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
2022

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. Dr. Banca 1 Prof. Dr. Banca 2
CIC/UnB CIC/UnB

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

Brasília, 06 de Agosto de 2022

Dedicatória

Dedico este trabalho a minha familia, 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, desta vez especificamente sobre o ano de 2020. 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. Foram obtidas evidências 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 têm grande 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 of the tests, this time specifically about the year 2020. 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. 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 have a great influence on the candidates' performance 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.

IBM International Business Machines Corporation.

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.

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 Educa- ção 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

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, além de provocar medo e outras preocupações nos candidatos que pretendiam realizar o exame. Os impactos da pandemia e consequentes decisões do governo fizeram com que esta edição tenha tido o maior número de abstenções da história do exame. É inevitável realizar comparações com os anos anteriores e se perguntar quais serão as consequências para o futuro dos candidatos. Assim, 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 e produção de evidências educacionais 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.

1.3 Objetivos

O objetivo deste trabalho é realizar uma análise nos microdados do Enem, dos anos de 2019 e 2020, para comparar qual foi impacto, em termos das variáveis disponíveis, da pandemia sobre os canditados que realizaram a prova no ano de 2020, e sobre a realização

do exame em geral neste mesmo ano. Também procura-se identificar se existem grupos ou perfis de candidatos que foram mais prejudicados.

Capítulo 2

Contextualização do Problema

2.1 Inep

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) foi criado pela Lei n.º 378, de 13 de Janeiro de 1937 [1], 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), e é responsável pelas evidências educacionais, atuando em três esferas: avaliações e exames educacionais; pesquisas estatísticas e indicadores educacionais; e gestão do conhecimento e estudos. [2] [3] Como exemplos tem-se o Censo escolar e o Enem, este último cujos dados serão importantes para este trabalho.

2.2 Enem

O Exame Nacional do Ensino Médio (Enem) 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 e as matrizes de avaliação foram reformuladas, e o exame passou a ser realizado em dois dias. Em 2014, o exame também passou a ser aceito em algumas universidades de Portugal. [4] [5]

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. [5] São aplicadas 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. [6]

Em 2020, devido a pandemia de COVID-19, as provas 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. [7]

Também em 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. Espera-se que este novo formato ganhe espaço e seja ampliado de forma progressiva. [7]

2.3 Pandemia de COVID-19 e seu impacto na educação

2.3.1 Histórico da Pandemia

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. [8]

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. [9] 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 [10]. No dia 22 de março de 2020, todas as unidades federativas do país já haviam notificado casos da doença. [9]

Diante do agravamento do cenário mundial de infecção pela doença, foram necessárias adotar medidas de enfrentamento, conforme previstas pela Lei 13.979, 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. [11] Essas medidas foram regulamentadas posteriormente por outros decretos, como o nº 10.282, de 20 de março de 2020, que definiu quais serviços seriam considerados essenciais e excluídos da possibilidade de interrupção. [12] 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 [13] do Ministério da Educação, de 17 de março de 2020, resolve autorizar, em carater 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 [14], 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. [14]

2.3.2 Discrepâncias no acesso ao ensino remoto

Segundo a pesquisa TIC Educação 2019 [15], edição de uma pesquisa realizada anualmente pelo Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informa-

ção (Cetic.br), 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 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 à Tecnologias da informação e comunicação (TIC). Segundo outra pesquisa, a TIC Domicílios [16], em 2019 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 [15], 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, 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. [15]

Ainda segundo a mesma pesquisa, 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 podem ter ficado sem condições de acessar a rede. [15]

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, 82% dos domicílios localizados em áreas rurais não possuíam computadores e 48% não contavam com acesso à Internet. [16] 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. [15]

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. [15]

Em setembro de 2021, foi divulgada a nota técnica nº 88 [17] 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 desempenho pior dos participantes

em relação aos anos anteriores, principalmente entre os canditados cercados por piores condições socioeconômicas, dado o impacto da pandemia e as dificuldades enfrentadas por todos, que afetou com maior ênfase essas parcelas da população.

2.4 Análise de dados educacionais

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. [18] 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. [19]

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 um aumento sobre o conhecimento dos estados emocionais dos alunos e sua relação com a aprendizagem. [20]

Por fim, é importante que sejam feitas análise 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 essas pessoas.

Capítulo 3

Referencial Teórico

3.1 Dados, Informação e Conhecimento

Para começar, é importante visitar a literatura para tentar esclarecer as definições de alguns termos que serão utilzados ao longo do 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 [21], dados são fatos brutos, que podem ser alfanuméricos, áudio, imagem ou vídeo. Sua definição concorda com a de Laudon e Laudon [22], 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 [22] informação é um dado moldado em formato significativo e útil para seres humanos. Já para Stair e Reynolds [21], 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 [23] 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 descrições, respostas para perguntas que começem 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 trasmissão de outro que já o possui, ou extraído pela experiência.

Valdemar W. Setzer [24] define dado como uma seqüência 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

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. [25]

Segundo Ralph Stair [21], 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.

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 [26]. Uma evolução natural da tecnologia da informação, com o objetivo de transformar as enormes quantidades de dados em informações, e utilizá-las como apoio na tomada de decisões. [25]

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 [22] 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 [27] 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 [28] 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 [26] concordam com estas definições.

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. [28]

A etapa de pré-processamento compreende atividades relacionadas à captação, à organização e ao tratamento dos dados. [26]. Segundo Han, Kamber e Pei [25], a etapa de preparação dos dados pode conter diversas atividades como:

1. Limpeza de dados: Para remover ruído e insconsistências 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, é realizada a busca por padrões e conhecimentos úteis no contexto da aplicação de KDD. [26] Na fase de avaliação 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. [28].

3.4 Técnicas de Mineração de dados

Todas 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. [26] No caso da segunda, ainda pode ser subdividida em predição, onde o sistema encontra padrões com a finalidade de prever o comportamento futuro de algumas entidades, e descrição, onde o sistema encontra padrões com a finalidade de apresentá-los a um usuário de forma compreensível para humanos. [28]

A seguir será brevemente descrito o objetivo de algumas técnicas de mineração de dados. Para cada uma delas, existem várias implementações de algoritmos.

3.4.1 Classificação

Classificação é o processo de encontrar um modelo ou função que descreva e diferencie classes de dados ou conceitos. [25]. De forma que seja possível mapear um conjunto de dados em categorias predefinadas, 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. [26] As funções ou modelos são derivados por meio da análise de um conjunto de dados de treinamento. [25]

3.4.2 Regressão

A técnica de regressão é parecida com a classificação, mas o mapeamento é feito para valores númericos 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. [26]

3.4.3 Sumarização

A sumarização consiste em encontrar descrições compactas para subconjuntos de dados. [28] 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. [26]

3.4.4 Clusterização

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 atribue rótulos para os grupos como é feito com as classes na segunda, e também não tenta classificar ou prever o valor da variável alvo. [29] Em muitos casos os rótulos de classes não existem inicialmente e a clusterização pode ser usada para gerá-los. [25]. Esse processo geralmente necessita que seja determinado qual o número de grupos a serem considerados para a segmentação dos dados. [26]

3.4.5 Detecção de desvios

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. [26] A maioria das técnicas de mineração de dados descarta esses valores atípicos como rúido ou exceções, mas em algumas aplicações como análise de fraudes, estes valores podem ser de grande interesse. Eles 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. [25]

3.4.6 Associação

O objetivo da técnica de associação é encontrar regras para quantificar o relacionamento entre dois ou mais atributos, usando parâmetros de suporte e confiança. [29] 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. [26]

3.5 Ferramentas de Mineração de dados

3.5.1 Python

Python 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. [30] As bibliotecas Pandas e Matplotlib são muito utilizados para atividades de mineração de dados.

