



MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO
TOCANTINS
CAMPUS ARAGUAÍNA
CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE
SISTEMAS

RENATO FERNANDES SANTOS SILVA

**ENEM e o capital humano e cultural: uma análise a partir da utilização de
técnica de inteligência artificial**

ARAGUAÍNA

2021

RENATO FERNANDES SANTOS SILVA

ENEM e o capital humano e cultural: uma análise a partir da utilização de técnica de inteligência artificial

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do Título de Tecnólogo do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal do Tocantins, Campus Araguaína.

Orientadora: Profa. Mestra Sônia Eduardo de Moraes
Coorientador: Prof. Mestre Walisson Pereira de Sousa

ARAGUAÍNA

2021

Dados Internacionais de Catalogação na Publicação (CIP)
Bibliotecas do Instituto Federal do Tocantins

F363e Fernandes Santos Silva, Renato
 ENEM e o capital humano e cultural : uma análise a partir da
 utilização de técnica de inteligência artificial / Renato Fernandes
 Santos Silva. – Araguaína, TO, 2021.
 73 p. : il. color.

Trabalho de Conclusão de Curso (Tecnólogo em Análise e
Desenvolvimento de Sistemas) – Instituto Federal de Educação,
Ciência e Tecnologia do Tocantins, Campus Araguaína, Araguaína,
TO, 2021.

Orientadora: Ma. Sônia Eduardo de Moraes
Coorientador: Me. Walisson Pereira de Sousa

1. Enem. 2. Agrupamento de dados. 3. K-Means. I. Eduardo de
Moraes, Sônia. II. Pereira de Sousa, Walisson. III. Título.

CDD 004.2

A reprodução total ou parcial, de qualquer forma ou por qualquer meio, deste documento é autorizada para fins
de estudo e pesquisa, desde que citada a fonte.

Elaborado pelo sistema de geração automática de ficha catalográfica do IFTO com os dados fornecidos
pelo(a) autor(a).

RENATO FERNANDES SANTOS SILVA

**ENEM e o capital humano e cultural: uma análise a partir da utilização de
técnica de inteligência artificial**

Trabalho de Conclusão de Curso apresentado como
requisito parcial para obtenção do Título de Tecnólogo do
Curso Superior de Tecnologia em Análise e
Desenvolvimento de Sistemas do Instituto Federal do
Tocantins, Campus Araguaína.

Data de aprovação: ____ / ____ / ____

Banca Examinadora

Prof. Msc. Sônia Eduardo de Moraes (Orientadora)
IFTO - Campus Araguaína

Prof. Msc. Walisson Pereira de Sousa (Co-Orientador)
IFTO - Campus Araguaína

Prof. Msc. Janderson Henrique Mota
IFTO - Campus Araguaína

MSc. Virgínia de Sousa Venega
IFTO - Campus Araguaína

Araguaína - TO, 2021.

DEDICATÓRIA

Aos meus pais (Rubens e Súsie) e minha irmã (Suzana Fernandes).

AGRADECIMENTOS

Agradeço aos meus pais, Rubens e Súsie, pelo incentivo e apoio aos estudos.

À minha irmã, Suzana Fernandes, agradeço pela leitura e sugestões de escrita.

À professora Sônia Eduardo de Moraes, minha orientadora, agradeço pelas diretrizes apresentadas à escrita de meu TCC, pelo incentivo institucional e emocional.

Ao professor Walisson Pereira de Sousa, meu co-orientador, agradeço pela paciência e pelas sugestões apresentadas quanto ao uso das técnicas de agrupamento de dados por meio da Inteligência Artificial.

Agradeço aos professores do IFTO, campus Araguaína-TO pelos ensinamentos compartilhados.

Finalmente, agradeço a Deus pela permissão em estar concluindo este curso.

RESUMO

O ingresso no ensino superior, a partir de 1998, seguiu as diretrizes instituídas pelo Enem. Executado sob o critério de larga escala, o agrupamento de seus dados a partir de técnicas de Inteligência Artificial possibilita o estudo das condicionalidades econômicas e sociais de seus participantes. O objetivo da pesquisa que realizamos concentrou-se em identificar e avaliar os padrões dos perfis de estudantes que obtiveram sucesso nas provas do Enem e, conseqüentemente, a nota que lhe deu o direito de ingresso no curso desejado. Realizamos uma pesquisa de abordagem bibliográfica e documental. No viés bibliográfico buscamos a compreensão da técnica *K-Means* com uma das ferramentas necessárias ao agrupamento de dados. No foco documental a aplicação dessa técnica enquanto ferramenta utilizada pela Inteligência Artificial para o agrupamento de dados e análise a partir dos conceitos da teoria do capital humano e capital cultural de Pierre Bourdieu. Diante do exposto, a pesquisa mostrou que os exames realizados em larga escala podem ser agrupados pelas técnicas de IA como meio para a discussão de seus resultados no contexto social.

Palavras-chaves: Enem. Inteligência Artificial. Agrupamento de dados. K-Means.

ABSTRACT

The entry into higher education, as of 1998, followed the guidelines established by Enem. Executed under the criterion of large scale, the grouping of its data using Artificial Intelligence techniques makes it possible to study the economic and social conditionalities of its participants. The objective of the research we carried out focused on identifying and evaluating the patterns of the profiles of students who were successful in the Enem tests and, consequently, the score that gave them the right to enter the desired course. We carried out a bibliographical and documental approach research. In the bibliographic bias, we seek to understand the K-Means technique as one of the tools necessary for data grouping. In the documental focus, the application of this technique as a tool used by Artificial Intelligence for data grouping and analysis based on the concepts of Pierre Bourdieu's human capital and cultural capital theory. Given the above, the research showed that examinations performed on a large scale can be grouped by AI techniques as a means to discuss their results in the social context.

Keywords: Enem. Artificial intelligence. Grouping of data. K-Means.

LISTA DE ILUSTRAÇÕES

Figura 1 - O histórico do Enem.....	16
Figura 2 - Competências Gerais da BNCC.....	19
Figura 3 - Os Quatro V's da Big Data	27
Figura 4 - O volume de dados gerados	28
Figura 5 - A variedade de dados gerados	29
Figura 6 - A velocidade dos dados	29
Figura 7 - A veracidade dos dados.....	30
Figura 8 - Agrupamentos de dados similares	34
Figura 9 – Dados do saneamento básico das unidades federativas do Brasil.....	35
Figura 10 - Agrupamento Hierárquico com base no saneamento de cada estado ...	36
Figura 11 - Exemplo de aplicação do <i>K-Means</i>	39

LISTA DE GRÁFICOS

Gráfico 1 - Porcentagem de participantes que possuem uma determinada renda média	43
Gráfico 2 - Método do Cotovelo (Elbow) para o K ideal	44
Gráfico 3 - Gráfico de dispersão com os indicadores do ano de conclusão e idade	45
Gráfico 4 - Gráfico de dispersão com os indicadores de renda média e idade	46
Gráfico 5 - Gráfico de dispersão com os indicadores da nota média e idade	47
Gráfico 6 - Gráfico de dispersão com os indicadores do ano de conclusão e renda média	48
Gráfico 7 - Gráfico de dispersão com os indicadores do ano de conclusão e nota média	49
Gráfico 8 - Gráfico de plotagem das variáveis nota e renda média.....	50
Gráfico 9 - Quantidade de participantes determinado por sexo	51
Gráfico 10 - Quantidade de pessoas que moram em uma residência.....	52
Gráfico 11 - A presença de empregado doméstico na residência	53
Gráfico 12 - Quantidade de banheiros na residência	54
Gráfico 13 - Quantidade de quartos na residência	55
Gráfico 14 - Quando de carros na residência.....	56
Gráfico 15 - Quantidade de motocicletas na residência	57
Gráfico 16 - Quantidade de geladeiras na residência	58
Gráfico 17 - Quantidade de Freezers na residência.....	59
Gráfico 18 - Quantidade de máquinas de lavar roupa na residência.....	60
Gráfico 19 - Quantidade de máquina de secar roupa na residência	61
Gráfico 20 - Quantidade de forno micro-ondas na residência	62
Gráfico 21 - A disponibilidade de um aspirador de pó na residência.....	63
Gráfico 22 - Quantidade de Televisores na residência.....	64
Gráfico 23 - A presença de aparelho DVD na residência	65
Gráfico 24 - A presença de TV por assinatura na residência	66
Gráfico 25 - A presença de telefone celular na residência	67
Gráfico 26 - A presença de telefone fixo na residência	68
Gráfico 27 - Quantidade de computadores na residência	69
Gráfico 28 - A disponibilidade de acesso à internet na residência	70

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizagem de Máquina
BNCC	Base Nacional Comum Curricular
Enem	Exame Nacional do Ensino Médio
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IFTO	Instituto Federal de Educação, Ciência e Tecnologia do Tocantins
MEC	Ministério da Educação e Cultura
TRI	Teoria de Resposta ao Item

SUMÁRIO

1. INTRODUÇÃO	12
2. A INSTITUCIONALIZAÇÃO DO ENEM E DA BNCC	14
2.1. A história do Enem.....	14
2.2. O Enem e a BNCC.....	17
2.3 O poder financeiro e o capital cultural.....	20
3. FUNDAMENTOS E TÉCNICAS DE IA PARA AGRUPAMENTO DE DADOS.....	23
3.1. Dado, Informação e Conhecimento	23
3.2. Big Data	26
3.3. Inteligência Artificial	31
3.4. Técnicas de Inteligência Artificial (IA)	32
3.4.1. Aprendizagem de Máquina	33
3.4.1.1. Agrupamento de Dados.....	34
3.4.1.1.1 Agrupamento de Dados Hierárquico	35
3.4.1.1.2 Agrupamento de dados particional.....	37
4. METODOLOGIA.....	40
5. RESULTADOS E DISCUSSÕES	42
CONSIDERAÇÕES FINAIS	71
• REFERÊNCIAS.....	72

1. INTRODUÇÃO

O ingresso no ensino superior, até meados de 1998, se dava apenas por meio da participação em exames vestibulares. Tradicionalmente, o vestibular era organizado pelas próprias instituições de ensino, o que, de certo modo, não possuía padronização nacional.

No objetivo de padronizar e deixar mais acessível o processo avaliativo de ingresso no ensino superior o Ministério da Educação e Cultura (MEC) instituiu o Exame Nacional do Ensino Médio (Enem), cujas avaliações verificavam o domínio das competências e habilidades dos estudantes em fase de conclusão do ensino médio, bem como daqueles não tinham sido aprovados em exames vestibulares.

A aplicação do Enem em larga escala foi tomada como objeto de estudo para a produção deste trabalho monográfico porque os dados fornecidos no ato das inscrições favorecem o agrupamento e a análise por meio das técnicas de Inteligência Artificial (IA). Isso se dá porque as técnicas de IA são instrumentos que possibilitam o agrupamento de dados organizados em diferentes categorias.

No contexto de que as técnicas de IA possibilitam o agrupamento de dados, nossa investigação científica pautou-se na seguinte problemática: De que forma o agrupamento de dados por meio de técnicas de IA podem contribuir para a melhoria dos resultados dos estudantes pela participação no Enem? Na intenção de discutirmos este questionamento sem respostas engessadas, traçamos a hipótese de que o agrupamento de dados por meio das técnicas de IA é um fator que permite às escolas da educação básica a adoção de estratégias voltadas para a abordagem de conteúdos relacionados às competências e habilidades exploradas nas avaliações do Enem e sistematizadas, atualmente, pela BNCC.

O ingresso dos participantes nas universidades através da realização do Enem, se dá diretamente pelo resultado obtido durante a realização do exame, que por sua vez é composto por 4 áreas do conhecimento segundo a BNCC, sendo elas: Linguagens e suas Tecnologias; Matemática e suas Tecnologias; Ciências da Natureza e suas Tecnologias; Ciências Humanas e Sociais Aplicadas. A nota que será utilizada no processo de disputa da vaga desejada, é avaliada conforme os pesos definidos para cada área do conhecimento, levando em conta que a nota pode ser

maior ou menor dependendo do curso em que a pessoa se inscreveu.

É por meio destes dados e da renda média por participante, que está diretamente relacionada aos meios que o participante teve para se preparar para o exame, que conseguimos traçar perfis para que fosse possível identificar aqueles que conseguiram obter um resultado melhor, acarretando assim em notas maiores que tiveram facilidade para o ingresso no ensino superior.

O objetivo da pesquisa que realizamos concentrou-se em identificar e avaliar os padrões dos perfis de estudantes que obtiveram sucesso nas provas do Enem e, conseqüentemente, a nota que lhe deu o direito de ingresso no curso desejado.

A justificativa de realização desta pesquisa está relacionada ao contexto de que as avaliações em larga escala, como é o caso do Enem, produzem dados que ao serem tratados pelas técnicas de IA podem contribuir para a geração de padrões de estudo de performance de alto rendimento de seus participantes.

Quanto a análise dos dados, ancoramo-nos teoricamente a partir dos conceitos teoria do capital humano e capital cultural de Pierre Bourdieu, em que empreende em sua abordagem, a relação de dominação da classe dominante, acerca dos bens produzidos pela humanidade, bem como seus conhecimentos e saberes. Essas relações, que dadas a nossa inscrição teórica, podem ser capitalizadas e a educação, especificamente, o nosso objeto de análise a relação entre a renda per capita e a média de notas do ENEM, na busca de estabelecer essa conexão, que se torna símbolo dessa representação para acesso ao capital cultural, que são os saberes e conhecimentos reconhecidos por diplomas e títulos que as boas notas do exame podem possibilitar acesso.

Além da Introdução, das Considerações Finais e das Referências, esta monografia está estruturada em três capítulos. No primeiro capítulo apresentamos um percurso sobre a institucionalização do Enem e da BNCC. No segundo capítulo apresentamos fundamentos e técnicas de IA para agrupamento de dados. No terceiro capítulo apresentamos e analisamos os dados agrupados a partir da aplicação das técnicas de Inteligência Artificial.

Diante do exposto, este trabalho monográfico aborda a importância de utilização das técnicas de IA para o agrupamento de dados como ferramenta que favorece a realização de estudos segundo a padronização das avaliações realizadas pelo Enem como forma de ingresso no ensino superior.

2. A INSTITUCIONALIZAÇÃO DO ENEM E DA BNCC

Neste capítulo, conforme destacamos na parte introdutória deste trabalho monográfico, apresentamos um percurso sobre a institucionalização do Enem (Exame Nacional do Ensino Médio) e da BNCC (Base Nacional Comum Curricular). Referenciamos o Enem em primeiro plano porque ele representa uma das principais ações realizadas pelo sistema de ensino brasileiro como mecanismo de ingresso no ensino superior. Expandimos nosso percurso pela BNCC porque se trata do documento oficial que regulamenta o ensino a partir da ação de competências e habilidades que os estudantes devem assimilar ao longo do percurso escolar.

2.1. A história do Enem

A política de institucionalização do Exame Nacional do Ensino Médio (Enem), segundo dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2020), teve sua primeira edição no dia 20 de agosto de 1998. Na ocasião, a participação de estudantes acima de 18 anos totalizou 53%. Os estudantes de escolas públicas totalizaram apenas 9% e os resultados foram utilizados como forma de ingresso no ensino superior.

Ao longo de sua trajetória, o Enem obteve credibilidade e passou a ser utilizado por 93 instituições de ensino superior em sua segunda edição, ou seja, em meados de 1999. Na sequência de suas edições, os investimentos e a adesão das instituições de ensino tiveram números expressivos. Assim, o acesso aos cursos de graduação incluiu também os estudantes portadores de necessidades especiais, sendo esta uma das principais conquistas do certame realizado no ano 2000.

A partir de 2001, as inscrições no Enem passaram a ser feitas pela Internet, com gratuidade aos estudantes concluintes do ensino médio. Essa inovação possibilitou a participação de mais de 1.600 mil estudantes, com aplicação das provas em 277 municípios.

Na edição de 2002 o exame foi aplicado em 600 municípios e mais de 50% dos participantes eram concluintes do ensino médio. Esse fator deu expressividade ao trabalho realizado pelo INEP, a ponto de em 2003 receber a inscrição de estudantes

“treineiros”, ou seja, aqueles que não tinham concluído o Enem, mas que resolveram participar da avaliação como meio de conhecimento do sistema avaliativo aplicado.

A criação do Programa Universidade para Todos (ProUni) em 2004 fez com que o Enem recebesse a adesão de novas universidades. No ano de 2005 o total de inscritos superou a casa de 3 milhões de estudantes, elevando-se ainda mais esse quantitativo nos anos de 2006 e 2007. Historicamente, a edição de 2008 marcou os 10 anos de criação do Enem, o qual seria utilizado, também, para a certificação do ensino médio. Essa novidade possibilitou a inscrição de mais de 4 milhões de estudantes.

As inovações realizadas no Enem entre os anos de 2009 e 2017 deram aos estudantes novas possibilidades, destacando-se: o uso da nota como requisito para o Fies, a expansão do atendimento especializado como mecanismo de acessibilidade, a ampliação da gratuidade do exame, adesão quase totalidade das instituições federais, a adesão de universidades de Portugal, uma maior participação de treineiros, maior segurança no sistema de aplicação das provas.

No ano de 2018, dez anos após a primeira edição, o Enem comemorou o sucesso de sua iniciativa. A criação de um documentário marcou a força de um exame que adquiriu aceitabilidade dos estudantes e das instituições brasileiras, além de um significativo número de universidades portuguesas.

Na Figura 1, a seguir, construímos um infográfico apresentando os principais pontos de destaque da trajetória do Enem ao longo de seus 10 anos. Com isso, destacamos que este exame marcou a experiência de aplicação de avaliações em larga escala. É, portanto, neste foco que este trabalho monográfico foi construído, ou seja, utilizamos a técnica de agrupamento de IA para analisarmos diferentes dados das condicionalidades sociais dos participantes do Enem para compreendermos sua relevância na vida das pessoas que concluíram ou que estão em fase de conclusão do ensino médio e que desejam ingressar em curso superior.

Figura 1 - O histórico do Enem



Fonte: <https://infograph.venngage.com/view/b82911ee-f680-446e-ada2-c940ba9a0b52>

As informações constantes no infográfico acima evidenciam o percurso e os principais avanços atingidos pelo Enem ao longo de sua existência. De modo geral, é um instrumento que obteve aceitação das diferentes instituições de ensino como ferramenta de acesso aos cursos superiores.

2.2. O Enem e a BNCC

A instrumentalização do Enem como ferramenta avaliativa tem como base principal as diretrizes instituídas pelo sistema educacional brasileiro. Atualmente, o documento oficial recebe o nome de BNCC. Por sua vez, é um documento que sistematiza as competências exigidas dos estudantes. De modo geral, a BNCC faz parte do método de avaliação do Enem porque aponta as 10 Competências Gerais que acompanham o desenvolvimento dos alunos desde a Educação Infantil até o Ensino Médio. Vejamos o detalhamento destas competências:

A primeira competência geral da BNCC diz respeito ao “conhecimento”. Por sua vez, é um tipo de competência que busca valorizar e utilizar os conhecimentos historicamente construídos sobre o mundo físico, social, cultural e digital para entender e explicar a realidade, continuar aprendendo e colaborar para a construção de uma sociedade justa, democrática e inclusiva (BRASIL, 2017).

O “pensamento científico, crítico e criativo” integra a segunda competência geral da BNCC. De modo geral, é uma competência que busca exercitar a curiosidade intelectual e recorrer à abordagem própria das ciências, incluindo a investigação, a reflexão, a análise crítica, a imaginação e a criatividade, para investigar causas, elaborar e testar hipóteses, formular e resolver problemas e criar soluções (inclusive tecnológicas) com base nos conhecimentos das diferentes áreas (BRASIL, 2017).

A terceira competência geral diz respeito ao “repertório cultural”, o qual busca valorizar e fruir as diversas manifestações artísticas e culturais, das locais às mundiais, e participar de práticas diversificadas da produção artístico-cultural (BRASIL, 2017).

Na quarta competência geral da BNCC o foco está voltado para a “comunicação”. Por isso, tem como objetivo utilizar diferentes linguagens – verbal (oral ou visual-motora, como Libras, e escrita), corporal, visual, sonora e digital –, bem como conhecimentos das linguagens artística, matemática e científica, para se expressar e partilhar informações, experiências, ideias e sentimentos em diferentes contextos, além de produzir sentidos que levem ao entendimento mútuo (BRASIL, 2017).

A quinta competência geral está alinhada ao campo da “cultura digital”, a qual busca compreender, utilizar e criar tecnologias digitais de informação e comunicação

de forma crítica, significativa, reflexiva e ética nas diversas práticas sociais (incluindo as escolares) para se comunicar, acessar e disseminar informações, produzir conhecimentos, resolver problemas e exercer protagonismo e autoria na vida pessoal e coletiva (BRASIL, 2017).

