UNIVERSIDADE FEEVALE

RAFAEL VINICIOS DO CARMO

# ANÁLISE DE CLUSTERS DOS PARTICIPANTES DO ENEM DE 2018 E 2019

Novo Hamburgo

2022

RAFAEL VINICIOS DO CARMO

# ANÁLISE DE CLUSTERS DOS PARTICIPANTES DO ENEM DE 2018 E 2019

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Ciência da Computação pela Universidade Feevale

Orientador: Juliano Varella de Carvalho

Novo Hamburgo

2022

RAFAEL VINICIOS DO CARMO

# ANÁLISE DE CLUSTERS DOS PARTICIPANTES DO ENEM DE 2018 E 2019

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Ciência da Computação pela Universidade Feevale

APROVADO EM: ___ / ___ / _____

JULIANO VARELLA DE CARVALHO
Orientador – Feevale

RODRIGO RAFAEL VILLARREAL GOULART
Examinador interno – Feevale

SANDRA TERESINHA MIORELLI
Examinador interno – Feevale

Novo Hamburgo
2022

## AGRADECIMENTOS

Primeiramente, gostaria de agradecer aos meus pais, Elci e Vildomar, que sempre me apoiaram e me motivaram a seguir em frente.
Gostaria de agradecer a minha namorada, Raiana, que sempre esteve ao meu lado, me motivando e me dando forças para o desenvolvimento do trabalho.
Gostaria de agradecer ao meu orientador, professor Dr. Juliano, pelo auxílio, disposição, paciência e parceria durante todo o desenvolvimento deste trabalho.
Também agradeço aos meus amigos Feevaleiros pela parceria de sempre.
Agradeço também a todos que, de alguma forma, contribuiram para a conquista deste objetivo.

## RESUMO

O Exame Nacional do Ensino Médio (ENEM) foi criado em 1998 com o propósito de avaliar o desempenho escolar dos estudantes ao término da educação básica. Em 2004, o ENEM se tornou popular quando foi possível usar a nota do exame para acesso a bolsas de instituições privadas. Conforme o decorrer dos anos, o número de participantes da prova vem aumentando e, com isso, são acumuladas grandes quantidades de dados sobre os participantes. Além das informações referentes às notas de cada estudante, são salvos também dados socioeconômicos daqueles que participaram do exame. Com esses extensos *datasets* é possível a aplicação de *Machine Learning* (ML) porém, junto a isso, existe o problema de limitação de hardware, pois uma máquina comum não suporta a quantidade de gigabytes de dados que o ENEM possui. Neste trabalho, após o pré-processamento dos dados, foi aplicado o algoritmo de clusterização *k-menas*, gerando 3 agrupamentos. O objetivo da clusterização neste trabalho é, a partir de uma análise nas visualizações geradas, identificar as características socioeconômicas de cada agrupamento, em relação a nota média atingida no ENEM de 2018 e 2019. Os *clusters* gerados indicam que as questões socioeconômicas tendem a se relacionar com a nota média atingida pelos participantes.



Palavras-chave: ENEM. Machine Learning. Clusterização. Cluster.

## ABSTRACT

*The Exame Nacional do Ensino Médio (ENEM) was created in 1998, with the purpose of evaluating the academic performance of students at the end of basic education. In 2004, ENEM became popular when it became possible to use the exam grade to access scholarships from private institutions. Over the years, the number of participants in the competition has increased and, as a result, large amounts of data about the participants are accumulated. In addition to information regarding each student's grades, socioeconomic data for those who took the exam are also saved. With these extensive datasets it is possible to apply Machine Learning (ML) but, along with that, there is the problem of hardware limitation, as a common machine does not support the amount of gigabytes of data that ENEM has. In this work, after pre-processing the data, the k-menas clustering algorithm was applied, generating 3 clusters. The objective of clustering in this work is, based on an analysis of the generated visualizations, to identify the socioeconomic characteristics of each grouping, in relation to the average score achieved in the 2018 and 2019 ENEM. The generated clusters indicate that socioeconomic issues tend to be related with the average score achieved by the participants.*



*Keywords: ENEM. Machine Learning. Clustering. Cluster.*

# LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| ENEM | Exame Nacional do Ensino Médio |
| IES | Instituições de ensino superio |
| INEP | Instituto Nacional de Estudos e Pesquisas Educacionais AnísioTeixeira |
| MEC | Ministério da Educação |
| ML | Machine learning |
| PICOC | População, interveção,comparação, resultados e contexto |
| ProUni | Programa Universidade para Todo |

**SUMÁRIO**

# 1 INTRODUÇÃO

O Exame Nacional do Ensino Médio (ENEM) tem o objetivo de avaliar o desempenho escolar dos estudantes ao término da educação básica. Instituído em 1998, o ENEM é aplicado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). O exame foi criado como alternativa aos exames de acesso aos cursos profissionalizantes pós-médio e ao ensino superior, além de possibilitar a participação de programas governamentais, como o Programa Universidade para Todos (ProUni). O INEP é uma autarquia federal, vinculada ao Ministério da Educação (MEC) e tem como objetivo financiar a criação de políticas educacionais, para a contribuição do desenvolvimento econômico e social do país. Além do ENEM, o INEP atua em outros exames e indicadores de educação, como por exemplo, o Exame Nacional de Desempenho de Estudantes (ENADE) (INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANíSIO TEIXEIRA, 2021).

Em 1998, em sua primeira edição, o ENEM teve 157,2 mil inscritos e 115,6 mil participantes. Na quarta edição, ocorrida em 2001, o número de participantes se tornou expressivo, tendo 1,6 milhão de inscritos e 1,2 milhão de participantes. No ano de 2004, o ENEM tornou-se popular, quando o MEC criou o ProUni e vinculou a concessão de bolsas em instituições de ensino superior (IES) privadas à nota obtida no exame. No ano seguinte à atuação do MEC, o exame chegou a 3 milhões de inscritos e 2,2 milhões de participantes. O ENEM, até 2008, era uma prova interdisciplinar de 63 questões, além da redação. Em 2009 houve uma reformulação da prova e a partir de então são quatro provas objetivas referentes às áreas Linguagens e códigos, Ciências humanas e suas tecnologias, Ciências da natureza e suas tecnologias e Matemática e suas tecnologias e uma redação dissertativa-argumentativa. Assim, o exame aborda diretamente o currículo do Ensino Médio e se torna comparável a outras edições do exame (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATíSTICA, 2021).

Os dados dos participantes de cada edição do ENEM são armazenados pelo INEP. Além das informações dos estudantes como idade, sexo, raça, etc, a base contém registros sobre a escola do inscrito, desempenho em cada uma das provas e redação, entre outras informações. Conforme o passar dos anos, o número de participantes do ENEM foi aumentando e, consequentemente, o número de dados em cada base também. A base de sua primeira edição, de 1998, tem o tamanho de 64,4 *megabytes*, possuindo 157.221 registros com 137 respostas de questões socioeconômicas, além do vetor de respostas das 63 questões da prova e das notas de cada competência. Na edição de 2009, ano em que foi feita a reformulação na prova, a base tem 3,4 *gigabytes*, com maior número de registros e de informações, comparando com o *dataset* de 1998.

*Machine Learning* (ML) é uma tecnologia de aprendizado, a qual usa uma ou mais base de dados para que aplicações de computadores consigam reconhecer padrões e tomem decisões de forma inteligente (HAN; PEI; KAMBER, 2011). O aprendizado de máquina surgiu na década de 1950, quando o pesquisador Arthur Lee Samuels desenvolveu um dos primeiros programas de *machine learning*, um programa de autoaprendizagem para jogar damas (HURWITZ; KIRSCH, 2018). Essa tecnologia é dividida em subcategorias, das quais as mais conhecidas são: aprendizado supervisionado; aprendizado não supervisionado; aprendizado semi supervisionado; e aprendizado ativo (HAN; PEI; KAMBER, 2011).

Clusterização (ou Agrupamento) é uma técnica da subcategoria de aprendizado não supervisionado. A técnica consiste em uma divisão de dados, baseado em suas características, gerando grupos separados chamados de *clusters*. Um *cluster* é composto por um grupo de objetos com características semelhantes (JAIN, 1988) e existem diversos algoritmos que permitem a geração desses *clusters*.

Moreira (2016) utilizou a base de dados do ENEM de 2010 para a identificação de perfis de participantes, utilizando a técnica de clusterização, relacionando a média da nota do exame com os dados socioeconômicos. O resultado da comparação foi de que os fatores socioeconômicos não têm grande relação com a nota final da prova.

Em 2020, Silva et al. (2020) aplicaram técnicas de mineração de dados, focadas na identificação de desigualdades sociais a partir do desempenho dos participantes do ENEM de 2019. Foram identificados dois grupos: um com as notas mais baixas e outro com as notas mais altas. No grupo de notas mais baixas existe um predomínio de características socioeconômicas semelhantes, com destaque na educação em rede estadual de ensino e a baixa renda familiar. No grupo de notas mais altas, é perceptível a diferença de desempenho dos alunos da rede privada e federal, em relação aos demais.

Com a melhoria na rede, facilidade de acesso às informações e melhorias de hardware, agora temos menos limitações físicas para o gerenciamento de grandes quantidades de dados. Com as tecnologias de *big data* e *machine learning*, as organizações são capazes de antecipar o futuro (HURWITZ; KIRSCH, 2018). O ENEM possui grande volume de dados, trazendo a possibilidade de estudo sobre os padrões de perfil dos estudantes em relação às notas alcançadas no exame, com a aplicação de *machine learning*. Desta forma, é proposto um estudo, comparando os diferentes *clusters* de participantes do ENEM, gerando visualizações, para evidenciar as diversidades e discutir os resultados encontrados.

Portanto, este trabalho, dividido em 5 capítulos, propõe a aplicação do algoritmo de clusterização *k-means*, nas bases de dados do ENEM de 2018 e 2019, a fim de identificar padrões entre as questões socioeconômicas dos participantes, com relação a nota média. O primeiro capitulo apresentou uma introdução deste trabalho. O segundo, apresenta uma revisão sistemática sobre os algoritmos de clusterização, atributos usados, ferramentas

e validações utilizadas em trabalhos relacionados ao ENEM. Uma revisão bibliográfica sobre *machine learning* é feita no terceiro capitulo. No quarto capitulo, é apresentada a metodologia do trabalho, além aplicação do algoritmo *k-means* e apresentação dos *clusters* gerados. Por fim, no sexto capitulo, são apresentadas as conclusões deste trabalho.

# 2 REVISÃO SISTEMÁTICA SOBRE TÉCNICAS DE CLUSTERIZAÇÃO APLICADAS EM BASE DE DADOS EDUCACIONAIS

O principal objetivo da proposta e execução dessa revisão sistemática é buscar na literatura especializada, técnicas de clusterização que foram aplicadas em base de dados semelhantes a do ENEM.

## 2.1 O PROTOCOLO DE REVISÃO SISTEMÁTICA

O protocolo foi criado conforme a metodologia PICOC (população, interveção, comparação, resultados e contexto), proposta por Kitchenham e Charters (2007).

### 2.1.1 O PROTOCOLO

O planejamento desta revisão sistemática foi realizado seguindo a estrutura:

a) Título

Revisão sistemática sobre técnicas de clusterização aplicadas em base de dados educacionais

b) Resumo

O protocolo desta revisão sistemática segue o protocolo da pesquisadora Kitchenham e Charters (2007) e visa encontrar na literatura técnicas, ferramentas e algoritmos a serem aplicadas na base de dados do ENEM.

c) Objetivo

Esta revisão sistemática busca encontrar artigos que abordam técnicas de clusterização, visando verificar na literatura as melhores técnicas a serem aplicadas na base de dados do ENEM.

#### 2.1.1.1 Formulação da pesquisa

a) Foco da questão

Encontrar técnicas, algoritmos, metodologias de aplicação de clusterização, que são mais utilizadas e possuem resultados promissores para auxiliar na geração de *clusters* de perfis dos estudantes.

b) Questões de interesse

- Técnicas utilizadas

- Algoritmos utilizados
- Ferramentas utilizadas
- Atributos relevantes
- Resultados apresentados

c) Palavras-chave

Agrupamento, Clusterização, Clustering, Educational Data Mining, Machine Learning.

d) Intervenção

Identificar técnicas de clusterização que são utilizadas em base de dados semelhantes a do ENEM.

e) Controle

Inexistente.

f) Efeito

Identificar oportunidades de pesquisa na área de aprendizagem de máquina com foco no agrupamento de perfis de estudantes.

g) Medida de resultado

Identificar quantitativamente nos estudos encontrados suporte na geração de embasamento teórico para o trabalho de conclusão de curso e escrita de artigos.

h) População de interesse

Pesquisadores, professores, desenvolvedores, gestores públicos e profissionais da área da computação e educação.

i) Aplicação

Esta pesquisa tem como foco pesquisadores, professores, desenvolvedores e profissionais da área da computação, pois tem como objetivo identificar técnicas, algoritmos e ferramentas para a o agrupamento de perfis de estudantes. Ela também poderá auxiliar pesquisadores, professores e gestores públicos da área educacional, a partir dos *insights* gerados.

j) Desenho do experimento

Não há.

k) Financiamento

Não há financiamento.

#### 2.1.1.2 Formulação de critérios

a) Definição de critérios de seleção das fontes de dados

A seleção das fontes de dados foi feita com base na indicação do orientador. Na busca de trabalhos na área da computação, serão utilizadas as bases *Web of Science*[1], Portal CAPES[2] e Google Acadêmico[3]. Essas bases podem conter materiais em português, trazendo maior probabilidade de encontrar artigos sobre o ENEM, o qual é um exame nacional.

b) Idiomas das fontes de dados

Serão considerados somente os materiais bibliográficos nos idiomas português e em inglês.

c) *String* de busca

Com base nas palavras-chave definidas anteriormente gerou-se a string de busca abaixo. Esta string é aplicada nos motores de buscas das bibliotecas definidas anteriormente.

*("clusterização" OR "clustering") AND "ENEM" AND ("data mining" OR "machine learning") AND ("perfil" OR "profile")*

d) Artigos de controle

Optou-se por não utilizar nenhum artigo de controle para esta revisão sistemática.

#### 2.1.1.3 Seleção dos estudos

1. Critérios para inclusão/exclusão dos resultados

   a) Ser um artigo escrito em português ou inglês.

   b) Artigos publicados de 2009 até 2021.

   c) Ser publicado em um *journal*.

   d) Possuir relação com os assuntos da pesquisa.

2. Procedimentos para seleção dos estudos

   Primeiramente, a *string* de busca foi aplicada nos motores de busca das bases de dados selecionadas. Após isso, os resultados foram exportados para a ferramenta StArt (LAPES, 2021). Por fim, foram executadas as 5 fases definidas a seguir.

3. Fase de seleção de artigos

[1] http://www.periodicos.capes.gov.br/
[2] http://www.periodicos.capes.gov.br/
[3] https://scholar.google.com.br/

a) Fase 1 - Validar os critérios de inclusão/exclusão;

b) Fase 2 - Identificar artigos duplicados;

c) Fase 3 - Leitura do título, palavras-chave e resumo;

d) Fase 4 - Leitura da introdução e conclusão;

e) Fase 5 - Leitura integral dos artigos e validação das respostas para os critérios de qualidade;

4. Critérios de qualidade

a) Quais foram os algoritmos de clusterização utilizados

b) Quais algoritmos foram utilizados?

c) Quais foram as técnicas utilizadas?

d) Quais ferramentas foram usadas?

e) Quais linguagens de programação foram utilizadas?

f) Quais atributos para clusterização foram utilizados?

g) Quais métodos de validação foram usados?

h) Quais os resultados apresentados?

i) Quais as métricas foram utilizadas para avaliar o resultado da aplicação dos algoritmos?