1. **Pandas** 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. [31]

2. **Matplotlib** 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. [32]

3.5.2 Projeto R

R é uma linguagem de programação e um ambiente para computação estatística e gráficos. Fornece uma grande variedade de métodos para manipulação de dados, cálculos e apresentação de gráficos, facilitando as análises e apresentação dos resultados. Também pode ser extendido via pacotes. [33]

3.5.3 Jupyter

Jupyter 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. [34]

3.5.4 Anaconda

Anaconda é uma plataforma que contem diversas ferramentas utilizadas em ciência de dados. 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. [35]

3.5.5 WEKA

Weka 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. [36]

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 Média Aritmética, Mediana e Moda

A média aritmética é a soma das observações dividas pelo número total de observações. [37] Pode ser representado pela seguinte fórmula:

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

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

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. [37]

A moda é definida como a observação mais frequente no conjunto de valores. [37]

3.6.2 Desvio Padrão e Variância

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. [37]

3.6.3 População e Amostra

Segundo Moretin e Bussab [37], 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 se tiver, 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.

3.7 CRISP-DM

CRISP-DM [38], 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. [38]

No modelo de referência [38], 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.

Como fonte de consulta adicional, vale conferir também o guia da *IBM Corporation* sobre o modelo. [39]

Capítulo 4

Metodologia

A produção da análise feita neste trabalho foi orientada pelo modelo CRISP-DM, apresentado na seção 3.7. Seguindo as etapas de sua metodologia, até o momento foi realizado o entendimento do negócio, descrito nos Capítulos 1 a 2. Neste capítulo, serão apresentados os dados e informações referentes a realização das demais etapas.

4.1 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 [40] [41]. 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. 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). [42]

Segundo o Inep, 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. [43] No entanto, em um estudo técnico formalizado no Termo de Execução Descentralizada (TED) 8750, 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, 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. [44]

Ainda segundo o estudo técnico, 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 quseidentificadores 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. [44]

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) sugeriu as seguintes alterações: Exclusão do código da escola de conclusão do ensino médio; Exclusão das informações referentes aos pedidos de atendimento especiazlizado 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. [45] 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 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. [46]

4.2 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 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. 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.

Com relação a reestruturação da apresentação dos microdados citada na seção 4.1, foram realizadas algumas mudanças, segundo os arquivos leia-me dos dados [43] [7]. 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.

Também foram excluídos do conjunto de dados, 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. [43] 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. [7]

4.3 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. Sendo assim, serão primeiramente listados os objetivos específicos e as informações que são esperadas como resultado do processo.

- 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 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)
- Dentre as características dos indivíduos e das escolas a serem consideradas estão cor/raça, UF de realização da prova, tipo de escola (pública ou privada) e localização da escola (urbana ou rural).
- Fatores do questionário socioeconômico a serem considerados: renda; condição de acesso à internet; escolaridade dos pais.
- O tamanho da amostra escolhido será igual ao total da população, ou seja, todo o conjunto de dados.

Para compreender melhor as taredas de preparação e análise dos dados, precisa-se conhecer também as ferramentas utilzadas, 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
- Microsoft Windows 10 Home 21H2

A correta instalação da distribuição Anaconda já instala também todas as outras ferramentas e deixa tudo pronto para ser utilizado, 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, 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. 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 caracter utilizado como separador na base de dados. No caso, o caracter ponto e vírgula (";"), informação disponível nos arquivos leia-me (Leia_Me_Enem_2019.pdf [43] e Leia_Me_Enem_2020.pdf [7]).

Após a abertura, foi utilizado um método 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.

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.

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 nos resultados para 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 23 colunas. As variáveis das colunas 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 e respostas das perguntas do questionário socioeconômico com relação a escolaridade dos pais, renda e acesso a celular, computador e internet. 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.

4.4 Análise e Resultados

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 4.1 e Figura 4.2, 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.

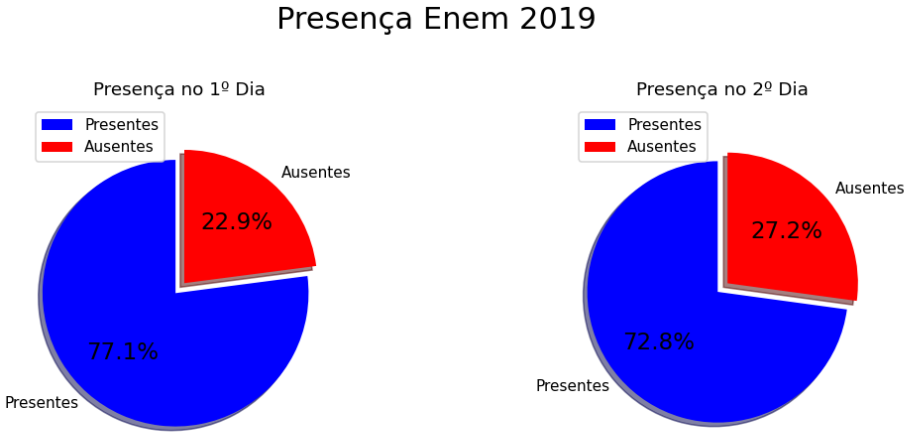

Figura 4.1: Presença no Enem de 2019.

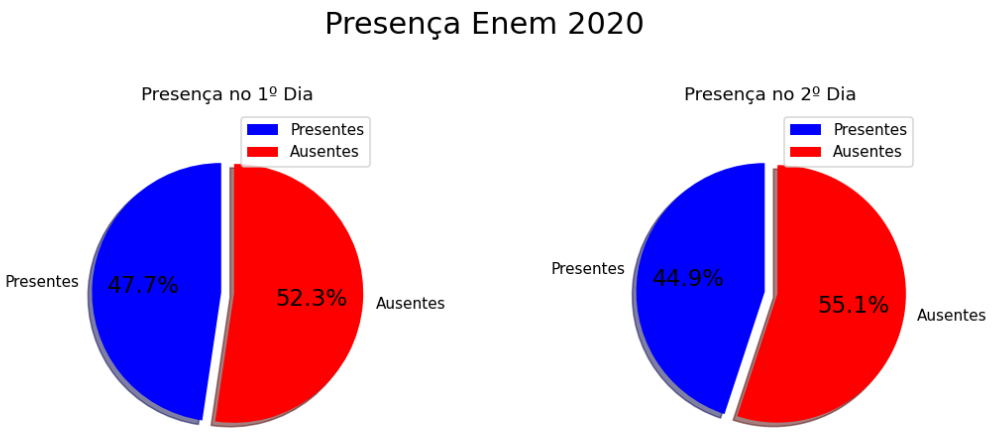

Figura 4.2: Presença no Enem de 2020.