Na sexta competência geral, o foco está voltado para o “trabalho e projeto de vida”, pois institui meios de valorizar a diversidade de saberes e vivências culturais, apropriar-se de conhecimentos e experiências que lhe possibilitem entender as relações próprias do mundo do trabalho e fazer escolhas alinhadas ao exercício da cidadania e ao seu projeto de vida, com liberdade, autonomia, consciência crítica e responsabilidade (BRASIL, 2017).

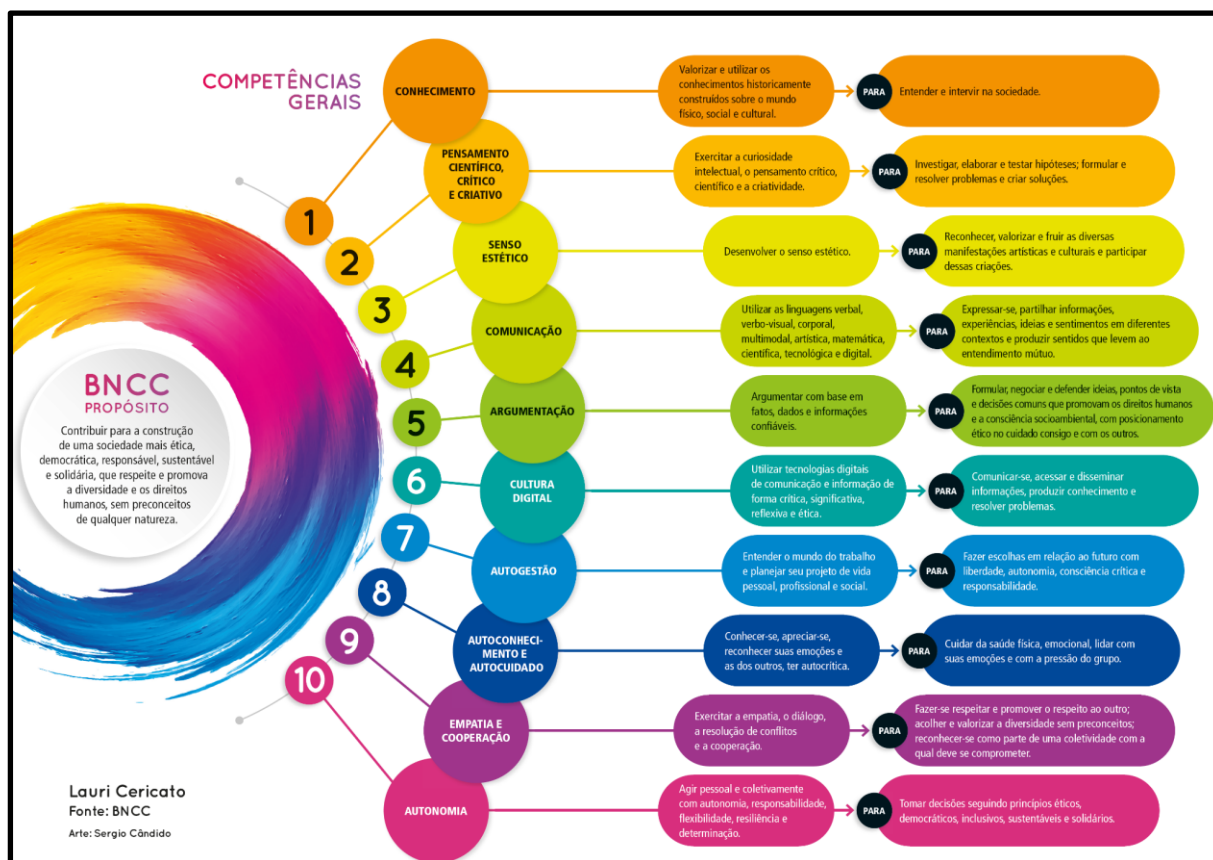
Na linha da “argumentação”, a BNCC instituiu sua sétima competência geral, qual objetiva argumentar com base em fatos, dados e informações confiáveis, para formular, negociar e defender ideias, pontos de vista e decisões comuns que respeitem e promovam os direitos humanos, a consciência socioambiental e o consumo responsável em âmbito local, regional e global, com posicionamento ético em relação ao cuidado de si mesmo, dos outros e do planeta (BRASIL, 2017).

O contexto sobre o “autoconhecimento e autocuidado” compuseram a oitava competência geral da BNCC. Por sua vez, diz respeito ao foco de conhecer-se, apreciar-se e cuidar de sua saúde física e emocional, compreendendo-se na diversidade humana e reconhecendo suas emoções e as dos outros, com autocrítica e capacidade para lidar com elas (BRASIL, 2017).

Na linha da “empatia e cooperação” a nova competência geral da BNCC foi instituída com o objetivo de exercitar a empatia, o diálogo, a resolução de conflitos e a cooperação, fazendo-se respeitar e promovendo o respeito ao outro e aos direitos humanos, com acolhimento e valorização da diversidade de indivíduos e de grupos sociais, seus saberes, suas identidades, suas culturas e suas potencialidades, sem preconceitos de qualquer natureza (BRASIL, 2017).

A décima competência geral da BNCC contemplou o foco da “responsabilidade e cidadania”, o que diz respeito aos atos de “agir pessoal e coletivamente com autonomia, responsabilidade, flexibilidade, resiliência e determinação, tomando decisões com base em princípios éticos, democráticos, inclusivos, sustentáveis e solidários (BRASIL, 2017).

Figura 2 - Competências Gerais da BNCC



Fonte: <https://www.passeidireto.com/arquivo/87209596/infografico-bncc-competencias-gerais>

Um dos fatores que faz com que o Enem tenha sua eficácia comprovada, é o fato de ter como principal aliado na correção das provas a Teoria de Resposta ao Item (TRI), que é uma metodologia de avaliação usada pelo Ministério da Educação que não se baseia em contabilizar apenas o número total de acertos no teste.

Segundo Araújo, Andrade e Bortolotti (2009, p.1001):

A TRI fornece modelos matemáticos para os traços latentes, propondo formas de representar a relação entre a probabilidade de um indivíduo dar uma certa resposta a um item, seu traço latente e características (parâmetros) dos itens, na área de conhecimento em estudo. A partir de um conjunto de respostas apresentadas por um grupo de respondentes a um conjunto de itens, a TRI permite a estimação dos parâmetros dos itens e dos indivíduos em uma escala de medida.

A forma como os participantes são avaliados através deste método, se baseia em um modelo que necessita de dois pressupostos relevantes, sendo eles a Curva Característica do Item (CCI), que determina de forma específica qual mecanismo de processo de resposta se encaixa melhor, e é utilizado a independência local ou dimensionalidade.

A Curva de Característica do Item consiste em descrever a forma como a mudança do traço latente relaciona-se com a mudança na probabilidade de uma resposta específica.

A independência local pode ser obtida a partir do momento em que o nível do traço latente dos itens do teste é independente. Desse modo, a probabilidade de que um item seja respondido é precisamente determinada pelo nível do traço latente que o respondente apresentou, e não apenas pelas respostas a outros itens que pertenciam ao mesmo conjunto. Ela é vista como uma consequência a partir de uma determinação correta da dimensionalidade dos dados.

A dimensionalidade dos dados consiste em um número de fatos que são responsáveis por expressar o traço latente, além de verificar por meio de uma análise fatorial que seja apropriada para os dados que foram categorizados.

Segundo BRASIL (2012) o item, que é avaliado nas correções dos participantes, pode ser qualificado de acordo com três parâmetros. Sendo o primeiro o poder de discriminação, que consiste na capacidade de um item distinguir os estudantes que têm a proficiência requisitada daqueles quem não a têm. O segundo é baseado no grau de dificuldade da questão que se está avaliando, e o terceiro é a possibilidade de acerto ao acaso, conhecido também como chute.

2.3 O poder financeiro e o capital cultural

O sistema educacional é sem dúvida um dos locais esperados pela grande maioria dos estudantes brasileiros, como espaço possível, para a ascensão social, para formação profissional e humana, e claro para aquisição de status quo na sociedade vigente.

Desse modo, o ENEM é uma das ferramentas utilizadas, conforme já foi explicitado, para se chegar aos níveis da educação superior no país. Observa-se que essa avaliação também, coloca-se na condição de categorizar, selecionar e excluir por classe social os acessos aos cursos educacionais, o que podemos exemplificar por cursos tidos mais elitizados, que exigem maior nota, como: Medicina, Engenharias, Direitos, e outros que dado a baixa procura, exige-se menor nota, como é o caso das licenciaturas.

Uma das variáveis que foi definida como principal durante esta pesquisa foi a renda média entre os participantes. O motivo da escolha se dá por meio de que o

poder aquisitivo, assim como a etnia do cidadão, são variáveis que acabam impactando no desempenho escolar dos estudantes da educação básica no Brasil, segundo o que apresenta a tese de Carvalho (2017). Essa variável é de grande importância, pois ela se divide entre 17 categorias, começando em 'Nenhuma renda', percorrendo até 'Mais de R\$ 19.960,00' reais.

Como podemos identificar nos resultados e discussões no Gráfico 1, a maior parte dos participantes se concentram entre as rendas médias nos valores de R\$ 998 até R\$ 2.245 reais, mostrando que estes estudantes possivelmente possuem uma educação mais simples em relação aos participantes que possuem uma renda média superior aos R\$ 7.446 reais.

Desta forma, ao analisarmos a realidade educacional brasileira por meio dos dados fornecidos pelo Enem, apoiamo-nos teoricamente nos conceitos do capital humano e capital cultural de Pierre Bourdieu que foram elaborados na década de 1960, em que destaca a relação do desempenho escolar e a classe social que o indivíduo pertence.

O capital humano fundamenta-se sobre o fato de que é um dos mais importantes a serem abordados em um contexto socioeconômico que busca reduzir a desigualdade, pois está relacionado diretamente ao nível de conhecimento e tempo de estudo que um indivíduo possui, podendo ainda considerar um dos principais indicadores da pobreza.

No Brasil, o capital cultural adquirido pelos cidadãos está altamente relacionado ao capital financeiro que um indivíduo possui, tendo em vista que para que se possam possuir itens que remetem a cultura brasileira, é preciso possuir um capital econômico significativo, visto que o Estado não disponibiliza de forma simplificada o acesso a este tipo de bem a população de forma geral.

Participantes das lutas que se opõem ao conjunto dos agentes ou das instituições que possuem quantidade significativa de um capital específico, que ocupam posições dominantes no interior de seus campos respectivos, utilizam-se da conservação ou a transformação da "taxa de câmbio" entre os diferentes tipos de capital e, a partir disto, o poder sobre as instâncias burocráticas que tem poder para alterá-las por meio de medidas administrativas, como a escassez de títulos escolares que dão acesso a posições consideradas dominantes, afetando assim o valor desses títulos e das posições que possuem correspondência. (BOURDIEU, 2008).

Bourdieu (2008) afirma que alguns dos detentores de um grande volume de capital global, como empresários, membros de profissões liberais e professores universitários, opõem-se globalmente àqueles considerados menos providos de capital econômico e de capital cultural, como os operários não-qualificados. A partir da perspectiva do peso relativo do capital econômico e do capital cultural no seu patrimônio, os professores, que ainda são considerados relativamente mais ricos em capital cultural do que em capital econômico, opõem-se de maneira nítida aos empresários que são relativamente mais ricos em capital econômico do que em capital cultural.

Ele cita ainda que:

o capital econômico e o capital cultural tem, nesse caso, um peso importante, o espaço social organiza-se de acordo com três dimensões fundamentais: na primeira dimensão, os agentes se distribuem de acordo com o volume global do capital possuído, aí incluídos todos os tipos; na segunda, de acordo com a estrutura desse capital, isto é, de acordo com o peso relativo do capital econômico e do capital cultural no conjunto de seu patrimônio; na terceira, de acordo com a evolução, no tempo, do volume e da estrutura de seu capital. (BOURDIEU, 2008, p.30).

Por fim, ele demonstra ainda que a reprodução de uma estrutura de distribuição com o capital cultural, se dá através da relação entre estratégias que as famílias possuem e a lógica específica para com cada instituição escolar. As famílias se situam como corpos animados, tendo ainda a tendência de perpetuar o seu ser social, contendo seus poderes e privilégios.

3. FUNDAMENTOS E TÉCNICAS DE IA PARA AGRUPAMENTO DE DADOS

3.1. Dado, Informação e Conhecimento

Segundo Valdemar Setzer (2001), dado é tudo aquilo que pode ser definido como uma sequência de símbolos quantificados ou quantificáveis. Quantificável é aquilo que pode ser quantificado e posteriormente reproduzido sem que seja perceptível uma diferença com o original. Isso é o que nos permite associar um texto a um dado. As letras do alfabeto, como conhecemos, são símbolos já quantificados, pois ele por si só, já pode representar uma base numérica. Como por exemplo a base de números hexadecimais utilizada pelos computadores, que é composta por 10 dígitos decimais e letras entre A e E.

Sendo assim, um dado é inevitavelmente uma unidade matemática, sendo considerado algo sintático. Desta forma, compreendemos que os dados podem ser escritos a partir de representações formais ou estruturais, levando em conta de que são dados quantificados ou quantificáveis, e que podem ser gravados em um computador e ainda interpretados por ele.

Esses dados, já armazenados em um computador, podem estar conectados virtualmente a outros trechos de dados por meio de contiguidade física, ou até mesmo por “ponteiros”, que são nada menos que endereços dentro da unidade de armazenamento - mais conhecida como HD (Hard Disk), ou Disco Rígido - utilizados para a formação de estruturas dos dados.

Ainda segundo Vitória Peres, é apresentado uma outra definição de dado, sendo ela:

Dados (mensagens ou artefatos), são meios concretos, suportes físicos codificados. Entretanto, mesmo dizer que ‘eles são materiais codificados’, exige alguém, um observador, que seja capaz de identificá-los como tal, eles não podem ser considerados em si mesmos, fora de um contexto, como contendo informação possível, e, portanto, como sendo um dado. (PERES, 2005, p. 4).

Quanto à Informação, esta pode ser considerada uma abstração informal, visto que ela não pode ser formalizada por meio de uma teoria através da lógica ou da matemática, pois geralmente uma informação está na mente de alguém, representando desta forma o significado de algo para aquela pessoa.

Temos então a exemplificação de como uma Informação pode ser formada, visto que ainda não se desenvolveu a definição da mesma. Como por exemplo, a frase

“A maçã está madura”, ao ser lida ou ouvida por alguém, a palavra “maçã” pode significar para esta pessoa que o fruto é produzido pela macieira, e que “está madura” signifique que a fruta esteja com seu desenvolvimento completo, sendo o mais apropriado para consumo.

Caso a representação e a Informação da frase seja dada por meio de dados, como citado no parágrafo acima, ela pode ser armazenada em um computador. Contudo, aquilo que é armazenado na máquina, não é de fato a informação em sua forma real, mas sim, uma representação em forma de dados, tendo em mente que esta informação pode ser transformada pela máquina a partir do momento que se encontra nela.

As transformações da informação, podem variar das mais singelas formas, como por exemplo, uma simples formatação do texto, sendo considerada uma transformação sintática. A máquina por sua vez, não tem autonomia de alterar o significado daquela informação inserida, pois isso até o momento, somente pode ser orquestrado por uma pessoa que possua determinada informação. Levemos em conta que a máquina pode, por si só, embaralhar a informação transcrita nela previamente, tornando-a incompreensível para a pessoa que a está recebendo.

Consideremos também que é possível transformar a forma como uma informação é interpretada pela máquina, para que ela seja contrária à forma original para o destinatário, como por exemplo, mudar a palavra “maçã” para “laranja”, sendo que a mudança foi apenas para o receptor, mas no computador em que a informação original se encontra, ela ainda permanece a mesma, sendo isso uma manipulação de dados matemáticos para a máquina.

Ainda segundo Setzer (2001):

Uma distinção fundamental entre dado e informação é que o primeiro é puramente sintático e a segunda contém necessariamente semântica (implícita nas palavras “significativo” e “significação” usada em sua caracterização). É interessante notar que é impossível introduzir e processar semântica em um computador, porque a máquina mesmo é puramente sintática (assim como a totalidade da matemática). (SETZER, 2001, p. 1-2).

Em geral, a Informação (SETZER, 2001) por sua vez, pode ser uma propriedade intelectual de uma pessoa ou grupo, podendo ainda ser algo que foi recebido de uma outra pessoa ou grupo. No primeiro caso, ela se forma a partir da sua interpretação ou percepção de algo, como por exemplo o sentimento de dor, quando uma pessoa sente seu pé doer, o corpo gera uma informação que é enviada

para o cérebro, que por sua vez interpreta tal sinal. Já no segundo caso, a informação pode não se dar apenas pelo meio simbólico dos dados, como textos, figuras, sons, animações etc.

Setzer (2001) caracteriza ainda o Conhecimento como uma abstração interior de algo que foi vivenciado por alguém. Complementa ainda que o conhecimento não é algo que pode ser descrito como os demais, mas apenas a informação sobre ele - desde que ela seja entendida pelo receptor - assim também como os dados repassados ao ouvinte.

Marta Lígia Valentim (2002), nos apresenta três tipos de conhecimentos, sendo: explícito, tácito e estratégico. Segundo ela, o conhecimento explícito consiste em um conjunto de informações que são apresentadas em algum tipo de conteúdo com suporte para os dados expostos em documentos, livros, artigos, etc. São estes conhecimentos que, por sua vez, caracterizam o saber sobre o tema tratado.

No que tange o conhecimento tácito, ele se baseia em acumular o saber sobre um assunto específico, agregando convicções, crenças, sentimentos, emoções e diversos outros fatores relacionados tanto à experiência, mas também à personalidade daquele que o detém. Já o conhecimento estratégico consiste na junção dos conhecimentos explícitos e tácitos, sendo formado a partir das informações que são cedidas por meios dos materiais de acompanhamento, complementando assim o conhecimento pertencente a especialistas (VALENTIM, 2002).

De modo geral, o conhecimento ainda pode ser inserido em um computador através de uma representação em forma de dados. Levando em conta quando se está armazenando na máquina, ele deixa de ser informação, pois está escrito de forma quantificada para que ela o compreenda.

Setzer (2001) explica ainda que por conta de o conhecimento não conseguir se sujeitar a representações ele não pode ser inserido em um computador, tornando totalmente equivocada a afirmação de que uma base de conhecimento pode estar armazenada em um computador. Por isso, o que se pode ter é apenas uma tradicional base - ou banco - de dados.

3.2. Big Data

O crescente uso das tecnologias na área educacional tem proporcionado o contato com grande quantidade de dados. Por isso, o termo *big data* passou a fazer parte dos estudos voltados à percepção de determinados interesses, como é o caso das avaliações aplicadas pelo ENEM.

A análise de dados a partir de grandes quantidades/volumes de informações é importante porque se consegue elaborar estratégias e inovações. Quando se trata do ensino na educação básica, e com vistas ao ingresso no ensino superior, as discussões apresentadas por Pasqueline Scaico, Ruy Queiroz e Alexandre Scaico (2014) destacam que a geração de padrões facilita a adoção de estratégias e aprimoramento do processo de ensino e aprendizagem.

Em razão de sua importância, o conceito de *big data* é percebido como um mecanismo que favorece o contato com grande quantidade de informações. A partir disso, os dados obtidos podem ser agrupados com a finalidade de filtragem em determinados campos/eixos.