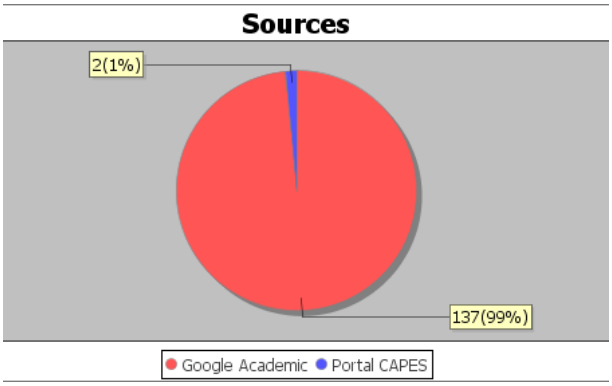
## 2.2 DESENVOLVIMENTO DA REVISÃO SISTEMÁTICA

Para o desenvolvimento da Revisão Sistemática (RS), foi utilizada a ferramenta StArt. Segundo (HECKLER, 2018), "foi desenvolvida para auxiliar os pesquisadores na aplicação da técnica de revisão sistemática, desde o cadastro do protocolo até a fase final da revisão. Também são gerados gráficos com informações sobre o andamento do trabalho".

Primeiramente, o protocolo foi cadastrado na ferramenta StArt. Após isso, foi executada a pesquisa nos motores das bases de dados com a *string* de busca já definida. As consultas foram feitas no dia 30 de abril de 2021. Após execução da busca, os resultados foram exportados para a ferramenta StArt.

A primeira etapa iniciou com a execução da *string* de busca na *Web of Science*, onde nenhum resultado foi encontrado. Em sequência, a busca foi feita na base Portal CAPES, na qual apenas 2 artigos foram encontrados. No Google Acadêmico, foram econtrados 137 artigos, utilizando como filtro o período de publicação entre 2009, ano em que o ENEM passou por uma reformulação, e 2021. Portanto, foram encontrados 139 artigos, conforme Figura 1.

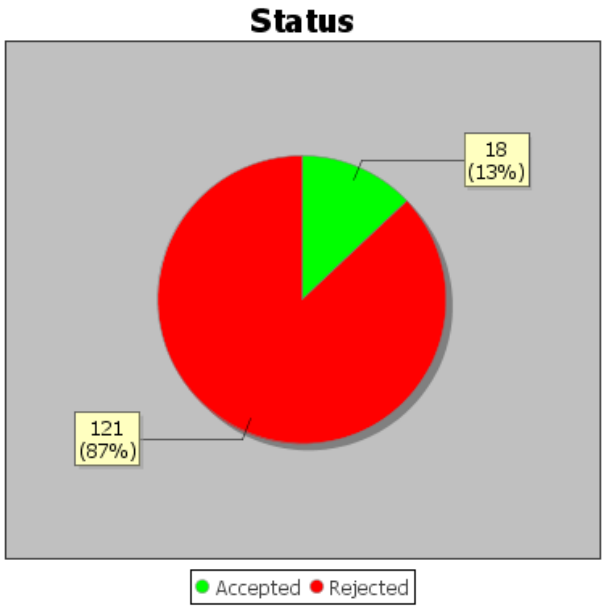
**Figura 1 – Total de artigos encontrados.**




**Fonte: elaborado pelo autor.**

A segunda etapa foi responsável pela eliminação de artigos duplicados. Nenhum artigo duplicado foi encontrado, fato que se deve ao Google Acadêmico unificar artigos iguais. Na terceira etapa foram lidos os títulos, palavras-chave e resumo de cada artigo, para a identificação daqueles que possuem relação com o tema do trabalho, ou seja, que aplicam clusterização nos dados do ENEM, ou de provas semelhantes. Além disso, nessa etapa foram descartados os artigos que são de um idioma diferente de português ou inglês. Ao final dessa etapa, 18 artigos foram selecionados e 121 excluídos, conforme Figura 2. Os dezoito artigos selecionados, são aqueles que aplicam *machine learning* nos *datasets* do ENEM.

**Figura 2 – Resultado após filtro da segunda e terceira etapa.**

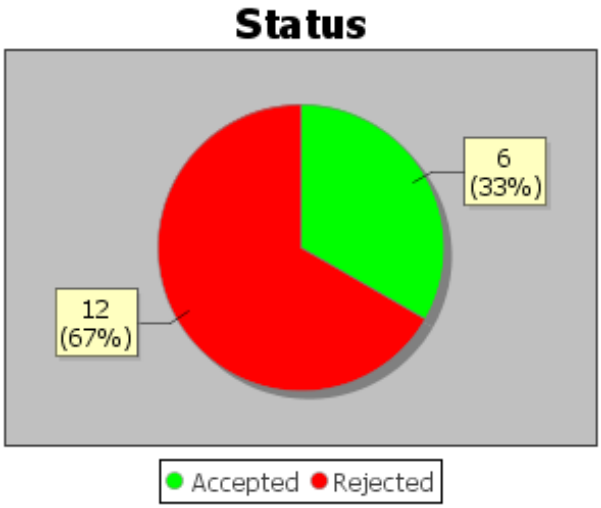



**Fonte: elaborado pelo autor.**

Na quarta etapa, foram lidas a introdução e conclusão dos artigos restantes. O critério de eliminação usado nessa etapa, foi a utilização de clusterização na base de dados do ENEM. Com isso, foram rejeitados 12 artigos, conforme Figura 3. A Tabela 1

contém os artigos selecionados e seus autores.

**Figura 3 – Resultado após seleção dos artigos.**




**Fonte: elaborado pelo autor.**

**Tabela 1 – Artigos selecionados e autores**

| Título do artigo | Autor(es)(as) |
|---|---|
| Analysis of ENEM's attendants between 2012 and 2017 using a clustering approach | Lima, Afonso Matheus Sousa<br>Florez, Alexander Ylnner Choquenaira<br>Lescano, Alexis Iván Aspauza<br>Novaes, João Victor de Oliveira<br>Martins, Natalia de Fatima<br>Junior, Caetano Traina<br>Sousa, Elaine Parros Machado<br>Junior, José Fernando Rodrigues<br>Cordeiro, Robson Leonardo Ferreira |
| Aplicação de Data Mining na Base de Dados do Processo Seletivo do Exame Nacional do Ensino Médio - ENEM 2010 | Carreira, Suely da Silva<br>Carreira, Manoel Francisco<br>Antonelli, Gilberto Clovis<br>Samed, Márcia Marcondes Altimari<br>Leal, Gislaine Camila Lapasini Leal |

| | |
|---|---|
| Desempenho das escolas públicas e privadas da região do vale Paraíba: Uma aplicação da técnica de agrupamentos Kmeans com base nas variáveis do ENEM 2015 | Leoni, Roberto Campos<br>Sampaio, Nilo Antonio de Souza |
| Detecção de atributos que melhor caracterizam perfis de inscritos do ENEM utilizando Redução de Dimensionalidade | Moreira, Natália Lionel |
| Pedagogical Recommendation to Improve the Quality of Writing: A Case Study In a Public School | Santos, Danilo Abreu<br>Bittencourt, Ig Ibert<br>Paiva, Ranilson Oscar A.<br>Dermeval, Diego |
| Mineração de Dados Abertos | Junior, Raimundo de Acacio Leonel<br>Junior, João Holanda Freires<br>Silva, Tércio Jorge<br>Silva, Ticiana Linhares Coelho<br>Magalhães, Regis Pires |

**Fonte: elaborado pelo autor.**

## 2.3 AVALIAÇÃO DOS CRITÉRIOS DE QUALIDADE

Na quinta fase, foi finalizada a leitura dos artigos selecionados, confirmando que os artigos selecionados utilizam algoritmos de clusterização em dados retirados do *dataset* do ENEM. As perguntas referentes aos critérios de qualidade serão apresentadas neste capítulo.

Referente a pergunta "Quais foram os algoritmos de clusterização utilizados?", identificou-se que, dentre os trabalhos selecionados, o algoritmo mais usado é o *k-means*, o qual aparece em 4 (LIMA et al., 2020) (MOREIRA, 2016) (SANTOS et al., 2018) (LEONI; SAMPAIO, 2017) dos 6 artigos. Leoni e Sampaio (2017) justificam a escolha do algoritmo por ser extremamente rápido, ou seja, facilmente aplicável em grandes conjuntos de dados. Outro algoritmo citado é o DBSCAN, o qual aparece em somente 1 artigo (SANTOS et al., 2018) e sem justificativa de uso. O algoritmo *x-means* é citado em 1 (JUNIOR et al., 2018) trabalho e seu uso foi justificado por não ser necessário determinar o número de *clusters*, como é necessário no *k-means*. Em apenas 1 (CARREIRA et al., 2012) dos 6 artigos não foi informado o algoritmo usado. As respostas para a pergunta são exibidas na Tabela 2.

**Tabela 2 – Algoritmos de clusterização encontrados na RS.**

| Artigos | Algoritmos utilizados |
|---|---|
| (LIMA et al., 2020) | *k-means* |
| (CARREIRA et al., 2012) | Não informado |
| (MOREIRA, 2016) | *k-means* |
| (SANTOS et al., 2018) | *k-means*<br>DBSCAN |
| (LEONI; SAMPAIO, 2017) | *k-means* |
| (JUNIOR et al., 2018) | *x-means* |

**Fonte: elaborado pelo autor.**

Além dos algoritmos de clusterização, foram analisados os algoritmos em geral, usados nos artigos. As respostas referente a pergunta "Quais algoritmos foram utilizados?" são apresentadas na Tabela 3. Carreira et al. (2012) citam os algoritmos J48 e Apriori, respectivamente, para aplicação de classificação e associação. Moreira (2016) também cita o algoritmo J48, o qual é usado antes da execução de 3 algoritmos de busca de subconjunto: RankerSearch, GeneticSearch e BestFirst. Além disso, nesse trabalho, o algoritmo J48 também é usado antes da execução dos algoritmos CFsSubsetEval e WrapperSubsetEval, os quais servem para a seleção de atributos. O algoritmo Normalize é usado em somente 1 (MOREIRA, 2016) artigo. Dentre os 6 artigos, 4 (LIMA et al., 2020) (SANTOS et al., 2018) (LEONI; SAMPAIO, 2017) (JUNIOR et al., 2018) não informaram o algoritmo usado.

**Tabela 3 – Algoritmos encontrados na RS.**

| Artigos | Algoritmos utilizados |
|---|---|
| (LIMA et al., 2020) | Não informado |
| (CARREIRA et al., 2012) | J48<br>Apriori |
| (MOREIRA, 2016) | CFsSubsetEval(CFs)<br>WrapperSubsetEval<br>BestFirst<br>GeneticSearch<br>RankerSearch<br>Normalize<br>C4.4 (J48) |
| (SANTOS et al., 2018) | Não informado |
| (LEONI; SAMPAIO, 2017) | Não informado |
| (JUNIOR et al., 2018) | Não informado |

**Fonte: elaborado pelo autor.**

As respostas referentes à pergunta "Quais foram as técnicas utilizadas?" são exibidas na Tabela 4. O método Elbow é usado para identificação de quantidade ideal de *clusters* a serem gerados. Esse método é citado em 2 artigos (LIMA et al., 2020) (MOREIRA, 2016). Moreira (2016) cita mais duas técnicas, *Filter* e *Wrapper*, as quais são

usadas para seleção de atributos. Neste trabalho, foi identificado que o *Filter* teve um resultado melhor. A técnica *Blocked Adaptive Computationally-Efficient Outlier Nominators* foi usada em 1 (LEONI; SAMPAIO, 2017) artigo, para identificação de dados multivariados discrepantes. Na metade dos artigos (LIMA et al., 2020) (CARREIRA et al., 2012) (JUNIOR et al., 2018) não foram informadas as técnicas usadas.

**Tabela 4 – Técnicas encontradas na RV.**

| Artigos | Técnicas utilizadas |
|---|---|
| (LIMA et al., 2020) | Elbow Method |
| (CARREIRA et al., 2012) | Não informado |
| (MOREIRA, 2016) | Elbow Method<br>Filter<br>Wrapper |
| (SANTOS et al., 2018) | Não informado |
| (LEONI; SAMPAIO, 2017) | Blocked Adaptive Computationally-Efficient Outlier Nominators |
| (JUNIOR et al., 2018) | Não informado |

**Fonte: elaborado pelo autor.**

Em relação a pergunta "Quais ferramentas foram usadas?", verificou-se que a mais citada foi uma ferramenta de aplicação de algoritmos de *machine learning*, WEKA, a qual aparece em 3 (CARREIRA et al., 2012) (MOREIRA, 2016) (SANTOS et al., 2018) artigos. Moreira (2016) também cita a ferramenta *Pentaho Data Integration* e *IDE Rstudio*. Outra ferramenta de aplicação de algoritmos de *machine learning*, o *RapidMiner - 5*, é citada em 1 artigo (JUNIOR et al., 2018), o qual também cita o *TextMaster Split & Join*. A Tabela 5 agrupa as respostas encontradas.

**Tabela 5 – Ferramentas encontradas na RV.**

| Artigos | Ferramentas utilizadas |
|---|---|
| (LIMA et al., 2020) | Não informado |
| (CARREIRA et al., 2012) | WEKA |
| (MOREIRA, 2016) | Pentaho Data Integration<br>WEKA<br>IDE Rstudio |
| (SANTOS et al., 2018) | WEKA |
| (LEONI; SAMPAIO, 2017) | Não informado |
| (JUNIOR et al., 2018) | TextMaster Split & Join<br>RapidMiner - 5 |

**Fonte: elaborado pelo autor.**

Na Tabela 6 estão as informações que respondem a pergunta "Quais linguagens de programação foram utilizadas?". Em 4 (LIMA et al., 2020) (CARREIRA et al., 2012) (SANTOS et al., 2018) (JUNIOR et al., 2018) dos 6 artigos, não foi informado qual linguagem de programação foi utilizada. Isso, pois em alguns casos, a ferramenta de aplicação

de *machine learning* usada, já é suficiente para o estudo. Em 1 artigo (MOREIRA, 2016) foi usada a linguagem de programação R, para o cálculo de correlação. Junior et al. (2018) utilizaram a linguagem de programação C++ para desenvolvimento de uma aplicação que retirava apenas as informações necessárias do *dataset*.

**Tabela 6 – Linguagens de programação encontradas na RV.**

| Artigos | Linguagens de programação utilizadas |
|---|---|
| (LIMA et al., 2020) | Não informado |
| (CARREIRA et al., 2012) | Não informado |
| (MOREIRA, 2016) | R |
| (SANTOS et al., 2018) | Não informado |
| (LEONI; SAMPAIO, 2017) | Não informado |
| (JUNIOR et al., 2018) | C++ |

**Fonte: elaborado pelo autor.**

Antes da aplicação do algoritmo de clusterização, é necessário selecionar os atributos que serão utilizados, para que o resultado obtido seja mais preciso. Os dados que respondem a pergunta "Quais atributos para clusterização foram utilizados?" estão na Tabela 7. Lima et al. (2020) usaram os atributos referentes as notas, regiões, atributos que determinam uma deficiência e outros não citados. Carreira et al. (2012) utilizaram os atributos grau de estudo do pai, acesso a internet e tipo de escola do participante. Moreira (2016) também utilizou o atributo de tipo de escola, além de atributos relacionados as questões socioeconômicas. (LEONI; SAMPAIO, 2017) usaram a nota de cada proficiência, da base de dados do ENEM, além da participação percentual de alunos, taxas de rendimento escolar de aprovação e indicador de formação docente, que foram retiradas de outras bases, como o Censo Escolar. (JUNIOR et al., 2018) utilizaram os atributos: acesso a internet, escolaridade do pai, renda familiar mensal, tipo de escola e nota. Somente 1 (SANTOS et al., 2018) dos 6 artigos não informou os atributos usados na clusterização.

