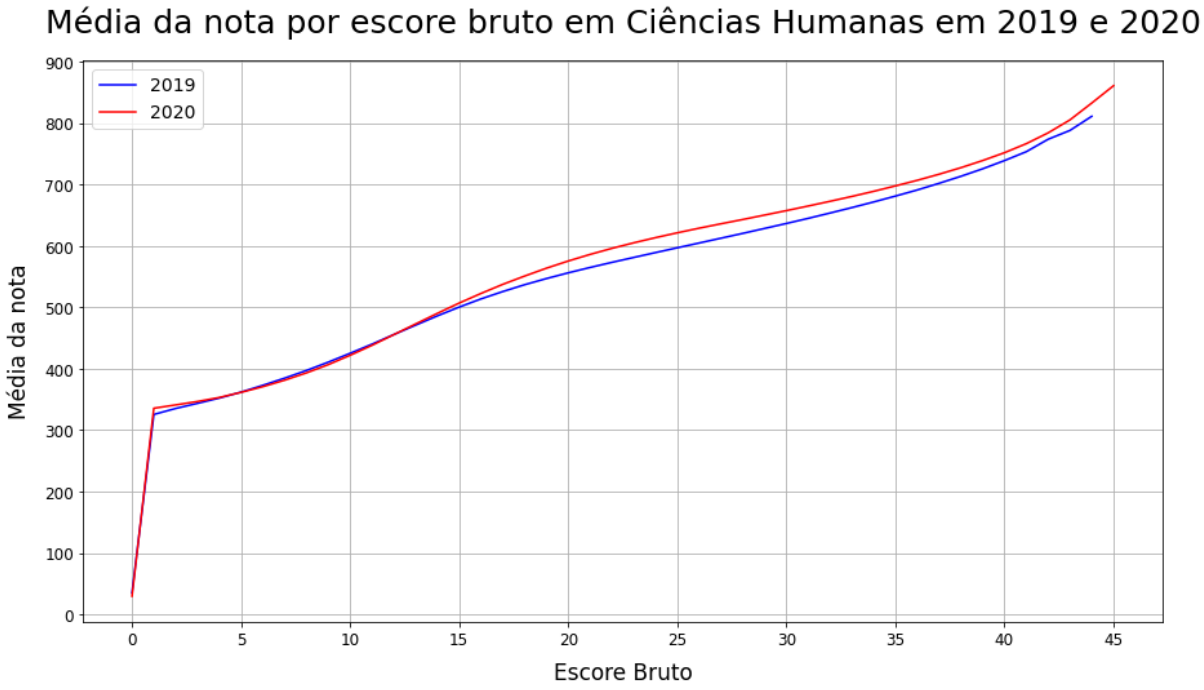

Figura 5.45: Média da nota na prova de Ciências Humanas por escore bruto em 2019 e 2020.

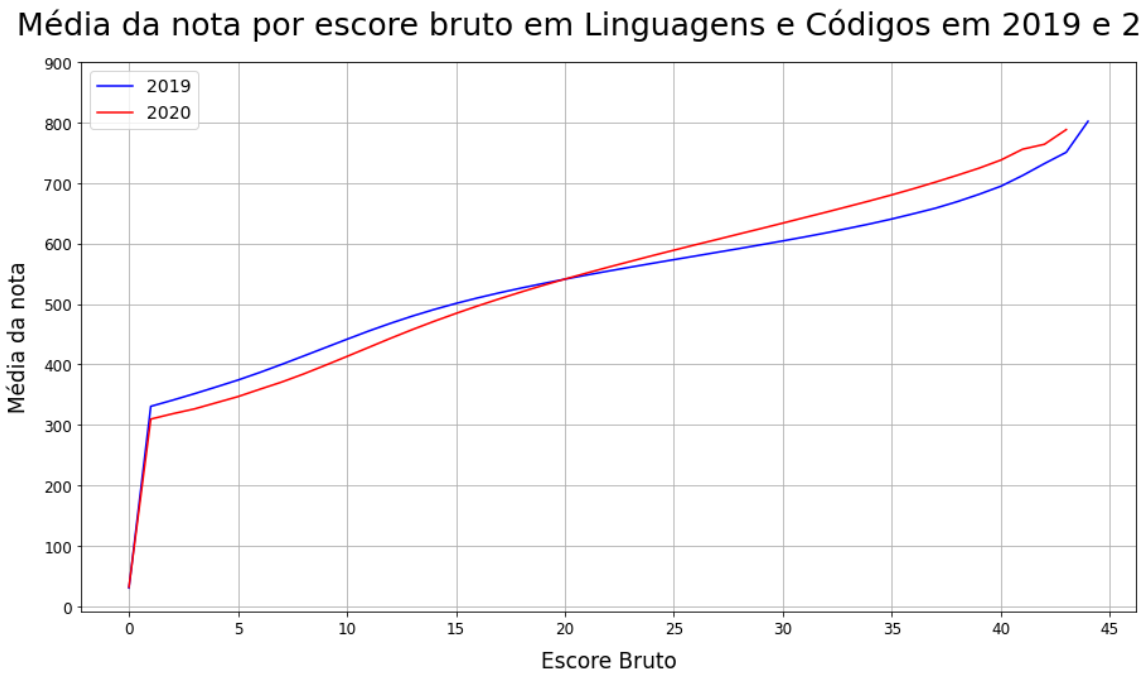

Figura 5.46: Média da nota na prova de Linguages e Códigos por escore bruto em 2019 e 2020.

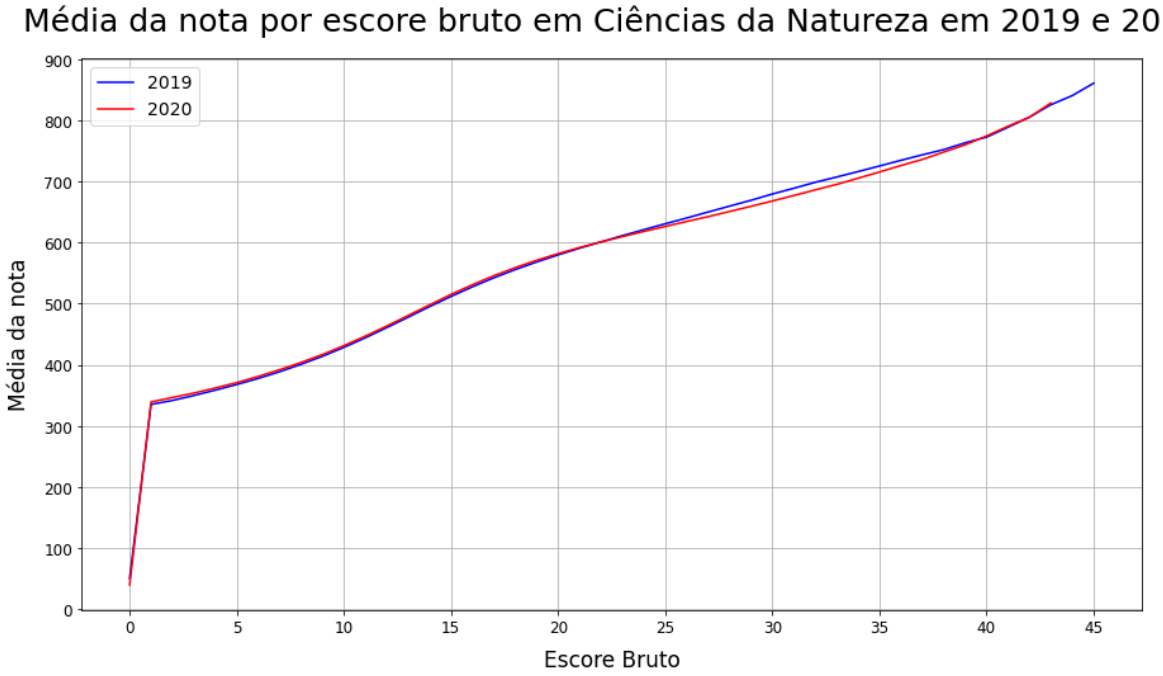

Figura 5.47: Média da nota na prova de Ciências da Natureza por escore bruto em 2019 e 2020.

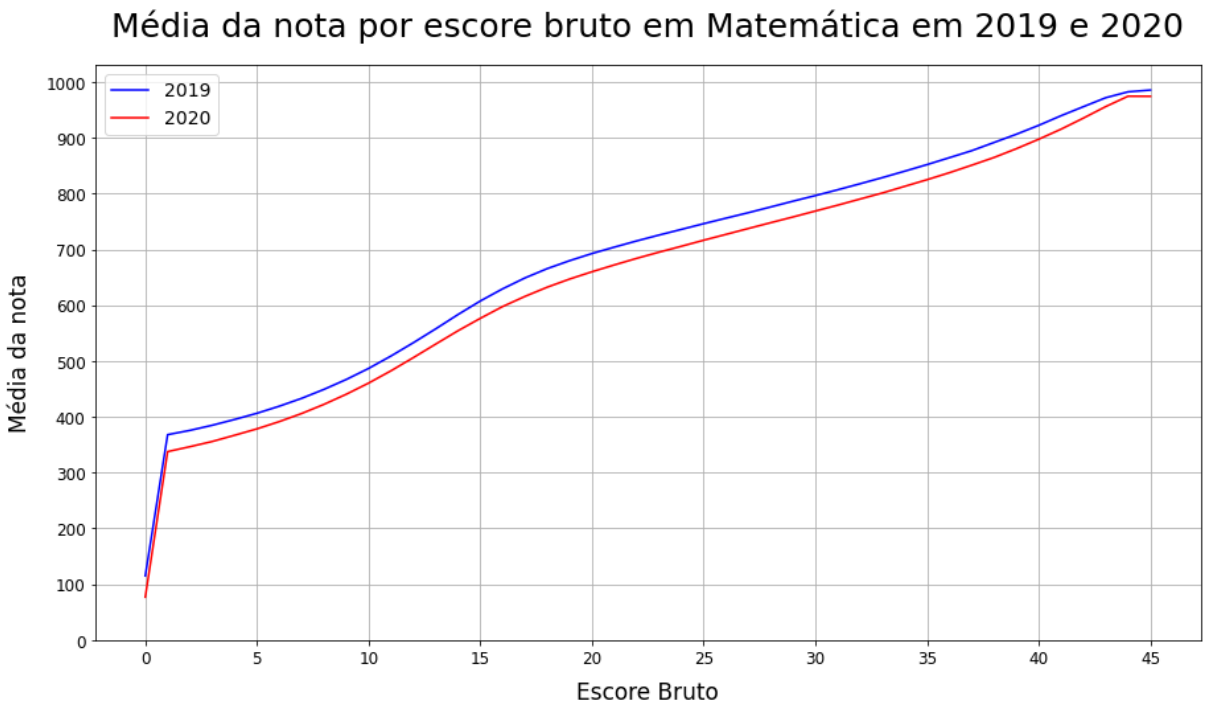

Figura 5.48: Média da nota na prova de Matemática por escore bruto em 2019 e 2020.