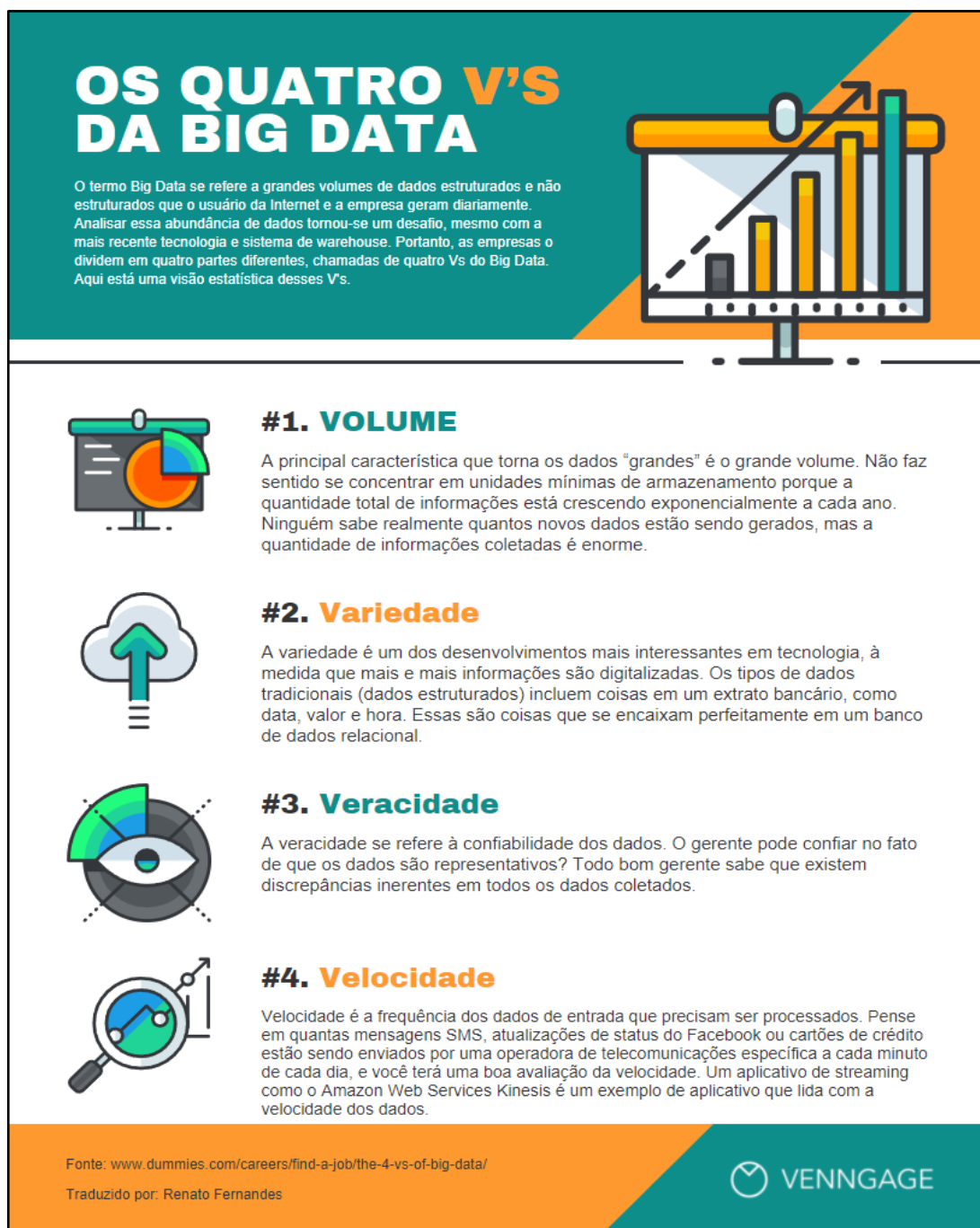
Segundo explicam Scaico, Queiroz e Scaico (2014), o sistema educacional pode obter avanços significativos se o uso de *big data* fizer parte da maioria de suas atividades. Nesse sentido, estes autores apontam:

O conceito de Big Data está relacionado à capacidade de processar e analisar grandes volumes de informação que permitem a extração de conhecimento úteis. No campo educacional, as técnicas relacionadas ao *big data* podem tornar mais eficazes os processos de avaliação, feedback e entrega do conteúdo (Scaico; Queiroz; Scaico, 2014, p. 328).

Conforme apresentado no Capítulo I, as avaliações aplicadas pelo ENEM podem ser objeto de estudo a partir do uso das técnicas da *Big Data* em razão da quantidade de dados informados pelo grande número de candidatos inscritos.

Na figura 3, trazemos alguns dos conceitos mais importantes que são utilizados na *Big Data*, fazendo com que se mantenham boas práticas, e principalmente dando um tratamento adequado a estes dados.

Figura 3 - Os Quatro V's da Big Data



Fonte: <https://infograph.venngage.com/view/9ef1c2b0-5c89-4a37-8d44-bdf9ded91027>

Os conceitos aqui apresentados estão divididos em quatro partes principais, sendo a primeira delas o volume de dados, como é apresentado na figura 4. Consistem em grandes quantidades de informações geradas pelas aplicações e programas que fazemos uso diariamente, sendo a maioria delas compostas por redes sociais, aplicativos de mensagens instantâneas, ou até mesmo sistemas bancários.

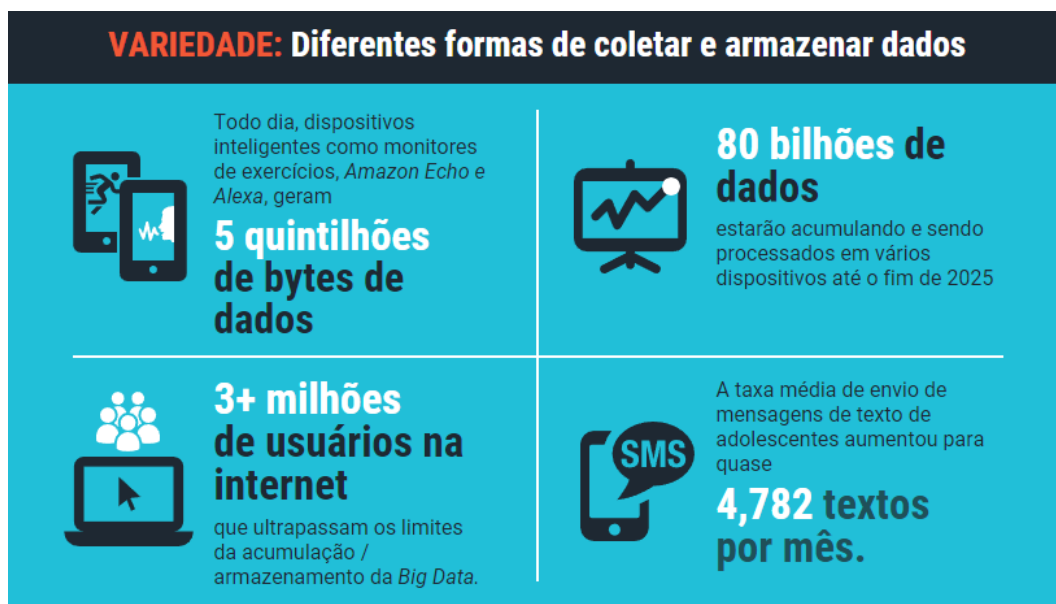
Figura 4 - O volume de dados gerados



Fonte: <https://infograph.venngage.com/view/17205488-f747-45cd-aaf6-7207bf00b015>

A segunda parte é a variedade destes dados, como apresenta a figura 5. Se baseia em analisar os mais diversos tipos de dados gerados pelas aplicações. Estes que por sua vez podem variar desde pequenos conjuntos de dados enviados por aparelhos inteligentes (robôs, televisores inteligentes, aspiradores de pó), e até mesmo grandes arquivos compartilhados através de serviços de streaming de arquivos, como vídeos ou arquivos de textos.

Figura 5 - A variedade de dados gerados



Fonte: <https://infograph.venngage.com/view/17205488-f747-45cd-aaf6-7207bf00b015>

A terceira parte se trata da velocidade em que estes dados são gerados e tratados, como mostra a Figura 6. Já é de conhecimento das pessoas da área da informática, que em grandes centros urbanísticos, as velocidades de internet estão aumentando cada vez mais, e isso não se deve somente ao fato de querer entregar informações de forma mais rápida, mas também de recebê-las.

Leva-se em conta que atualmente uma boa parte das empresas se beneficiam dos dados compartilhados pelos usuários para que seja possível realizar a exibição de anúncios e serviços que possam o interessar. Isso ocorre graças a grande capacidade de processamento e armazenamento que as empresas possuem.

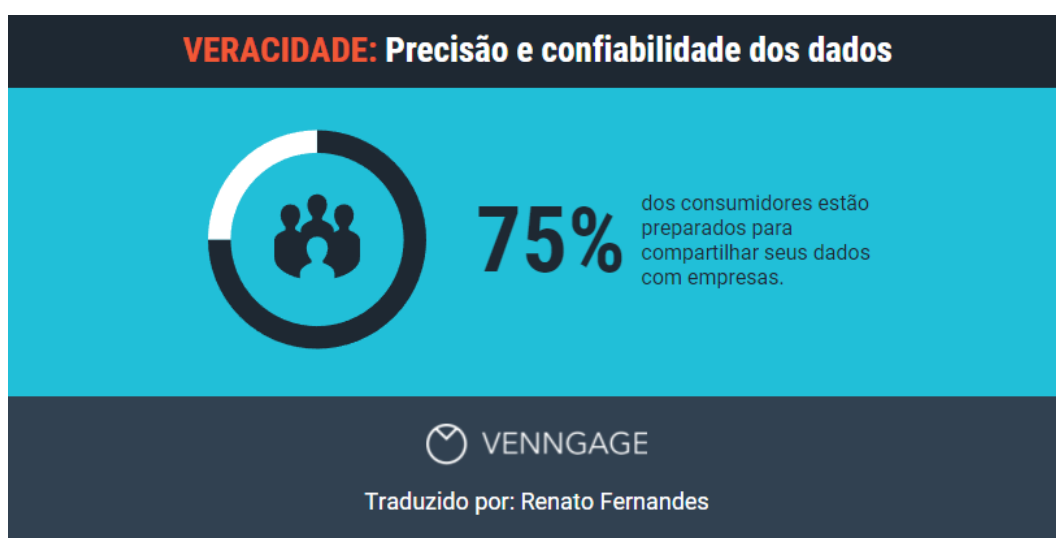
Figura 6 - A velocidade dos dados



Fonte: <https://infograph.venngage.com/view/17205488-f747-45cd-aaf6-7207bf00b015>

Por fim temos a quarta parte, sendo a mais importante dentre todas as outras, pois estamos falando da veracidade dos dados e a forma como são tratados, como mostra a figura 7. Uma das maiores preocupações de grandes empresas como Google, Microsoft, e empresas bancárias, é manter as informações de seus clientes da forma mais original possível em comparação à quando foram inseridas ou criadas, pois qualquer divergência nos dados poderia gerar conflitos, ou até mesmo grandes prejuízos para os clientes e empresas.

Figura 7 - A veracidade dos dados



Fonte: <https://infograph.venngage.com/view/17205488-f747-45cd-aaf6-7207bf00b015>

Moisés Lima Dutra e Douglas Dyllon Jeronimo de Macedo (2016), enfatizam ainda que:

A característica da veracidade em cenários Big Data está relacionada à necessidade de se ter um grau de certeza sobre a confiabilidade e a consistência dos dados que se possui, bem como com relação a autenticidade e precisão dos dados. (DUTRA; DE MACEDO, 2016, p. 17).

Sendo assim, não deve ser levado em conta apenas se uma informação está sendo armazenada para poder ser visualizada ou requisitada posteriormente, mas também analisar a segurança do local em que ela será armazenada, para que ela se mantenha íntegra e da forma mais fidedigna à original.

A grande quantidade de dados gerada continuamente, conforme apresentada pela Figura 5, necessita de estratégias ou técnicas para organização, processamento, entendimento e apresentação desses dados. Diante disso, a Inteligência Artificial (IA) tem se mostrado uma grande aliada para esse fim.

3.3. Inteligência Artificial

De acordo com Norvig e Russell (2004), a Inteligência Artificial (IA) é um campo do conhecimento que engloba conceitos das áreas de Filosofia, Matemática, Economia, Neurociência, Psicologia, Linguística, entre outras. Basicamente, a IA busca reproduzir o comportamento humano no que diz respeito ao pensamento, aprendizagem e tomada de decisão. Além disso, com o avanço da tecnologia e, consequentemente com o aumento do poder computacional, surgiram abordagens baseadas em IA que auxiliam em diversos contextos da sociedade.

Essas abordagens acontecem das mais diversas formas, sendo que algumas delas estão bastante presentes no nosso cotidiano, que muitas vezes não levamos em conta que fazem parte de um processo de IA. Alguns dos equipamentos mais comuns que costumam fazer esta abordagem, são os dispositivos inteligentes que temos com o papel de auxiliar ou monitorar o nosso dia a dia, os carros autônomos que estão se popularizando cada vez mais. Além disso, existem também os sistemas de recomendações de conteúdo que são principalmente utilizados em serviços de streaming e redes sociais, os telefones celulares, aparelhos televisores inteligentes etc.

Na área da saúde, Luiz Carlos Lobo (2017) formulou um estudo sobre o uso de técnicas baseadas em IA para a detecção precoce de Melanoma (câncer de pele) através do uso de imagens de dermatológicas, apresentando um desempenho superior ao de especialistas, sendo esse desempenho 76% versus 70,5%, com uma especificidade de 62% versus 59%, e uma sensibilidade de 82%.

A IA também tem sido amplamente utilizada na área econômica, principalmente na bolsa de valores, em que Kaupa (2013) realizou um estudo sobre a crescente demanda por investimentos considerados de alto risco com Bolsas de Valores, com o auxílio de robôs na tomada de decisões, decisões essas que eram praticadas e aperfeiçoadas através de uma Rede Neural Artificial (RNA).

Nesta área os robôs investidores tomam decisões de negociação de ativos de forma automática, através da análise dos gráficos, entrando ou saindo de posições a partir de regras e estratégias específicas ensinadas aos robôs. Essas posições são estudadas em cima de regras que são definidas para os robôs, sendo elas o momento

certo para comprar e vender as ações, o limite aceitável de perda (conhecido por stop), e a meta de lucro das operações realizadas.

Na área da educação, Arianne Sarmento Torcate, José Carlos Feliz Barbosa e Cleyton Mário de Oliveira Rodrigues (2020), utilizam técnicas de agrupamento para analisar a dificuldade de aprendizagem na educação básica entre conteúdos que utilizavam da matemática nos alunos do 6º ano do Ensino Fundamental de um colégio público. Eles realizaram o levantamento desses dados através do desenvolvimento de jogos digitais que fossem referentes a cada conteúdo alvo, e através das informações individuais de interação de cada aluno, conseguiram confeccionar um conjunto de dados para que fosse feito a análise.

Roberto Campos Leoni e Nilo Antonio de Souza Sampaio (2017) apresentam também a utilização de técnicas de agrupamento para analisar o desempenho de escolas públicas e privadas da região do Vale do Paraíba em relação ao Enem de 2015. Eles realizaram esta análise a partir de variáveis indicadoras do exame, para poder classificá-las de acordo com seu desempenho, separando-as em grupos similares e determinando assim quais escolas obtiveram melhor resultado de acordo com os alunos participantes.

A Ciência de Dados segundo Fernando Amaral (2016), pode ser definida como processo, modelos e tecnologias que estudam a forma como os dados funcionam durante o seu ciclo de vida, este que ainda por sua vez pode ser separado em 5 partes, sendo elas a produção, o armazenamento, a transformação, a análise e o descarte destes dados. É nesse momento em que a inteligência artificial se diferencia, pois faz o uso de técnicas para que possa utilizar o máximo possível daquele dado, enquanto ele está no seu ciclo de vida.

3.4. Técnicas de Inteligência Artificial (IA)

A IA pode ser subdividida em diversas outras áreas, sendo que cada uma delas ainda se subdividem entre outras. Nesta seção, serão apresentadas algumas das técnicas fundamentadas nos conceitos da IA.

O Processamento da Linguagem Natural (PLN) pode ser definido como:

Uma área de Ciência da Computação que estuda o desenvolvimento de programas de computador que analisam, reconhecem e/ou geram textos em linguagens humanas, ou linguagens naturais. O PLN não é uma tarefa trivial

devido à rica ambiguidade da linguagem natural. Essa ambiguidade torna o PLN diferente do processamento das linguagens de programação de computador, as quais são formalmente definidas evitando, justamente, a ambiguidade. (VIEIRA e LOPES, 2010, p. 183).

Backes e Junior (2019) mostram que a Visão Computacional, derivada da IA, pode ser definida como uma área que utiliza um conjunto de técnicas de processamento de imagem para permitir que as máquinas consigam visualizar formas distintas. Além disso, engloba aquisição e melhoria de imagens, sobretudo a detecção de objetos de interesse presentes em uma cena, para então relacionar com outras formas, possibilitando assim que os computadores “enxerguem” um ou mais elementos.

A Lógica Nebulosa (LN) foi proposta por Zadeh em 1965 que, “por meio do princípio da incerteza, apresentou pela primeira vez as noções da lógica dos conceitos vagos onde é admissível um conjunto com valores não precisos”. A LN dissocia-se do conceito da lógica booleana, em que só existem duas opções como resposta: 0 ou 1; certo ou errado; sim ou não, possibilitando atuação em contextos de incerteza, chamadas de regiões de nebulosidade. Na prática, as respostas informadas por sistemas baseados em lógica nebulosa possuem um grau de certeza, utilizando intervalos (de 0 a 1, por exemplo) ao invés de respostas binárias. (AGUADO e CANTANHEDE, 2010).

3.4.1. Aprendizagem de Máquina

Wittek (2014) discorre que a Aprendizagem de Máquina (AM) é uma subárea da IA que busca entender o que é aprendizagem e como os algoritmos de aprendizagem funcionam. O autor considera que esses algoritmos são o cerne de técnicas de visualização e mineração de dados. As aplicações baseadas na AM são incontáveis, podendo ser utilizadas desde a criação de robôs na área econômica, com foco na bolsa de valores, como cita Kaupa (2013) em seu trabalho, ou até mesmo para o reconhecimento de placas de sinalização, como apresenta Fabrício Augusto Rodrigues (2002) em sua dissertação.

A AM pode ainda ser subdividida em Aprendizagem: Supervisionada, Semisupervisionada e não Supervisionada. A Aprendizagem Supervisionada é aquela em que:

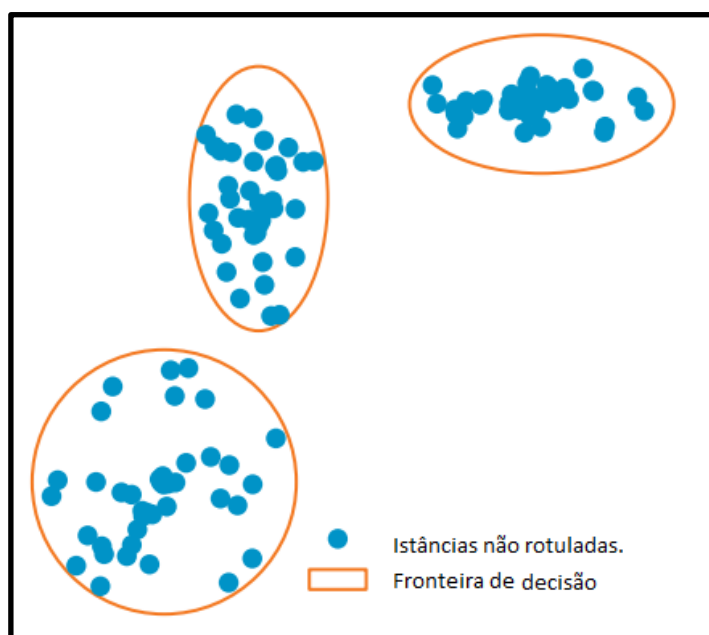
exemplos são fornecidos ao sistema com suas respectivas saídas desejadas. A cada novo conjunto de dados fornecidos, o sistema deve classificá-los com base no que foi estipulado. O desafio é encontrar conjuntos de exemplos satisfatórios. (VALDATI, p. 39, 2020).

Por sua vez, a Aprendizagem não supervisionada não utiliza modelos prévios para entendimento dos dados, ou seja, essa abordagem recebe esse nome porque não é necessário previamente ter uma coleção de dados identificada. (CAVALCANTI, p. 37-38, 2011). Por último, a Aprendizagem Semissupervisionada, de acordo com Valdati (2020), mescla técnicas de Aprendizagem Supervisionada e não Supervisionada.

3.4.1.1. Agrupamento de Dados

Conforme supracitado, a Aprendizagem não Supervisionada é utilizada em cenários em que não há conhecimento prévio dos dados. Desse modo, a aprendizagem ocorre por meio da extração dos próprios dados a serem analisados. Algoritmos de redução de dimensionalidade e de agrupamento de dados se enquadram nesta categoria. A ideia de agrupar dados refere-se a encontrar grupos de dados com atributos similares. (WITTEK, p. 17, 2014).

Figura 8 - Agrupamentos de dados similares



A Figura 8 apresenta a forma em que os dados foram separados, utilizando uma técnica de Aprendizagem não supervisionada, com características ou elementos em comum em relação aos demais, gerando assim grupos que satisfazem as condições comuns, apresentando assim padrões para que aquele elemento seja incluso em um grupo que atende às condições. Dentre as técnicas de agrupamento destacam-se o hierárquico e o particional.

3.4.1.1.1 Agrupamento de Dados Hierárquico

O agrupamento hierárquico consiste em uma representação simplificada que, ao invés de ser necessário determinar um número fixo de grupos, o número muda em cada interação. Desta forma, se o número está aumentando, nós falamos sobre a divisão do agrupamento, em que todas as instâncias de dados iniciam em um grupo e as divisões são executadas em cada interação, resultando assim em uma hierarquia de agrupamentos. Por outro lado, a aglomeração desses grupos consiste em uma abordagem ascendente, ou seja, cada instância é um grupo no início e os grupos são mesclados em cada interação.