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). 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 distribuição considerando todas as notas das provas finalizadas regularmente e maiores que zero. A Tabela 4.1 mostra o resultado para o ano de 2019 e a Tabela 4.2 para o ano de 2020.

Tabela 4.1: 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
Nota mínima	315.9	322.0	327.9	359.0	40.0
Nota máxima	835.1	801.7	860.9	985.5	1000.0
1º quartil (25%)	448.2	483.6	417.8	435.2	500.0
2º quartil (50%)	510.8	526.2	470.3	501.1	580.0
3º quartil (75%)	566.7	565.3	533.2	597.8	680.0
Desvio padrão	80.1	62.5	75.9	108.8	155.3

Tabela 4.2: 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
Nota mínima	313.7	288.7	323.9	327.1	40.0
Nota máxima	862.6	801.1	854.8	975.0	1000.0
1º quartil (25%)	435.7	478.2	427.1	425.8	480.0
2º quartil (50%)	512.7	530.0	483.7	505.2	580.0
3º quartil (75%)	580.7	576.4	548.7	602.3	700.0
Desvio padrão	93.8	73.0	79.6	116.9	176.3

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 4.3 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. É interessante também calcular a redução percentual para cada grupo. 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.

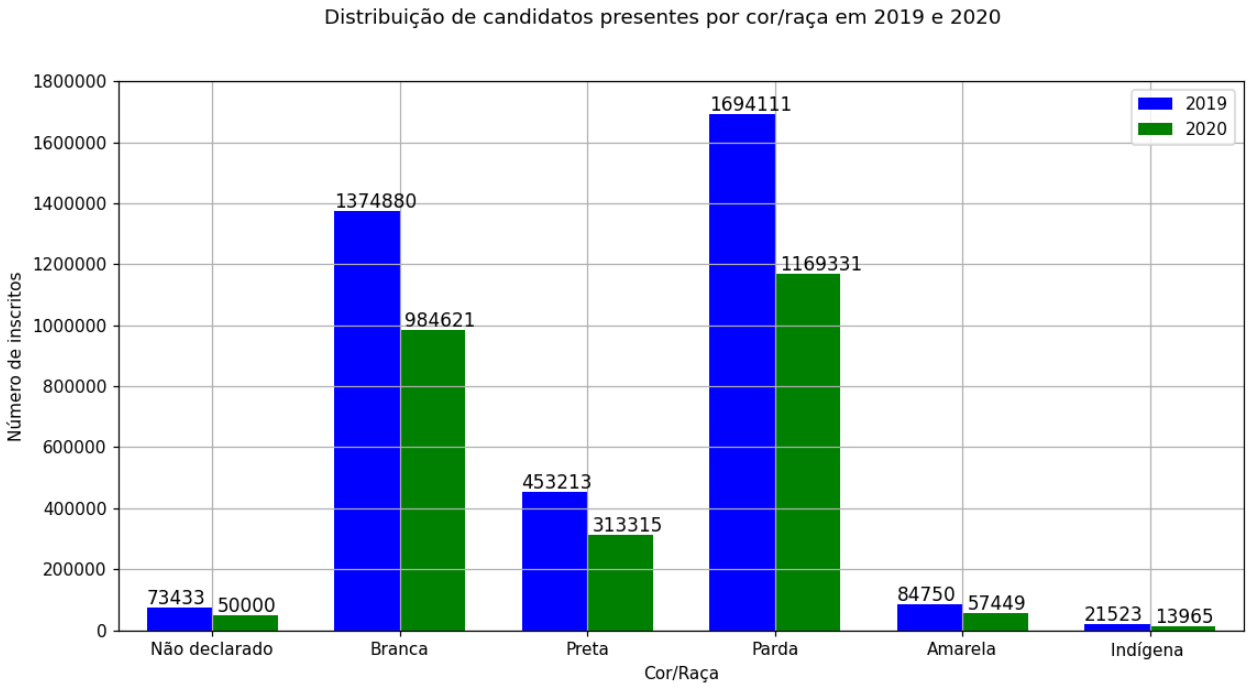

Figura 4.3: Distribuição de indivíduos presentes nas duas provas por cor/raça em 2019 e 2020.

A respeito da distribuição de candidatos presentes por categorias de renda nas duas edições do exame, observando a Figura 4.4 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 eram 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). Entretando, em 2020, apesar das classes B e C ainda possuirem 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. Somando-se os número das clases 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.

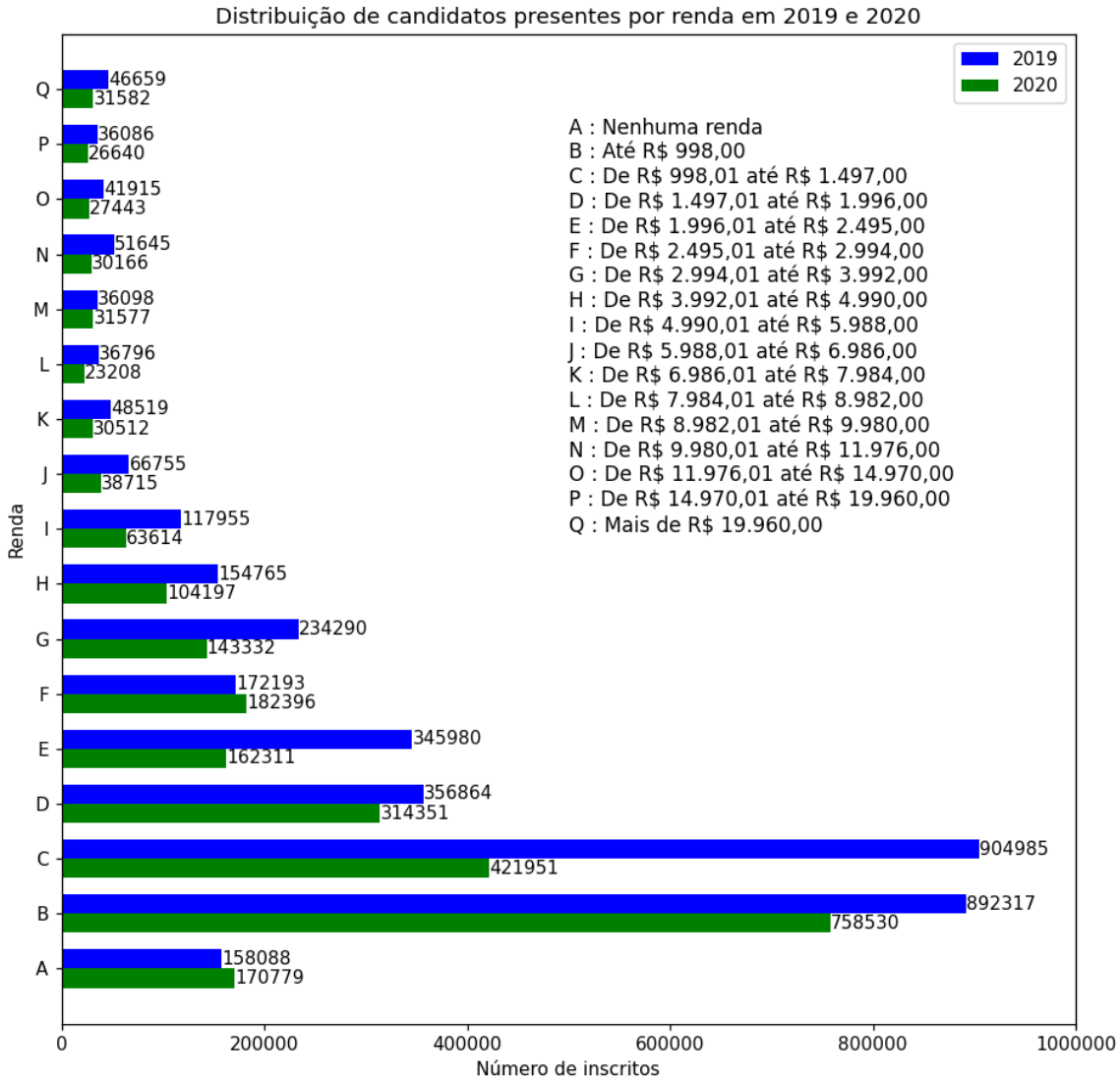

Figura 4.4: Distribuição de indivíduos presentes nas duas provas por renda em 2019 e 2020.