**Tabela 7 – Atributos para clusterização encontrados na RV.**

| Artigos | Atributos para clusterização utilizados |
|---|---|
| (LIMA et al., 2020) | Notas, regiões, atributos que determinam uma deficiência, entre outros não citados. |
| (CARREIRA et al., 2012) | Grau de estudo do pai<br>Acesso a internet<br>Tipo de escola |

| | |
|---|---|
| (LEONI; SAMPAIO, 2017) | Proficiência média em ciências da natureza e suas tecnologias ($\text{MEDIA}_{CN}$)<br>Proficiência média em ciências humanas e suas tecnologias ($\text{MEDIA}_{CH}$)<br>Proficiência média em linguagens, códigos e suas tecnologias ($\text{MEDIA}_{LC}$)<br>Proficiência média em matemática e suas tecnologias ($\text{MEDIA}_{MT}$)<br>Indicador de formação docente ($\text{IND}_{FORM_DOCENTE}$)<br>Taxas de rendimento escolar de aprovação ($\text{TAXA}_{APROVACAO}$)<br>Participação percentual de alunos ($\text{TAXA}_{PARTICIPACAO}$) |
| (JUNIOR et al., 2018) | Acesso a internet<br>Escolaridade do pai<br>Renda familiar mensal<br>Região (urbana, rural ou indigena)<br>Tipo de escola<br>Nota |

| | |
|---|---|
| (MOREIRA, 2016) | Quantas pessoas moram com você?<br>Qual é o nível de escolaridade do seu pai?<br>Qual é o nível de escolaridade da sua mãe?<br>Somando a sua renda com a renda das pessoas que moram com você, quanto é, aproximadamente, a renda familiar mensal?<br>Qual a sua renda mensal, aproximadamente?<br>A casa onde você mora é?(Cedida, Alugada, Própria)<br>Sua casa está localizada em?<br>Você trabalha ou já trabalhou?<br>Testar meus conhecimentos<br>Prosseguir os estudos no Ensino Superior<br>Obter a certificação do Ensino Médio ou acelerar meus estudos<br>Conseguir uma bolsa de estudos (ProUni, outras)<br>Quantos anos você levou para concluir o ensino fundamental?<br>Você deixou de estudar durante o Ensino Fundamental?<br>Em que tipo de escola você cursou o Ensino Fundamental?<br>Quantos anos você levou para concluir o Ensino Médio?<br>Você deixou de estudar durante o Ensino Médio?<br>Em que tipo de escola você cursou o Ensino Médio?<br>Média |
| (SANTOS et al., 2018) | Não informado |

**Fonte: elaborado pelo autor.**

Os dados que respondem a pergunta “Quais métodos de validação foram usados?” são apresentados na Tabela 8. Em 1 dos artigos (SANTOS et al., 2018) é usado o método Wilcoxon, para validar se houve diferença estatística nas amostras. Leoni e Sampaio (2017) usou os métodos *wss*, *silhouette* e *gap_statistic*, para validar o número de *clusters* gerados. Dentre os 6 artigos, 4 (LIMA et al., 2020) (CARREIRA et al., 2012) (MOREIRA, 2016) (JUNIOR et al., 2018) não informaram a validação.

**Tabela 8 – Métodos de validação encontrados na RV.**

| Artigos | Métodos de validação utilizados |
|---|---|
| (LIMA et al., 2020) | Sem validação |
| (CARREIRA et al., 2012) | Sem validação |
| (MOREIRA, 2016) | Sem validação |
| (SANTOS et al., 2018) | Wilcoxon |
| (LEONI; SAMPAIO, 2017) | Wss<br>Silhouette<br>Gap_statististic |
| (JUNIOR et al., 2018) | Não informado |

**Fonte: elaborado pelo autor.**

Para a última pergunta “Que resultados teve?”, foi gerada a Tabela 9. Lima et al. (2020) geraram 3 *clusters*, formados por: performance baixa, performance média e performance alta. Carreira et al. (2012) destacam 2 *clusters* gerados, em um deles o participante do ENEM é de escola pública, com acesso a internet e com a escolaridade do pai sendo “1 grau”. Esse cluster obteve um desempenho “regular”. O outro *cluster* destacado, é o participante que é de escola pública, não tem acesso a internet e com pai, cuja escolaridade é “1 grau”. Esse *cluster* obteve um desempenho “insatisfatório”. Moreira (2016) concluiu que os fatores socioeconômicos não são capazes de determinar o rendimento do participante. Santos et al. (2018) propuseram uma ferramenta auxiliar, que apresenta o conteúdo educacional de acordo com os *clusters* gerados. Concluiu-se que a ferramenta trouxe melhora na aprendizagem. Leoni e Sampaio (2017) identificaram padrões de desempenho similares suficientemente significativos para permitir afirmar a existência de agrupamentos naturais dentre os tipos de escolas. Por fim, Junior et al. (2018) concluíram que características como renda familiar, escolaridade dos pais, escolaridade do inscrito, lugar de residência e acesso a internet, impactam na nota final do inscrito.

**Tabela 9 – Resultados obtidos em cada artigo selecionado na RV.**

| Artigos | Resultados obtidos |
|---|---|
| (LIMA et al., 2020) | 3 *clusters*: Performance baixa Performance média Performance alta |

| | |
|---|---|
| (CARREIRA et al., 2012) | *Cluster* 1: escola publica; tem acesso a internet; escolaridade "1 grau" do pai. Obteve-se desempenho "regular".<br>*Cluster* 2: escola publica; sem acesso a internet; escolaridade "1 grau" do pai; Obteve-se desempenho "insatisfatório". |
| (MOREIRA, 2016) | Conclui-se que os fatores socioeconômicos não são capazes de determinar o rendimento do inscrito. |
| (SANTOS et al., 2018) | Ferramenta gerada mostrou resultados positivos. |
| (LEONI; SAMPAIO, 2017) | Identificou padrões de desempenho similares suficientemente significativos para permitir afirmar a existência de agrupamentos naturais dentre os tipos de escolas. |
| (JUNIOR et al., 2018) | Características como renda familiar, escolaridade dos pais, escola do inscrito, lugar de residência e se possui internet, impactam na nota final do inscrito. |

**Fonte: elaborado pelo autor.**

## 2.4 AVALIAÇÃO DOS RESULTADOS DOS ARTIGOS SELECIONADOS

Todos os artigos selecionados obtiveram resultados. Dentre os 6, um criou uma ferramenta de recomendação, baseada em *clusters* gerados. Nos demais trabalhos foram gerados *clusters* do perfil de participantes do ENEM. Foram usados diferentes algoritmos de *machine learning* e diferentes ferramentas de aplicação de algoritmos.

Lima et al. (2020) propuseram uma análise do perfil de participantes do ENEM, considerando o intervalo de tempo entre 2012 e 2017. A abordagem do artigo foi a utilização do algoritmo *k-means*, para agrupar alunos com base nas notas (em cada área de conhecimento e dissertação), além de agrupar por regiões.

Lima et al. (2020) geraram 3 *clusters*: baixo desempenho, médio desempenho e alto desempenho. Afirma-se que, para todos os períodos de tempo considerados e para todas

as regiões, a maioria dos alunos foi agrupada nos grupos de baixo e médio desempenho, enquanto apenas um pequeno número de alunos foi agrupado como de alto desempenho.

Carreira et al. (2012) tem como foco associar o desempenho dos participantes, na prova objetiva do ENEM, com as situações socioeconômicas. Na análise, foram usados os dados dos estados: Rio Grande do Sul, Santa Catarina e Paraná. O processo de clusterização resultou em grupos distintos para os 3 estados, porém sem diferenças significativas. O *cluster*, mais assertivo, é constituído pelo seguinte perfil: estudante de escola pública, com acesso a internet, filho de pais com escolaridade de "primeiro grau", desempenho no ENEM "regular".

Moreira (2016) teve como objetivo a aplicação de métricas de seleção de atributos, a fim de identificar quais são os mais relevantes na identificação de perfis de inscritos, na base de dados do ENEM de 2010. O trabalho apresenta a relevância da etapa de seleção de atributos em bases de alta dimensionalidade e afirma que a seleção de forma aleatória pode gerar um resultado impreciso ou inútil.

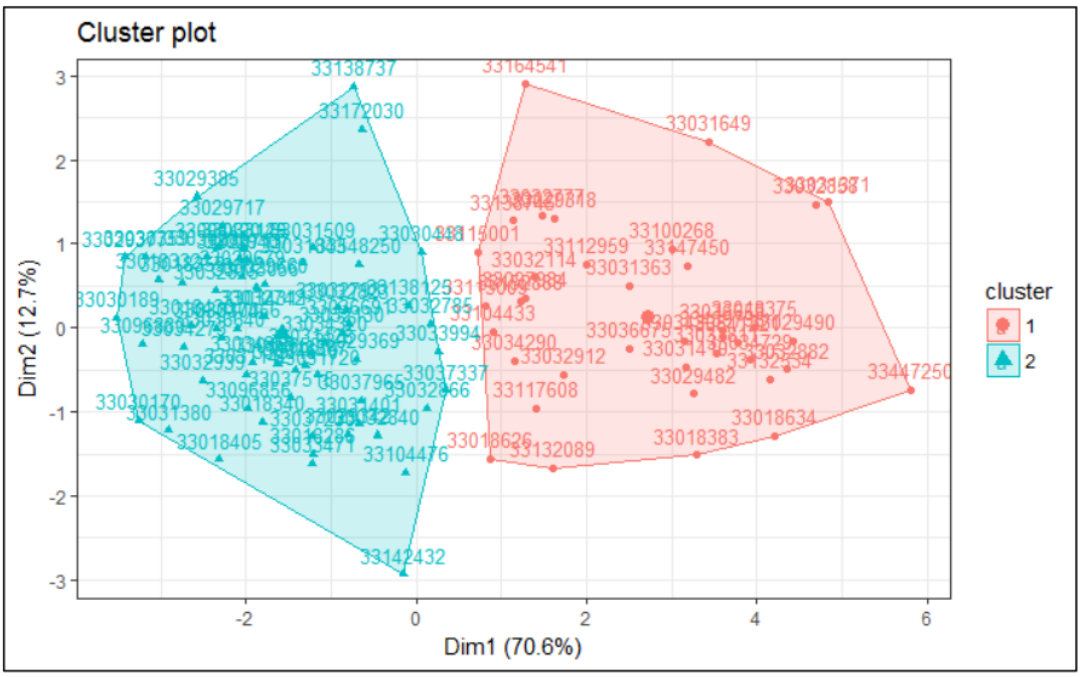
Foram comparadas duas abordagens, *filter* e *wrapper*, de seleção de atributos. Essa comparação foi executada com o software WEKA. Com os resultados obtidos, afirma-se que a abordagem *filter* é a melhor, entre as duas, para ser aplicada na base de dados do ENEM. Após a análise dos resultados da clusterização, conclui-se que os dados socioeconômicos não são capaz de determinar o rendimento do inscrito.

Santos et al. (2018) propuseram um motor de recomendações pedagógicas, para melhorar a qualidade dos textos em um sistema de aprendizagem online. Nesse estudo, foram gerados 20 *clusters*, porém, não foi informado quais atributos foram usados. Ao final do estudo, concluiu-se que a ferramenta trouxe resultados positivos.

Leoni e Sampaio (2017) avaliam o desempenho dos alunos de escolas públicas e privadas da região sul fluminense (Angra dos Reis, Barra do Piraí, Barra Mansa, Itatiaia, Paraty, Piraí, Porto Real, Resende, Três Rios e Volta Redonda) no ENEM. A avaliação é feita por meio da técnica de clusterização *k-means*.

Após a aplicação do algoritmo de agrupamento, foram formados 2 grupos de escolas. A Figura 4 ilustra a formação dos grupos, um com 38 escolas e outro com 65.

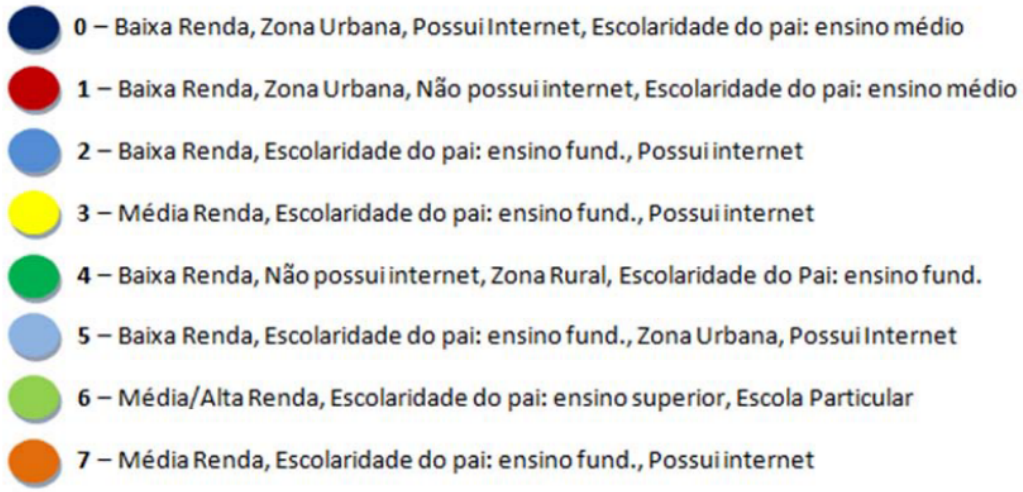
**Figura 4 – Grupos resultantes da aplicação do algoritmo *k-means*.**




**Fonte: Leoni e Sampaio (2017).**

Junior et al. (2018) aplicaram técnicas de mineração de dados para descobrir o perfil dos alunos que prestaram o exame no ano de 2011. A Figura 5 apresenta os agrupamentos encontrados após a aplicação do algoritmo de clusterização *x-means*.

**Figura 5 – Agrupamentos encontrados.**




**Fonte: Junior et al. (2018).**

A partir desta revisão sistemática, foi identificada possibilidade de pesquisa na aplicação de clusterização nos *datasets* do ENEM. Poucos algoritmos foram usados, além de que, em nenhum dos artigos, foram utilizados os dados de 2019.