Figura 9 – Dados do saneamento básico das unidades federativas do Brasil

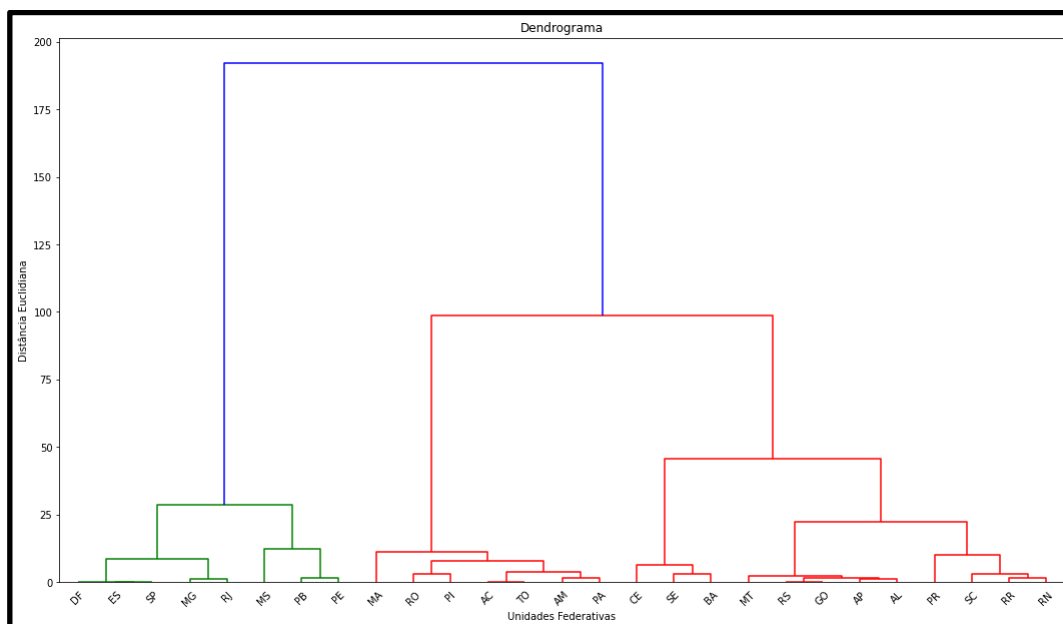
	UF	QTD_Municipios	Agua_Tratada(%)	Esgoto_Tratado(%)
0	RO	52	98	20
1	AC	22	100	14
2	AM	62	98	13
3	RR	15	100	47
4	PA	144	97	14
5	AP	16	100	38
6	TO	139	100	14
7	MA	217	100	7
8	PI	224	98	17
9	CE	184	100	67
10	RN	167	99	46
11	PB	223	98	86
12	PE	185	99	87
13	AL	102	99	38

Para fins de exemplificação, utilizamos os dados de saneamento básico do ano de 2017 das Unidades Federativas (UF) do Brasil (Figura 9), retirados do *site* do Instituto Brasileiro de Geografia e Estatística (IBGE), com as seguintes informações:

- UF: Unidade Federativa;
- QTD_Municipios: quantidade de municípios da respectiva UF;
- Agua_Tratada(%): percentual de municípios da UF que possuem serviço de água tratada;
- Esgoto_Tratado(%): percentual de municípios da UF que possuem serviço de coleta e tratamento de esgoto sanitário.

O Algoritmo hierárquico necessita de características (variáveis) para realizar o agrupamento. Dessa maneira, foram escolhidos os percentuais de água tratada (Agua_Tratada(%)) e esgotamento sanitário (Esgoto_Tratado(%)). Além disso, é necessário utilizar uma métrica de similaridade, cuja opção foi a distância euclidiana.

Figura 10 - Agrupamento Hierárquico com base no saneamento de cada estado



Conforme pode ser visto na Figura 10, o algoritmo hierárquico agrupa as Unidades Federativas de acordo com as variáveis escolhidas, comparando-as de forma que apenas as que tenham dados próximos estejam dentro do mesmo grupo, comparando ainda os grupos formados e montando outros grupos até que todos tenham sido analisados e separados hierarquicamente. É possível verificar pelo

dendograma da Figura 10 que, no grupo em verde, as UFs DF, ES e SP possuem distância euclidiana igual ou próxima de zero. Isso demonstra que essas 3 UFs possuem características muito similares em relação ao percentual de municípios com serviços de água tratada e esgotamento sanitário. O mesmo pode ser analogamente inferido para as outras UFs.

3.4.1.1.2 Agrupamento de dados particional

Diferente do agrupamento hierárquico, o agrupamento particional necessita informar a quantidade de grupos em que os dados serão particionados, não levando em consideração a hierarquia desses grupos. Algoritmos dessa categoria permitem trabalhar com grandes quantidades de dados, além de executarem mais rapidamente quando comparados aos algoritmos hierárquicos.

O K-means é um dos algoritmos mais populares quando se trata de técnicas de agrupamento de dados. É devido a facilidade de aprendizagem e a quantidade de conteúdo disponível na internet, que ele possui essa fama. Ele se baseia em um algoritmo de AM não supervisionado utilizado para particionar grandes massas de dados. Nesse contexto, K é o número de grupos que serão gerados ao final da execução do algoritmo. É um método que funciona apenas com variáveis numéricas, não sendo possível utilizá-lo em variáveis categóricas. Nesse sentido, é importante entender os seguintes conceitos:

- centróide: ponto central associado aos pontos do grupo.
- variáveis: campos de dados serão utilizados pelo algoritmo para cálculo.

Na equação 1 é possível visualizar a fórmula utilizada para calcular os centróides do *K-Means*:

$$C = \frac{1}{N_c} \sum_{j=1}^{N_c} X_j \quad (1)$$

Em que:

- c: centróide
- N_c : número de vetores (linhas) do subconjunto.
- j: posição do dado a ser calculado.
- X: subconjunto dos dados.

Outro conceito importante é a dissimilaridade, que diz respeito o quão distintos os dados são. O *K-Means* agrupa dados tentando reduzir essa dissimilaridade. Para isso, utiliza-se uma medida de distância, sendo a métrica mais comum utilizada a distância euclidiana. Depois de atribuído o centróides de cada grupo, o algoritmo calcula a distância de cada ponto para os centróides, conforme apresentado pela equação 2.

$$D(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2)$$

O K-Means pode ser sucintamente compreendido em 5 etapas:

- **etapa 1:** Escolha do valor de K (quantidade de grupos) e inicialização dos centróides;
- **etapa 2:** calcular a distância entre cada ponto do grupo e os centróides gerados;
- **etapa 3:** agrupar os pontos de acordo com as menores distâncias dos centróides;
- **etapa 4:** cálculo dos novos centróides de acordo com os atributos dos dados do grupo;
- **etapa 5:** executar as etapas 02, 03 e 04 até que o algoritmo consiga agrupar todos os dados de acordo com a métrica escolhida.

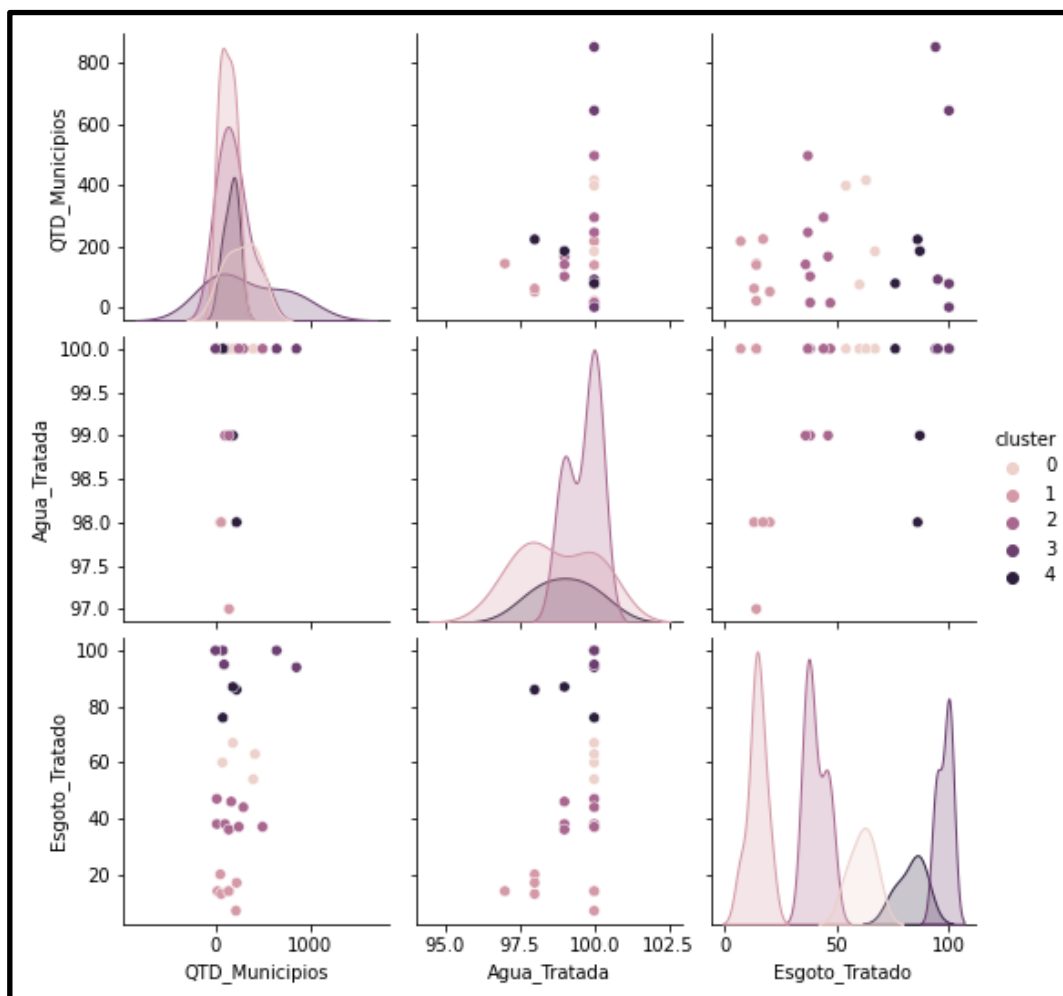
Supomos que escolhemos particionar os dados em 3 grupos (K=3). Cada ponto terá a distância euclidiana calculada para cada um dos 3 centróides inicializados. Os pontos são atribuídos ao grupo de acordo com a menor distância entre os centróides, que determinará a qual grupo cada ponto pertencerá.

Com os dados da Figura 9, exemplificamos sucintamente como o k-means separa esses dados. Utilizando as mesmas variáveis do exemplo anterior (percentuais de água tratada (Agua_Tratada(%)) e esgotamento sanitário (Esgoto_Tratado(%))) e a quantidade de grupos igual a 5 (K=5), foram gerados os gráficos apresentados na Figura 11.

Neste algoritmo, cada dado da tabela é visto como um ponto no plano cartesiano, sendo que esses pontos estão ainda divididos em grupos (clusters), que indicam a semelhança desses dados baseados nas variáveis escolhidas. Ainda, na

Figura 11, é possível visualizar a correlação das variáveis juntamente com os respectivos grupos gerados.

Figura 11 - Exemplo de aplicação do *K-Means*



O K-means tem se mostrado bastante útil no entendimento de grandes quantidades de dados, conforme discorrido na Seção 3.3. Frente a isso, levando em consideração a fluidez e versatilidade de sua utilização, optou-se pela adoção deste método para análise dos dados neste trabalho.

4. METODOLOGIA

A metodologia da pesquisa se baseou em uma abordagem de contexto bibliográfico e documental. Pelo bibliográfico analisamos a contextualização do agrupamento de dados por meio da Inteligência Artificial. Pelo documental coletamos informação para a discussão dos resultados. Em modo geral, o processo metodológico foi aplicado por meio da técnica denominada *K-Means*, que é detalhada no Segundo Capítulo, para que possamos analisar o perfil social dos participantes do Enem da edição do ano de 2019 que residem no município de Araguaína - TO, sendo este o local em que o orientando reside.

Por meio da utilização do *K-Means*, foi possível analisar os dados de 8665 participantes da região de Araguaína-TO, formando assim perfis separados por grupos conforme a renda média de cada um. A partir destas informações foi possível verificar quais pessoas obtiveram melhores resultados de acordo com a condição social que possuem.

Os dados que foram analisados neste trabalho foram adquiridos através do *site* do MEC, estando disponibilizados na aba de microdados, junto a diversos outros que podem ser de avaliações ou exames realizados pelo INEP. Todos os dados do Enem vêm em um arquivo compactado, que por sua vez foi extraído, possibilitando o acesso à planilha onde ficam armazenados os microdados de cada participante.

Através de um código UNIX utilizado no Linux foi possível separar apenas os participantes que residem e que realizaram a prova no município de Araguaína-TO. Após encontrar todas as informações das pessoas que atendiam a condição de filtro, foi gerado um arquivo reduzido para que fosse utilizado dentro do Google Collab, plataforma escolhida para executar os códigos.

Foram utilizadas as seguintes variáveis na execução do algoritmo:

- NU_IDADE: idade do participante;
- TP_ANO_CONCLUIU: ano de conclusão do Ensino Médio;
- MEDIA: média aritmética calculada com as notas das provas de Linguagens, Códigos e suas Tecnologias, Ciências Humanas e suas Tecnologias, Ciências da Natureza e suas Tecnologias, Matemática e suas Tecnologias e Redação;

- RENDA: renda familiar média calculada de acordo com as faixas de variação. Isso foi necessário porque o Enem utiliza o conceito de faixa de renda, no entanto, o K-means necessita de valores numéricos.

Além disso, removemos também dados de participantes que por algum motivo possuíam algum campo do formulário socioeconômico em branco, além de participantes faltosos ou que acabaram tirando zero no exame.

Após todos os filtros terem sido aplicados, definimos através do método do cotovelo (THE SCIKIT-YB DEVELOPERS, 2019), que é apresentado no Gráfico 2, qual seria a quantidade ideal de grupos para que os dados dos participantes pudessem ser separados. O cálculo apresentou como ponto ideal a separação em três grupos, sendo assim, começamos a montagem dos gráficos 3 à 8, utilizando como método de identificação o valor de cluster que foi atribuído para cada participante.

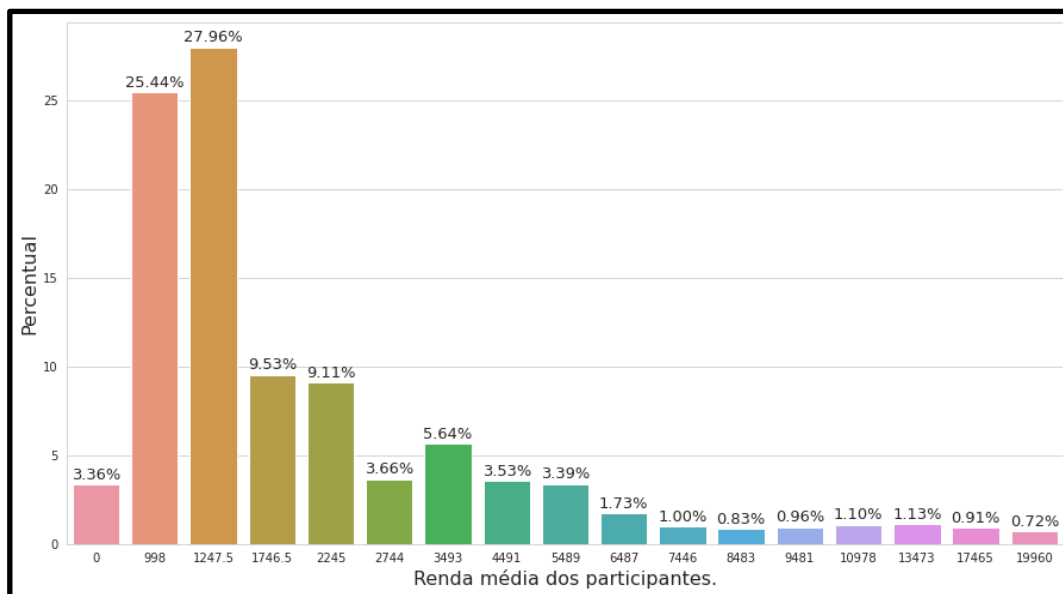
5. RESULTADOS E DISCUSSÕES

Neste capítulo, apresentamos e analisamos, registrados em gráficos, os dados obtidos do agrupamento de dados resultantes da aplicação de técnica de Inteligência Artificial denominada *K-Means*. Aplicando essa técnica de IA na base de dados do Enem conseguimos identificar diferentes contextos da situação econômica, acadêmica e social dos estudantes. Nesse sentido, e conforme constam nos gráficos abaixo, nossa análise contemplou todas as variáveis que pertencem ao formulário socioeconômico, além de variáveis estabelecidas durante o processo de filtragem dos dados que foram utilizados.

Contemplamos também durante a análise dos gráficos, algumas das concepções de Bourdieu (2008) sobre questões que remetem ao capital humano e cultural, apresentando ainda como tais questões estão relacionadas com o tipo de assunto abordado neste trabalho.

Conforme consta no Gráfico 1, abaixo, o agrupamento de dados obtidos pela técnica *K-Means* de IA revelou que a maioria dos estudantes participantes da edição de 2019 do Enem possuíam renda média acima de R\$1.200,00. Isso significa que uma porção significativa dos participantes do Enem pertencem à classe econômica de baixo poder aquisitivo. Nesse sentido, deduz-se que estes estudantes integram o grupo dos que fizeram o ensino médio em escolas públicas.

Gráfico 1 - Porcentagem de participantes que possuem uma determinada renda média



A maior concentração populacional de participantes do Enem corresponde a estudantes, cujas renda média familiar está entre R\$ 990,00 (representando cerca de 25%) e R\$ 1.200,00 (representando quase 28%). Isso denota que a política de institucionalização desse processo avaliativo para ingresso no ensino superior está voltada para pessoas que possuem baixo poder aquisitivo.

Os estudantes que pertencem a famílias com renda média acima de R\$ 4.491,00 até R\$ 6.487,00 representam proporção decrescente de 3.39% para 1.73%. A elevação da renda familiar acima de R\$ 7.446,00 faz com que o número de inscritos no Enem seja inferior a 1%. Esse número chega à casa de 0.72% para os estudantes cuja renda familiar está acima de R\$ 19 mil.

Bourdieu (2008) nos apresenta de forma objetiva como o capital influencia no conhecimento cultural do indivíduo. Vejamos sua concepção:

Os agentes são distribuídos, na primeira dimensão, de acordo com o volume global de capital (desses dois tipos diferentes) que possuam e, na segunda dimensão, de acordo com a estrutura de seu capital, isto é, de acordo com o peso relativo dos diferentes tipos de capital, econômico e cultural, no volume global de seu capital. (BOURDIEU, 2008, p. 18).

De modo geral, os indicadores registrados no Gráfico 1 apontam que as escolas públicas têm papel relevante na formação da maior parte dos estudantes do ensino médio, pois são eles que terão possibilidade de acesso ao ensino superior pelas instituições superiores públicas e gratuitas.

Os demais indicadores agrupados pela técnica *K-means* de Inteligência Artificial apontaram, segundo gráficos adiante, que os participantes do Enem na edição de 2019 correspondiam a cidadãos de poder aquisitivo alinhado à renda básica brasileira, cerca de apenas 1 salário mínimo.

Para que pudéssemos obter a quantidade ideal de grupos que serão observados, utilizamos o método de Elbow, conhecido também como método do cotovelo. Este método se baseia em determinar a quantidade ideal de grupos que podem estar em um agrupamento, sem conhecer previamente nenhum dos dados utilizados. Na aplicação ao nosso trabalho, foi feita a análise destes dados e foi possível identificar que a quantidade de grupos ideal para nosso caso seria 3, como é apresentado no Gráfico 2.