5.4 Análise das Notas por Cor/Raça e Classe de Renda

Ao comparar as médias das notas agrupadas pela renda e raça dos indivíduos, percebeu-se que as médias continuam aumentando conforme a renda aumenta, tanto numa perspectiva geral, considerando todas as categorias de cor/raça, quanto numa perspectiva local, considerando cada tipo de cor/raça individualmente. É interessante perceber que as médias das notas de cada grupo de cor/raça têm valores próximos entre grupos de cor/raça diferentes, para faixas de renda próximas. A Figura 5.49 apresenta os valores de referência desta análise. Nas Figuras 5.50 a 5.54 têm-se os gráficos para as provas de Linguagens e Códigos, Ciências Humanas, Ciências da Natureza, Matemática e Redação, respectivamente. Observa-se que as categorias de cor/raça Branca e Amarela possuem as maiores médias nas provas, seguidas das categorias de cor/raça Parda, Preta e Indígena (exceto na maior faixa de renda para as provas de Ciências Humanas e Linguagens e Códigos).

Cor/raça	Renda	Ciências da Natureza	Ciências Humanas	Linguagens e Códigos	Matemática	Redação
Branca	A	464.4	483.9	500.4	476.3	524.9
Branca	B	472.2	494.8	512.3	490.7	552.1
Branca	C	489.0	515.3	530.5	518.9	575.5
Branca	D	499.1	526.8	538.7	535.0	593.2
Branca	E	510.6	539.3	548.2	554.8	616.2
Branca	F	518.2	547.7	554.1	565.2	632.5
Branca	G	530.4	560.6	562.8	585.3	655.3
Branca	H	540.8	571.2	569.5	601.5	674.7
Branca	I	548.9	579.5	574.8	613.6	691.5
Branca	J	554.7	586.1	579.4	622.7	704.4
Branca	K	559.4	590.9	582.5	630.5	711.0
Branca	L	563.2	594.8	584.6	635.6	716.0
Branca	M	565.5	596.7	586.0	641.8	721.9
Branca	N	569.9	601.6	589.3	649.5	729.2
Branca	O	574.4	607.8	592.8	657.7	738.5
Branca	P	578.6	612.2	595.5	667.1	747.4
Branca	Q	583.7	618.8	599.9	681.5	750.4

Cor/raça	Renda	Ciências da Natureza	Ciências Humanas	Linguagens e Códigos	Matemática	Redação
Preta	A	445.0	460.9	480.4	446.6	488.9
Preta	B	454.9	473.3	495.4	461.6	517.0
Preta	C	469.8	492.7	513.4	485.2	536.8
Preta	D	478.4	503.8	522.6	498.1	550.4
Preta	E	488.9	516.4	532.8	518.8	572.8
Preta	F	495.1	524.0	538.1	523.9	588.3
Preta	G	506.0	537.0	547.2	541.6	609.2
Preta	H	518.2	548.5	554.7	559.0	631.3
Preta	I	525.0	558.5	561.3	568.1	646.8
Preta	J	530.6	565.0	565.1	577.1	665.4
Preta	K	535.8	573.9	569.2	586.3	666.3
Preta	L	536.5	570.5	569.3	582.5	669.6
Preta	M	540.0	571.2	569.8	589.0	673.6
Preta	N	549.3	587.7	579.9	608.0	698.2
Preta	O	549.7	587.8	577.9	611.8	696.9
Preta	P	558.7	594.9	583.4	616.8	714.5
Parda	Q	557.8	595.6	583.3	627.3	704.1

Cor/raça	Renda	Ciências da Natureza	Ciências Humanas	Linguagens e Códigos	Matemática	Redação
Parda	A	445.1	459.3	477.1	449.2	494.7
Parda	B	456.3	473.4	492.3	467.3	524.3
Parda	C	473.6	494.4	512.2	493.8	549.2
Parda	D	483.8	507.0	522.3	509.9	565.5
Parda	E	495.1	519.8	532.5	530.0	586.7
Parda	F	502.3	528.0	538.7	538.3	605.0
Parda	G	514.5	541.9	548.3	557.9	629.3
Parda	H	526.3	553.8	555.8	574.7	653.4
Parda	I	533.6	562.8	562.3	586.7	671.3
Parda	J	539.5	570.1	567.1	595.2	680.8
Parda	K	545.1	574.9	570.4	605.2	690.9
Parda	L	548.7	578.6	571.4	609.1	698.7
Parda	M	552.6	582.2	574.5	614.2	702.2
Parda	N	556.9	587.3	578.8	625.3	713.6
Parda	O	561.8	593.6	582.0	632.4	719.3
Parda	P	567.5	599.7	586.1	643.3	728.2
Amarela	Q	571.8	604.4	588.6	658.0	737.5

Cor/raça	Renda	Ciências da Natureza	Ciências Humanas	Linguagens e Códigos	Matemática	Redação
Amarela	A	447.5	459.9	480.1	451.3	499.7
Amarela	B	458.9	473.4	495.8	468.9	530.1
Amarela	C	475.9	493.9	514.1	496.9	554.9
Amarela	D	488.3	507.7	524.6	518.9	578.1
Amarela	E	507.0	528.0	537.2	547.7	598.8
Amarela	F	515.6	536.3	543.9	564.1	621.6
Amarela	G	532.1	555.8	557.3	589.1	655.3
Amarela	H	546.8	567.4	564.9	614.2	669.2
Amarela	I	559.1	580.9	572.0	635.6	699.5
Amarela	J	565.1	586.4	575.8	644.4	700.7
Amarela	K	574.2	597.6	582.4	663.1	712.8
Amarela	L	577.8	599.6	581.7	666.1	727.5
Amarela	M	582.5	603.2	584.3	678.7	718.7
Amarela	N	585.4	608.0	590.2	681.0	730.1
Amarela	O	586.4	607.7	588.9	683.2	726.7
Amarela	P	594.6	614.3	591.7	693.0	727.7
Amarela	Q	602.7	630.7	605.6	718.1	752.9

Cor/raça	Renda	Ciências da Natureza	Ciências Humanas	Linguagens e Códigos	Matemática	Redação
Indígena	A	430.2	442.0	453.7	431.8	435.7
Indígena	B	443.3	456.4	472.7	447.3	480.1
Indígena	C	459.4	476.1	495.5	473.4	505.5
Indígena	D	469.9	488.0	503.6	484.0	530.0
Indígena	E	476.0	489.8	504.0	498.3	541.1
Indígena	F	484.0	508.1	516.3	500.9	561.6
Indígena	G	496.1	514.0	529.2	525.5	582.1
Indígena	H	501.5	526.3	532.6	540.5	594.1
Indígena	I	513.7	527.8	536.1	537.3	614.2
Indígena	J	524.3	552.4	554.5	555.4	639.7
Indígena	K	532.2	541.3	542.0	554.1	639.4
Indígena	L	520.1	546.2	538.0	573.2	670.0
Indígena	M	509.8	526.2	531.9	577.9	572.8
Indígena	N	500.6	527.0	536.4	530.3	634.5
Indígena	O	533.7	558.7	550.5	573.5	625.9
Indígena	P	544.9	579.0	571.7	606.8	680.0
Indígena	Q	561.6	603.4	585.8	661.1	666.0

Figura 5.49: Média das notas agrupadas por cor/raça e renda em 2020.