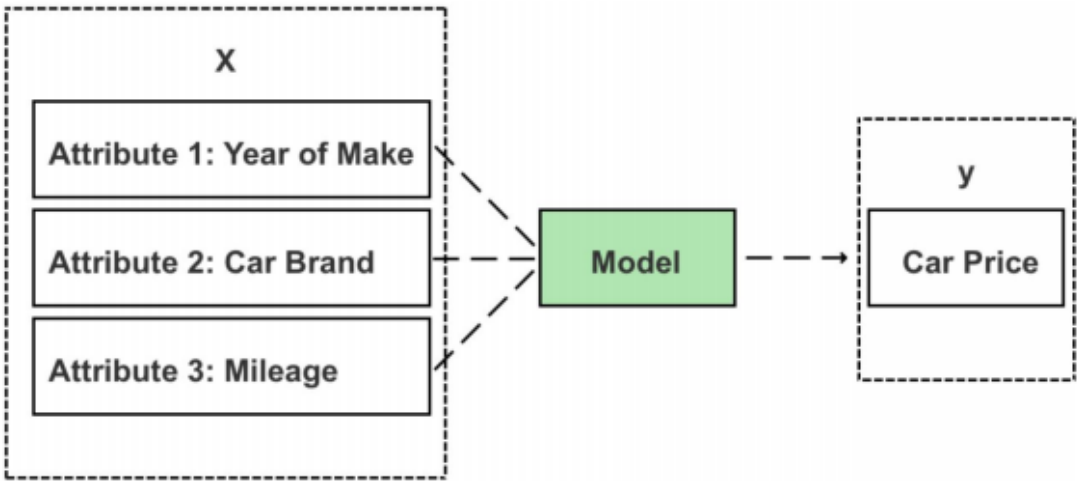
# 3 MACHINE LEARNING

Segundo Burkov (2019), *Machine Learning* (ML) é uma área da ciência da computação que tem como foco a construção de algoritmos, que dependem de uma coleção de exemplos. Para Hurwitz e Kirsch (2018), ML é um conjunto de tecnologias, que auxiliam a compreensão de dados em grande volume. A aprendizagem pode ser supervisionada, semi-supervisionada, não supervisionada e de reforço (BURKOV, 2019).

## 3.1 APRENDIZADO SUPERVISIONADO

Na aprendizagem supervisionada, o *dataset* é uma coleção de exemplos rotulados. Algoritmos de ML supervisionados são usados para, a partir de um *dataset*, criar um modelo que receba um objeto como entrada e gere informações que possibilitem rotular esse objeto (BURKOV, 2019). A Figura 6 ilustra a entrada de dados, sendo rotulados após a aplicação do modelo.

**Figura 6 – Exemplo de aprendizado supervisionado.**




**Fonte: Theobald (2017).**

Esse método de aprendizagem tem como objetivo encontrar padrões nos dados (HURWITZ; KIRSCH, 2018). A Figura 6 exemplifica a aprendizagem supervisionada, especificamente, a regressão, que é uma técnica desse método. Conforme a Figura 6 exemplifica, o 'X' é a entrada de dados, nesse caso o objeto de um carro, contendo 3 atributos: ano de fabricação, marca do carro, e a quilometragem. Após a aplicação do modelo de aprendizagem supervisionado, é possível chegar no valor 'Y', que nesse caso é o preço do carro.

As duas principais técnicas de aprendizado supervisionado são classificação e regressão. A classificação é usada para prever rótulos classes, ou seja, classificar atributos.

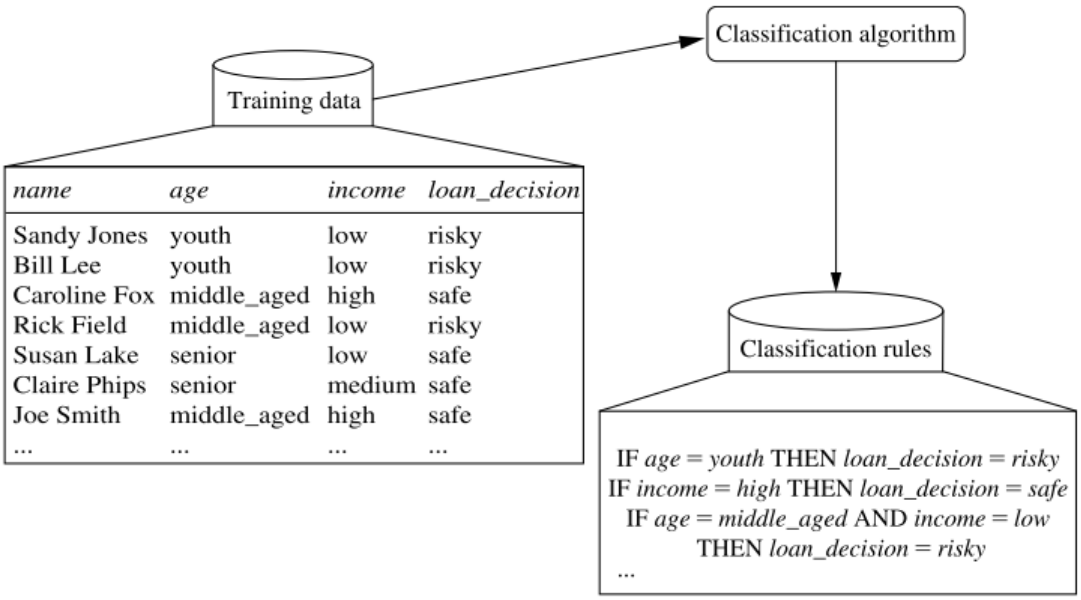
A regressão é usada para prever rótulos numéricos (TAN et al., 2018).

### 3.1.1 Classificação

A classificação é uma técnica de aprendizado supervisionado que, conforme Han, Pei e Kamber (2011) é composta por duas etapas. A primeira é a etapa de aprendizado, onde o modelo de classificação é construído. A segunda é a classificação, que é constituída pela aplicação do modelo de classificação gerado. A classificação é usada para prever rótulos de classe de dados fornecidos. Um rótulo de classe é um rótulo (atributo) que representa uma categoria.

A Figura 7 representa a primeira etapa da classificação, que consiste na geração de regras de classificação, ou seja, geração do modelo. Nesta figura, o *traning data* representa um *dataset* onde os objetos possuem os seguintes atributos: nome, idade, renda e decisão de empréstimo. Também é ilustrada nessa figura a aplicação da técnica de classificação, para geração das regras de classificação, que, nesse exemplo, tem o objetivo de prever o rótulo classe 'decisão de empréstimo'.

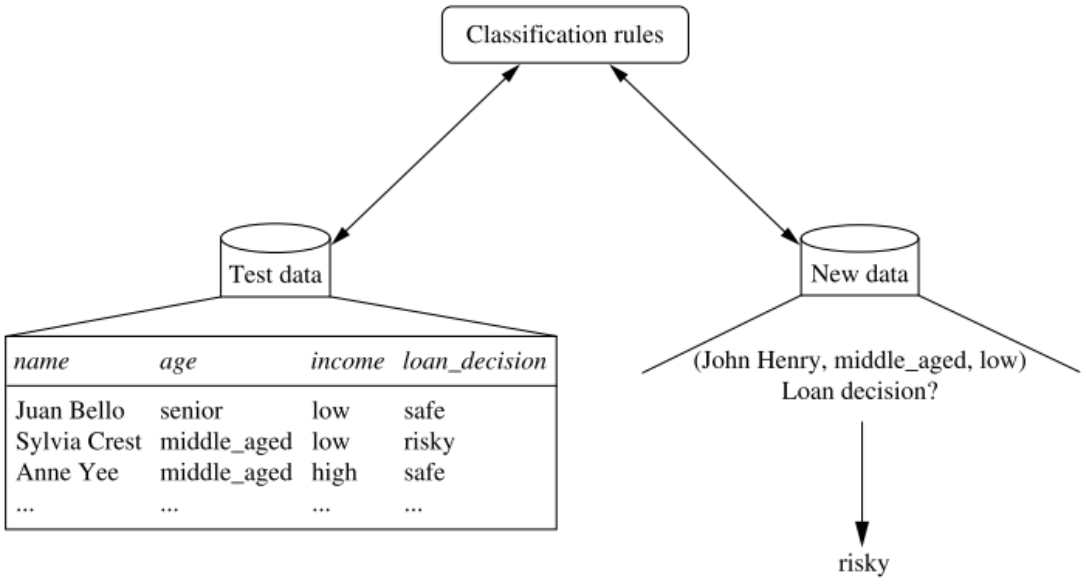
**Figura 7 – Primeira etapa da classificação.**




**Fonte: Han, Pei e Kamber (2011).**

A Figura 8 representa a segunda etapa da classificação, onde são usadas as regras de classificação geradas, para rotular novos dados.

**Figura 8 – Segunda etapa da classificação.**




**Fonte: Han, Pei e Kamber (2011).**

### 3.1.2 Regressão

A regressão também é uma técnica de aprendizado supervisionado que, diferente da classificação, é utilizada para prever um rótulo numérico (BURKOV, 2019). Um exemplo de regressão, é a estimativa de valor de uma residência, com base nas suas características, como número de quartos, localização, área, entre outras (BURKOV, 2019).

## 3.2 APRENDIZADO NÃO SUPERVISIONADO

Os algoritmos de aprendizagem não supervisionada para Hurwitz e Kirsch (2018):

> *segmentam dados em grupos de exemplos (clusters) ou grupos de recursos. Os dados não rotulados criam os valores dos parâmetros e a classificação dos dados. Em essência, esse processo adiciona rótulos aos dados para que sejam supervisionados. A aprendizagem não supervisionada pode determinar o resultado quando há um grande quantidade de dados.*

Para Han, Pei e Kamber (2011), o aprendizado não supervisionado é um sinônimo de clusterização.

Essa aprendizagem concentra-se em analisar as relações entre as variáveis de entrada e busca encontrar padrões ocultos. Um cenário exemplo é na área de análise de fraude, onde a aprendizagem não supervisionada é aplicada para detecção de fraudes não conhecidas (THEOBALD, 2017).

Algoritmos de aprendizagem não supervisionada colaboram com a compreensão de grande volume de dados não rotulados. Assim como os algoritmos de aprendizado supervisionado procuram padrões nos dados. A diferença, é que os dados ainda não foram compreendidos (HURWITZ; KIRSCH, 2018).

### 3.2.1 Clusterização

Com o objetivo de identificar padrões em diferentes grupos de estudantes, que fizeram o ENEM de 2018 e 2019, este trabalho utilizará da abordagem de aprendizado não supervisionado, especificamente, a clusterização.

A clusterização tem como objetivo agrupar um conjunto de dados, cujas características se assemelham (SHAI, 2014). Diferente da classificação e da regressão, que analisam dados rotulados de classe, a clusterização analisa os dados sem consultar os rótulos de classe (HAN; PEI; KAMBER, 2011).

Han, Pei e Kamber (2011) afirmam que "o agrupamento de objetos são formados de forma que objetos dentro de um agrupamento tenham alta similaridade em comparação com os outros, mas bastante diferentes de objetos em outros agrupamentos". Cada *cluster* formado pode ser visto como uma classe de objetos, ou seja, um rótulo classe.

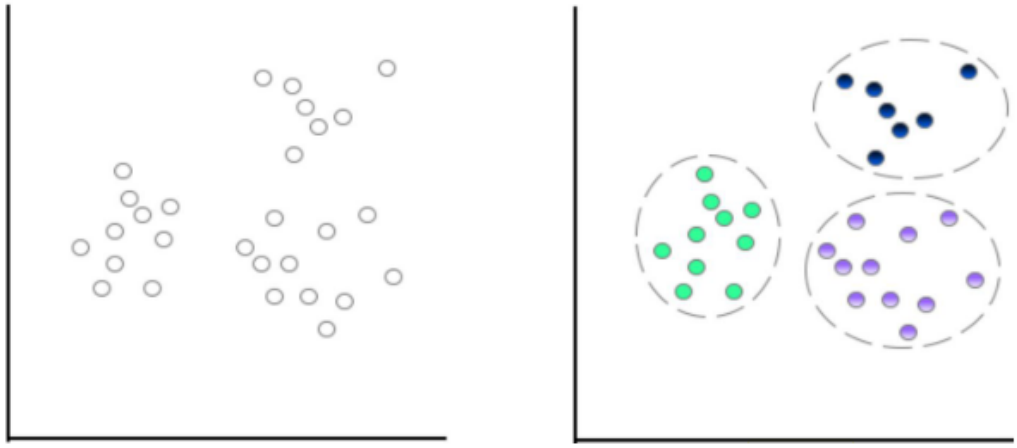
Existem muito algoritmos de clusterização, que, em geral, podem ser classificados da seguinte forma: *hierarchical methods*, *density-based methods* e *grid-based methods* (HAN; PEI; KAMBER, 2011). Nesse trabalho, serão abordados os algoritmos mais citados na revisão sistemática, sendo eles *k-means* e *DBSCAN*.

#### 3.2.1.1 K-means

O algoritmo *k-means*, conforme Theobald (2017), tenta dividir os dados em um número 'K' de grupos discretos. A Figura 9 ilustra a aplicação do algoritmo *k-means*, para formação de 3 *clusters*. O valor de 'K', nesse caso, é 3.

Moreira (2016) explica o algoritmo da seguinte forma:

**Figura 9 – Comparação dos dados originais com o agrupamento dos dados usando *k-means*.**



**Fonte: Theobald (2017).**

> *Para gerar os clusters e determinar seus elementos, o algoritmo realiza comparações entre cada elemento e cada centróide por meio de uma função de distância ou de (dis)similaridade. O elemento é designado ao cluster cujo centroide é mais similar (ou de menor distância). A função de distância é calculada utilizando medidas de dissimilaridade, que determinam o quão diferente são dois elementos em um grupo. Em seguida, o algoritmo recalcula os centroides para cada um dos grupos baseado nos seus elementos. Esse procedimento se repete várias vezes até que alguma condição de parada seja atingida. Exemplos de condições de paradas podem ser: (i) número máximo de iterações ou (ii) até que a soma do erro quadrático total (entre cada centroide e os elementos do cluster) se estabilize entre uma iteração e outra.*

A Figura 10 mostra as interações durante a execução do algoritmo *k-means*, com o 'K' tendo o valor 3. Em (a), os centroides '+' são gerados na primeira iteração. Em (b), (c) e (d), dois centroides movem-se para baixo, nas sucessivas iterações, onde encontram, cada um, um grupo.

#### 3.2.1.2 DBSCAN

Tan et al. (2018) definem o algoritmo DBSCAN como um algoritmo baseado em densidade, que localiza regiões de alta densidade, as quais são separadas uma das outras por regiões de baixa densidade. A densidade de um objeto pode ser medida com base no número de objetos próximos.

O algoritmo DBSCAN seleciona pontos de forma aleatória, e, caso o número de pontos vizinhos desse ponto selecionado seja igual ou maior que o número definido (parâmetro) ao utilizar o algoritmo, esses pontos são considerados pontos *core*. Pontos que

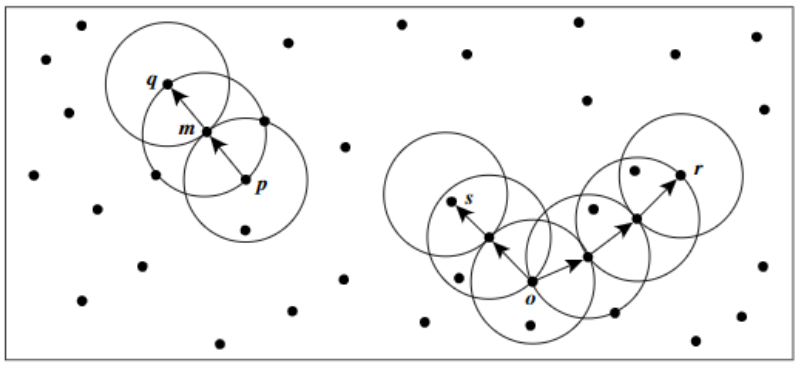
**Figura 10 – Ilustração da aplicação de *k-means*, quando 'K' igual 3.**

(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

**Fonte: Tan et al. (2018).**

possuem o número de vizinhos menor do que o definido são considerados pontos não *core*. Para identificar se um ponto é vizinho ou não, do ponto selecionado, é considerado se esse ponto próximo está dentro de um raio, *epsilon*, definido (parâmetro), ou seja, apenas os pontos próximos, dentro do raio, são considerados vizinhos (HAN; PEI; KAMBER, 2011). A Figura 11 ilustra a aplicação do algoritmo em alguns pontos, afim de identificar se são pontos *core* ou não.