Gráfico 2 - Método do Cotovelo (Elbow) para o K ideal

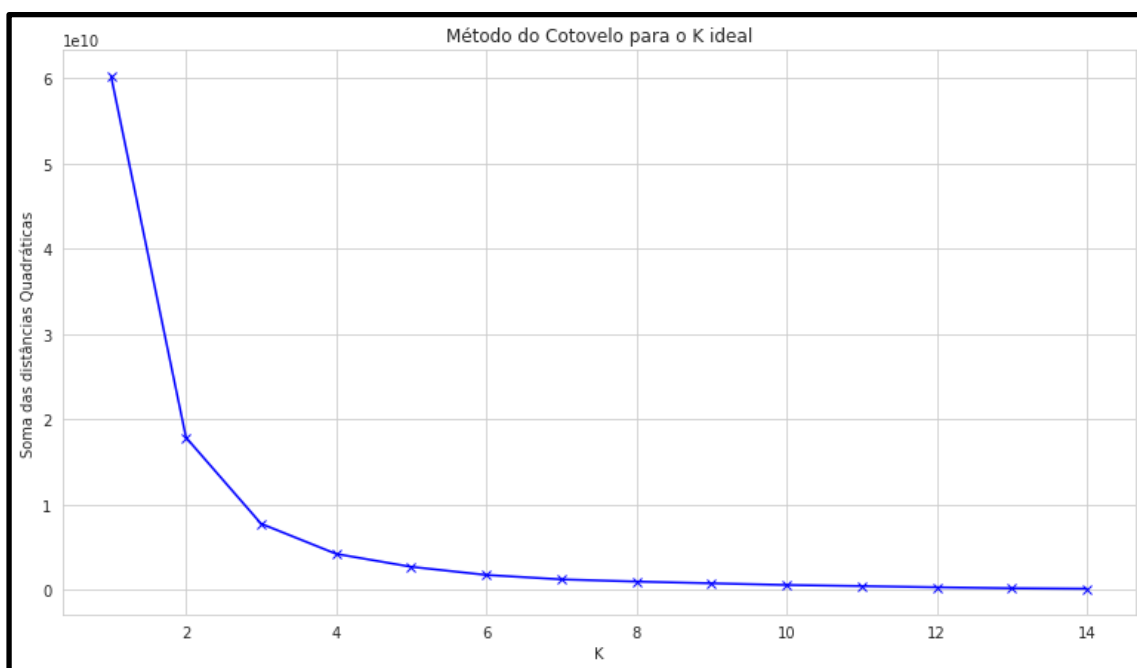
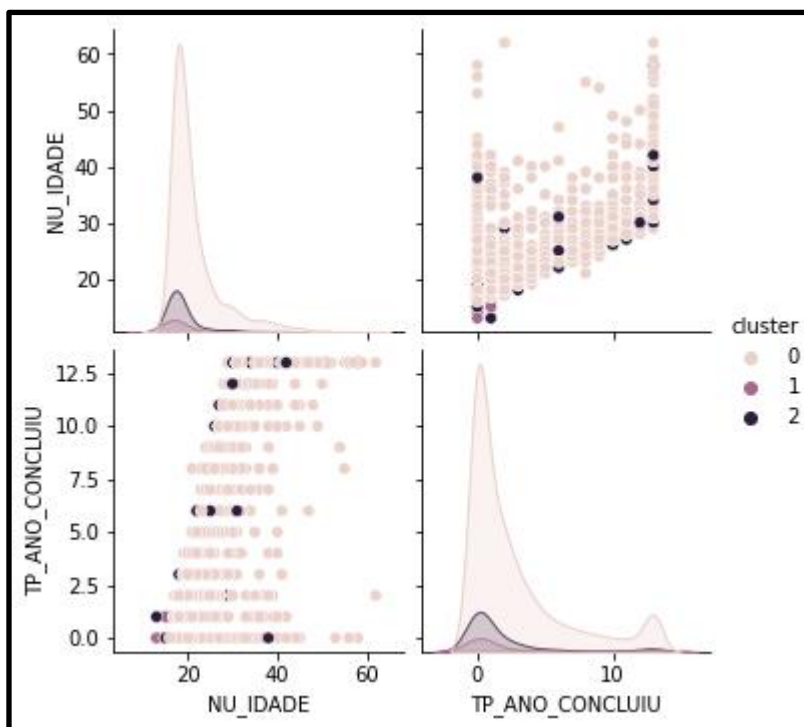


Gráfico 3 - Gráfico de dispersão com os indicadores do ano de conclusão e idade

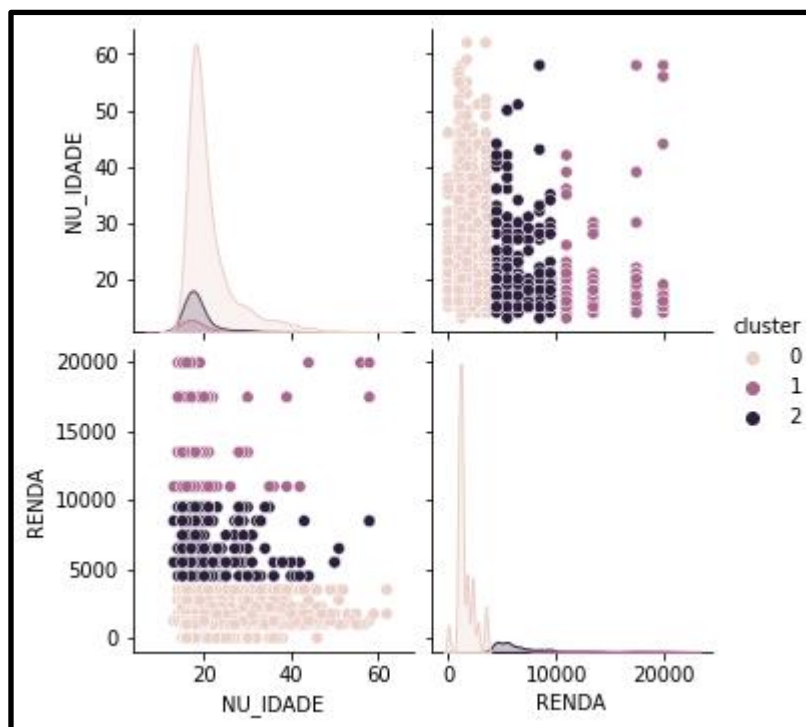


O gráfico 3 apresenta um quadro de dispersão, relacionando a idade dos participantes ao ano de conclusão do ensino médio. Ele está dividido em três grupos, sendo o primeiro deles pertencente ao cluster 0, e identificados na cor mais clara. O segundo grupo, pertence ao cluster 1, que é identificado através da cor roxa. E por fim o cluster 2, podendo ser identificado a partir da cor escura.

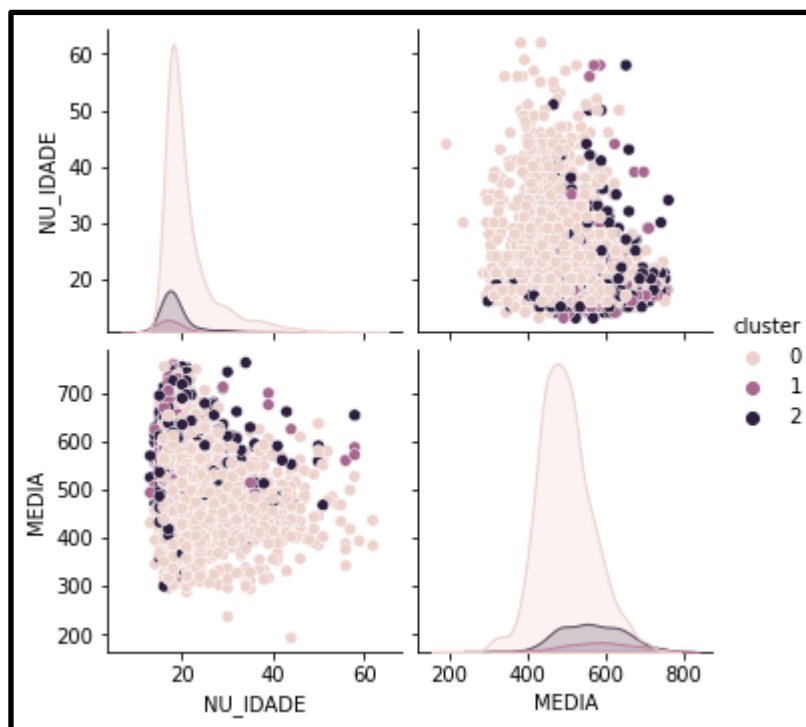
Através das variáveis TP_ANO_CONCLUIU e NU_IDADE que estão presentes no gráfico, é possível averiguar que uma boa parte dos participantes que possuem menos de 20 anos de idade, ainda não concluíram, ou tem no máximo 5 anos de conclusão do Ensino Médio. Enquanto a maior parcela dos que irão realizar o exame, se encontra entre 21 e 40 anos de idade.

O interesse tardio em realizar o exame, pode ocorrer principalmente pela necessidade de se obter uma qualificação para atuação no mercado de trabalho, ou porque muitos veem a faculdade como um desafio complexo e exaustivo, entendendo sua real importância apenas após perceber que o conhecimento obtido em algum curso superior pode aumentar significativamente sua renda.

Gráfico 4 - Gráfico de dispersão com os indicadores de renda média e idade

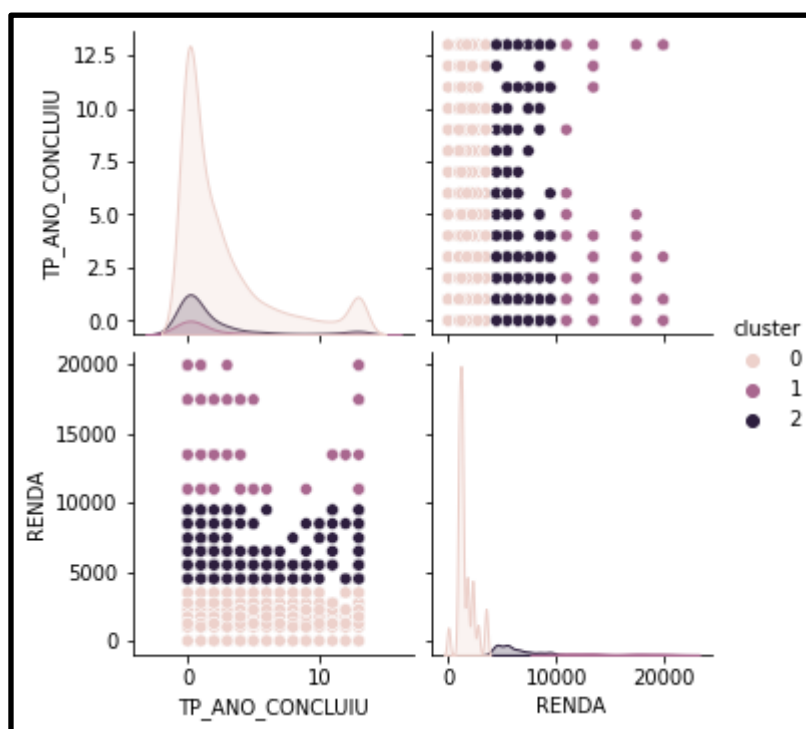


O gráfico 4 nos apresenta o relacionamento entre as variáveis NU_IDADE e RENDA, referindo-se a renda média dos participantes e idade que possuem. Podemos perceber que a maioria das pessoas se concentram entre os 15 e 32 anos de idade, indo desde a renda zero, até o valor máximo de R\$ 19.960 reais.

Gráfico 5 - Gráfico de dispersão com os indicadores da nota média e idade

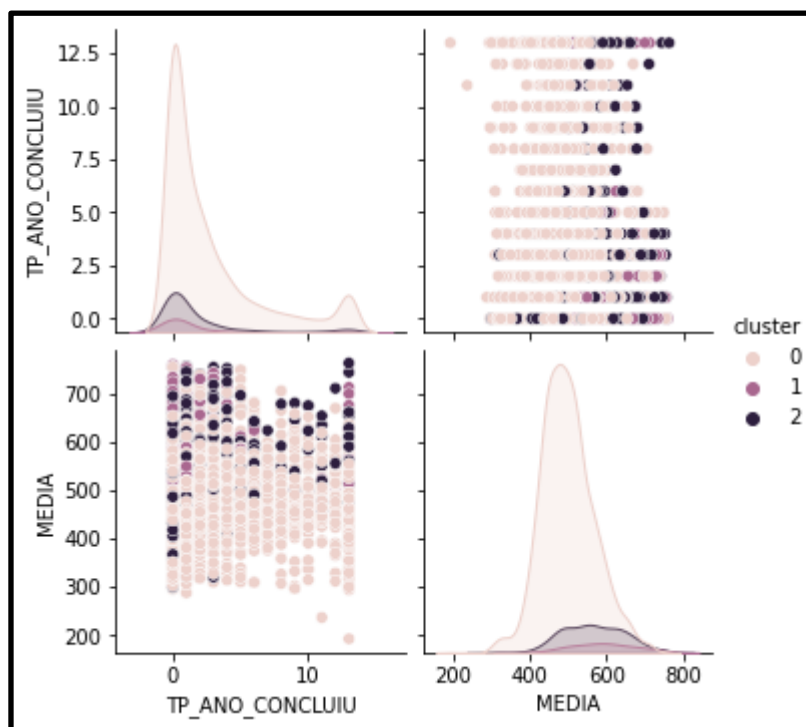
O gráfico 5 apresenta a forma em como a idade dos participantes está relacionada com a nota média. Ele mostra que a maior parte dos participantes com nota média acima de 600 pontos são indivíduos que têm de 15 a 35 anos.

Gráfico 6 - Gráfico de dispersão com os indicadores do ano de conclusão e renda média

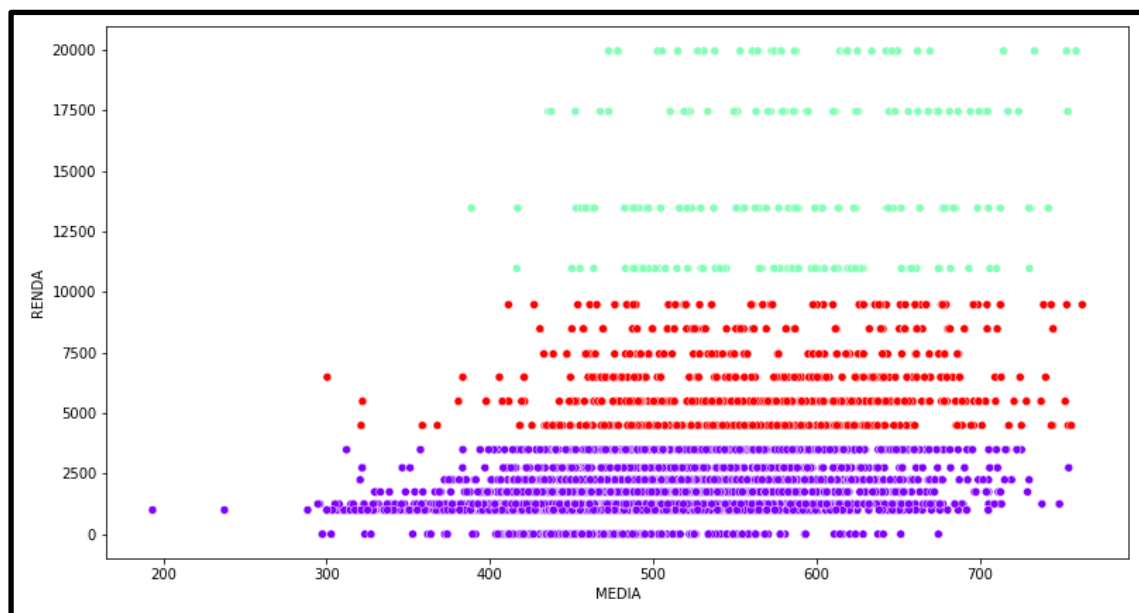


No gráfico 6 analisamos a forma em que a renda média e o ano de conclusão estão relacionados. Neste gráfico podemos observar que os participantes estão bem distribuídos, ainda que possuam alguns indivíduos com as rendas médias acima de dez mil reais. Nota-se que uma boa parte dos que informam ter uma renda média elevada, ainda não concluíram o Ensino Médio ou tem menos de 5 anos de conclusão.

Gráfico 7 - Gráfico de dispersão com os indicadores do ano de conclusão e nota média



No gráfico 7 podemos observar a correlação entre a nota média e o ano de conclusão dos participantes. Os resultados mostram-se mais satisfatórios entre os indivíduos que ainda não concluíram, ou que possuem até 5 anos desde a conclusão do Ensino Médio, mas ainda demonstrando pontuações relevantes dentro os que concluíram a mais de 5 anos.

Gráfico 8 - Gráfico de plotagem das variáveis nota e renda média

No gráfico 8, podemos perceber a quantidade de pessoas que obtiveram resultado acima de 600 pontos. Apesar de percebermos um grande número de participantes que alcançaram notas que podem lhes dar o direito ao ingresso no ensino superior, ainda temos um grande número de participantes com resultados mais baixos. Conforme a renda escala, percebemos que o número de notas consideradas satisfatórias avança ainda mais dentre os indivíduos que estão naquela faixa de renda média, sendo cada linha do gráfico a representativa das faixas de renda informada no formulário socioeconômico.

Isso coincide com a fala de Bourdieu (2008) sobre o fato de quanto maior o capital econômico de um indivíduo, maior será a facilidade de se ter acesso ao capital humano e capital cultural, agregando assim conhecimento e possibilidade de se obter uma nota mais elevada na realização do exame.

O gráfico 9, apresenta a quantidade de participantes no exame, conforme o sexo informado durante a inscrição. A maioria dos participantes na região pertencem ao grupo feminino (3811). Já no grupo masculino, esse quantitativo totalizou 2558 participantes.

Empiricamente, os homens preferem trabalhar desde cedo que ingressar em uma instituição de ensino superior. Isso acontece porque eles entendem que precisam ganhar dinheiro para ter condições de realizar seus sonhos e objetivos de vida, além de contribuir com o sustento financeiro de suas famílias. Isso também acontece porque a sociedade impõe a eles que precisam prover o sustento e segurança de seus familiares.

Gráfico 9 - Quantidade de participantes determinado por sexo

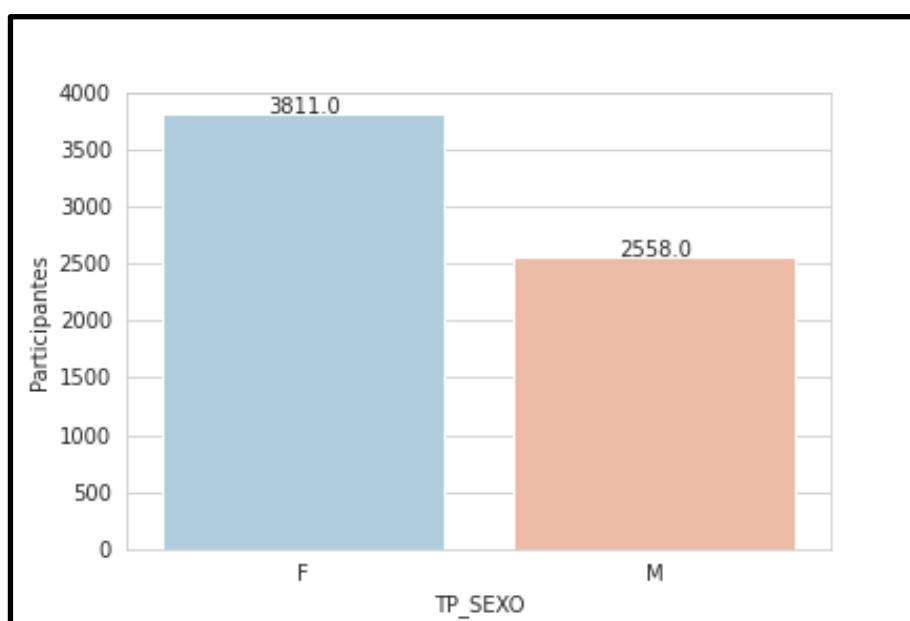
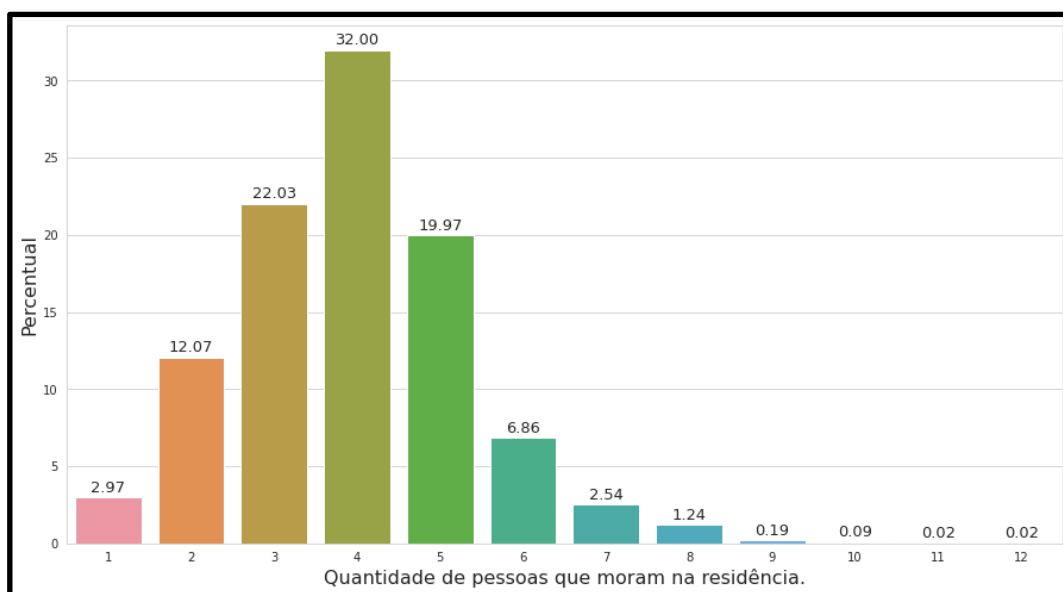
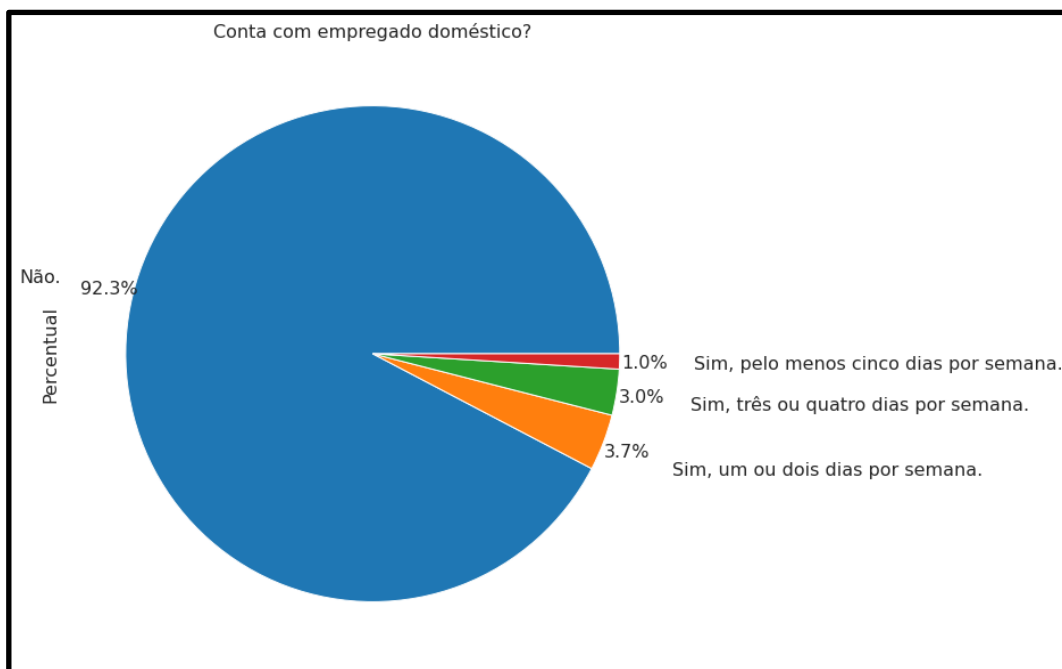