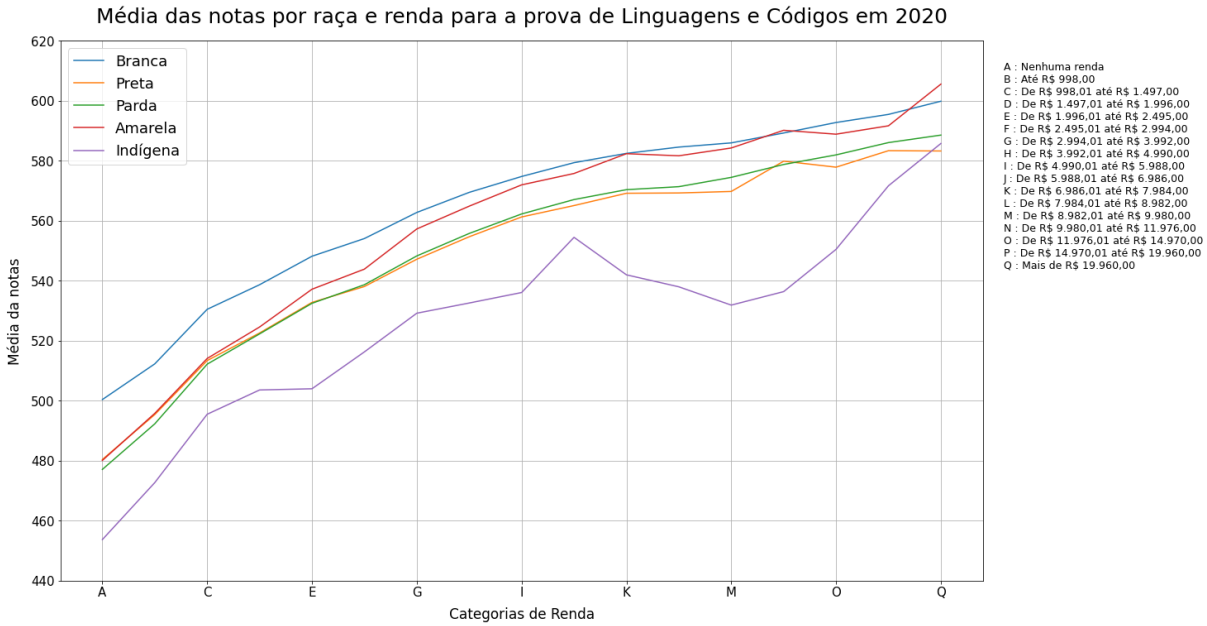

Figura 5.50: Média das notas de Linguagens e Códigos por renda e raça em 2020.

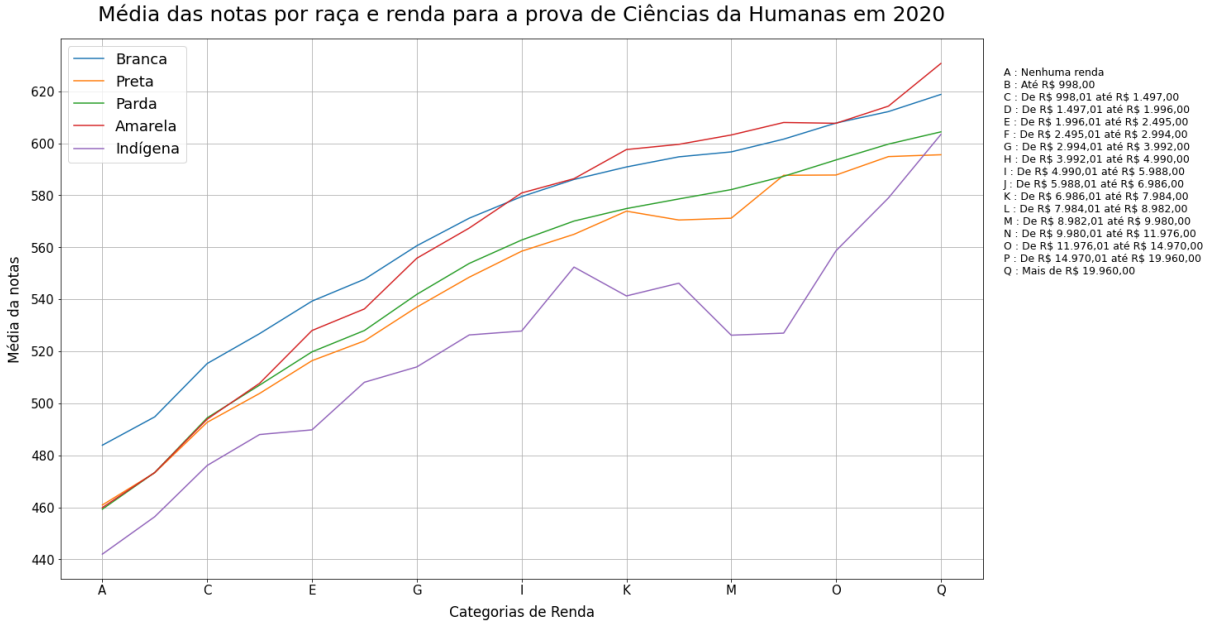

Figura 5.51: Média das notas de Ciências Humanas por renda e raça em 2020.

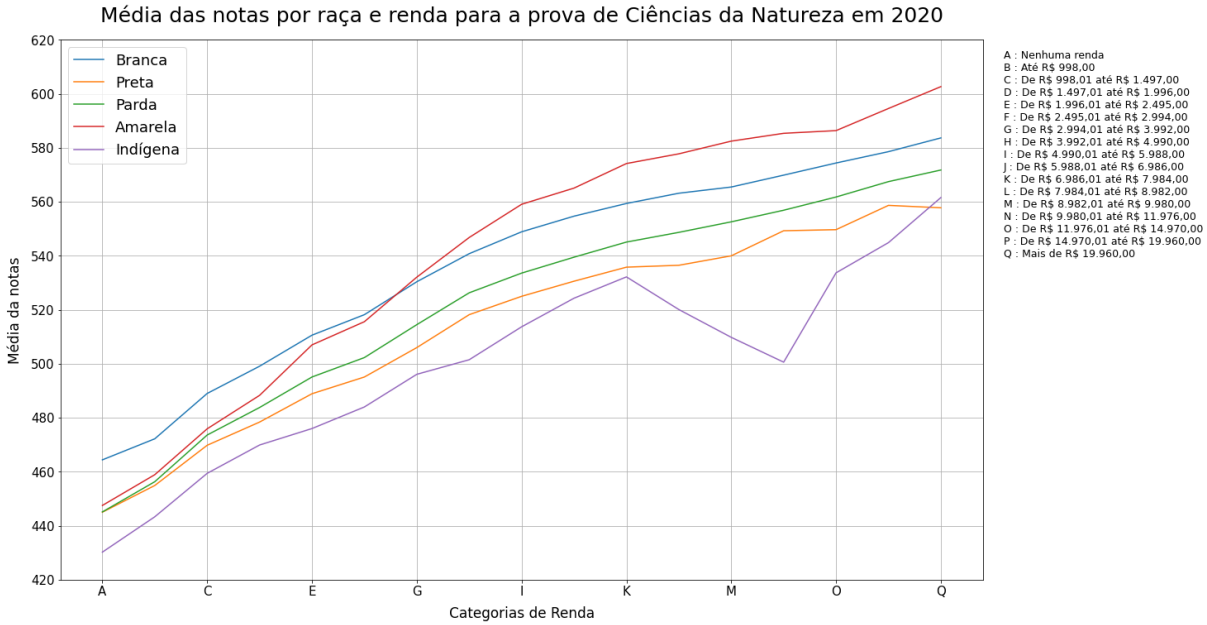

Figura 5.52: Média das notas de Ciências da Natureza por renda e raça em 2020.

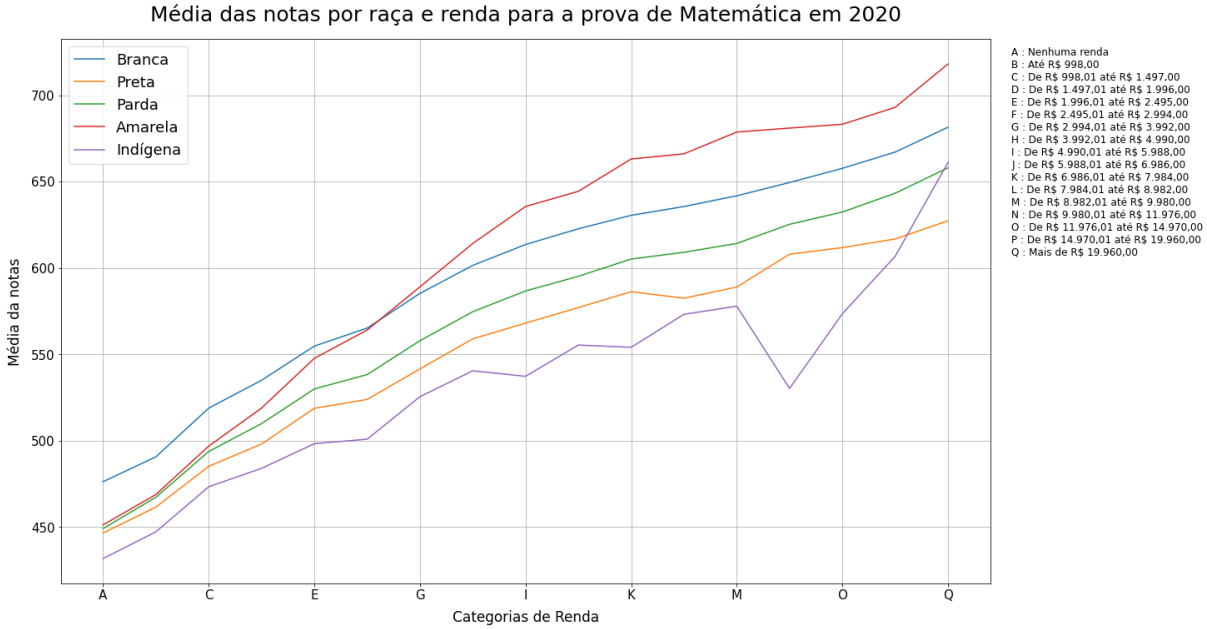

Figura 5.53: Média das notas de Matemática por renda e raça em 2020.