**Figura 11 – Ilustração da aplicação do algoritmo.**



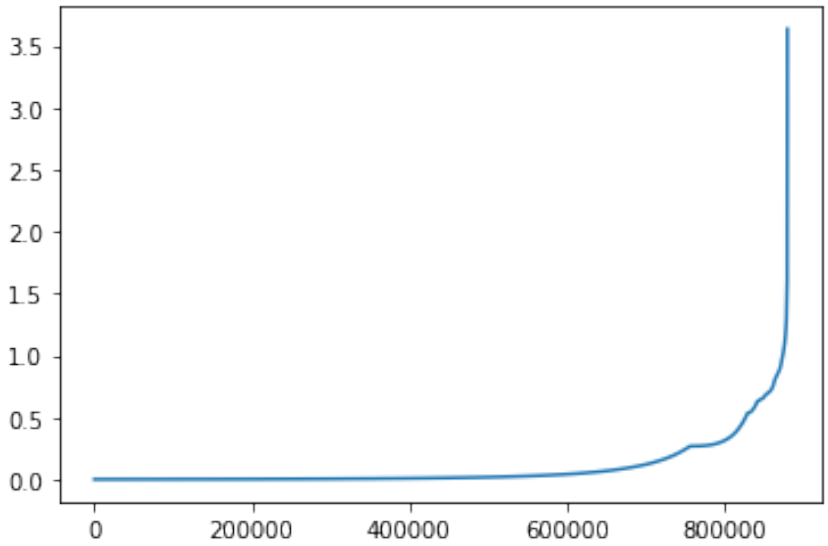

**Fonte: Han, Pei e Kamber (2011).**

Após a identificação dos pontos *core*, é selecionado 1 ponto *core*, de forma aleatória, para a criação de um cluster. Então, os vizinhos desse ponto *core* também fazem parte do mesmo *cluster*. O mesmo processo é feito para cada ponto vizinho, que entrou para o agrupamento. Importante destacar que, um ponto não *core*, pode fazer parte do agrupamento, caso seja vizinho de um ponto *core*. Um ponto *core* que é vizinho apenas de outros pontos não *core*, não é inserido no agrupamento. Dessa forma, todos os pontos próximos, conforme o parâmetro definido, serão considerados partes de um mesmo agrupamento. O pontos que ficaram de fora desse *cluster*, passam pelo mesmo processo, gerando novos *clusters*. Por isso, diferente do algoritmo *k-means*, no DBSCAN não é necessário definir o

número de agrupamentos (HAN; PEI; KAMBER, 2011).

Na seleção do valor ideal do número de vizinhos ao redor do ponto, para ser considerado um ponto *core* ou não, é utilizada a dimensão do *dataset*. O valor é calculado multiplicando a dimensão, da base de dados, por 2. Já, para a identificação do valor ideal do *epsilon*, é necessário observar a distância de cada ponto e seu vizinho mais próximo. Essas distâncias devem ser ordenadas em ordem crescente e plotadas. O ponto do gráfico, onde existir uma maior variação, é considerado o valor ideal para o *epsilon* (HAN; PEI; KAMBER, 2011). A Figura 12 ilustra essa visualização.

**Figura 12 – Ilustração da observação do valor de *epsilon***




**Fonte: elaborado pelo autor.**

Esse capitulo aborda os métodos supervisionado e não supervisionado de ML. Com base nas definições, neste trabalho será utilizando o método de ML não supervisionada, afim de analisar as relações dos atributos socioeconômicos, com relação a nota média dos participantes do ENEM de 2018 e 2019.

# 4 APLICAÇÃO DE CLUSTERIZAÇÃO

Com base na revisão sistemática, foi selecionado o algoritmo *k-means* para ser aplicado, afim de gerar agrupamentos de participantes do ENEM de 2018 e 2019. A ferramenta mais citada nos artigos foi o WEKA, porém o mesmo não foi usado.

Após a seleção do algoritmo e do filtro de ferramentas a serem usadas para a aplicação do *k-means*, iniciou-se o processo de pré-processamento de dados. Após a redução do volume de dados e dos ajustes aplicados em cada dado, o algoritmo *k-means* foi aplicado e, em seguida, foram geradas visualizações para as questões socioeconômicas em relação a nota média.

## 4.1 METODOLOGIA

Os *datasets* do ENEM estão disponíveis no INEP. Atualmente, no portal do INEP, estão disponíveis os dados de 1998 até 2019. Para este trabalho, foram selecionadas as bases de dados dos anos 2018 e 2019.

Para os *datasets* selecionados, serão utilizados atributos relacionados ao perfil socioeconômico dos participantes, escolhidos de forma manual, para que seja possível uma análise comparativa entre os resultados gerados nessas duas bases de dados, com os resultados dos anos anteriores, de 2012 até 2017. Não será possível utilizar os mesmos atributos do artigo de (LIMA et al., 2020), pois, nesse trabalho, não foram citados os atributos utilizados. Será aplicado o algoritmo de clusterização *k-means*, pois é o algoritmo mais citado dentre os trabalhos da revisão sistemática.

Para trabalhar com um grande volume de dados, foi utilizada a ferramenta Colaboratory[1] (Colab), que é um produto da Google Research[2]. Colab é uma ferramenta que possibilita a escrita e execução de código em Python[3] através do navegador (GOOGLE RESEARCH, 2021). Conforme Google Research (2021), "Colab é um serviço de blocos de notas alojados do Jupyter[4] que não requer nenhuma configuração para utilizar e que, ao mesmo tempo, oferece acesso gratuito a recursos informáticos como GPUs".

Colab é uma ferramenta gratuita, porém possui algumas limitações de recursos, memória RAM e armazenamento, nessa versão. Devido a isso, foi utilizada, para este trabalho, a versão paga do Colab, o Colab Pro. Nessa versão os recursos são maiores, possibilitando o manuseio de grandes volumes de dados.

[1] https://colab.research.google.com/
[2] https://research.google.com/
[3] https://www.python.org/
[4] https://www.jupyter.org/

Jupyter é uma aplicação *server-client*, que permite editar e executar *notebooks* via navegador. *Notebooks* são documentos de texto, que possibilitam a escrita de códigos de programação e textos informativos (textos explicativos, figuras, gráficos) (JUPYTER, 2021).

Além disso, o Google Drive também será utilizado. Essa ferramenta será usada para armazenamento dos *datasets* do ENEM. O Colab possui integração com o Google Drive, possibilitando a leitura e manipulação dos dados que já estão na nuvem. Para este trabalho, foi feita a aquisição do plano com 100 *gigabytes* (GB) de armazenamento no Google Drive.

Para o desenvolvimento, será utilizada a linguagem de programação Python, especificamente a versão 3. Na revisão sistemática, destacaram-se as linguagens R e C++, porém, Python também é uma opção viável, visto que, em outro trabalho de análise de participantes do ENEM (CARMO; HECKLER; CARVALHO, 2020), ela foi usada.

Para a aplicação de algoritmos de seleção de atributos e de clusterização, será utilizada a biblioteca *scikit-learn*[5]. Essa biblioteca é de código aberto e possui algoritmos para aprendizado supervisionado e não supervisionado. Além disso, fornece ferramentas para pré-processamento de dados, seleção e validação de modelos, entre outros.

## 4.2 VOLUME DE DADOS

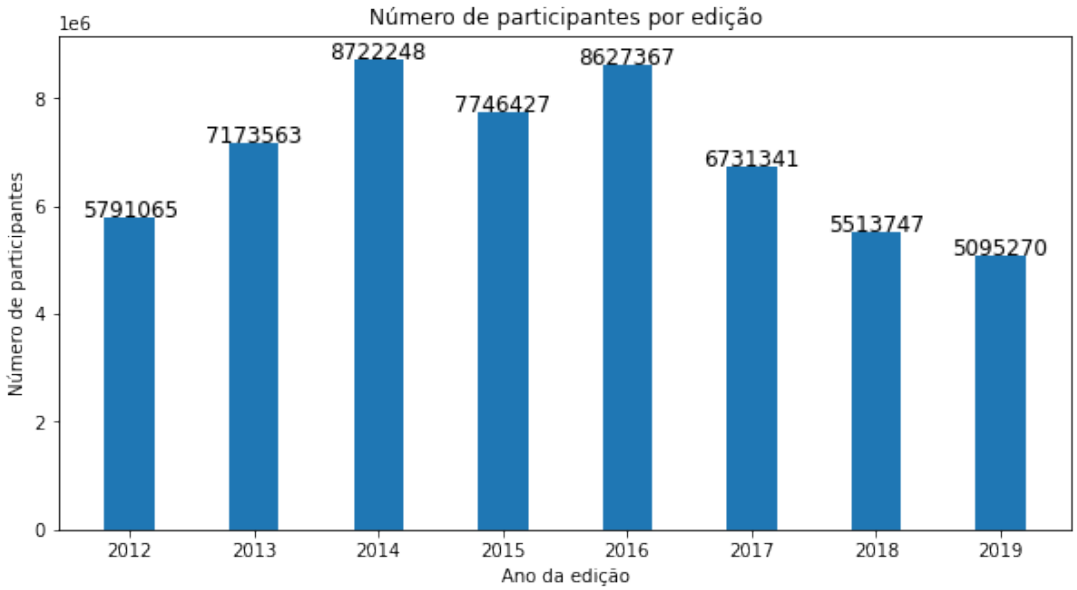
O INEP disponibiliza os dados em arquivos zipados, contendo, além do *dataset*, PDF's com as questões, respostas das questões, dicionário de dados. O arquivo de dados do ano de 2018, possui 3.352.699 *Kilobytes* (kB), contendo 5.513.747 linhas e 137 colunas. O arquivo do ano de 2019, possui 3.123.980 KB, contendo 5.095.270 linhas e 136 colunas. O número de linhas equivale ao número de participantes do ENEM. O número de colunas, representa o número de atributos de cada participante.

Comparando o volume dados dos anos de 2018 e 2019 com os anos anteriores, até 2012, identifica-se que os últimos dois anos possuem menor número de participantes. A Figura 13 ilustra essa diferença.

Os dados do ENEM estão divididos em 9 seções. A primeira seção é constituída pelos dados sobre o participante, como raça, idade e sexo. A segunda possui dados da escola, como a localização, tipo dependência administrativa e situação de funcionamento. A terceira, quarta e quinta seções apresentam dados sobre atendimento especializado para estudantes com deficiência, gestantes e idosos, respectivamente. A sexta seção é composta por dados sobre o local de aplicação da prova. Na sétima e oitava seções estão os dados sobre as provas objetivas e sobre a redação. Por fim, na nona seção, estão as respostas

[5] https://scikit-learn.org/

**Figura 13 – Gráfico com os números de participantes do ENEM por edição.**

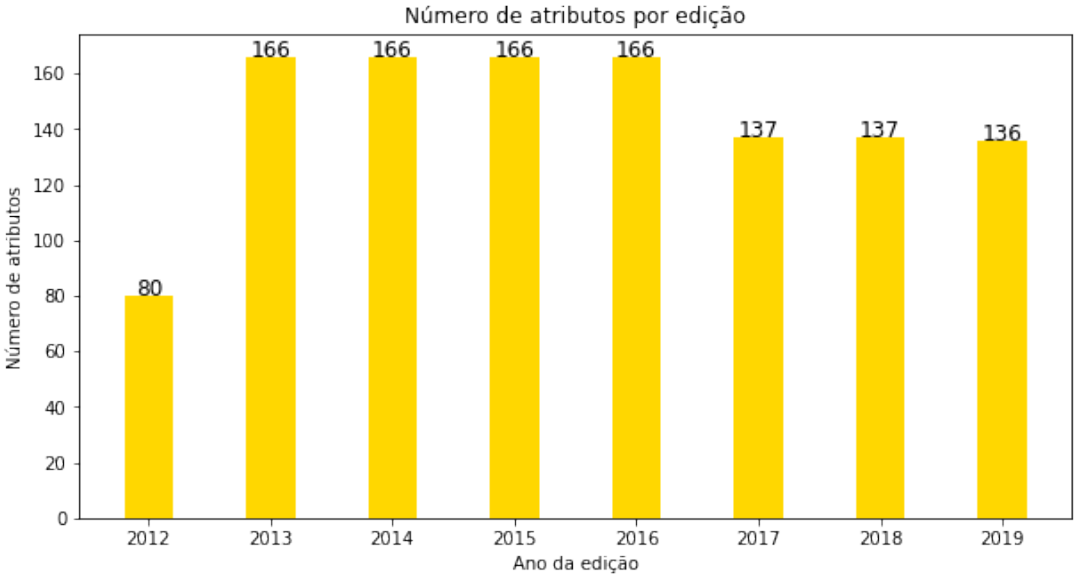



**Fonte: elaborado pelo autor.**

para um questionário socioeconômico, que mapeia o perfil dos estudantes, desde a renda familiar mensal até o grau de estudo dos pais.

O número de atributos pode variar em cada edição do ENEM. A Figura 14 mostra a diferença no número de atributos ao decorrer dos anos de 2012 até 2019.

**Figura 14 – Gráfico com os números de atributos por edição do ENEM.**




**Fonte: elaborado pelo autor.**

## 4.3 PRÉ-PROCESSAMENTO

Primeiramente, foi feita uma redução no volume de dados, a partir de uma limpeza dos dados inconsistentes. Foram removidos os registros de participantes de ambas edições, 2018 e 2019, que não possuíssem alguma das notas: nota da prova de Ciências Humanas, nota da prova de Ciências da Natureza, nota da prova de Linguagens e Códigos, nota da prova de matemática e nota da redação.

Para a seleção manual de atributos a serem usados para geração dos *clusters*, foi feito, primeiramente, um filtro por seção nas bases de dados do ENEM. Com isso, foram descartados os atributos referentes aos dados de local de aplicação da prova, de atendimento especializado, dos pedidos de atendimento especializado e de pedidos de recursos especializados e específicos para realização das provas. Nessas seções removidas, não constam informações socioeconômicas de relevância, além de não possuir nenhuma informação básica do candidato, como sexo, raça e idade.

A seção de dados do questionário socioeconômico foi quase inteiramente utilizada. Após uma análise dos dicionários de dados do ENEM, dos anos de 2018 e 2019, verificou-se que, no ano de 2019, existem duas questões socioeconômicas a menos. Uma das questões removidas é a "Q026" que corresponde a pergunta "Você já concluiu ou está concluindo o Ensino Médio?". A outra questão removida é a "Q027" que corresponde a pergunta "Em que tipo de escola você frequentou o Ensino Médio?". Essa segunda pergunta possui relevância, pois é possível identificar a relação desse atributo em diferentes desempenhos no trabalho de Carmo, Heckler e Carvalho (2020). Devido a isso, será utilizado o atributo "TP_DEPENDENCIA_ADM_ESC", para que essa resposta ainda seja usada.

A Tabela 10 contém os atributos escolhidos, de forma manual e com base em uma análise de um trabalho anterior (CARMO; HECKLER; CARVALHO, 2020).