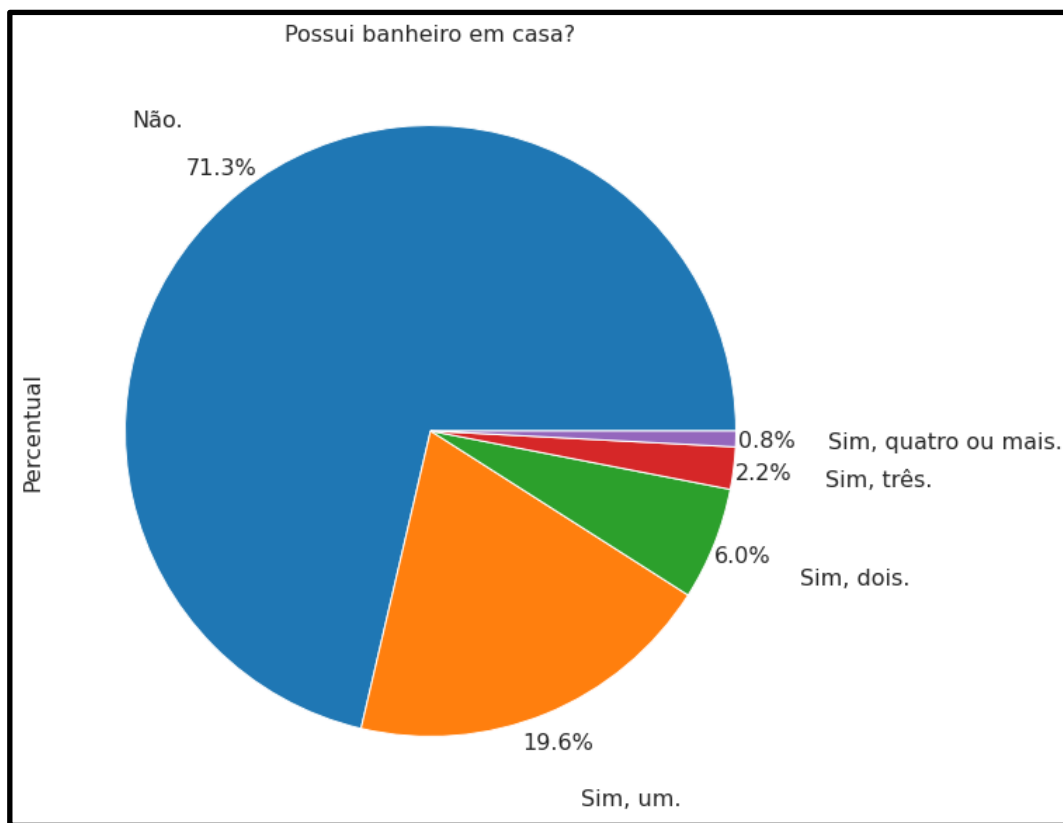
Gráfico 10 - Quantidade de pessoas que moram em uma residência

A situação familiar dos participantes do Enem, conforme dados registrados no Gráfico 10, indica que a maioria das residências têm 4 pessoas, o que corresponde a 32%. As famílias com mais de 7 pessoas na mesma residência representam cerca de 2.5%. De modo geral, a maior concentração de participantes do Enem está entre famílias que têm de 3 a 5 pessoas na mesma residência.

Gráfico 11 - A presença de empregado doméstico na residência

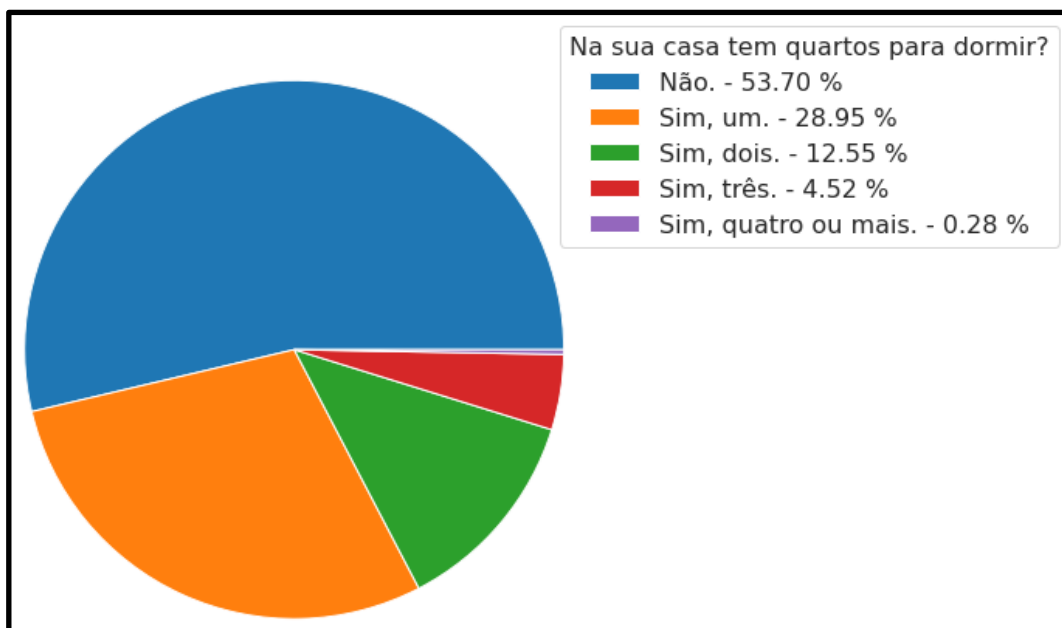
Os dados quantitativos identificados a respeito das famílias dos participantes do Enem que têm empregado doméstico apontam que a maior parte (92.3%) de suas famílias não usufruem desse serviço. Nesse sentido, e conforme apresentado no Gráfico 11, os serviços de empregado doméstico são prestados regularmente apenas para 1.0% das famílias do público do Enem.

Em linhas gerais, há a percepção de que esse exame representa a oportunidade de ingresso no ensino superior para cidadãos de baixa renda. Isso se dá porque eles não apresentam as condições necessárias ao ingresso em cursos superiores de instituições de ensino privadas.

Gráfico 12 - Quantidade de banheiros na residência

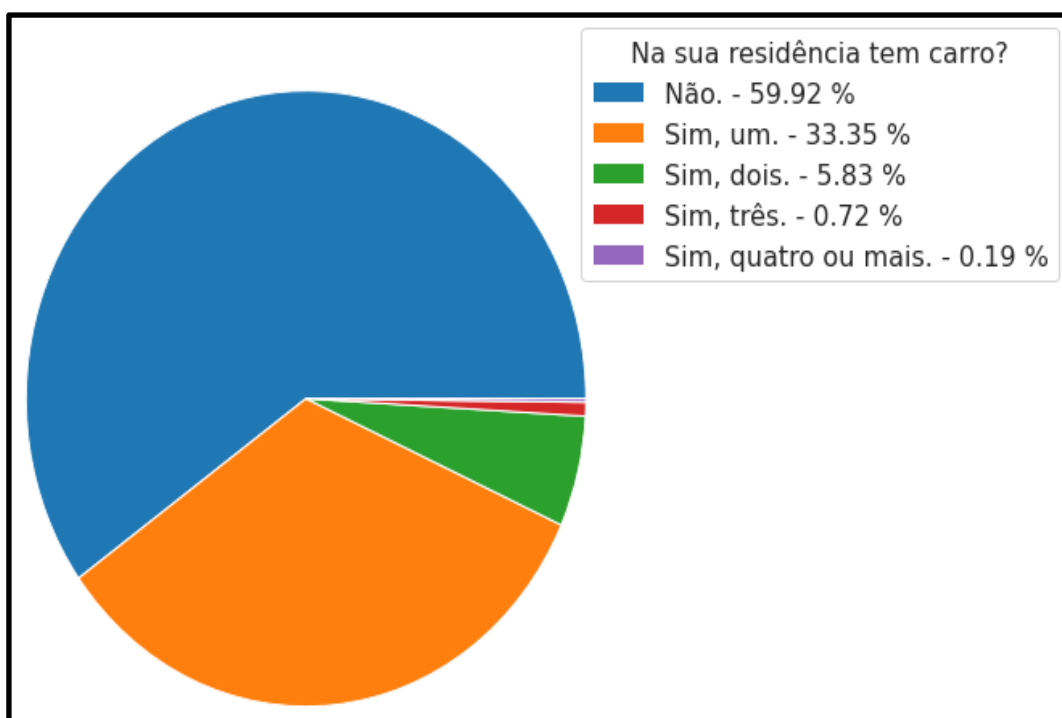
A abordagem a respeito do quantitativo de banheiros nas residências dos participantes do Enem revelou, segundo o Gráfico 12, acima, um dado preocupante socialmente. A afirmativa de não possuir banheiro em casa foge do sentido da vida, pois todas as pessoas precisam de banheiros para suas necessidades fisiológicas.

O quantitativo de que 71.3% das famílias não têm banheiro, salvo informação errônea ou omissão, pode significar, também, que parte delas possuem banheiro na parte externa de suas casas, o que é sensato. Apesar disso, o dado de 19.6% gera o conflito em relação à primeira resposta.

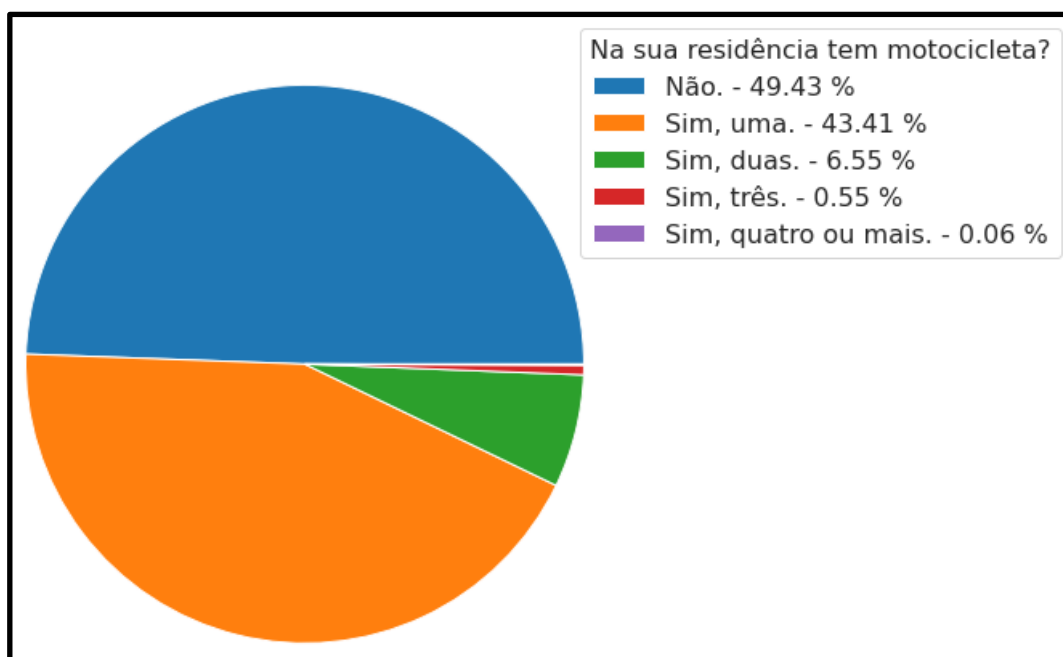
Gráfico 13 - Quantidade de quartos na residência

A informação de que 53.70% das famílias não possuem quarto para dormir assume a mesma inquietação apresentada na primeira resposta do Gráfico 13. No entanto, não se descarta a possibilidade de que algumas residências apresentem estruturas sem quartos.

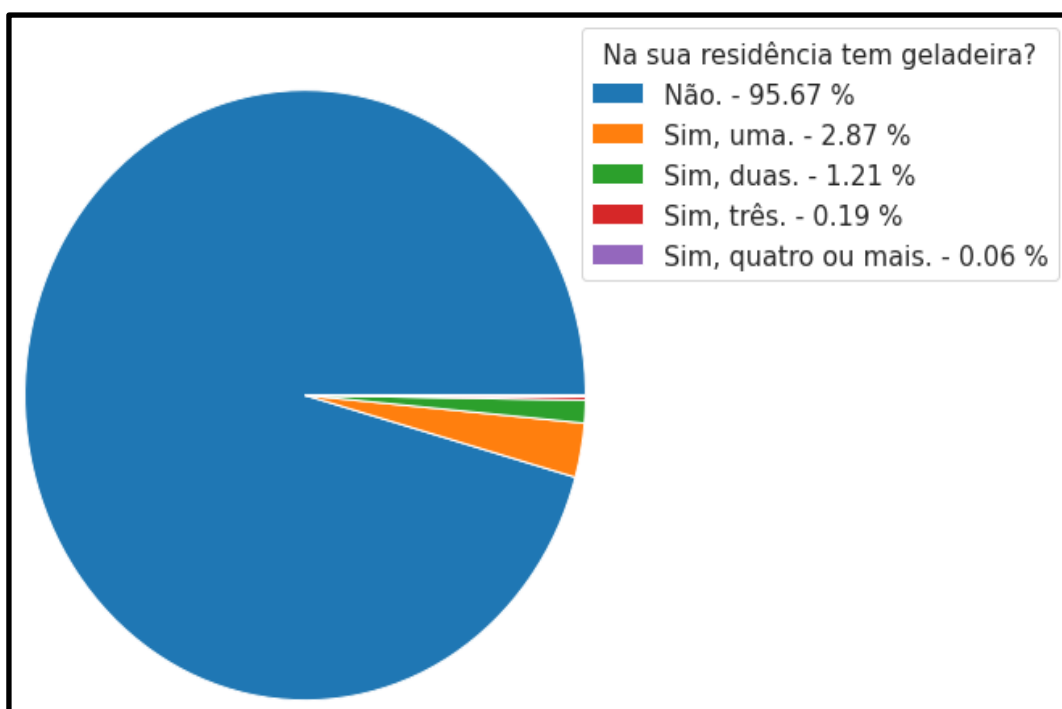
Diante do contexto da renda financeira apresentada no Gráfico 1, as residências dos participantes do Enem, segundo dados do Gráfico 13, apresentam estrutura mínima. Isso se constata no indicativo de existência de apenas um quarto para mais de 28% das famílias. No mesmo sentido, as famílias que possuem casas com mais de quatro quartos representam menos de 0.5%.

Gráfico 14 - Quando de carros na residência

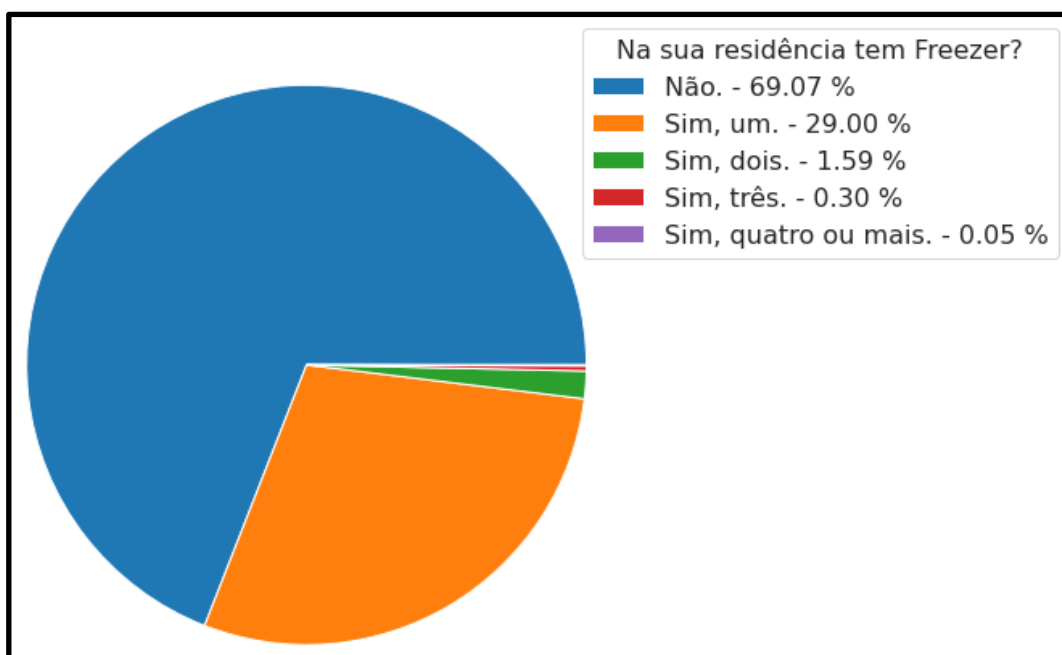
O poder aquisitivo das famílias dos participantes do Enem pode ser observado também em relação aos bens patrimoniais. Quase 60% delas não possui carros, sugerindo que a sua locomoção seja feita a partir de transporte público ou, na maioria das vezes, a pé. Essa percepção se acentua no quantitativo de que apenas 0.19% possui carro próprio.

Gráfico 15 - Quantidade de motocicletas na residência

O Gráfico 15, apresenta que 49.43% dos participantes do Enem não possuem motocicleta; já 43.41% afirmam possuir ao menos uma. Esse número é reduzido para 6.55% dos participantes quando se relata ter duas. No caso de três motocicletas existe uma porcentagem de 0.55%, e somente 0.06% dos entrevistados afirmam ter 4 ou mais motocicletas.