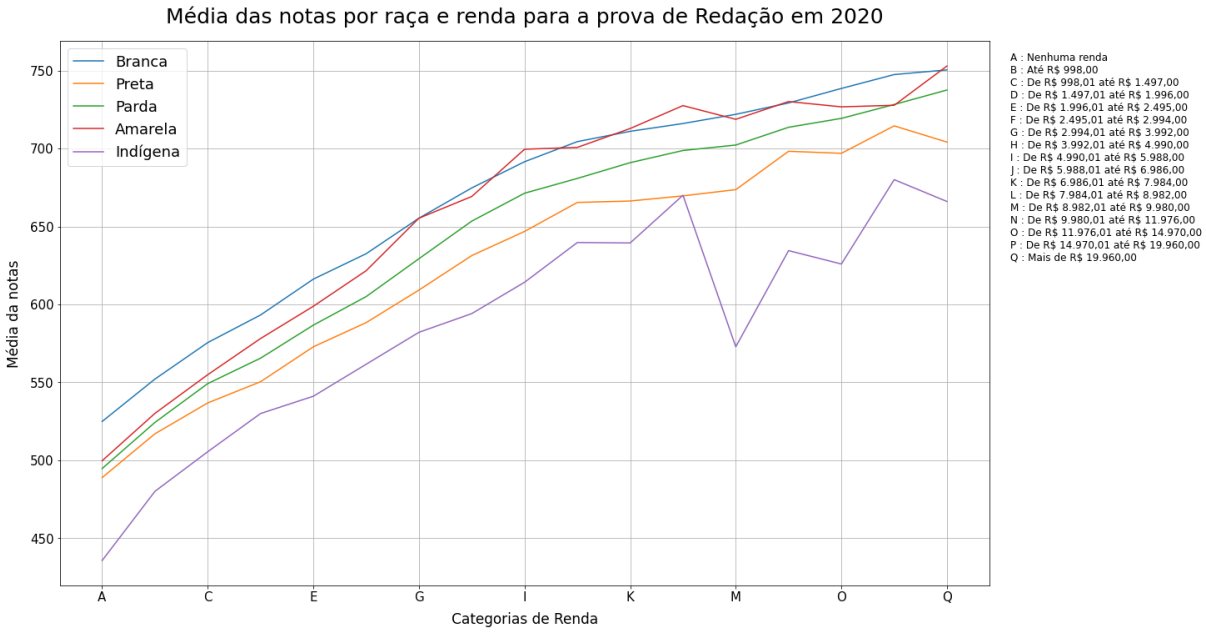

Figura 5.54: Média das notas de Redação por renda e raça em 2020.

5.5 Efeito do Índice de Desenvolvimento Humano Municipal na Nota do Enem

As próximas análises são referente as médias das notas de candidatos presentes nas duas provas em 2020, levando em conta a Unidade da Federação em que a prova foi realizada. Para isso, foi utilizada a plataforma Atlas do Desenvolvimento Humano no Brasil. Esta plataforma é uma ferramenta de divulgação de informações sobre o desenvolvimento humano no país, oferecendo informações estatísticas que evidenciam características e desigualdades sociais no território brasileiro. Nela é possível consultar o Índice de Desenvolvimento Humano Municipal (IDHM), que é uma medida composta pelas mesmas três dimensões (longevidade, educação e renda) do índice de Desenvolvimento Humano global, mas com uma metodologia aplicada ao contexto brasileiro e à disponibilidade de indicadores nacionais. A última versão dos dados disponível no site é a de 2017, por isso foi escolhida para a comparação. [2]

A Figura 5.55 mostra as médias das notas por estado, e a Figura 5.56 apresenta os dados do IDHM dos estados em 2017. O Índice de Desenvolvimento Humano Municipal é considerado muito alto se está entre 0,800 e 1,000; alto, se está entre 0,700 e 0,799; médio, para valores entre 0,600 e 0,699; baixo, na faixa de 0,500 a 0,599; e muito baixo, entre os valores de 0,000 e 0,499. Comparando as duas figuras, observa-se que muitos estados que possuem valores de Índice de Desenvolvimento Humano Municipal mais altos, tem também médias das notas no Enem 2020 maiores que os estados que possuem valores de IDHM mais baixos. Observando os seis primeiros estados do IDHM, têm-se Distrito Federal, São Paulo, Santa Catarina, Rio de Janeiro, Paraná e Minas Gerais. As médias das notas nesses estados foram todas acima de 500 pontos. A relação não é linear, as médias das notas não obedecem exatamente as posições no ranking do IDHM, mas nota-se que existe uma relação. A Figura 5.57 apresenta um gráfico que relaciona as médias das notas das provas de 2020 com o IDHM dos estados. É possível observar a tendência descrita acima, com médias maiores para valores de IDHM mais altos, apesar de existirem algumas exceções. Interessante observar que nos pontos que não seguiram essa tendência, as médias de todas as provas foram afetadas de maneira semelhante, tanto para valores maiores quanto menores. Valores mais altos do IDHM significam que o estado possui indicadores altos a respeito da expectativa de vida, renda per capita e escolaridade. Considerando que foi visto uma influência positiva nas notas de maiores faixas de renda e grau de escolaridade dos pais, faz sentido que o IDHM e as notas dos estados sejam comparáveis.

UF	IDHM (2017)	Ciências da Natureza	Ciências Humanas	Linguagens e Códigos	Matemática	Redação
AC	0.719	464.8	483.5	502.9	473.4	549.5
AL	0.683	471.9	489.7	506.1	492.9	581.7
AM	0.733	467.4	492.0	492.4	482.3	555.5
AP	0.740	462.4	483.5	495.8	467.7	551.8
BA	0.714	476.6	495.9	511.6	494.1	577.3
CE	0.735	481.3	504.0	516.4	513.8	600.8
DF	0.850	503.6	531.2	541.6	534.7	597.8
ES	0.772	505.2	528.8	534.7	540.6	608.1
GO	0.769	491.3	513.7	526.5	517.9	599.8
MA	0.687	462.7	481.0	497.0	475.7	562.3
MG	0.787	508.1	535.9	542.3	550.6	624.3
MS	0.766	488.4	510.2	524.3	513.4	571.8
MT	0.774	483.0	505.1	516.9	506.0	571.9
PA	0.698	467.9	486.2	499.2	478.3	571.7
PB	0.722	477.9	498.0	511.1	501.9	591.0
PE	0.727	482.0	500.8	518.4	510.6	587.9
PI	0.697	471.2	489.3	502.4	492.1	584.6
PR	0.792	508.1	536.9	542.0	547.0	590.1
RJ	0.796	501.5	530.6	542.7	537.7	617.9
RN	0.731	486.9	507.6	521.0	515.4	597.2
RO	0.725	476.6	493.6	507.2	492.3	553.7
RR	0.752	478.5	500.6	513.3	492.1	553.3
RS	0.787	497.4	530.8	541.3	537.0	597.6
SC	0.808	509.9	539.3	542.9	552.1	602.3
SE	0.702	481.5	498.7	510.3	498.2	605.8
SP	0.826	512.7	542.5	552.0	558.4	611.1
TO	0.743	469.4	487.8	503.7	488.9	561.7

Figura 5.55: Média das notas por estado em 2020.

Territorialidade	Posição IDHM	IDHM	Posição IDHM Renda	IDHM Renda	Posição IDHM Educação	IDHM Educação	Posição IDHM Longevidade	IDHM Longevidade
Distrito Federal	1	0,85	1	0,89	2	0,804	1	0,859
São Paulo	2	0,826	5	0,854	1	0,828	2	0,796
Santa Catarina	3	0,808	3	0,866	3	0,779	4	0,783
Rio de Janeiro	4	0,796	4	0,858	6	0,763	6	0,769
Paraná	5	0,792	9	0,843	5	0,764	5	0,771
Minas Gerais	6	0,787	2	0,875	8	0,753	10	0,741
Rio Grande do Sul	6	0,787	7	0,849	12	0,729	3	0,787
Mato Grosso	7	0,774	10	0,825	7	0,758	9	0,742
Espírito Santo	8	0,772	6	0,85	11	0,732	11	0,74
Goiás	9	0,769	11	0,822	9	0,74	8	0,747
Mato Grosso do Sul	10	0,766	8	0,847	15	0,71	7	0,748
Roraima	11	0,752	22	0,781	4	0,771	12	0,706
Tocantins	12	0,743	16	0,811	13	0,727	14	0,696
Amapá	13	0,74	13	0,82	15	0,71	15	0,695
Ceará	14	0,735	14	0,818	14	0,717	21	0,676
Amazonas	15	0,733	20	0,786	10	0,735	18	0,682
Rio Grande do Norte	16	0,731	7	0,849	19	0,677	19	0,68
Pernambuco	17	0,727	12	0,821	17	0,685	18	0,682
Rondônia	18	0,725	23	0,776	16	0,703	13	0,699
Paraíba	19	0,722	17	0,809	20	0,671	16	0,694
Acre	20	0,719	12	0,821	18	0,682	22	0,664
Bahia	21	0,714	15	0,812	23	0,654	17	0,685
Sergipe	22	0,702	18	0,799	24	0,64	20	0,677
Pará	23	0,698	19	0,788	22	0,661	24	0,654
Piauí	24	0,697	24	0,771	21	0,666	23	0,66
Maranhão	25	0,687	25	0,764	18	0,682	26	0,623
Alagoas	26	0,683	21	0,783	25	0,636	25	0,639

Figura 5.56: IDHM dos estados em 2017. Fonte: Atlas Brasil, 2022 [2].