Sobre a escolaridade dos pais, os dados são divididos entre escolaridade do pai e da mãe. Na Figura 4.5 e na Figura 4.6 tem-se a distribuição obtida para os inscritos presentes nas duas provas em 2019 e 2020 com relação a escolaridade do pai e da mãe, respectivamente.

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. Em 2020, essa proporção foi a que teve o maior aumento, 1,57%, subindo para 28,75%. A maior redução percentual foi a dos inscritos com pais que não completaram a 4ª série/5º ano do ensino fundamental, com 1,46% a menos. O percentual que ficou mais estável foi o de inscritos com pais que completaram a 8ª série/9º ano do ensino fundamental, mas não completaram o ensino médio, com uma variação de apenas 0,15%.

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, aumento de 1,06%. Dessa vez este não foi o maior aumento proporcional, que ficou com os inscritos com mães que completaram a pós-graduação, 1,8% a mais em 2020. A maior redução foi dos inscritos com mães que completaram a 8ª série/9º ano do ensino fundamental, mas não completaram o ensino médio. Lembrando que estes números são sobre a proporção de inscritos da categoria em relação a outros valores do mesmo ano e não sobre os números absolutos de 2019 e 2020. Outra informação observada é que dentre os inscritos que responderam "Não sei"sobre a escolaridade dos pais, quase três vezes mais não sabem a escolaridade do pai em relação a da mãe.

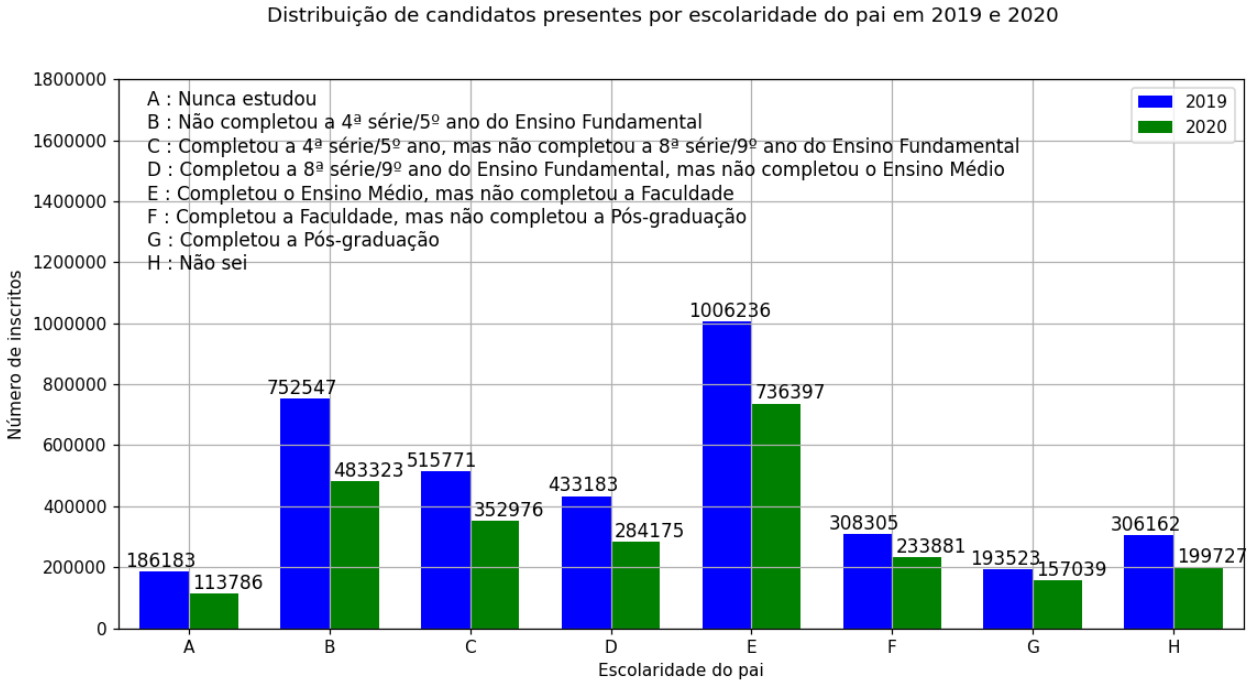

Figura 4.5: Distribuição de indivíduos presentes nas duas provas por escolaridade do pai em 2019 e 2020.

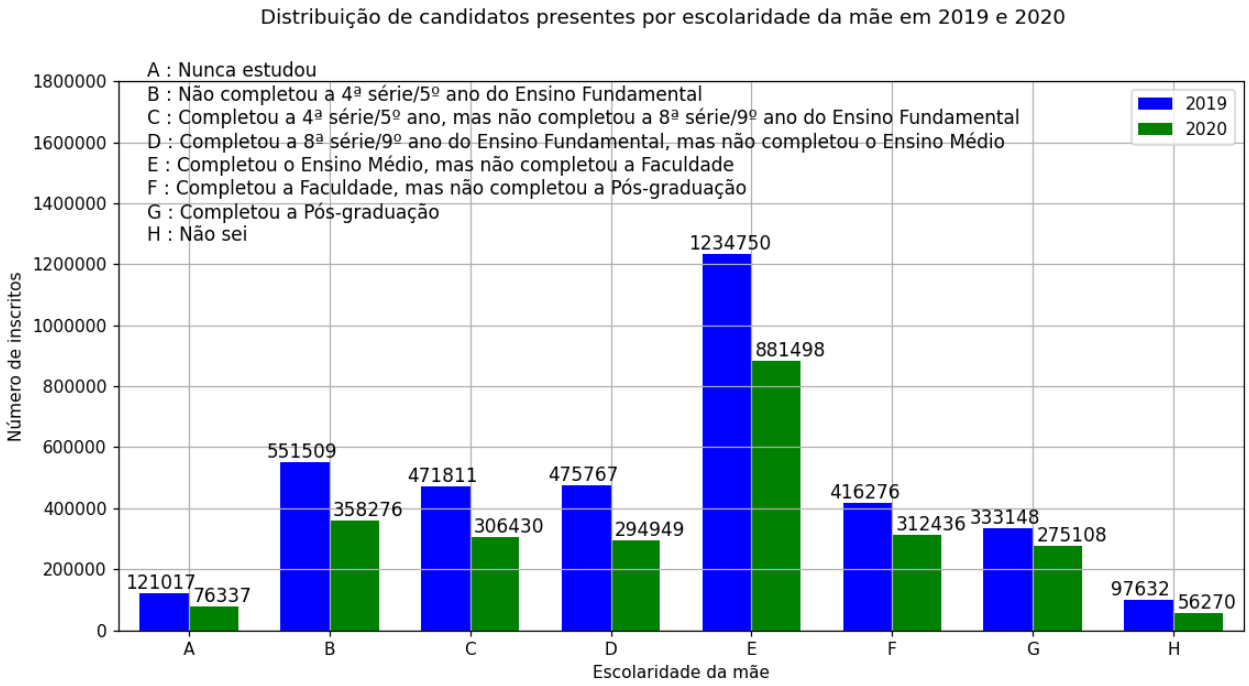

Figura 4.6: Distribuição de indivíduos presentes nas duas provas por escolaridade da mãe em 2019 e 2020.