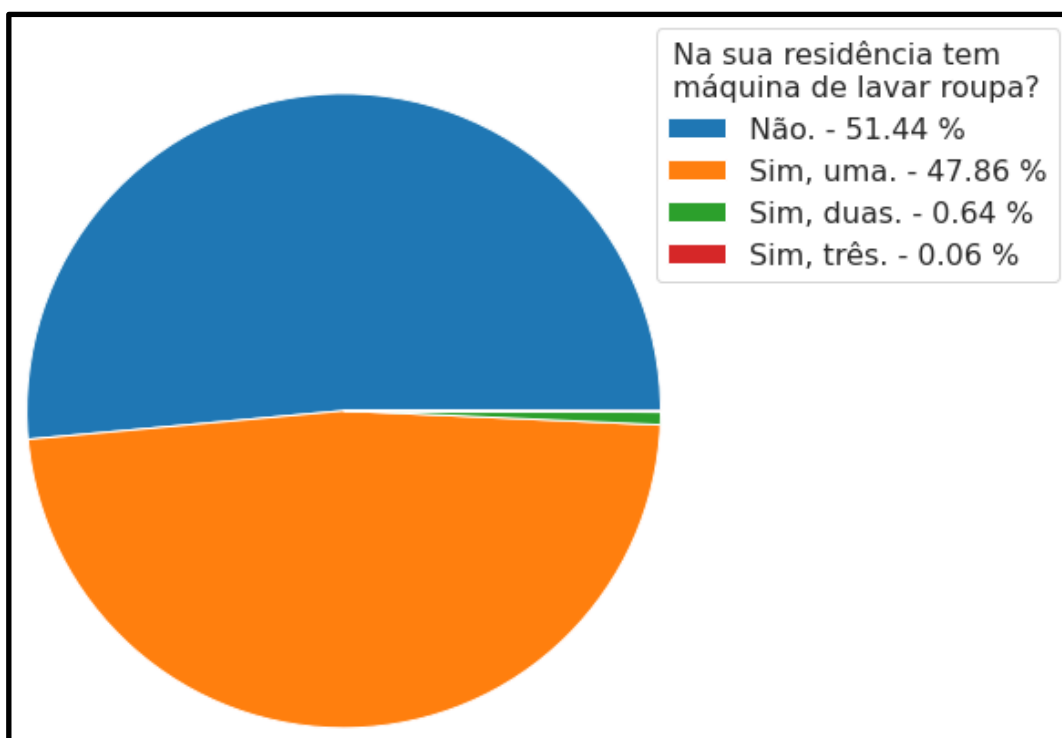
Gráfico 16 - Quantidade de geladeiras na residência

No Gráfico 16 é demonstrado que 95.67% dos participantes do Enem não dispõem de geladeiras em suas residências, o que é um dado preocupante, pois está diretamente relacionado à qualidade dos alimentos que são armazenados, e posteriormente ingeridos por estas famílias. No caso dos que possuem ao menos uma, temos a porcentagem de 2.87%. Enquanto 1.21% possui mais de duas, há um quantitativo menor ainda, ou seja, há apenas 0.19% que têm três, e por fim aqueles com quatro ou mais geladeiras totalizam 0.06% do total.

Gráfico 17 - Quantidade de Freezers na residência

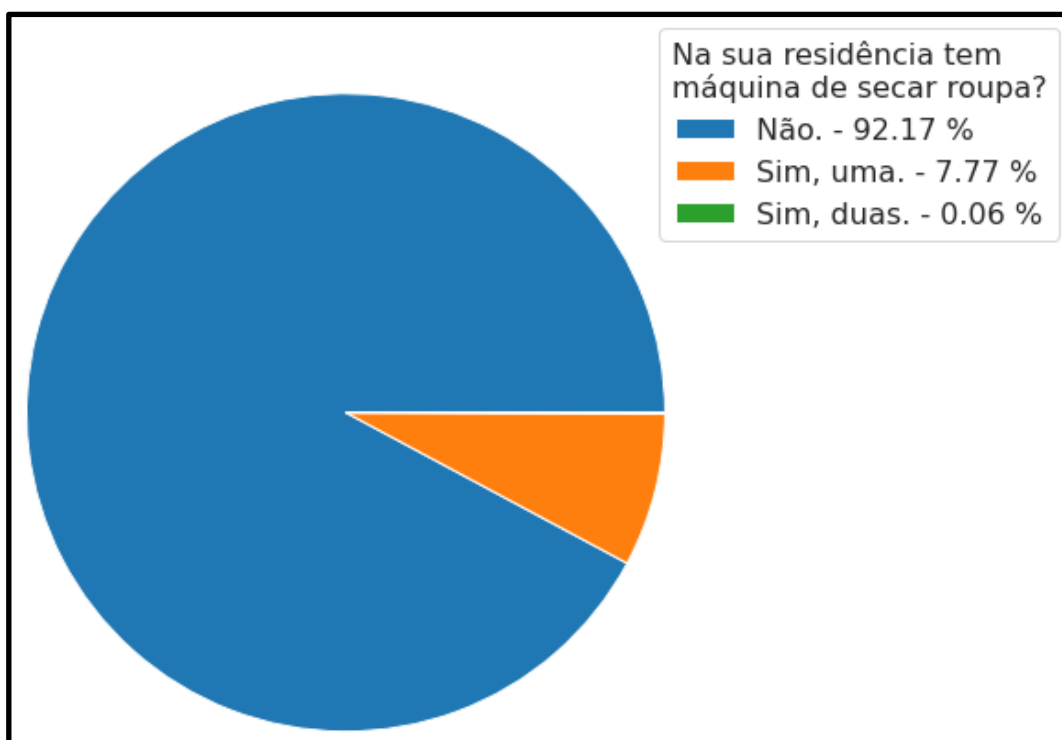
O Gráfico 17 demonstra que a maior parte dos participantes, 69.07% não contam com um Freezer em casa, à medida que 29.00% possuem ao menos um. No caso dos que afirmaram possuir mais de um temos: 1.59% com dois, 0.30% com três e 0.05% com quatro ou mais.

Gráfico 18 - Quantidade de máquinas de lavar roupa na residência

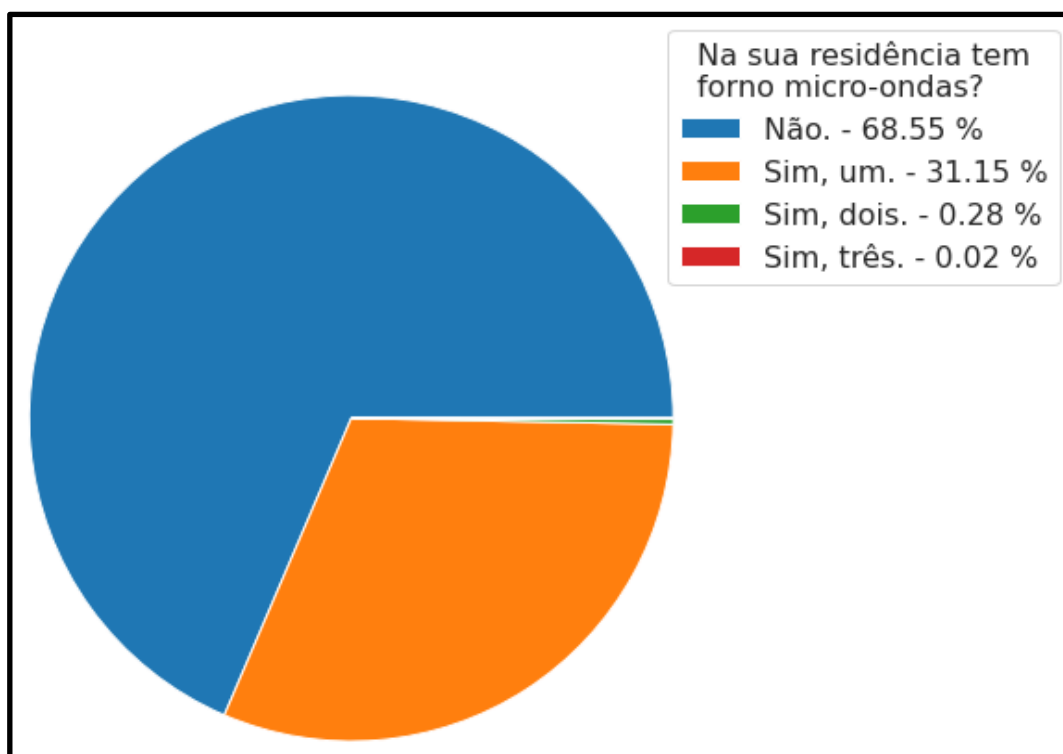


Sobre o Gráfico 18, observa-se que em 51.44% das respostas constam possuir nenhuma máquina de lavar em sua residência, enquanto 47.86% dos participantes relatam ter ao menos uma, 0.64% contam com duas e 0.06% do total possuem três.

Gráfico 19 - Quantidade de máquina de secar roupa na residência

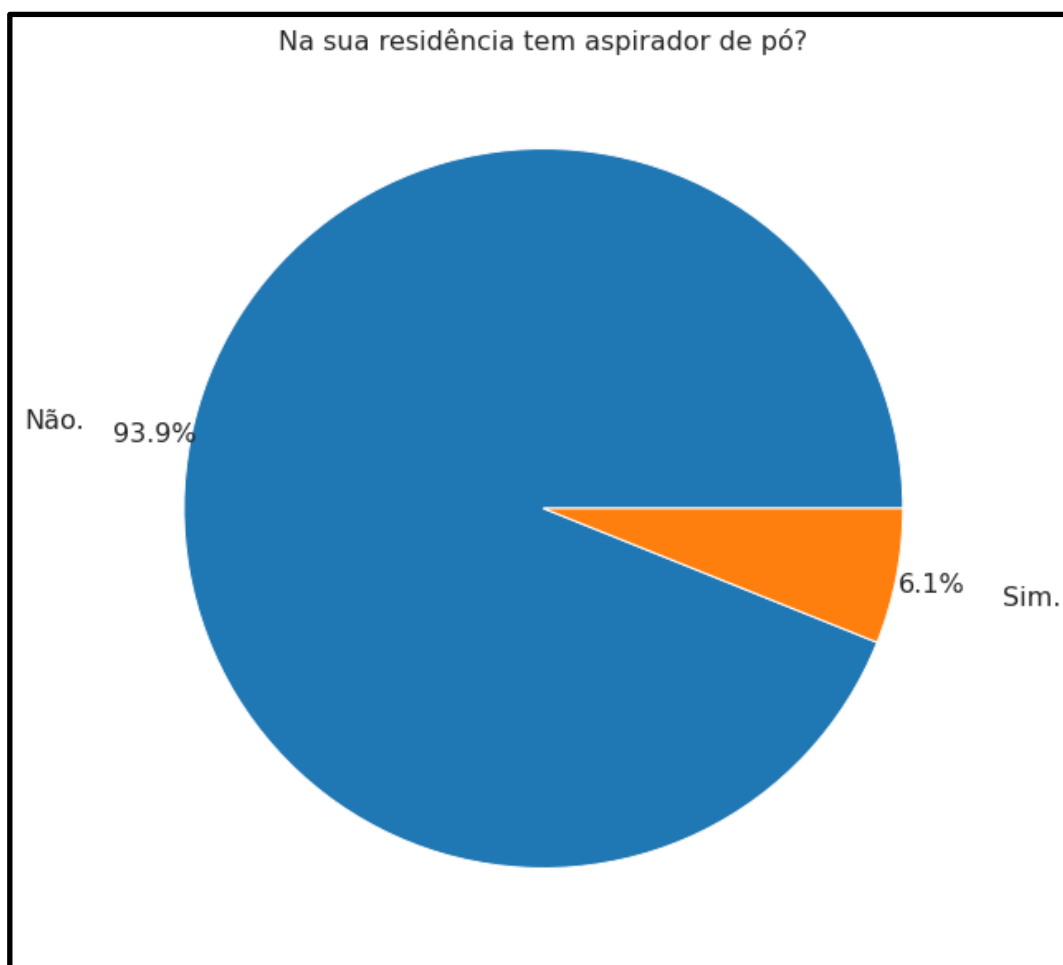


O Gráfico 19, que apresenta o quantitativo quanto a existência de máquinas de secar nas residências dos participantes, expõe que 92.17% responderam não ter nenhuma, enquanto 7.77% afirmaram que possuem ao menos uma, e 0.06% dentre todos, pelo menos duas máquinas de lavar.

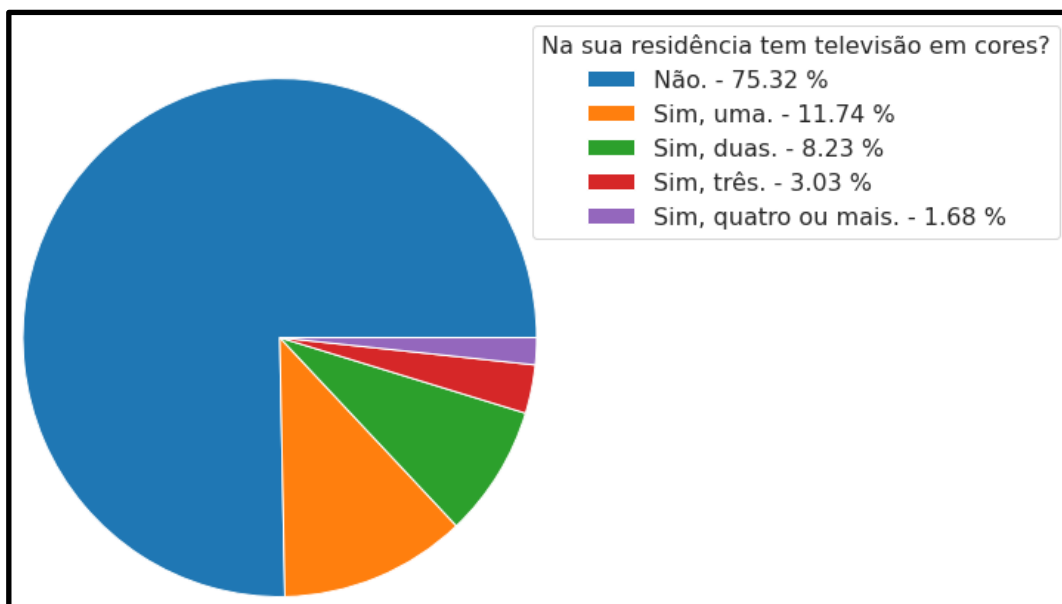
Gráfico 20 - Quantidade de forno micro-ondas na residência

O Gráfico 20 mostra que 68% dos participantes, de forma geral, não possuem nenhum forno micro-ondas, enquanto 31.15% afirmam ter ao menos um em sua residência. Esse número reduz ainda mais quando se trata de possuir de dois ou até três micro-ondas, sendo eles, respectivamente, 0.28% e 0.02%.

Gráfico 21 - A disponibilidade de um aspirador de pó na residência

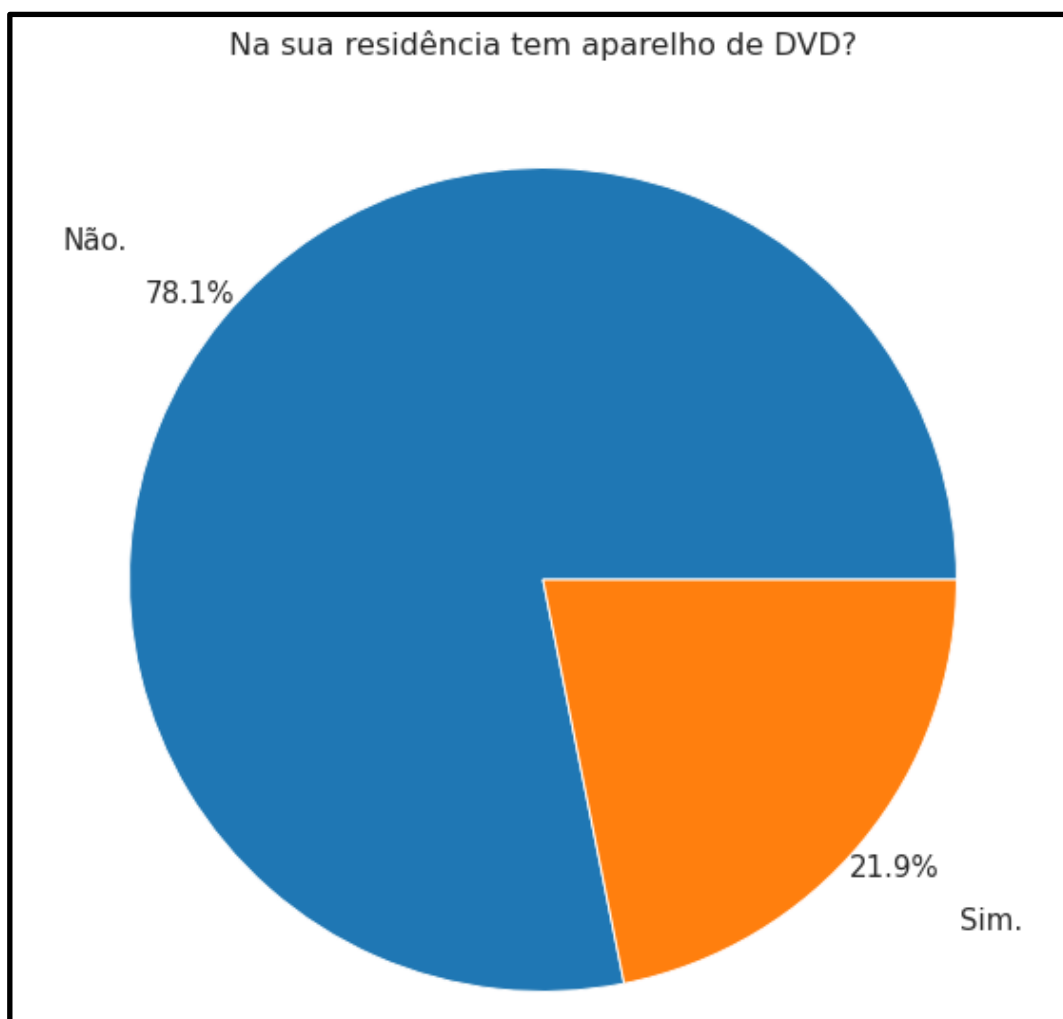


No Gráfico 21 é demonstrado que 93.9% dos participantes responderam não possuir nenhum aspirador de pó em suas residências, enquanto 6.1% dos participantes afirmam possuir ao menos um.

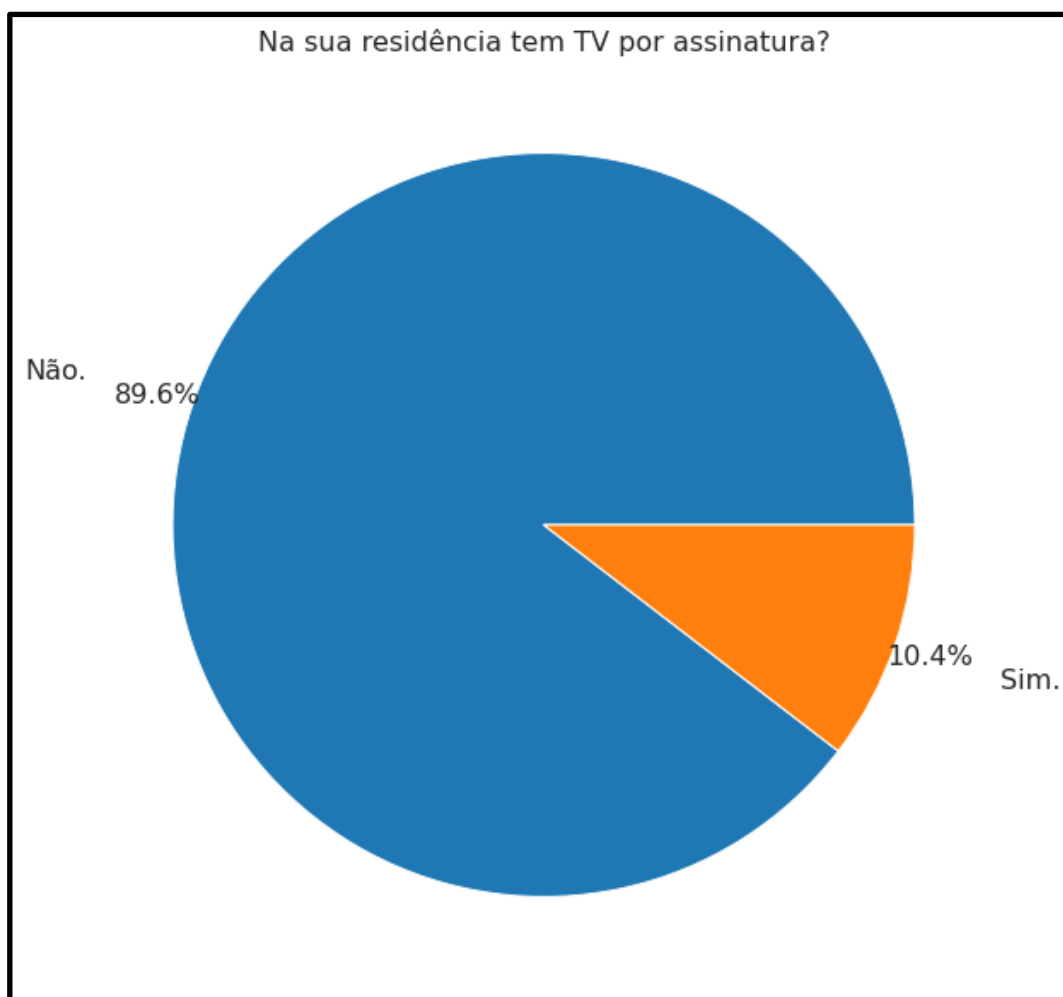
Gráfico 22 - Quantidade de Televisores na residência

O Gráfico 22 apresenta que 75.32% dos participantes do Enem não possuem televisão em cores em casa, já 11.74% afirmam possuir ao menos uma. Esse número é reduzido para 8.23% dos participantes quando se relata ter duas. No caso de três, apenas 3.03% e somente 1.68% do total demonstra ter quatro ou mais aparelhos em casa.