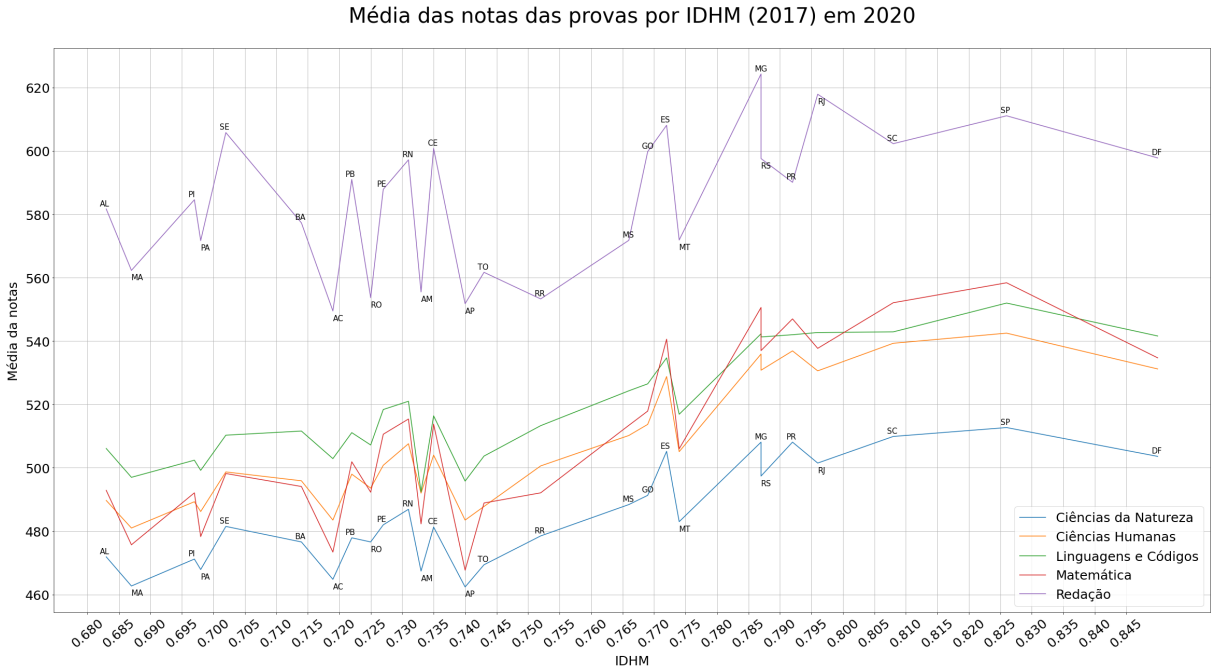

Figura 5.57: Média das notas das provas por IDHM (2017) em 2020.

Observando os valores do IDHM nos pontos que parecem não seguir a tendência na Figura 5.57 e fazendo uma consulta à Figura 5.55, nota-se que alguns desses valores pertencem a estados que compartilham a mesma região do país. Então, foi feita uma análise das médias das notas nas provas de 2020 considerando cada região do Brasil. Os resultados podem ser vistos na Figura 5.58. Percebe-se que as regiões Sul e Sudeste possuem as maiores médias nas provas, seguidas da região Centro-Oeste. Logo depois vem a região Nordeste, e por último a região Norte, que tem as menores médias do país.

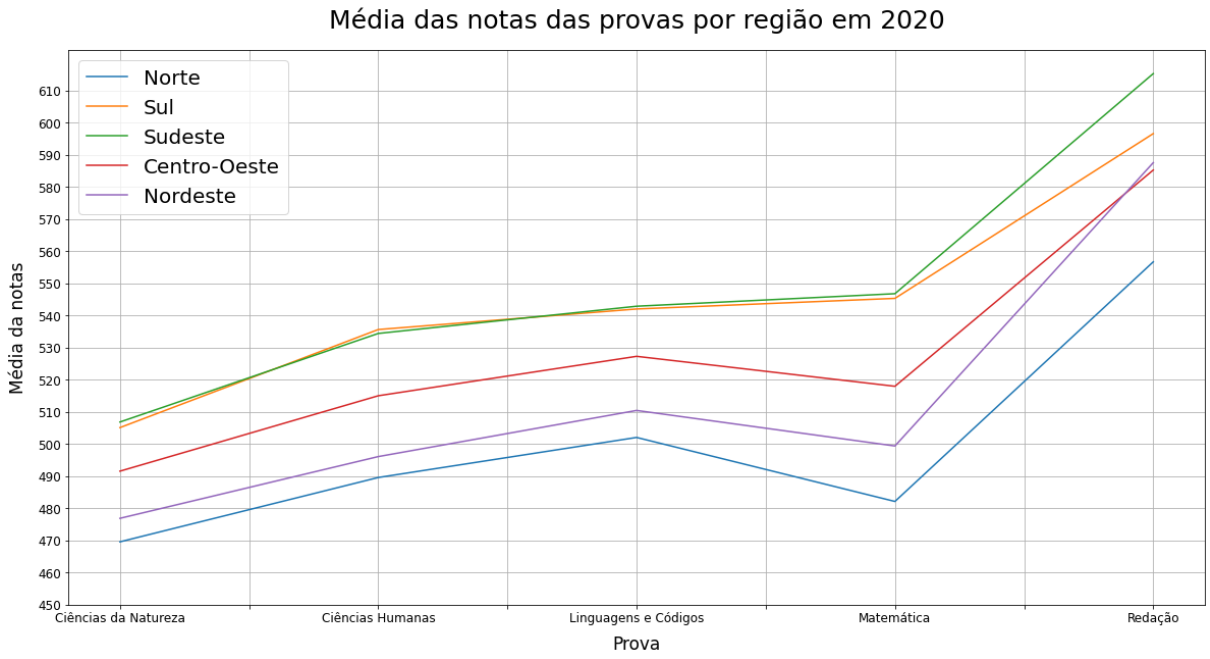

Figura 5.58: Média das notas das provas por região em 2020.

Analisando as informações obtidas até o momento, pode-se levantar a hipótese de que o aumento nas médias de algumas notas e no escore bruto de 2020, deve-se a redução de grande parte da população que costuma participar do Enem. Principalmente, considerando que grande parte desses ausentes são de grupos socioeconômicos que têm a tendência de ter um pior desempenho no exame, pois já tem menos oportunidades como visto no Capítulo 2.

5.6 Classificação com o Algoritmo JRip

Para a seguinte análise foi utilizado o algoritmo de classificação JRip, do Weka, em uma amostra dos microdados de 2020, para tentar entender quais variáveis podem influenciar na decisão de se ausentar em um dia de prova. O JRip foi escolhido por produzir regras fáceis de visualizar e entender, que podem ser utilizadas para políticas públicas. Após as modificações da etapa de pré-processamento, foi selecionada a classe de predição e executado o algoritmo. Como o objetivo era tentar entender quais variáveis podem influenciar na decisão de não ir um dia de prova, a classe de predição escolhida foi "TP_PRESENCA_MT", que diz respeito à presença na prova de Matemática. Ao final da execução, foram geradas 14 regras, que são listadas e traduzidas a seguir:

1. (TP_FAIXA_ETARIA = 2) => TP_PRESENCA_MT = 1 (1280.0/398.0)

- Se o inscrito tem 17 anos, então ele esteve presente na prova de Matemática.

2. (TP_ANO_CONCLUIU = 0) *and* (Q003 = D) *and* (TP_ST_CONCLUSAO = 2) => TP_PRESENCA_MT = 1 (205.0/62.0)

    - Se o inscrito não informou o ano de conclusão do ensino médio, e a ocupação do seu pai (ou homem responsável) é da categoria "D" (ver Figura 5.59. Por exemplo: Professor (de ensino fundamental ou médio), policial, militar de baixa patente, microempresário, trabalhador autônomo, entre outros.), e o inscrito está cursando e concluirá o Ensino Médio em 2020, então ele esteve presente na prova de Matemática.

3. (TP_FAIXA_ETARIA = 3) *and* (Q004 = D) => TP_PRESENCA_MT = 1 (267.0/81.0)

    - Se o inscrito tem 18 anos, e a ocupação da sua mãe (ou mulher responsável) é da categoria "D" (ver Figura 5.59. Por exemplo: Professora (de ensino fundamental ou médio), policial, militar de baixa patente, microempresária, trabalhadora autônoma ou por conta própria, entre outros.), então ele esteve presente na prova de Matemática.

4. (TP_ESTADO_CIVIL = 1) *and* (TP_FAIXA_ETARIA = 1) => TP_PRESENCA_MT = 1 (596.0/202.0)

    - Se o estado civil do inscrito é solteiro, e ele tem menos de 17 anos, então ele esteve presente na prova de Matemática

5. (TP_FAIXA_ETARIA = 3) *and* (Q010 = B) *and* (Q005 = 5) => TP_PRESENCA_MT = 1 (71.0/20.0)

    - Se o inscrito tem 18 anos, e na sua residência tem um carro, e nela moram 5 pessoas, então ele esteve presente na prova de Matemática.

6. (TP_FAIXA_ETARIA = 3) *and* (Q002 = E) => TP_PRESENCA_MT = 1 (392.0/164.0)

    - Se o inscrito tem 18 anos, e sua mãe completou o Ensino Médio mas não completou a Faculdade, então ele esteve presente na prova de Matemática.

7. (TP_FAIXA_ETARIA = 3) *and* (SG_UF_PROVA = BA) => TP_PRESENCA_MT = 1 (43.0/15.0)

- Se o inscrito tem 18 anos, e fez a prova na Bahia, ele esteve presente na prova de Matemática.

8. (TP_ESTADO_CIVIL = 1) *and* (TP_FAIXA_ETARIA = 4) *and* (Q004 = D) => TP_PRESENCA_MT = 1 (202.0/77.0)

    - Se o estado civil do inscrito é solteiro, e ele tem 19 anos, e a ocupação da sua mãe (ou mulher responsável) é da categoria "D" (ver Figura 5.59. Por exemplo: Professora (de ensino fundamental ou médio), policial, militar de baixa patente, microempresária, trabalhadora autônoma ou por conta própria, entre outros.), então ele esteve presente na prova de Matemática.