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 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 nas provas de 2020. Conhecer as características desses indivíduos é tão importante quanto a dos presentes. Foram analisados os dados de inscritos ausentes por cor/raça, renda, escolaridade dos pais e acesso a celular, computador e internet. Começando pelos resultados por cor/raça, 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, sendo este o maior valor percentual entre as categorias, seguido da cor/raça Preta com 56,25%. Os de cor/raça branca tiveram o menor valor de candidatos ausentes. A Figura 4.7 mostra a distribuição em números absolutos.

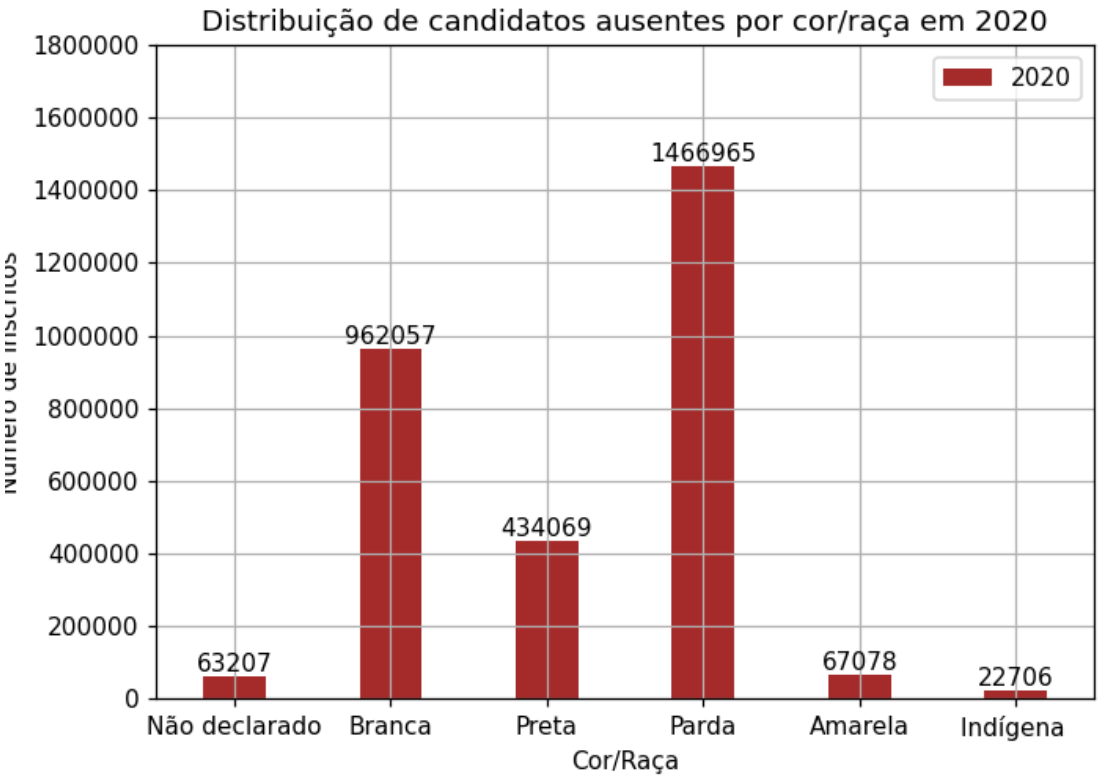

Figura 4.7: Distribuição de indivíduos ausentes nas duas provas por cor/raça em 2020.

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. O gráfico da Figura 4.8 apresenta os números.

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. Para a escolaridade da mãe, candidatos com mães que completaram o ensino médio, mas não completaram a faculdade, foram o maior número com 29,07%. Os números podem ser vistos na Figura 4.9 e na Figura 4.10

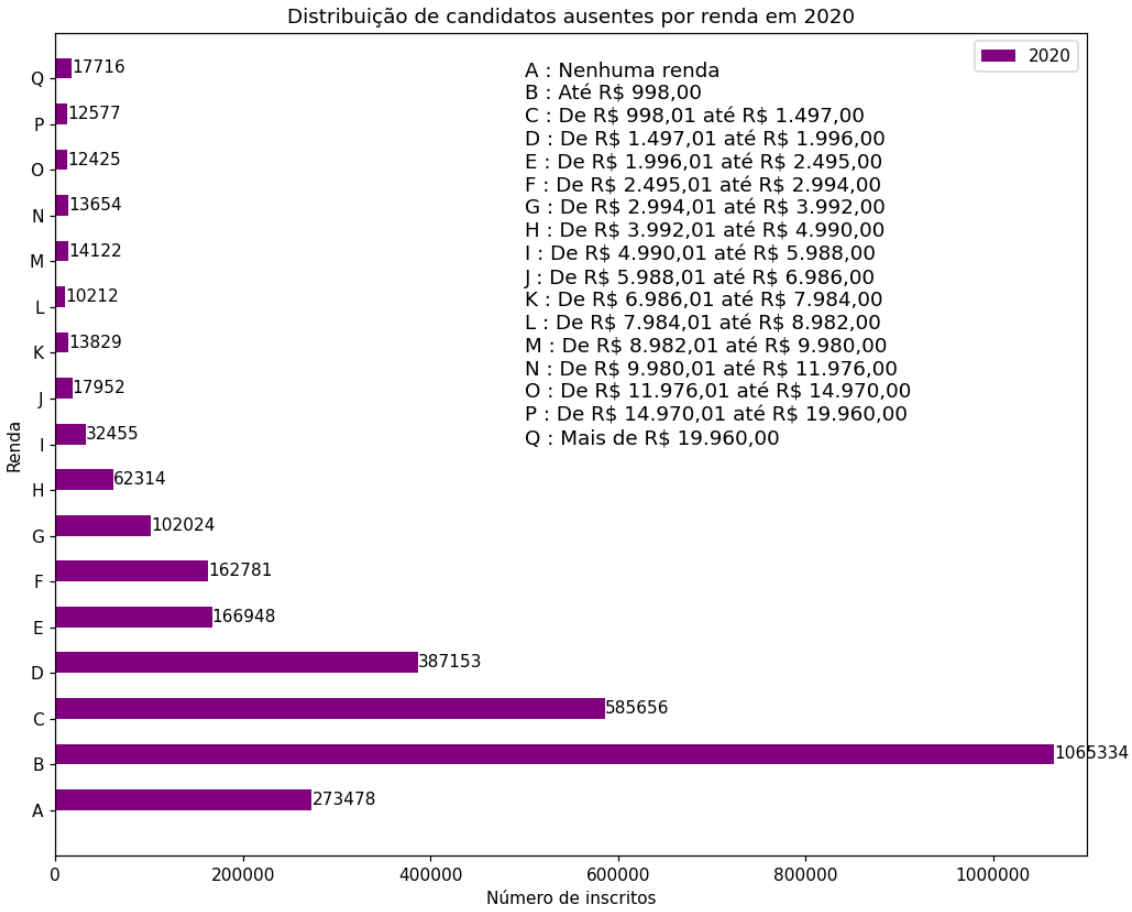

Figura 4.8: Distribuição de indivíduos ausentes nas duas provas por renda em 2020.