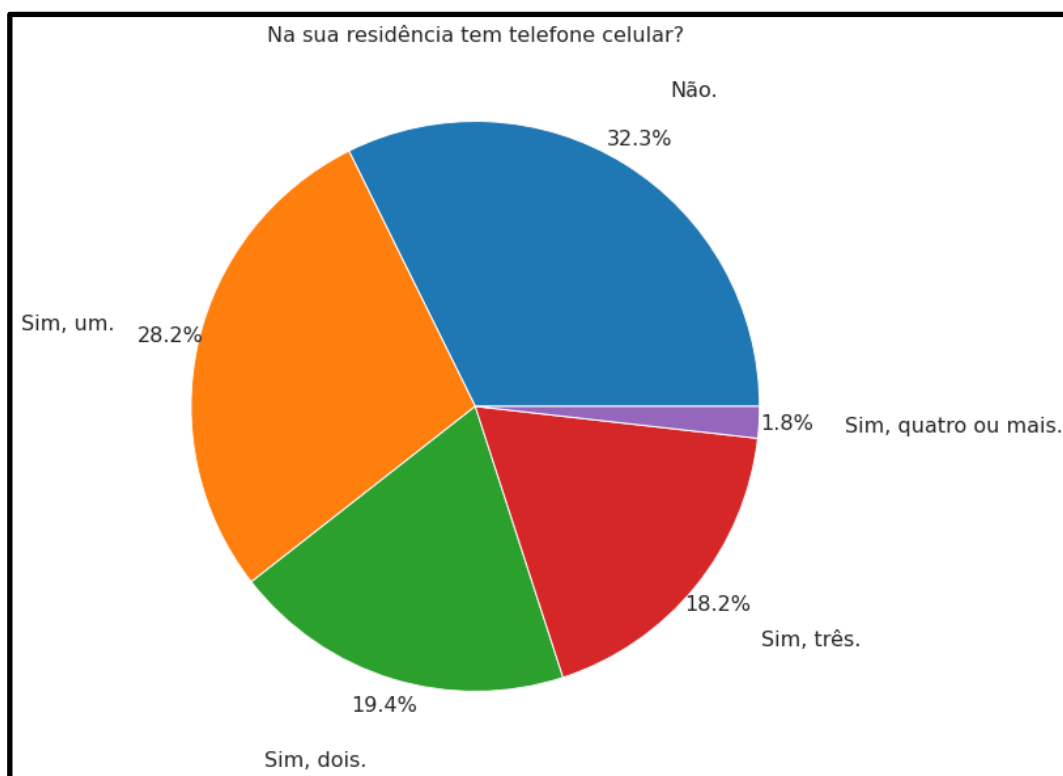
Gráfico 23 - A presença de aparelho DVD na residência



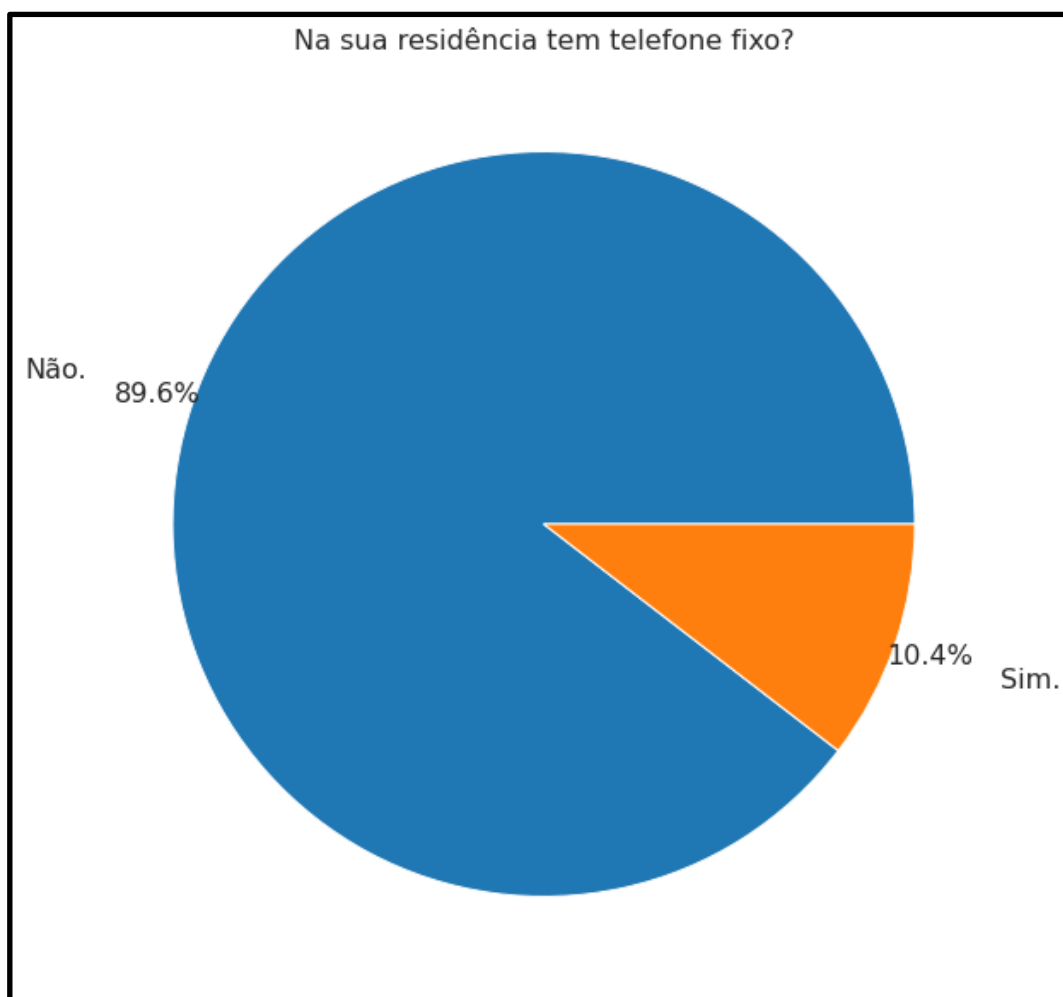
O Gráfico 23 apresenta que 78.1% dos participantes responderam não possuir nenhum aparelho de DVD na residência, enquanto 21.9% destes afirmam possuir ao menos um.

Gráfico 24 - A presença de TV por assinatura na residência

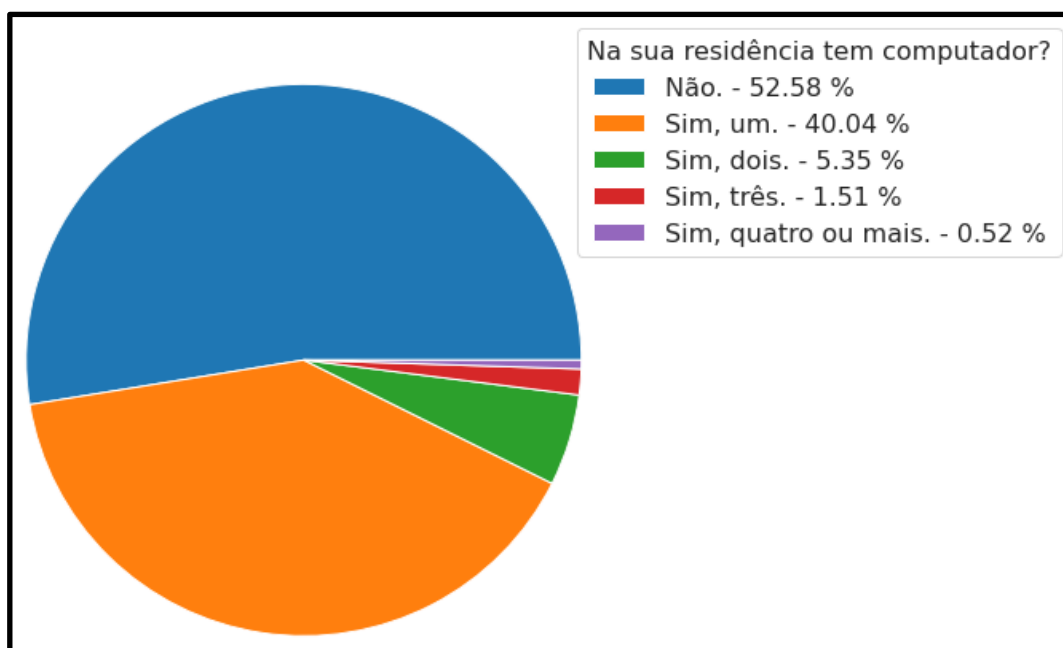
No Gráfico 24, quanto a presença de TV por assinatura na residência, 89.6% dos participantes responderam que não possuem TV por assinatura, ao passo que 10.4% respondeu que sim, que possuem TV por assinatura.

Gráfico 25 - A presença de telefone celular na residência

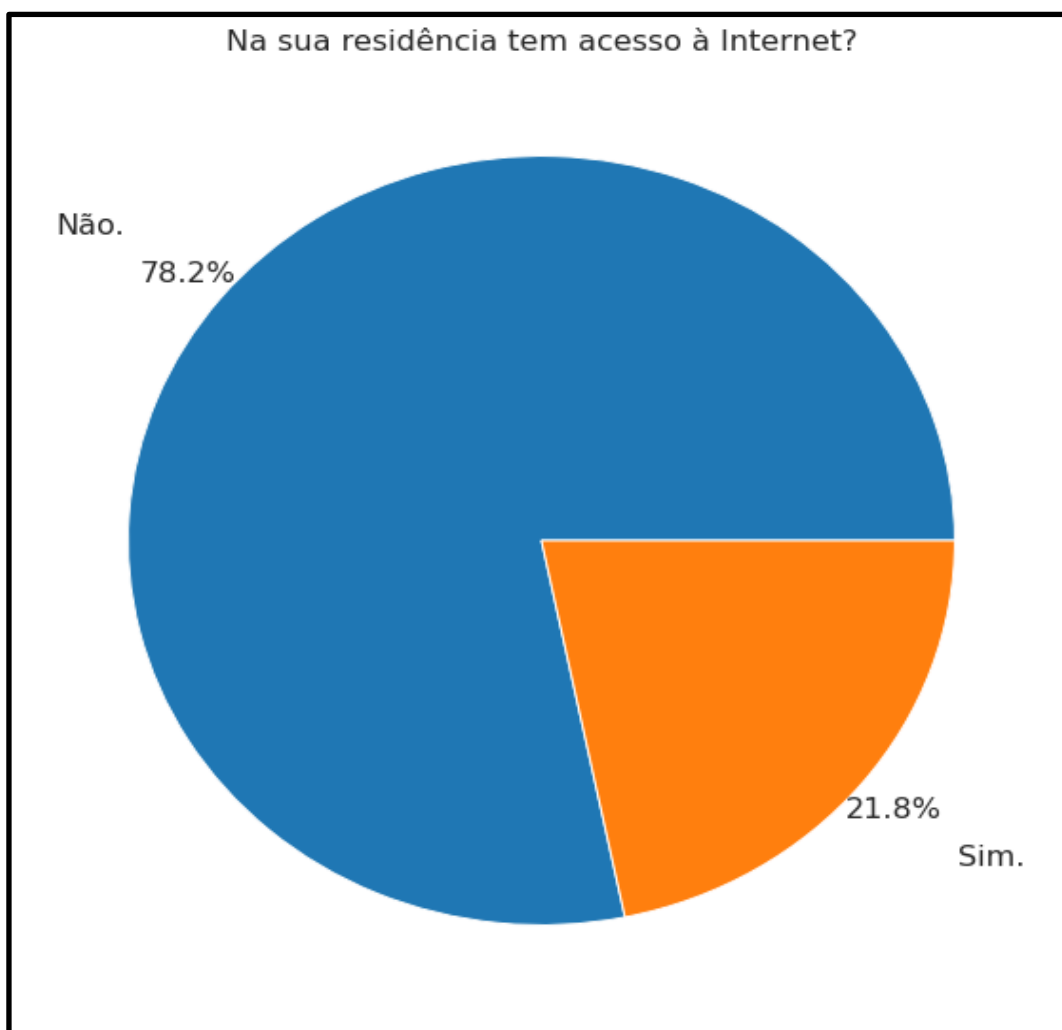
No gráfico 25, é possível percebermos que os aparelhos celulares estão presentes em boa parte das famílias, mas mesmo com estes dados, vemos que um grupo de 32.3% dos participantes afirmou não ter nenhum dispositivo celular em sua residência. Já as pessoas que afirmaram ter ao menos um dispositivo consistem em 28.2%, enquanto 19.4% informam possuir apenas dois dispositivos, e 18.2% ao menos três. Por fim, uma minoria dos participantes disse possuir quatro ou mais dispositivos, representando 1.8% do total.

Gráfico 26 - A presença de telefone fixo na residência

O Gráfico 26 apresenta que 89.6% dos participantes responderam não possuir nenhum telefone fixo em suas residências, enquanto 10.4% dos participantes afirmam possuir ao menos um.

Gráfico 27 - Quantidade de computadores na residência

O Gráfico 27 apresenta que 52.58% dos participantes do Enem não possuem computadores em casa, já 40.04% afirmam possuir ao menos um. Esse número é reduzido para 5,35% dos participantes quando se relata ter duas, no caso de três apenas 1.51%, e somente 0.52% do total demonstra ter quatro ou mais aparelhos em casa.

Gráfico 28 - A disponibilidade de acesso à internet na residência

O Gráfico 28, quanto a disponibilidade de acesso à internet na residência, 78.2% dos participantes responderam que não, não possuem acesso à internet, ao passo que 21.8% responderam que possuem acesso à internet em suas residências.

CONSIDERAÇÕES FINAIS

O Exame Nacional do Ensino Médio foi instituído pelo Ministério da Educação e Cultura como um instrumento de acesso ao Ensino Superior. Sua primeira edição foi realizada no ano de 1998 e deu início à realização de exames em larga escala.

Com o objetivo de identificar e avaliar os padrões dos perfis dos participantes do Enem no ano de 2019, no município de Araguaína-TO, este trabalho monográfico utilizou técnicas de Inteligência Artificial (*K-Means*) para o agrupamento dos dados.

A metodologia utilizada seguiu a abordagem bibliográfica e documental. A bibliográfica possibilitou a compreensão teórica e o uso da Inteligência Artificial. A documental, a partir da técnica *K-Means*, favoreceu a análise dos dados coletados dos participantes do Enem no ano de 2019 no município de Araguaína-TO. Com a teoria conceitual de Bourdieu, com cunho sociológico, fizemos uma análise crítica dessa realidade, evidenciando o quanto é longo o caminho entre as classes sociais e que tais demandas incidem em nossas vidas, a partir dos bens culturais, conhecimentos e saberes a que acessamos.

De modo geral, a pesquisa revelou que a IA, orientada pela técnica denominada *K-Means*, favorece o agrupamento de dados relacionados a exames de larga escala, como é o caso do Enem. Além disso, e no teor qualitativo, possibilita a análise de suas informações relacionadas a conjunturas econômicas, sociais e acadêmicas.

Concluimos, portanto, que esta pesquisa foi relevante para nossos estudos. Por conseguinte, o agrupamento de dados a partir de técnicas de IA apresenta possibilidades de novas pesquisas em razão da facilidade em analisar grandes quantidades de dados.

REFERÊNCIAS

AGUADO, Alexandre Garcia; CANTANHEDE, Marco André. **Lógica fuzzy**. Artigo sem, 2010.

AMARAL, Fernando. **Introdução à Ciência de Dados**: mineração de dados e big data. Alta Books Editora, 2016.

ARAUJO, Eutalia Aparecida Candido de; ANDRADE, Dalton Francisco de; BORTOLOTTI, Silvana Ligia Vincenzi. **Teoria da resposta ao item**. Revista da Escola de Enfermagem da USP, v. 43, p. 1000-1008, 2009.

BACKES, André Ricardo; JUNIOR, Jarbas Joaci de Mesquita Sá. **Introdução à visão computacional usando Matlab**. Alta Books Editora, 2019.

BOURDIEU, Pierre. **Razões práticas**: Sobre a teoria da ação. Tradução: Mariza Corrêa - Campinas, SP; Papirus, 2008.

BRASIL. MINISTÉRIO DA EDUCAÇÃO. **Base Nacional Comum Curricular**. 2017. Disponível em: <<http://basenacionalcomum.mec.gov.br/>>. Acesso em: 27 de nov. 2021.

BRASIL. MINISTÉRIO DA EDUCAÇÃO. **Teoria de resposta ao item avalia habilidade e minimiza o “chute” de candidatos**. 2011. Disponível em: <<http://portal.mec.gov.br/ultimas-noticias/389-ensino-medio-2092297298/17319-teoria-de-resposta-ao-item-avalia-habilidade-e-minimiza-o-chute/>>. Acesso em: 28 de nov. 2021.

CARVALHO, Francisca Eliane Dias de. **Fatores socioeconômicos associados ao desempenho dos estudantes na prova de redação do Exame Nacional do Ensino Médio (ENEM)**. 2017.

CAVALCANTI, Diana Cabral. **Uma abordagem não supervisionada para classificação de opinião usando o recurso léxico SentiWordNet**. 2011. Dissertação de Mestrado. Universidade Federal de Pernambuco.

DE OLIVEIRA, Vitória Peres. **Uma informação tácita**. DataGramZero. Revista de Ciência da Informação, v. 6, n. 3, 2005.

DUTRA, Moisés Lima; DE MACEDO, Douglas Dyllon Jeronimo. **Curadoria digital**: proposta de um modelo para curadoria digital em ambientes big data baseado numa abordagem semi-automática para a seleção de objetos digitais. Informação & Informação, v. 21, n. 2, p. 143-169, 2016.

FACELI, Katti et al. **Algoritmos de agrupamento de dados**. 2005.

KAUPA, Paulo Henrique et al. **Aplicação de técnicas da inteligência artificial na seleção de ações para investimento na bolsa de valores de São Paulo**. 2013.

LEONI, Roberto Campos; DE SOUZA SAMPAIO, Nilo Antonio. **Desempenho das escolas públicas e privadas da região do vale do paraíba**: Uma aplicação da técnica de agrupamentos kmeans com base nas variáveis do enem 2015. Cadernos do IME-Série Estatística, v. 42, p. 31, 2017.

LOBO, Luiz Carlos. Inteligência artificial e medicina. **Revista Brasileira de Educação Médica**, v. 41, p. 185-193, 2017.

NORVIG, Peter; RUSSELL, Stuart. **Inteligência artificial**. Editora Campus, v. 20, 2004.

RODRIGUES, Fabrício Augusto et al. **Localização e reconhecimento de placas de sinalização utilizando um mecanismo de atenção visual e redes neurais artificiais**. 2002.

SCAICO, Pasqueline Dantas. QUEIROZ, Ruy José G. B. de. SCAICO, Alexandre. O conceito big data na educação. In: **3º Congresso Brasileiro de Informática na Educação**. XX Workshop de Informática na Escola. Anais do WIE. Universidade Federal da Paraíba, 2014, p. 328-336.

SETZER, V.W. **Os Meios Eletrônicos e a Educação: Uma Visão Alternativa**. São Paulo: Editora Escrituras, Coleção Ensaios Transversais Vol. 10, 2001.) <
<https://www.ime.usp.br/~vwsetzer/dado-info.html/>> Acesso em: 02 de nov. 2021.

TORCATE, Arianne Sarmiento; BARBOSA, José Carlos Felix; DE OLIVEIRA RODRIGUES, Cleyton Mário. **Utilizando o Learning Analytics com o K-Means para Análise de Dificuldades de Aprendizagem na Educação Básica**. In: Anais do XXVI Workshop de Informática na Escola. SBC, 2020. p. 31-40.

VALENTIM, Marta Lígia Pomim et al. **Inteligência competitiva em organizações: dado, informação e conhecimento**. DataGramZero, Rio de Janeiro, v. 3, n. 4, p. 1-13, 2002.

VIEIRA, Renata; LOPES, Lucelene. **PROCESSAMENTO DE LINGUAGEM NATURAL E O TRATAMENTO COMPUTACIONAL DE LINGUAGENS CIENTÍFICAS**. EM CORPORA, p. 183, 2010.

WITTEK, Peter. **Quantum machine learning: what quantum computing means to data mining**. Academic Press, 2014.