9. (TP_ESTADO_CIVIL = 1) *and* (TP_FAIXA_ETARIA = 4) *and* (Q002 = E) => TP_PRESENCA_MT = 1 (303.0/145.0)

    - Se o estado civil do inscrito é solteiro, e ele tem 19 anos, e sua mãe completou o Ensino Médio mas não completou a Faculdade, então ele esteve presente na prova de Matemática.

10. (TP_ESTADO_CIVIL = 1) *and* (TP_FAIXA_ETARIA = 3) *and* (Q002 = F) => TP_PRESENCA_MT = 1 (51.0/15.0)

    - Se o estado civil do inscrito é solteiro, e ele tem 18 anos, e sua mãe completou a Faculdade mas não completou a Pós-graduação, então ele esteve presente na prova de Matemática.

11. (TP_ESTADO_CIVIL = 1) *and* (Q022 = D) *and* (TP_FAIXA_ETARIA = 3) *and* (TP_COR_RACA = 3) => TP_PRESENCA_MT = 1 (63.0/26.0)

    - Se o estado civil do inscrito é solteiro, e tem três telefones celulares em sua residência, e tem 18 anos, e tem cor/raça parda, então ele esteve presente na prova de Matemática.

12. (TP_ESTADO_CIVIL = 1) *and* (Q002 = E) *and* (TP_FAIXA_ETARIA = 6) => TP_PRESENCA_MT = 1 (203.0/99.0)

    - Se o estado civil do inscrito é solteiro, e sua mãe completou o Ensino Médio mas não completou a Faculdade, e tem 21 anos, então ele esteve presente na prova de Matemática.

13. (TP_ESTADO_CIVIL = 1) *and* (TP_FAIXA_ETARIA = 5) *and* (Q005 = 3) *and* (Q003 = B) => TP_PRESENCA_MT = 1 (56.0/22.0)

- Se o estado civil do inscrito é solteiro , e ele tem 20 anos, e moram três pessoas em sua residência, e e a ocupação do seu pai (ou homem responsável) é da categoria “B” (ver Figura 5.59. Por exemplo: Empregado doméstico, motorista particular, faxineiro, vigilante, porteiro, carteiro, atendente de loja, auxiliar administrativo, entre outros.), então ele esteve presente na prova de Matemática.

14. => TP_PRESENCA_MT = 0 (6273.0/2090.0)

    - Se nenhuma das outras regras são satisfeitas, então o inscrito não esteve presente na prova de Matemática.

Dados da execução:

- Número total de instâncias: 10.005;
- Atributos: 20;
- Número de regras: 14;
- Instâncias corretamente classificadas 64,9075%;
- Instâncias incorretamente classificadas 35,0925%.

Q003	A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você. (Se ele não estiver trabalhando, escolha uma ocupação pensando no último trabalho dele).	A	Grupo 1: Lavrador, agricultor sem empregados, bóia fria, criador de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultor, pescador, lenhador, seringueiro, extrativista.
		B	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.
		C	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.
		D	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.
		E	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.
		F	Não sei.
Q004	A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da sua mãe ou da mulher responsável por você. (Se ela não estiver trabalhando, escolha uma ocupação pensando no último trabalho dela).	A	Grupo 1: Lavradora, agricultora sem empregados, bóia fria, criadora de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultora, pescadora, lenhadora, seringueira, extrativista.
		B	Grupo 2: Diarista, empregada doméstica, cuidadora de idosos, babá, cozinheira (em casas particulares), motorista particular, jardineira, faxineira de empresas e prédios, vigilante, porteira, carteira, office-boy, vendedora, caixa, atendente de loja, auxiliar administrativa, recepcionista, servente de pedreiro, repositora de mercadoria.
		C	Grupo 3: Padeira, cozinheira industrial ou em restaurantes, sapateira, costureira, joalheira, torneira mecânica, operadora de máquinas, soldadora, operária de fábrica, trabalhadora da mineração, pedreira, pintora, eletricista, encanadora, motorista, caminhoneira, taxista.
		D	Grupo 4: Professora (de ensino fundamental ou médio, idioma, música, artes etc.), técnica (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretora de imóveis, supervisora, gerente, mestre de obras, pastora, microempresária (proprietária de empresa com menos de 10 empregados), pequena comerciante, pequena proprietária de terras, trabalhadora autônoma ou por conta própria.
		E	Grupo 5: Médica, engenheira, dentista, psicóloga, economista, advogada, juíza, promotora, defensora, delegada, tenente, capitã, coronel, professora universitária, diretora em empresas públicas ou privadas, política, proprietária de empresas com mais de 10 empregados.
		F	Não sei.

Figura 5.59: Descrição dos possíveis valores para os atributos Q003 e Q004. Fonte: Microdados Enem 2020 [3].

Um primeiro ponto importante a se observar nas regras é que a predição de classe foi feita em função da presença na prova de Matemática (TP_PRESENCA_MT = 1), e não

da ausência. Isso aconteceu devido ao maior número de abstenções em relação a presenças no ano de 2020, resultando em mais instâncias com TP_PRESENCA_MT = 0 do que TP_PRESENCA_MT = 1. Então, o algoritmo JRip gerou um modelo com regras que tentam primeiro classificar uma instância como presente na prova, e caso nenhuma das regras sejam satisfeitas, a instância é classificada como ausente (representado pela última regra).

A ordem das regras indicam sua relevância no modelo gerado. Os números entre parênteses ao final das regras indicam sua taxa de acertos na classificação de instâncias. O primeiro valor representa a quantidade de instâncias que satisfizeram as condições do antecedente e foram corretamente classificadas. Já o segundo valor, indica quantas instâncias também satisfizeram as condições mas foram incorretamente classificadas.

Analisando algumas das regras geradas, pode-se levantar a hipótese de que muitos dos participantes que estiveram presentes no segundo dia de provas, eram jovens entre 17 e 19 anos, solteiros. Algumas regras como as de número 1, 2, 4, 6, 7, 10 e 11 sugerem que se tratam de pessoas que estão no final do ciclo do Ensino Médio. A regra número 5 pode indicar que ter um veículo na residência favorece a presença na prova. Interessante observar algumas regras como a número 2, 3 e 8, que associam além do estado civil e da idade, a classe "D" como ocupação dos pais ou responsáveis.

Dessa forma, as regras parecem apontar indiretamente componentes socioeconômicos. A regra 1 (de maior relevância segundo o JRip) apresenta a idade de 17 anos como indicativo de que o candidato esteve presente na prova de Matemática. Em geral, estes são candidatos que provavelmente puderam cursar o Ensino Médio de maneira regular e sem reprovações. Isto é mais comum nas classes de renda mais altas que nas mais baixas, devido à grande influência de fatores socioeconômicos no desempenho dos estudantes [37]. Além disso, as regras 2 e 3 (que são as próximas na ordem de relevância segundo o JRip) apresentam a profissão dos pais como indicativo de presença na prova de Matemática, o que tem relação com a escolaridade e classe de renda dos pais, e portanto sendo um forte indicador socioeconômico.

Entretanto, observando-se a matriz de confusão do modelo gerado, na Tabela 5.3, é possível perceber que as instâncias de inscritos ausentes no segundo dia de provas (TP_PRESENCA_MT = 0) foram melhor classificadas que as dos inscritos presentes. Para os inscritos ausentes (TP_PRESENCA_MT = 0), 4.051 foram corretamente classificados como ausentes, enquanto 1.455 foram incorretamente classificados como presentes e 1 incorretamente classificado como eliminado. Já para os inscritos presentes (TP_PRESENCA_MT = 1), 2.433 foram corretamente classificados como presentes, 2.049 foram incorretamente classificados como ausentes, e 1 incorretamente classificado como eliminado.

Tabela 5.3: Matriz de Confusão para o modelo gerado pelo algoritmo JRip.

Classificado como –>	**Ausente**	**Presente**	**Eliminado**
Ausente (0)	4051	1455	1
Presente (1)	2049	2443	1
Eliminado (2)	3	2	0

5.7 Considerações Finais

Nesta seção, será feito um resumo com os principais resultados obtidos nas análises. Foi visto que a edição de 2020 do Enem realmente teve um número de abstenções muito grande, com mais de 50% dos inscritos ausentes em cada um dos dia de provas. Com relação à raça, grande parte destas abstenções foram de inscritos de cor/raça Indígena (60%) e Preta (56,25%). Para as classes de renda, houve grande redução de candidatos das categorias de renda "C" (de R$ 998,01 ate R$ 1497) e "E" (de R$ 1996,01 ate R$ 2495,00), diminuição de 53,37% e 53,09%, respectivamente. Além de um aumento de 8% no total de inscritos sem renda em 2020. Sobre a escolaridade dos pais, se observou que a maioria dos ausentes eram candidatos com pais que não completaram a 4ª série/5º ano do ensino fundamental (27,17% para escolaridade do pai e 22,21% para escolaridade da mãe) e dos que completaram o ensino médio mas não a faculdade (22,49% para escolaridade do pai e 29,08% para escolaridade da mãe). Também foi visto que participantes com acesso à Internet possuem médias maiores em relação aos que não possuem. Decidiu-se não analisar dados relativos às escolas, devido à grande quantidade de dados ausentes.

Observou-se que a renda tem grande influência sobre as médias das notas dos participantes. Pode-se perceber que as médias das notas das provas têm a tendência de aumentarem conforme maior a faixa de renda do indivíduo. Comparando-se as diferenças nas médias das notas de 2019 e 2020, notou-se que as médias das notas das categorias de renda mais altas se distanciaram ainda mais de categorias de renda mais baixas em 2020, o que pode ser um indicativo de que as categorias de renda mais baixas foram mais prejudicadas que as mais altas. Como exemplo, observou-se que a diferença nas notas da prova de Matemática entre a classe que não possui nenhuma renda (classe "A") e a classe com maior renda (classe "Q") aumentou 12,99 pontos em 2020, com relação a 2019.