**Tabela 10 – Atributos selecionados manualmente.**

| Nome do atributo no dicionario | Significado do atributo |
|---|---|
| Q001 | Até que série seu pai, ou o homem responsável por você, estudou? |
| Q002 | Até que série sua mãe, ou a mulher responsável por você, estudou? |

| | |
|---|---|
| Q003 | A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você. (Se ele não estiver trabalhando, escolha uma ocupação pensando no último trabalho dele). |
| Q004 | A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da sua mãe ou da mulher responsável por você. (Se ela não estiver trabalhando, escolha uma ocupação pensando no último trabalho dela). |
| Q005 | Incluindo você, quantas pessoas moram atualmente em sua residência? |
| Q006 | Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.) |
| Q007 | Em sua residência trabalha empregado(a) doméstico(a)? |
| Q008 | Na sua residência tem banheiro? |
| Q009 | Na sua residência tem quartos para dormir? |
| Q010 | Na sua residência tem carro? |
| Q011 | Na sua residência tem motocicleta? |
| Q012 | Na sua residência tem geladeira? |
| Q013 | Na sua residência tem freezer (independente ou segunda porta da geladeira)? |

| | |
|---|---|
| Q014 | Na sua residência tem máquina de lavar roupa? (o tanquinho NÃO deve ser considerado) |
| Q015 | Na sua residência tem máquina de secar roupa (independente ou em conjunto com a máquina de lavar roupa)? |
| Q016 | Na sua residência tem forno micro-ondas? |
| Q017 | Na sua residência tem máquina de lavar louça? |
| Q018 | Na sua residência tem aspirador de pó? |
| Q019 | Na sua residência tem televisão em cores? |
| Q020 | Na sua residência tem aparelho de DVD? |
| Q021 | Na sua residência tem TV por assinatura? |
| Q022 | Na sua residência tem telefone celular? |
| Q023 | Na sua residência tem telefone fixo? |
| Q024 | Na sua residência tem computador? |
| Q025 | Na sua residência tem acesso à Internet? |
| SG_UF_RESIDENCIA | Sigla da Unidade da Federação de nascimento |
| NU_IDADE | Idade |
| TP_SEXO | Sexo |
| TP_COR_RACA | Cor/raça |
| TP_NACIONALIDADE | Nacionalidade |
| TP_DEPENDENCIA_ADM_ESC | Dependência administrativa (Escola) |
| NU_NOTA_CN | Nota da prova de Ciências da Natureza |

| | |
|---|---|
| NU_NOTA_CH | Nota da prova de Ciências Humanas |
| NU_NOTA_LC | Nota da prova de Linguagens e Códigos |
| NU_NOTA_MT | Nota da prova de Matemática |
| NU_NOTA_REDACAO | Nota da redação |

**Fonte: elaborado pelo autor.**

Foi criado um rótulo, para ambas bases de dados (2018 e 2019), chamado de "MEDIA_GERAL", a partir das notas de cada prova, somando com a nota da redação, dividido por 5 (número de notas). As notas de cada prova podem variar entre os valores 0 e 1000.

Após a seleção dos atributos, as dimensões dos *datasets* da edição de 2018 e 2019, respectivamente, foram reduzidas a 5513747x31 e 5095270x31. Em seguida, foram removidas as linhas que tinham algum valor nulo. Antes da remoção de participantes sem alguma dessas notas, o número de participantes, do ENEM de 2018 era 5.513.747. Com a remoção, o número de participantes do ENEM de 2018 caiu para 1.025.790. Já a base de dados do ENEM de 2019, que possui 5.095.270 participantes, após o filtro por aqueles que possuem todas as notas, o número caiu para 953.889.

Além da remoção das linhas que continham qualquer um dos valores nulos, foram removidos os participantes que responderam "Não sei", nas questões socioeconômicas "Q001", "Q002", "Q003" e "Q004". Com isso, o número restantes de participante da edição de 2018 e 2019 é, respectivamente, 880.680 e 743.490.

**Figura 15 – Exemplo de colunas com valores não numéricos.**

| Q012 | Q013 | Q014 | Q015 | Q016 | Q017 | Q018 | Q019 | Q020 | Q021 | Q022 | Q023 | Q024 | Q025 | SG_UF_RESIDENCIA | NU_IDADE | TP_SEXO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B | A | B | A | A | A | A | C | A | B | D | A | B | B | TO | 25.0 | F |
| B | B | B | A | B | A | A | B | A | A | C | B | B | B | MG | 22.0 | F |
| B | A | B | A | B | A | A | B | A | A | E | A | B | B | MT | 37.0 | M |
| B | A | A | A | A | A | A | B | A | A | B | A | A | A | BA | 22.0 | F |
| B | B | B | A | B | A | A | B | A | B | C | B | B | B | SP | 17.0 | M |

**Fonte: elaborado pelo autor.**

Além disso, para a aplicação do algoritmo de clusterização, foram alterados o tipo de alguns dos atributos selecionados de *string* para numérico. Entre os atributos de questionário socioeconômico, somente o atributo "Q05" não foi alterado, pois esse já possui como valor de resposta números inteiros. As demais questões, de "Q001" até "Q025", além de "SG_UF_RESIDENCIA", "TP_SEXO" tinham a resposta em formato não numérico, variando entre "A", "B", "C", "D", "E" e "F". Após a alteração de tipo dos

atributos, todos os valores respostas se tornaram numéricos. A Figura 15 mostra algumas colunas e seus valores não numéricos. Já a Figura 16 ilustra 5 registros aleatórios de forma numérica.

**Figura 16 – Exemplo de colunas com valores numéricos.**

| ... | Q022 | Q023 | Q024 | Q025 | SG_UF_RESIDENCIA | NU_IDADE | TP_SEXO | TP_COR_RACA | TP_DEPENDENCIA_ADM_ESC | MEDIA_GERAL |
|---|---|---|---|---|---|---|---|---|---|---|
| ... | 1 | 0 | 0 | 0 | 5 | 19.0 | 1 | 3 | 2.0 | 417.16 |
| ... | 1 | 0 | 0 | 0 | 16 | 54.0 | 0 | 3 | 2.0 | 448.02 |
| ... | 3 | 1 | 1 | 1 | 25 | 17.0 | 0 | 1 | 2.0 | 473.72 |
| ... | 2 | 0 | 0 | 1 | 11 | 18.0 | 0 | 1 | 2.0 | 617.44 |
| ... | 1 | 0 | 0 | 1 | 7 | 17.0 | 1 | 2 | 2.0 | 456.38 |

**Fonte: elaborado pelo autor.**

As bases de dados do ENEM possuem atributos com magnitude muito maior que outros, como por exemplo as notas em comparação com as questões socioeconômicas que, pós a conversão dos atributos não numéricos para numéricos, a magnitude varia de 2 até 20. Conforme Pedregosa et al. (2011), é comum o uso de algoritmos de padronização em *machine learning* em *datasets*, cujos dados não possuem uma magnitude semelhante. Devido a isso, foi aplicado o algoritmo *StandardScaler*. A Figura 17 mostra algumas colunas e seus valores padronizados.

**Figura 17 – Exemplo de colunas com valores padronizados.**

| ... | Q022 | Q023 | Q024 | Q025 | SG_UF_RESIDENCIA | NU_IDADE | TP_SEXO | TP_COR_RACA | TP_DEPENDENCIA_ADM_ESC | MEDIA_GERAL |
|---|---|---|---|---|---|---|---|---|---|---|
| ... | -1.392712 | -0.696936 | -0.946009 | -1.799295 | -1.247307 | 0.349679 | 1.166333 | 0.906722 | -0.481638 | -1.323502 |
| ... | -1.392712 | -0.696936 | -0.946009 | -1.799295 | 0.180611 | 12.694131 | -0.857388 | 0.906722 | -0.481638 | -0.970143 |
| ... | 0.426480 | 1.434851 | 0.239546 | 0.555773 | 1.348906 | -0.355718 | -0.857388 | -1.038294 | -0.481638 | -0.675869 |
| ... | -0.483116 | -0.696936 | -0.946009 | 0.555773 | -0.468443 | -0.003019 | -0.857388 | -1.038294 | -0.481638 | 0.969779 |
| ... | -1.392712 | -0.696936 | -0.946009 | 0.555773 | -0.987685 | -0.355718 | 1.166333 | -0.065786 | -0.481638 | -0.874418 |

**Fonte: elaborado pelo autor.**

## 4.4 APLICANDO CLUSTERIZAÇÃO NO ENEM DE 2018 E 2019

Para a definição do número de *clusters* a ser utilizado como parâmetro nas funções da biblioteca *sklearn*, foi executado o seguinte trecho de código:

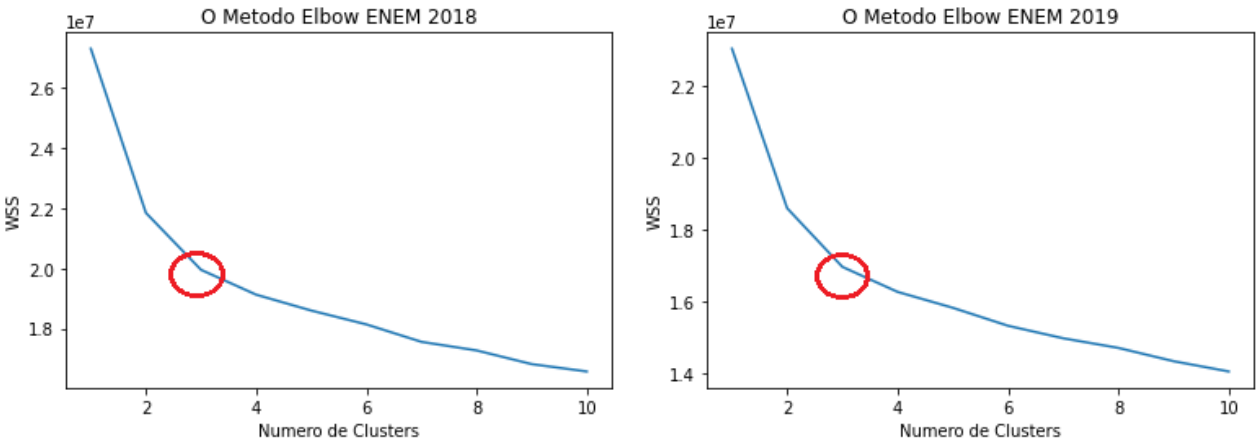
```
1       wcss = []
2
3       for i in range(1, 11):
4           kmeans = KMeans(n_clusters = i, init = 'random')
5           kmeans.fit(dftratado.values)
6           print(i,kmeans.inertia_)
7           wcss.append(kmeans.inertia_)
8           plt.plot(range(1, 11), wcss)
```

```
9           plt.title('O Metodo Elbow')
10          plt.xlabel('Numero de Clusters')
11          plt.ylabel('WSS')
12          plt.show()
```

Com isso, foram geradas as visualizações do Método *Elbow* para ambos os anos. A ideia do Método *Elbow* é rodar o algoritmo *k-means* várias vezes, com diferentes números de *clusters* a serem gerados, a fim de encontrar o número ideal de *clusters* (HAN; PEI; KAMBER, 2011). Para encontrar o número ideal de agrupamentos, é observada a variação intra-*clusters* de cada *cluster* (*within-clusters sum-of-squares*) (HAN; PEI; KAMBER, 2011).

A partir da visualização gerada, foi possível identificar o número de *clusters* a ser utilizado como parâmetro para a execução do algoritmo. A Figura 18 ilustra a visualização do Método *Elbow*, onde foi possível identificar que, para ambos os anos, o número de *clusters* a ser usado é 3.

**Figura 18 – Visualização do *elbow*, dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

Após a identificação do número ideal de *clusters*, foi aplicado o algoritmo *k-means*, utilizando os dados padronizados em formato numérico. Em seguida, foram atrelados os *clusters* gerados para sua respectiva linha (estudante). O trecho de código abaixo exemplifica a geração dos *clusters* e a atribuição:

```
1       kmeans = KMeans(n_clusters= 3)
2
3       clusters_labels = kmeans.fit_predict(dfWithStandardScaler.
            values)
4
5       dfnumeric['clusters'] = clusters_labels
```

Durante cada etapa do pré-processamento e aplicação do algoritmo de clusterização, foram registrados os tempos de execução de cada etapa. A Tabela 11 apresenta o

tempo necessário para execução de cada etapa.

**Tabela 11 – Tempo de execução de cada etapa.**

| Etapa | Tempo na base de 2018 | Tempo na base de 2019 |
|---|---|---|
| Leitura do *dataset* | 2 minutos e 34 segundos | 2 minutos e 8 segundos |
| Conversão de atributos não numéricos para numéricos | 4,82 segundos | 4,01 segundos |
| Aplicação da validação WSS | 6 minutos | 4 minutos e 9 segundos |
| Clusterização | 9,50 segundos | 9,43 segundos |
| Geração das visualizações | Média de 30 segundos | Média de 30 segundos |

**Fonte: elaborado pelo autor.**

Foram geradas visualizações para cada questão socioeconômica, para os *datasets* de 2018 e 2019. Todos os gráficos foram gerados com o mesmo formato. Para construção do Y é usado o atributo "MEDIA_GERAL" e o X, que varia em cada visualização, foram usadas as questões socioeconômicas.

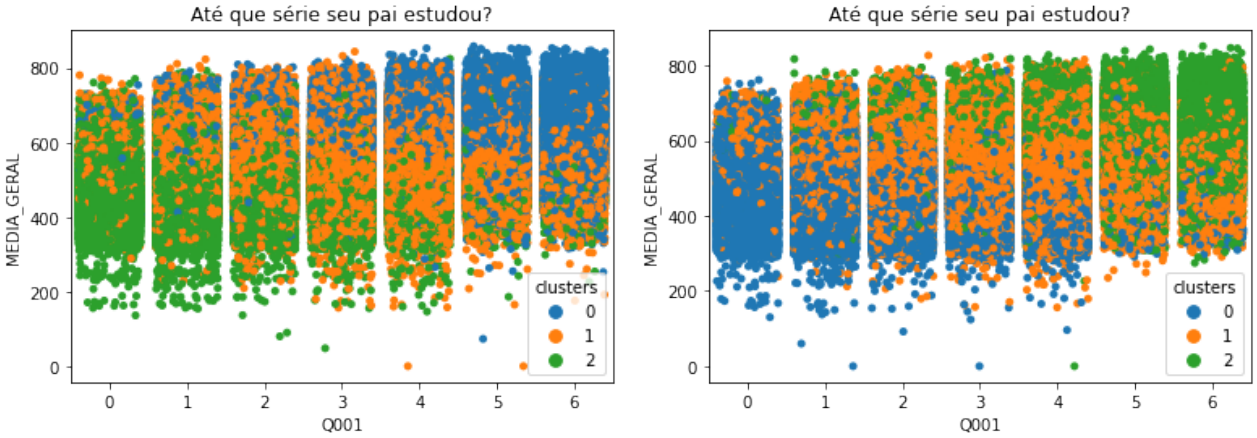
Para ambas edições do ENEM, 2018 e 2019, foram gerados 3 *clusters*, porém, as características dos agrupamentos 0 (azul) e 2 (verde), em ambas edições, são diferentes. Enquanto o *cluster* 2 (verde), da base de dados de 2018, compõem, em geral, as menores notas das visualizações, o *cluster* 2 (verde), da base de dados de 2019, compõem, em geral, as maiores notas das visualizações. Consequentemente, o mesmo ocorre com os agrupamentos 0 (azul), de 2018, e 2 (verde), de 2019.

### 4.4.1 Q001

As primeiras visualizações geradas, Figura 19, foram da questão socioeconômica "Q001", equivalente a pergunta "Até que série seu pai estudou?".