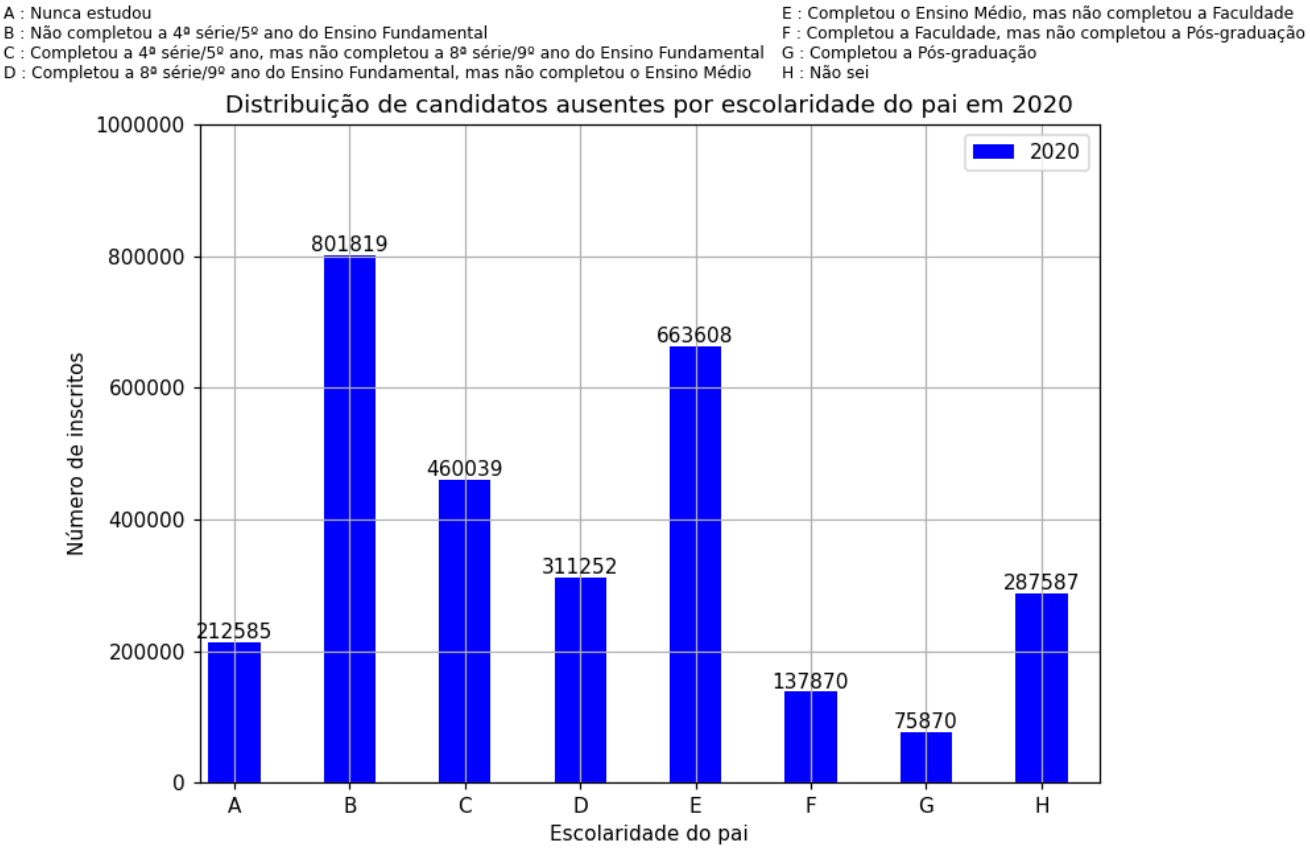

Figura 4.9: Distribuição de indivíduos ausentes nas duas provas por escolaridade do pai em 2020.