Analisando as notas com relação ao escore bruto nos anos de 2019 e 2020, foi possível observar o efeito da TRI nas notas de cada edição. Destaca-se que na prova de Matemática, para as mesmas quantidades de acertos, as notas de 2019 foram maiores que as de 2020.

Outra informação observada foi que a cor/raça também influencia no desempenho das provas. Além disso, quando se analisou a cor/raça junto da classe de renda, notou-se que as médias das notas de cada grupo de cor/raça têm valores próximos entre grupos de

cor/raça diferentes, para faixas de renda próximas.

Também foi visto que estados que possuem IDHM maiores tem a tendência de possuírem maiores médias nas provas. Além disso, as regiões Norte e Nordeste tiveram, em média, os piores resultados no exame.

Por fim, as regras obtidas pelo algoritmo de classificação JRip, apontaram indiretamente influência de componentes socioeconômicos na ausência na prova de Matemática em 2020.

Capítulo 6

Conclusão

O Exame Nacional do Ensino Médio (Enem) é o principal meio de acesso à educação superior atualmente. Os microdados do Enem, disponibilizados anualmente pelo Inep após a aplicação do exame, são muito úteis para avaliar a qualidade do ensino no país e auxiliar ações que tenham impacto nesta área. A realização da edição de 2020 do Enem sofreu com vários problemas devido à necessidade de adaptação de todos às novas situações causadas pela pandemia de COVID-19. Este trabalho teve como objetivo realizar uma análise comparativa utilizando os microdados do Enem dos anos de 2019 e 2020, para pesquisar o impacto da pandemia na realização do exame.

Assim, foram observadas nas análises indicativos de que a pandemia pode ter prejudicado muitos estudantes. O recorde de abstenções no Enem chama a atenção para as desigualdades sociais no país, e reforça a necessidade de ações nesta direção para reduzi-las. Notou-se que muitos dos inscritos que se ausentaram das provas, eram de parcelas da população menos favorecidas. Comparações das variações nas notas das categorias de renda entre os anos de 2019 e 2020 mostraram um aumento na diferença entre as notas de classes com maior renda em relação às com menor renda, o que pode ser um indicativo de que a pandemia tenha prejudicado mais as classes com menores rendas. No modelo de classificação gerado pelo algoritmo JRip para a presença na prova de Matemática, parece existir influência de fatores socioeconômicos nos casos em que o candidato não comparece à prova.

Os resultados permitiram conhecer melhor o perfil dos indivíduos que fizeram parte das edições de 2019 e 2020 do Enem. Foram encontradas evidências de que o desempenho dos candidatos tem grande influência de fatores socioeconômicos, e que parcelas da população estão em desvantagem. Embora algumas das análises não evidenciem com clareza uma piora no desempenho de algumas populações, a simples redução da quantidade de inscritos destas no exame indica um grave problema que terá consequências futuras para a sociedade. A falta ou o atraso da oportunidade de ingressar no ensino superior

e a evasão escolar podem aumentar as desigualdades sociais. Portanto, faz-se necessário considerar possibilidades de ações para auxiliar estas populações mais afetadas, de forma que minimizem esses impactos negativos e promovam maior acesso à oportunidades. É necessário que todas as parcelas da sociedade ocupem os espaços de maneira igualitária, para que suas vozes sejam ouvidas e suas necessidades levadas em conta.

Observou-se também a importância de adotar uma metologia para guiar o processo de KDD e suas atividades. O modelo CRISP-DM foi de grande ajuda na realização do trabalho. Entender o contexto do problema e dos dados é crucial para se obter resultados de qualidade nas análises.

Como sugestão de trabalhos futuros, pode ser interessante a construção de ferramentas ou plataformas interativas que utilizem estes dados abertos disponibilizados pelo Inep, não somente do Enem mas também de outras avaliações de desempenho da educação brasileira. Um ambiente *on-line* interativo, com os dados de outros anos da série histórica das avaliações, em que possam ser combinadas as variáveis de maneira simples e interativa, com retorno visual das ações, tornaria as análises muito mais agradáveis e acessíveis para diversos públicos. Algumas limitações dos dados impediram que outras variáveis fossem analisadas, como tipo e localização das escolas. É sugerido então que sejam feitos novos trabalhos com os microdados do Enem, avaliando e combinando novas variáveis. As mudanças que estão sendo feitas no formato de apresentação nos dados podem limitar ou trazer novas possibilidades de análise para o futuro.

Os dados educacionais proporcionam muitas oportunidades para se compreender melhor o cenário da educação, os contextos em que vivem os estudantes e melhores maneiras de se conduzir os processos de ensino e aprendizagem. É necessário que trabalhos e pesquisas sobre estes dados continuem sendo realizados, pois podem ser muito úteis para tomar melhores decisões a respeito da educação no Brasil.

Referências

[1] *Matplotlib quick start guide*. https://matplotlib.org/stable/tutorials/introductory/quick_start.html, acesso em 28/02/2023. ix, 34

[2] *Atlas do desenvolvimento humano no brasil*. http://www.atlasbrasil.org.br/ranking, acesso em 06/10/2022. xii, 81, 83

[3] INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA: *Microdados do enem 2020*, 2020. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem, acesso em 29/09/2022. xii, 20, 88

[4] *Boletim observatório covid-19 Fiocruz traz análise de seis meses da pandemia no brasil*. https://portal.fiocruz.br/noticia/boletim-observatorio-covid-19-fiocruz-traz-analise-de-seis-meses-da-pandemia-no-brasil-0, acesso em 26/02/2023. 1

[5] *Legislação informatizada - constituição de 1988 - publicação original*. https://www2.camara.leg.br/legin/fed/consti/1988/constituicao-1988-5-outubro-1988-322142-publicacaooriginal-1-pl.html, acesso em 26/02/2023. 1

[6] *Plano nacional de educação - lei n° 13.005/2014*. https://pne.mec.gov.br/18-planos-subnacionais-de-educacao/543-plano-nacional-de-educacao-lei-n-13-005-2014, acesso em 26/02/2023. 1

[7] World Bank: *The covid-19 pandemic : Shocks to education and policy responses*. 2020. https://openknowledge.worldbank.org/handle/10986/33696, acesso em 2023-02-27. 1

[8] Albernaz, Ângela, Francisco Ferreira e Creso Franco: *Qualidade e eqüidade na educação fundamental brasileira*. junho 2002. 2, 10

[9] Agência Senado: *Datasenado mostra que maioria dos brasileiros apoia adiamento do enem*. https://www12.senado.leg.br/noticias/materias/2020/05/22/datasenado-mostra-que-maioria-dos-brasileiros-apoia-adiamento-do-enem, acesso em 18/01/2023. 2

[10] *Enem: Os obstáculos extras dos estudantes na prova de 2021, que tem 2ª etapa neste domingo*. https://www.bbc.com/portuguese/brasil-59425040, acesso em 18/01/2023. 2

[11] *Candidatos ao enem 2020 contam que pandemia de covid virou pressão extra para a hora da prova.* https://g1.globo.com/educacao/enem/2020/noticia/2021/01/14/candidatos-ao-enem-2020-contam-que-pandemia-de-covid-virou-pressao-extra-para-a-hora-da-prova.ghtml, acesso em 18/01/2023. 2

[12] *Jovens sem estrutura para estudar em casa temem por 'desvantagem' no enem: 'atrapalhou tudo'.* https://g1.globo.com/educacao/enem/2020/noticia/2020/04/15/jovens-sem-estrutura-para-estudar-em-casa-temem-por-desvantagem-no-enem-atrapalhou-tudo.ghtml, acesso em 18/01/2023. 2

[13] *Enem 2020: Defensoria pública da união entra com pedido para adiar provas do exame marcadas para janeiro.* https://g1.globo.com/educacao/enem/2020/noticia/2021/01/08/enem-2020-defensoria-publica-da-uniao-entra-com-pedido-para-adiar-a-prova-marcada-para-17-e-24-de-janeiro.ghtml, acesso em 18/01/2023. 2

[14] *Enem 2020: a menos de 7 dias da prova, ação judicial e entidades questionam se medidas adotadas contra a covid são suficientes.* https://g1.globo.com/educacao/enem/2020/noticia/2021/01/11/enem-2020-a-menos-de-7-dias-da-prova-acao-judicial-e-entidades-questionam-se-medidas-adotadas-contra-a-covid-sao-suficientes.ghtml, acesso em 18/01/2023. 2

[15] Agência Senado: *Ministro da educação pode ser convocado pelo senado para prestar esclarecimentos sobre falhas no enem.* https://www12.senado.leg.br/noticias/audios/2021/01/ministro-da-educacao-pode-ser-convocado-pelo-senado-para-prestar-esclarecimentos-sobre-falhas-no-enem, acesso em 18/01/2023. 2

[16] *Enem 2020 acumula recordes de abstenções.* https://guiadoestudante.abril.com.br/enem/enem-2020-acumula-recordes-de-abstencoes/, acesso em 18/01/2023. 2

[17] Agência Senado: *Mais da metade dos inscritos não comparece ao primeiro dia do enem 2020.* https://www12.senado.leg.br/noticias/audios/2021/01/mais-da-metade-dos-inscritos-nao-comparece-ao-primeiro-dia-do-enem-2020, acesso em 18/01/2023. 2