Com base nas visualizações geradas, é possível identificar um padrão nos dois *datasets*. Em ambos, mesmo que os *clusters* de mesma identificação (0, 1 e 2) e cores não estejam nas mesmas posições nos gráficos, é evidente a divisão existente dentro das 7 possíveis respostas. A Tabela 12 contém as possíveis respostas da questão "Q001".

**Figura 19 – Visualização dos *clusters*, pela média geral e "Q001" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

**Tabela 12 – Questão "Q001" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Nunca estudou. |
| 1 | Não completou a 4ª série/5º ano do Ensino Fundamental. |
| 2 | Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental. |
| 3 | Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio. |
| 4 | Completou o Ensino Médio, mas não completou a Faculdade. |
| 5 | Completou a Faculdade, mas não completou a Pós-graduação. |
| 6 | Completou a Pós-graduação. |

**Fonte: elaborado pelo autor.**

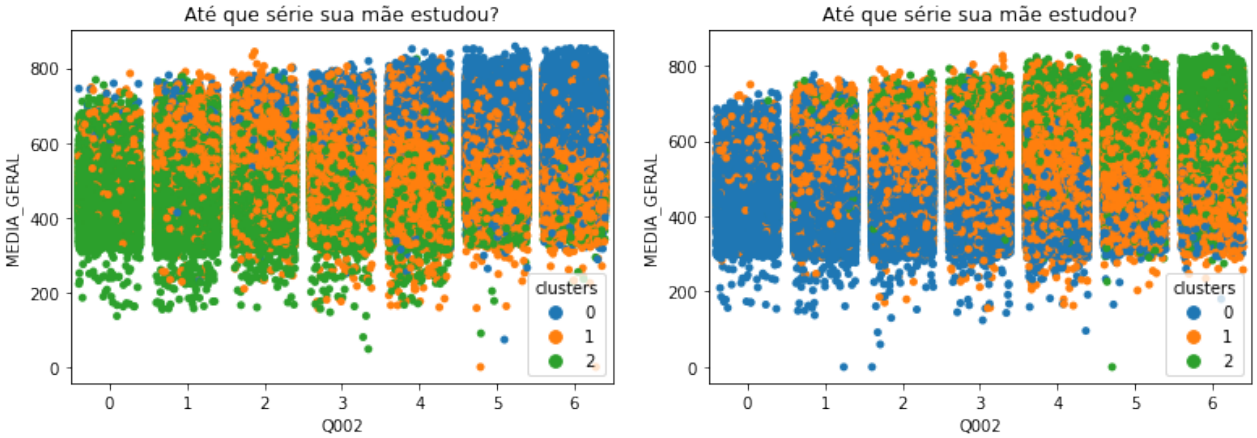
É possível identificar, a partir da visualização, a predominância de cada *cluster*, de ambos *datasets* em certos intervalos de respostas. Nas respostas "Nunca estudou" (0) e "Não completou a 4ª série/5º ano do Ensino Fundamental" (2), a predominância na base de dados do ENEM 2018 é do *cluster* 2 (verde) e, na base de dados do ENEM 2019, o *cluster* 0 (azul). Entre as respostas 2, 3, 4 e 5, o agrupamento mais visível é o laranja (1). Na resposta 6, é onde mais visualizamos o *cluster* 0 (azul) no *dataset* de 2018 e *cluster* 2 (verde) no *dataset* de 2019.

Além disso, é importante destacar que, os participantes cujos pais tem maior grau de estudo, tendem a pertencer a agrupamentos de alunos com médias mais altas.

### 4.4.2 Q002

A Figura 20, referente a questão socioeconômica "Q002", equivalente a "Ate que serie sua mãe estudou?", apresenta um padrão similar ao destacado na Figura 19.

**Figura 20 – Visualização dos *clusters*, pela média geral e "Q002" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

A Tabela 13 apresenta as respostas em formato numérico e seus respectivos valores em formato de texto. A predominância dos *clusters* em cada resposta também é similar a visualização da questão "Q001".

**Tabela 13 – Questão "Q002" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Nunca estudou. |
| 1 | Não completou a 4ª série/5º ano do Ensino Fundamental. |
| 2 | Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental. |
| 3 | Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio. |
| 4 | Completou o Ensino Médio, mas não completou a Faculdade. |
| 5 | Completou a Faculdade, mas não completou a Pós-graduação. |
| 6 | Completou a Pós-graduação. |

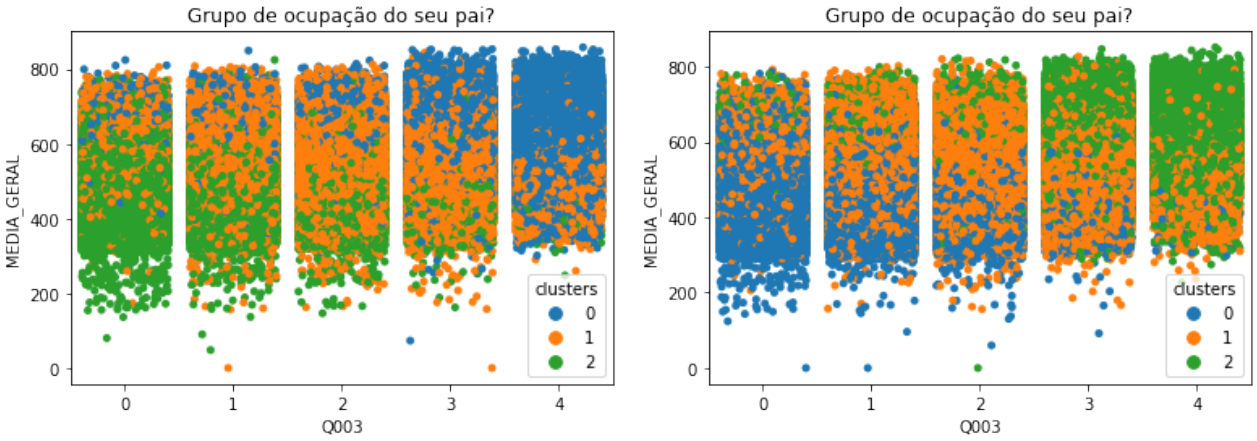
**Fonte: elaborado pelo autor.**

Nessas visualizações também destaca-se o fato de que os participantes cujas mães têm maior grau de estudo, tendem a pertencer a agrupamentos de alunos com médias mais altas.

### 4.4.3 Q003

A Figura 21 apresenta a distribuição de *clusters* em relação a questão socioeconômica "Q003", correspondente a "Grupo que contempla a ocupação mais próxima da ocupação do seu pai?".

**Figura 21 – Visualização dos *clusters*, pela média geral e "Q003" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

A Tabela 14 apresenta as respostas em formato numérico e seu respectivo valor em formato de texto.

**Tabela 14 – Questão "Q003" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Grupo 1: Lavrador, agricultor sem empregados, bóia fria, criador de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultor, pescador, lenhador, seringueiro, extrativista. |
| 1 | Grupo 2: Diarista, empregado doméstico, cuidador de idosos, babá, cozinheiro (em casas particulares), motorista particular, jardineiro, faxineiro de empresas e prédios, vigilante, porteiro, carteiro, office-boy, vendedor, caixa, atendente de loja, auxiliar administrativo, recepcionista, servente de pedreiro, repositor de mercadoria. |

| | |
|---|---|
| 2 | Grupo 3: Padeiro, cozinheiro industrial ou em restaurantes, sapateiro, costureiro, joalheiro, torneiro mecânico, operador de máquinas, soldador, operário de fábrica, trabalhador da mineração, pedreiro, pintor, eletricista, encanador, motorista, caminhoneiro, taxista. |
| 3 | Grupo 4: Professor (de ensino fundamental ou médio, idioma, música, artes etc.), técnico (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretor de imóveis, supervisor, gerente, mestre de obras, pastor, microempresário (proprietário de empresa com menos de 10 empregados), pequeno comerciante, pequeno proprietário de terras, trabalhador autônomo ou por conta própria. |
| 4 | Grupo 5: Médico, engenheiro, dentista, psicólogo, economista, advogado, juiz, promotor, defensor, delegado, tenente, capitão, coronel, professor universitário, diretor em empresas públicas ou privadas, político, proprietário de empresas com mais de 10 empregados. |

**Fonte: elaborado pelo autor.**

A partir da visualização, referente a questão socioeconômica “Q003”, é possível afirmar que o *cluster* 2 (verde), do ENEM de 2018, e 0 (azul), do ENEM de 2019, são compostos, na maior parte, por estudantes, cujos pais pertencem ao grupo de ocupação 1 e 2. Esses 2 agrupamentos também aparecem nas outras respostas de grupo de ocupação, mas em proporção menor. Já o *cluster* 1 (laranja), do ENEM 2018 e 2019, a ocupação

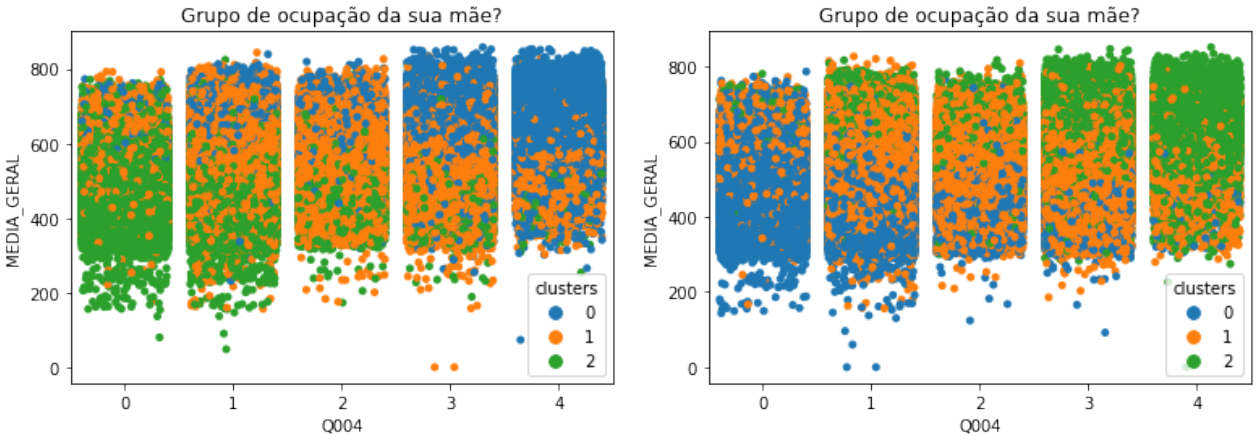
dos pais desses estudantes, em maioria, variam entre os grupos 2, 3, 4 e 5. Por fim, os agrupamentos 0 (azul), do ENEM de 2018, e 2 (verde), do ENEM de 2019, são mais visíveis na resposta 4 da questão "Grupo que contempla a ocupação mais próxima da ocupação do seu pai?".

É visível nos dois gráficos, de 2018 e 2019, que os participantes cujos pais pertencem ao grupo de ocupação 4 e 5, tendem a pertencer a agrupamentos de alunos com médias mais altas.

### 4.4.4 Q004

A Figura 22, referente a questão socioeconômica "Q004", equivalente a "Grupo que contempla a ocupação mais próxima da ocupação da sua mãe", apresenta um padrão similar ao destacado na Figura 21.

**Figura 22 – Visualização dos *clusters*, pela média geral e "Q004" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

A Tabela 15 apresenta as respostas em formato numérico e seus respectivos valores em formato de texto. A predominância dos *clusters* em cada respostas também é similar a visualização da questão "Q003".

**Tabela 15 – Questão "Q004" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Grupo 1: Lavrador, agricultor sem empregados, bóia fria, criador de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultor, pescador, lenhador, seringueiro, extrativista. |

| | |
|---|---|
| 1 | Grupo 2: Diarista, empregado doméstico, cuidador de idosos, babá, cozinheiro (em casas particulares), motorista particular, jardineiro, faxineiro de empresas e prédios, vigilante, porteiro, carteiro, office-boy, vendedor, caixa, atendente de loja, auxiliar administrativo, recepcionista, servente de pedreiro, repositor de mercadoria. |
| 2 | Grupo 3: Padeiro, cozinheiro industrial ou em restaurantes, sapateiro, costureiro, joalheiro, torneiro mecânico, operador de máquinas, soldador, operário de fábrica, trabalhador da mineração, pedreiro, pintor, eletricista, encanador, motorista, caminhoneiro, taxista. |
| 3 | Grupo 4: Professor (de ensino fundamental ou médio, idioma, música, artes etc.), técnico (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretor de imóveis, supervisor, gerente, mestre de obras, pastor, microempresário (proprietário de empresa com menos de 10 empregados), pequeno comerciante, pequeno proprietário de terras, trabalhador autônomo ou por conta própria. |

| | |
|---|---|
| 4 | Grupo 5: Médico, engenheiro, dentista, psicólogo, economista, advogado, juiz, promotor, defensor, delegado, tenente, capitão, coronel, professor universitário, diretor em empresas públicas ou privadas, político, proprietário de empresas com mais de 10 empregados. |

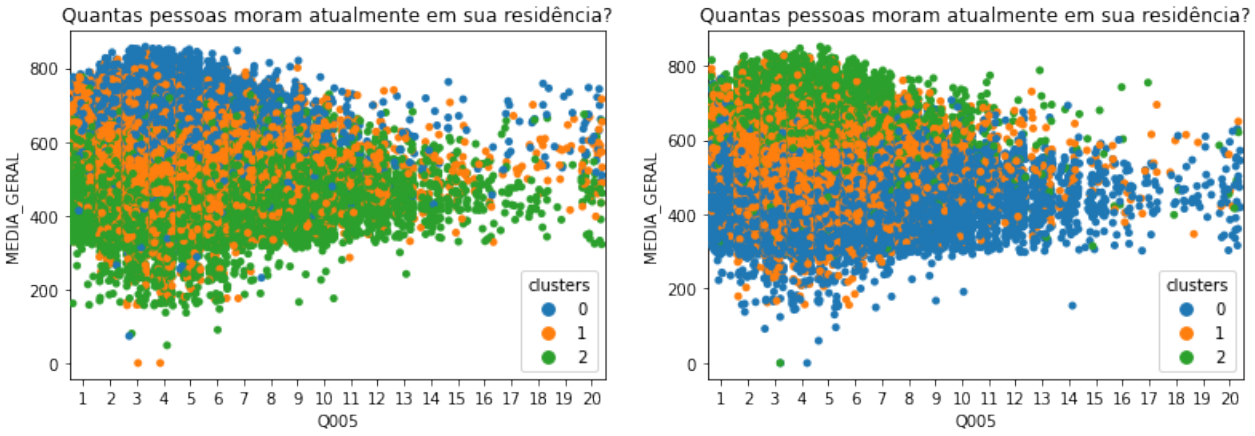
**Fonte: elaborado pelo autor.**

Os participantes cujas mães pertencem ao grupo de ocupação 4 e 5, tendem a pertencer a agrupamentos de alunos com médias mais altas.

### 4.4.5 Q005

Os *clusters* na Figura 23 estão, em geral, menos distribuídos que nas visualizações anteriores. A questão socioeconomia "Q005" é equivalente a pergunta "Incluindo você, quantas pessoas moram atualmente em sua residencia?".