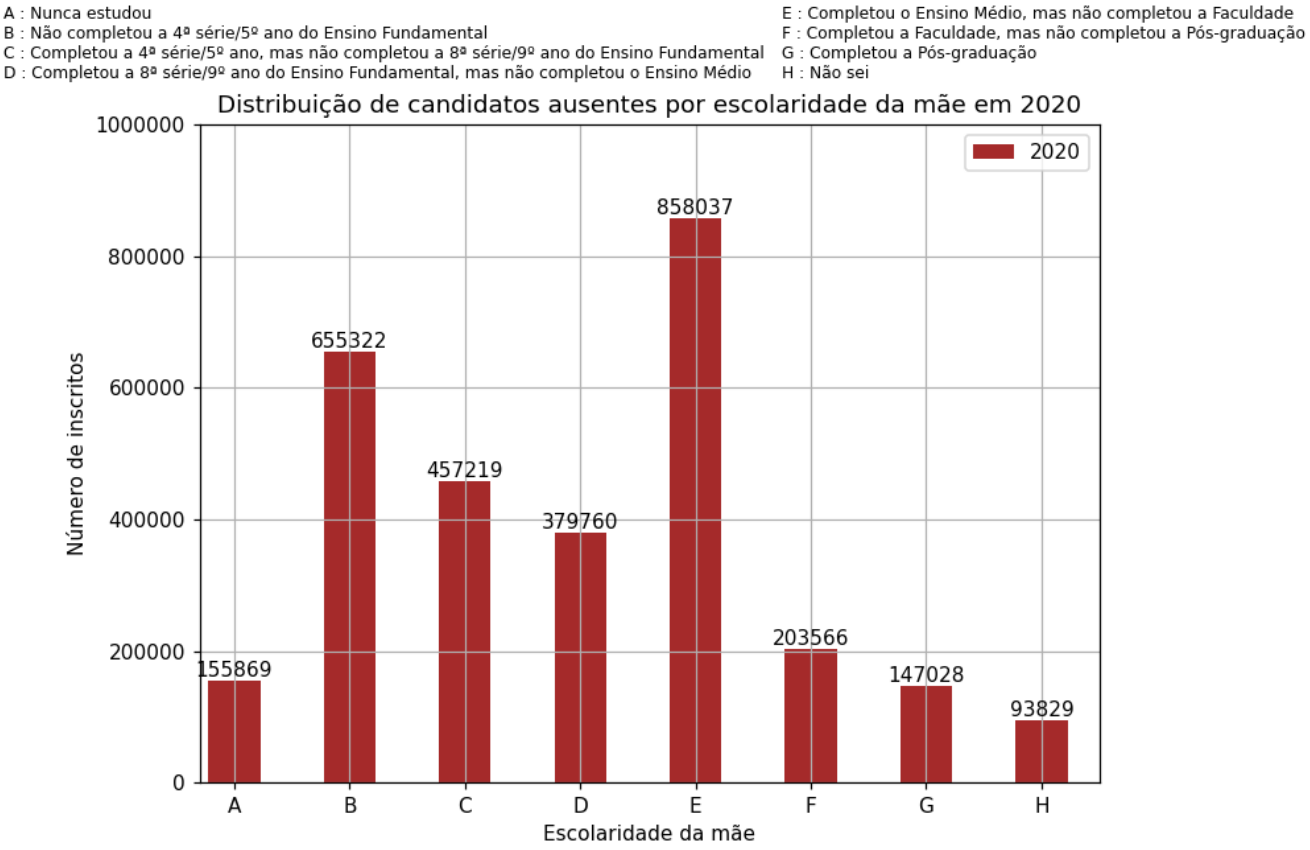

Figura 4.10: Distribuição de indivíduos ausentes nas duas provas por escolaridade da mãe em 2020.

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 4.11 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 [15], a grande maioria possuía celular.

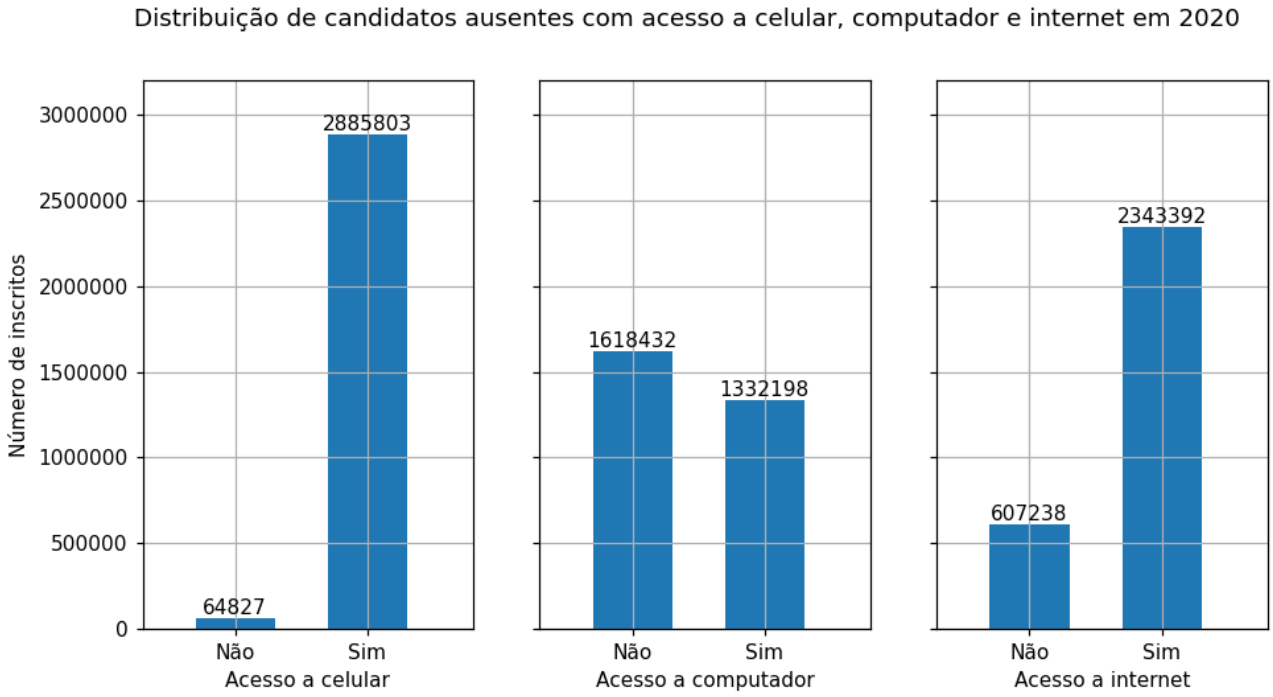

Figura 4.11: Distribuição de indivíduos ausentes nas duas provas por acesso a celular, computador e internet em 2020.

Também foram feitas análises combinando algumas variáveis. Para isso, consideraram-se os participantes que estiveram presentes nas duas provas no ano de 2020. A Figura 4.12 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 4.12: Média das notas em 2020 agrupadas por cor/raça e acesso à internet.

A renda também mostrou ter influência sobre as médias das notas dos participantes. Na Figura 4.13 pode-se perceber que as médias das notas das provas tem 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. Foi realizado também um teste de correlação entre essas variáveis. Para isso, os valores das faixas de renda foram convertidos de texto para valores numéricos, utilizando o limite superior do intervalo de renda como referência para cada categoria. A biblioteca Pandas possui um método que calcula um coeficiente de correlação entre colunas de um dataframe, permitindo escolher um método de correlação. Aplicando este método e escolhendo o coeficiente de correlação de Pearson, obteve-se como resultado que a renda possui uma correlação positiva e fraca (força com que uma variável aumenta a medida que a outra também aumenta) com as médias das notas das provas (índices de correlação entre 0.3 e 0.5).

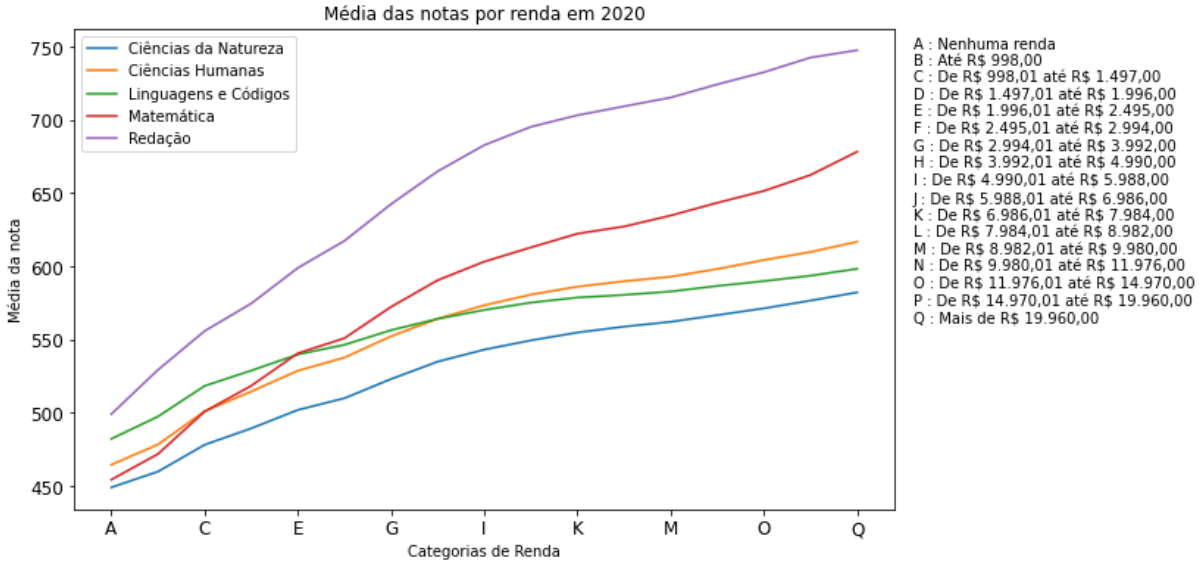

Figura 4.13: Média das notas por renda em 2020.