[18] *Abstenção do enem 2020 é de 55,3%; pedido de reaplicação deve ser feito a partir desta segunda.* https://g1.globo.com/educacao/enem/2020/noticia/2021/01/24/abstencao-do-enem-2020-e-de-553percent-24-milhoes-foram-aos-locais-de-prova-neste-domingo.ghtml, acesso em 18/01/2023. 2

[19] *Legislação informatizada - lei nº 378, de 13 de janeiro de 1937 - publicação original.* https://www2.camara.leg.br/legin/fed/lei/1930-1939/lei-378-13-janeiro-1937-398059-publicacaooriginal-1-pl.html, acesso em 08/09/2022. 5

[20] *História do Inep.* https://www.gov.br/inep/pt-br/acesso-a-informacao/institucional/historia, acesso em 08/09/2022. 5

[21] *Sobre o Inep*. https://www.gov.br/inep/pt-br/acesso-a-informacao/institucional/sobre, acesso em 08/09/2022. 5

[22] *História do Enem*. https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/historico, acesso em 09/09/2022. 5

[23] *Exame nacional do ensino médio (Enem)*. https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem, acesso em 08/09/2022. 5, 6

[24] *Microdados do Enem*. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem, acesso em 13/09/2022. 6

[25] Inep: *Microdados do enem 2019 leia-me*, 2019. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem, acesso em 29/09/2022. 6, 21, 23, 24, 25

[26] Inep: *Microdados do enem 2020 leia-me*, 2020. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem, acesso em 29/09/2022. 6, 23, 24, 25

[27] *Histórico da pandemia de COVID-19*. https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19, acesso em 16/09/2022. 7

[28] Ministério da Saúde: *Saúde Brasil 2020-2021: uma análise da situação de saúde diante da pandemia de covid-19, doença causada pelo coronavírus SARS-CoV-2*. 2022. https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/publicacoes-svs/vigilancia/saude-brasil-2020-2021_situacao-de-saude-diante-da-covid-19.pdf/view. 7

[29] Brasil: *Portaria nº 454, de 20 de março de 2020. declara, em todo o território nacional, o estado de transmissão comunitária do coronavírus (covid-19)*. Diário Oficial da República Federativa do Brasil, 2020. https://www.in.gov.br/en/web/dou/-/portaria-n-454-de-20-de-marco-de-2020-249091587, acesso em 2022-09-19. 7

[30] Brasil: *Lei nº 13.979, de 6 de fevereiro de 2020. dispõe sobre as medidas para enfrentamento da emergência de saúde pública de importância internacional decorrente do coronavírus responsável pelo surto de 2019*. Diário Oficial da República Federativa do Brasil, 2020. https://www.in.gov.br/en/web/dou/-/lei-n-13.979-de-6-de-fevereiro-de-2020-242078735, acesso em 2022-09-20. 7

[31] Brasil: *Decreto nº 10.282, de 20 de março de 2020. regulamenta a lei nº 13.979, de 6 de fevereiro de 2020, para definir os serviços públicos e as atividades essenciais.* Diário Oficial da República Federativa do Brasil, 2020. https://www2.camara.leg.br/legin/fed/decret/2020/decreto-10282-20-marco-2020-789863-publicacaooriginal-160165-pe.html, acesso em 2022-09-20. 7

[32] Brasil: *Portaria nº 343, de 17 de março de 2020. dispõe sobre a substituição das aulas presenciais por aulas em meios digitais enquanto durar a situação de pandemia do novo coronavírus - covid-19.* Diário Oficial da República Federativa do Brasil, 2020.

https://www.in.gov.br/en/web/dou/-/portaria-n-343-de-17-de-marco-de-2020-248564376, acesso em 2022-09-20. 7

[33] Brasil: *Parecer cne/cp nº: 5/2020*. Diário Oficial da República Federativa do Brasil, 2020. http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=145011-pcp005-20&category_slug=marco-2020-pdf&Itemid=30192, acesso em 2022-09-20. 8

[34] Comitê Gestor da Internet no Brasil: *Pesquisa sobre o uso das tecnologias de informação e comunicação nas escolas brasileiras : TIC Educação 2019*. Cetic.br, 1ªª edição, 2020. 8, 9, 61

[35] Comitê Gestor da Internet no Brasil: *Pesquisa sobre o uso das tecnologias de informação e comunicação nos domicílios brasileiros : TIC Domicílios 2019*. Cetic.br, 1ªª edição, 2020. 8, 9

[36] Nascimento, Paulo, Daniela Lima, Almeida Melo e Remi Castioni: *Nota técnica 88: Acesso domiciliar À internet e ensino remoto durante a pandemia*. 88, setembro 2020. 10

[37] Soares, José: *Qualidade e eqüidade na educação básica brasileira: fatos e possibilidades*. janeiro 2005. 10, 89

[38] Shaun, Ryan, Joazeiro Baker, Seiji Isotani, Adriana Maria e Joazeiro Carvalho: *Mineração de dados educacionais: Oportunidades para o brasil*. Revista Brasileira de Informática na Educação, 19:3–13, janeiro 2011. 10

[39] Stair, Ralph M. e George W. Reynolds: *Princípios de Sistemas de Informação*. Cengage Learning, 2015. Tradução da 11ª edição norte-americana. 12, 13

[40] Laudon, Kenneth C. e Jane P. Laudon: *Sistemas de Informação Gerenciais*. Prentice Hall, 2010. Tradução da 9ª edição. 12, 13

[41] Ackoff, R L: *From data to wisdom*. Journal of Applied Systems Analysis, 16:3–9, 1989. 12

[42] Setzer, Valdemar W.: *Dado, informação, conhecimento e competência*. https://www.ime.usp.br/~vwsetzer/dado-info.html, acesso em 04/09/2022. 13

[43] Han, Jiawei, Micheline Kamber e Jian Pei: *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., 3ªª edição, 2011. 13, 14, 15, 16, 17

[44] Goldschmidt, Ronaldo e Emmanuel Passos: *Data Mining: Um guia prático*. Elsevier, 2005. 4ª Tiragem. 13, 14, 15, 16

[45] Aggarwal, Charu C.: *Data Mining: The Textbook*. Springer Publishing Company, Incorporated, 2015. 14

[46] Fayyad, Usama, Gregory Piatetsky-Shapiro e Padhraic Smyth: *Knowledge discovery and data mining: Towards a unifying framework*. Em *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, página 82–88. AAAI Press, 1996. 14, 15

[47] Larose, Daniel T. e Chantal D. Larose: *Discovering Knowledge in Data*. John Wiley Sons, Ltd, 2ª edição, 2014. 15, 16

[48] Foundation, Python Software: *Python 3.10.6 documentation*. https://docs.python.org/3/, acesso em 06/09/2022. 16

[49] team the pandas development: *pandas documentation*. https://pandas.pydata.org/docs/index.html, acesso em 06/09/2022. 16

[50] team, The Matplotlib development: *Matplotlib: Visualization with python*. https://matplotlib.org, acesso em 06/09/2022. 16

[51] *Jupyterlab: A next-generation notebook interface*. https://jupyter.org, acesso em 06/09/2022. 16

[52] *Getting started with anaconda*. https://docs.anaconda.com/anaconda/user-guide/getting-started/, acesso em 06/09/2022. 16

[53] Frank, Eibe, Mark A. Hall e Ian H. Witten: *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 4ª edição, 2016. 17

[54] *Class jrip*. https://weka.sourceforge.io/doc.dev/weka/classifiers/rules/JRip.html, acesso em 28/02/2023. 17

[55] Cohen, William W.: *Fast effective rule induction*. Em *Machine Learning Proceedings 1995*. Morgan Kaufmann, 1995, ISBN 978-1-55860-377-6. https://www.sciencedirect.com/science/article/pii/B9781558603776500232. 17

[56] Fürnkranz, Johannes e Gerhard Widmer: *Incremental reduced error pruning*. Em Cohen, William W. e Haym Hirsh (editores): *Machine Learning Proceedings 1994*. Morgan Kaufmann, 1994. https://www.sciencedirect.com/science/article/pii/B9781558603356500179. 17

[57] Moretin, Pedro A. e Wilton de O. Bussab: *Estatística básica*. Saraiva, 7ª edição, 2012. 17, 18

[58] Chapman, Peter, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas P. Reinartz, Colin Shearer e Richard Wirth: *Crisp-dm 1.0: Step-by-step data mining guide*. 2000. 19

[59] INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA: *Microdados do enem 2019*, 2019. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem, acesso em 29/09/2022. 20

[60] *Enem e enade têm novo conjunto de microdados publicados*. https://www.gov.br/inep/pt-br/assuntos/noticias/institucional/enem-e-enade-tem-novo-conjunto-de-microdados-publicados, acesso em 28/09/2022. 20

[61] Departamento de Ciência da Computação da Universidade Federal de Minas Gerais: *Ted 8750 - price privacidade nos censos educacionais. termo de execução descentralizada entre o instituto nacional de estudos e pesquisas educacionais anísio teixeira e a universidade federal de minas gerais*. https://download.inep.gov.br/microdados/TED_8750-UFMG.pdf, acesso em 29/09/2022. 21

[62] Controladoria-Geral da União (CGU): *Nota técnica nº 1136/2022/cgat/dtc/stpc*, 2022. https://download.inep.gov.br/institucional/nota_tecnica_CGU_1136_2022.pdf, acesso em 29/09/2022. 21

[63] *Posicionamento público de entidades sobre exclusão de dados do censo escolar pelo inep*. https://www.anped.org.br/news/posicionamento-publico-de-entidades-sobre-exclusao-de-dados-do-censo-escolar-pelo-inep, acesso em 29/09/2022. 22

[64] Conroy, Ronán M.: *Sample size a rough guide*. 38