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 tem valores próximos entre grupos de cor/raça diferentes, para faixas de renda próximas. A Figura 4.14 apresenta os valores de referência desta análise.

As últimas 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 utilzada 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 evidênciam 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. [47]

A Figura 4.15 mostra as médias das notas por estado, e a Figura 4.16 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,

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

entre os valores de 0,000 e 0,499. Comparando as duas figuras, observa-se que 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, tem-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 se percebe que existe uma relação. Valores mais altos desse índice 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	Ciências da Natureza	Ciências Humanas	Linguagens e Códigos	Matemática	Redação
AC	464.8	483.5	502.9	473.4	549.5
AL	471.9	489.7	506.1	492.9	581.7
AM	467.4	492.0	492.4	482.3	555.5
AP	462.4	483.5	495.8	467.7	551.8
BA	476.6	495.9	511.6	494.1	577.3
CE	481.3	504.0	516.4	513.8	600.8
DF	503.6	531.2	541.6	534.7	597.8
ES	505.2	528.8	534.7	540.6	608.1
GO	491.3	513.7	526.5	517.9	599.8
MA	462.7	481.0	497.0	475.7	562.3
MG	508.1	535.9	542.3	550.6	624.3
MS	488.4	510.2	524.3	513.4	571.8
MT	483.0	505.1	516.9	506.0	571.9
PA	467.9	486.2	499.2	478.3	571.7
PB	477.9	498.0	511.1	501.9	591.0
PE	482.0	500.8	518.4	510.6	587.9
PI	471.2	489.3	502.4	492.1	584.6
PR	508.1	536.9	542.0	547.0	590.1
RJ	501.5	530.6	542.7	537.7	617.9
RN	486.9	507.6	521.0	515.4	597.2
RO	476.6	493.6	507.2	492.3	553.7
RR	478.5	500.6	513.3	492.1	553.3
RS	497.4	530.8	541.3	537.0	597.6
SC	509.9	539.3	542.9	552.1	602.3
SE	481.5	498.7	510.3	498.2	605.8
SP	512.7	542.5	552.0	558.4	611.1
TO	469.4	487.8	503.7	488.9	561.7

Figura 4.15: 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 4.16: IDHM dos estados em 2017.

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

Capítulo 5

Conclusão

Com as análises realizadas nos microdados Enem neste trabalho, foi possível observar que a pandemia realmente prejudicou muitos estudantes. O recorde de abstenções no Enem chama a atenção para as desigualdades sociais no país e reforça a necessidade de ações neste direção para reduzí-las. Foi visto que muitos dos inscritos que se ausentaram das provas, eram de parcelas da população menos favorecidas. A pesquisa científica é de grande importância nesta questão, formulando hipóteses, encontrando evidências e oferecendo soluções.

Considera-se que os objetivos deste trabalho foram atingidos. 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 naturalmente 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.

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. Este estudo reforça que a exploração destes dados utilizando técnicas de mineração de dados pode produzir muito conhecimento importante para a área da educação.

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 a 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. Também é necessário que trabalhos e pesquisas sobre dados educacionais continuem sendo realizados, pois podem ser muito úteis para tomar melhores decisões a respeito da educação no Brasil.

Referências

[1] *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. 3

[2] *História do Inep.* https://www.gov.br/inep/pt-br/acesso-a-informacao/institucional/historia, acesso em 08/09/2022. 3

[3] *Sobre o Inep.* https://www.gov.br/inep/pt-br/acesso-a-informacao/institucional/sobre, acesso em 08/09/2022. 3

[4] *História do Enem.* https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/historico, acesso em 09/09/2022. 3

[5] *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. 3, 4

[6] *Microdados do Enem.* https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem, acesso em 13/09/2022. 4

[7] 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. 4, 20, 22

[8] *Histórico da pandemia de COVID-19.* https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19, acesso em 16/09/2022. 5

[9] 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. 5

[10] 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. 5

[11] 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. 5

[12] 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. 5

[13] 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. 5

[14] 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. 5

[15] 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. 5, 6, 7, 34

[16] 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. 6, 7

[17] 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. 7

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

[19] Soares, José: *Qualidade e eqüidade na educação básica brasileira: fatos e possibilidades.* janeiro 2005. 8

[20] 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. 8

[21] Stair, Ralph M. e George W. Reynolds: *Princípios de Sistemas de Informação.* Cengage Learning, 2015. Tradução da 11ª edição norte-americana. 9, 10

[22] Laudon, Kenneth C. e Jane P. Laudon: *Sistemas de Informação Gerenciais*. Prentice Hall, 2010. Tradução da 9ª edição. 9, 10

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

[24] Setzer, Valdemar W.: *Dado, informação, conhecimento e competência*. `https://www.ime.usp.br/~vwsetzer/dado-info.html`, acesso em 04/09/2022. 10

[25] Han, Jiawei, Micheline Kamber e Jian Pei: *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., 3ªª edição, 2011. 10, 11, 12, 13

[26] Goldschmidt, Ronaldo e Emmanuel Passos: *Data Mining: Um guia prático*. Elsevier, 2005. 4ª Tiragem. 10, 11, 12, 13

[27] Aggarwal, Charu C.: *Data Mining: The Textbook*. Springer Publishing Company, Incorporated, 2015. 10

[28] 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. 11, 12

[29] Larose, Daniel T. e Chantal D. Larose: *Discovering Knowledge in Data*. John Wiley Sons, Ltd, 2ªª edição, 2014. 12, 13

[30] Foundation, Python Software: *Python 3.10.6 documentation*. `https://docs.python.org/3/`, acesso em 06/09/2022. 13

[31] team the pandas development: *pandas documentation*. `https://pandas.pydata.org/docs/index.html`, acesso em 06/09/2022. 13

[32] team, The Matplotlib development: *Matplotlib: Visualization with python*. `https://matplotlib.org`, acesso em 06/09/2022. 13

[33] *What is r?* `https://www.r-project.org/about.html`, acesso em 06/09/2022. 14

[34] *Jupyterlab: A next-generation notebook interface*. `https://jupyter.org`, acesso em 06/09/2022. 14

[35] *Getting started with anaconda*. `https://docs.anaconda.com/anaconda/user-guide/getting-started/`, acesso em 06/09/2022. 14

[36] 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, 4ªª edição, 2016. 14

[37] Moretin, Pedro A. e Wilton de O. Bussab: *Estatística básica*. Saraiva, 7ªª edição, 2012. 15

[38] 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. 15, 16

[39] IBM Corporation: *Ibm spss modeler crisp-dm guide*, 2011. https://inseaddataanalytics.github.io/INSEADAnalytics/CRISP_DM.pdf. 16

[40] 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. 17

[41] 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. 17

[42] *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. 17

[43] 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. 17, 20, 22

[44] 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. 18

[45] 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. 18

[46] *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 acesso em 29/09/2022. 19

[47] *Atlas do desenvolvimento humano no brasil.* http://www.atlasbrasil.org.br/ranking, acesso em 06/10/2022. 36