**Figura 23 – Visualização dos *clusters*, pela média geral e "Q005" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

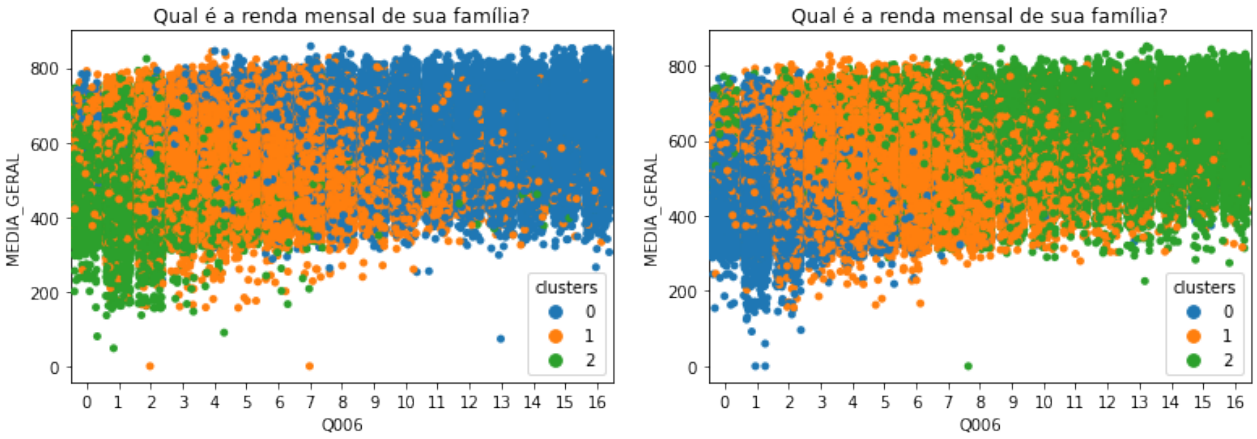
Nessa visualização, os 3 *clusters*, para ambos *datasets*, tem em comum a grande presença nas respostas 2 ate 9. Os agrupamentos 2 (verde), do ENEM de 2018, e 0 (azul), do ENEM de 2019, aparecem em maior quantidade também nas respostas 10 até 14.

### 4.4.6 Q006

A Figura 24 é referente a pergunta "Qual é a renda mensal de sua família?", equivale à questão socioeconômica "Q006". Nessa visualização a segmentação dos *clusters*

é mais evidente que as figuras citadas até então.

**Figura 24 – Visualização dos *clusters*, pela média geral e "Q006" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

A questão "Q006" possui 16 respostas possíveis. A equivalência, em formato de texto, para cada resposta, está representada na Tabela 16.

**Tabela 16 – Questão "Q006" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Nenhuma renda. |
| 1 | Até R$ 954,00. |
| 2 | De R$ 954,01 até R$ 1.431,00. |
| 3 | De R$ 1.431,01 até R$ 1.908,00. |
| 4 | De R$ 1.908,01 até R$ 2.385,00. |
| 5 | De R$ 2.385,01 até R$ 2.862,00. |
| 6 | De R$ 2.862,01 até R$ 3.816,00. |
| 7 | De R$ 3.816,01 até R$ 4.770,00. |
| 8 | De R$ 4.770,01 até R$ 5.724,00. |
| 9 | De R$ 5.724,01 até R$ 6.678,00. |
| 10 | De R$ 6.678,01 até R$ 7.632,00. |
| 11 | De R$ 7.632,01 até R$ 8.586,00. |
| 12 | De R$ 8.586,01 até R$ 9.540,00. |
| 13 | De R$ 9.540,01 até R$ 11.448,00. |
| 14 | De R$ 11.448,01 até R$ 14.310,00. |
| 15 | De R$ 14.310,01 até R$ 19.080,00. |
| 16 | Mais de R$ 19.080,00. |

**Fonte: elaborado pelo autor.**

Nos *clusters* gerados a partir da base de dados do ENEM de 2018, o agrupamento 2 (verde) é mais presente nas respostas 0, 1 e 2. Já o *cluster* 1 (laranja), aparece em maior proporção nas respostas 3, 4, 5, 6, 7, e 8. O agrupamento 0 (azul), é mais evidente das respostas 10 ate 16. Nos agrupamentos gerados a partir da base de dados do ENEM de

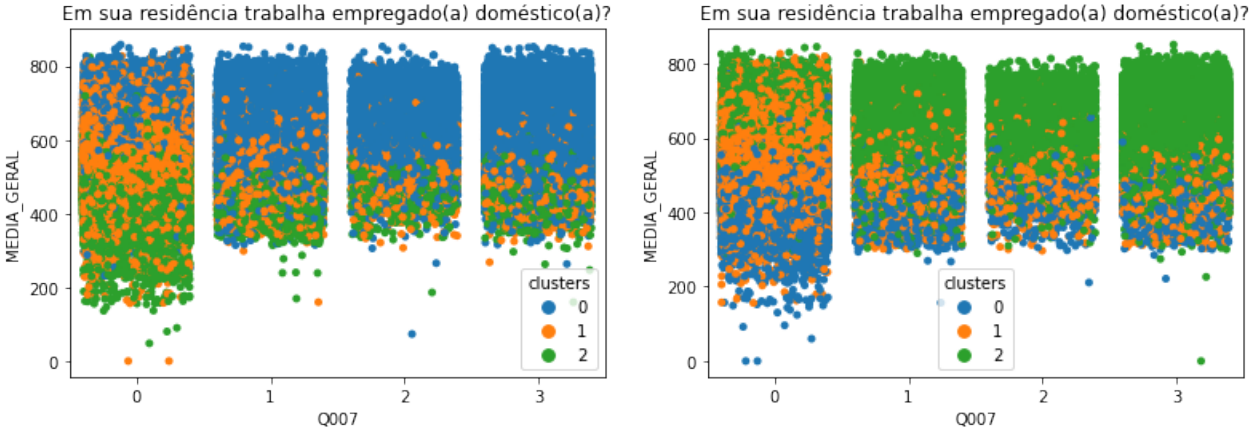
2019, a proporção dos *clusters* em cada resposta é muito similar aos gerados a partir do *dataset* de 2018.

As visualizações sobre a questão socioeconômica "Q006" indicam que, há uma tendência de quanto maior a renda familiar dos participantes, maior a média de nota.

### 4.4.7 Q007

A Figura 25 é referente a pergunta "Em sua residencia trabalha empregado(a) domestico(a)?", equivalente à questão socioeconômica "Q007".

**Figura 25 – Visualização dos *clusters*, pela média geral e "Q007" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

A questão "Q007" possui 4 respostas possíveis. A equivalência em formato de texto, para cada resposta, está representada na Tabela 17.

**Tabela 17 – Questão "Q007" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um ou dois dias por semana. |
| 2 | Sim, três ou quatro dias por semana. |
| 3 | Sim, pelo menos cinco dias por semana. |

**Fonte: elaborado pelo autor.**

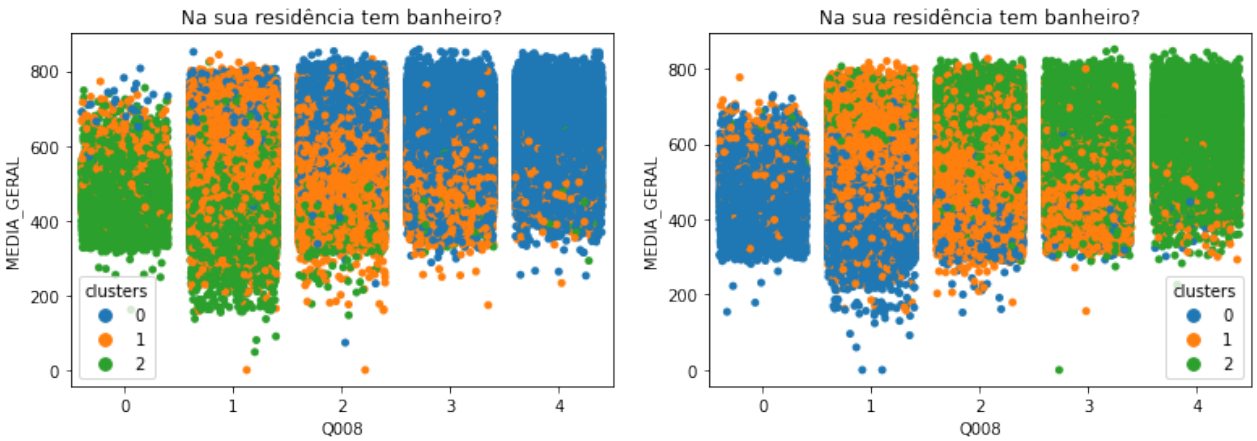
Na resposta 0, os 3 *clusters* são evidentes em ambos *datasets*. Já nas respostas 1, 2 e 3, onde o participante possui um(a) ou mais empregados(as) domésticos(as), o agrupamento 0 (azul), do ENEM de 2018, e 2 (verde), do ENEM de 2019, são mais presentes.

A partir das visualizações, é possível afirmar que os participantes que possuem um(a) ou mais empregados(das) doméstico(a), tendem a pertencer ao *cluster* que possui a média de nota mais alta.

### 4.4.8 Q008

A Figura 26 é referente a questão socioeconômica "Q008", equivalente a "Na sua residencia tem banheiro?". A questão "Q008" possui 4 respostas possíveis. A equivalência em formato de texto, para cada resposta, está representada na Tabela 18.

**Figura 26 – Visualização dos *clusters*, pela média geral e "Q008" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

Os agrupamentos 2 (verde), do *dataset* de 2018, e 0 (azul), do *dataset* de 2019, são mais presentes nas respostas "Não." e "Sim, um.". O *cluster* 1 (laranja), de ambas bases de dados, é mais visível nas respostas 1, 2 e 3. Já os *clusters* 0 (azul), do ENEM de 2018, e 2 (verde), do ENEM de 2019, tem como respostas mais comuns "Sim, três." e "Sim, quatro ou mais.".

**Tabela 18 – Questão "Q008" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

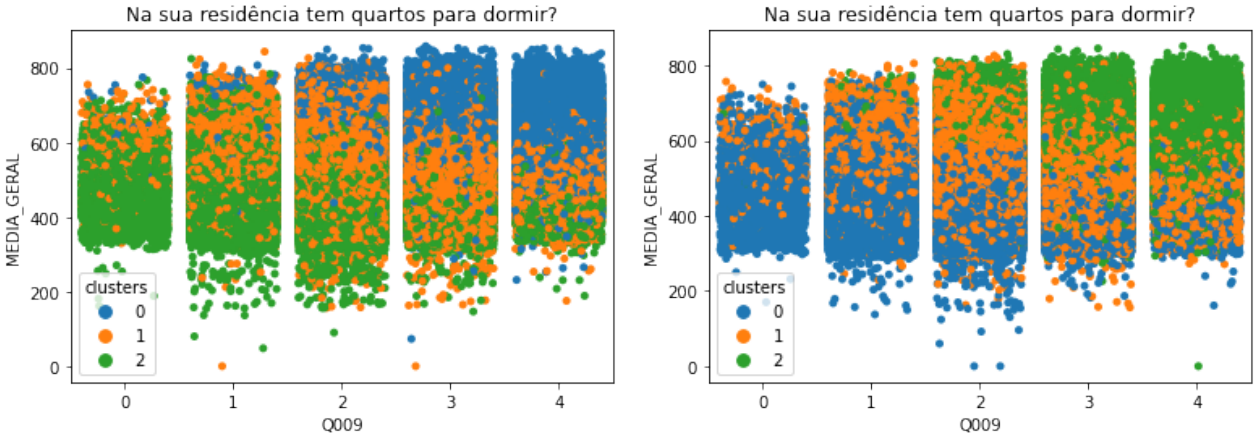
**Fonte: elaborado pelo autor.**

É importante destacar que, os participantes que responderam "Não." tendem a não pertencer aos *clusters* que possuem média de nota mais alta.

### 4.4.9 Q009

A Figura 27 é referente a questão socioeconômica “Q009”, equivalente a “Na sua residencia tem quantos quartos para dormir?”. A questão “Q009” possui 4 respostas possíveis. A equivalência em formato de texto, para cada resposta, está representada na Tabela 19.

**Figura 27 – Visualização dos *clusters*, pela média geral e “Q009” dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

Na Figura 27, os agrupamentos 2 (verde), da base de dados de 2018, e 0 (azul), da base de dados de 2019, estão presentes em todas as respostas, porém, em maior proporção nas 0, 1 e 2. O *cluster* 1 (laranja), das duas edições do ENEM, também aparece em todas as repostas, sendo as 2 e 3 em maior proporção. Os agrupamentos 0 (azul), do ENEM de 2018, e 2 (verde), do ENEM de 2019, aparecem mais nas respostas 3 e 4. Nessa visualização, também é importante destacar que os participantes que responderam a questão socioeconômica com a resposta “Não.”, tendem a não pertencer ao agrupamentos que tiveram médias de notas altas.

**Tabela 19 – Questão “Q009” e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

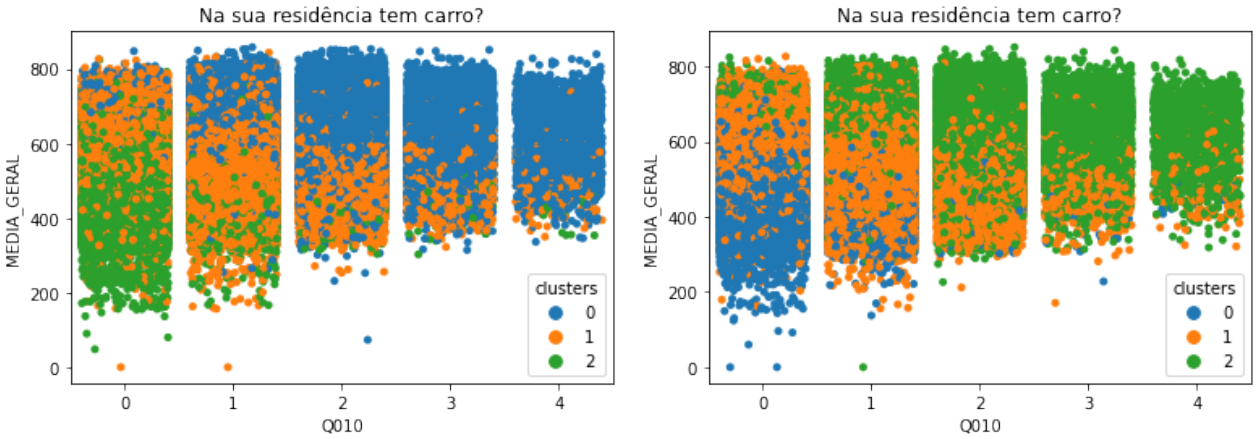
**Fonte: elaborado pelo autor.**

### 4.4.10 Q010

A Figura 28 apresenta os agrupamentos em relação as respostas da questão socioeconômica “Q010”, equivalente a pergunta “Na sua residencia tem carro?”. As respostas,

referentes a essa questão podem ser visualizadas na Tabela 20.

**Figura 28 – Visualização dos *clusters*, pela média geral e "Q010" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

A visualização apresenta um agrupamento semelhante, comparando as edições do ENEM de 2018 e 2019. O agrupamento 0 (verde), da edição de 2018, e 2 (azul), da edição de 2019, aparecem em maior proporção na resposta 0. O *cluster* 1 (laranja), de ambas base de dados, aparece em maior proporção nas respostas 0, 1 e 2. Já os *clusters* 0 (azul), do ENEM de 2018, e 2 (verde), do ENEM de 2019, estão bastante presentes nas respostas 2, 3 e 4. Destaca-se o fato de que, os participantes que responderam a questão com "Sim, três." e "Sim, quatro ou mais." não tiveram a média de notas baixas, em comparação com as outras respostas.

**Tabela 20 – Questão "Q010" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

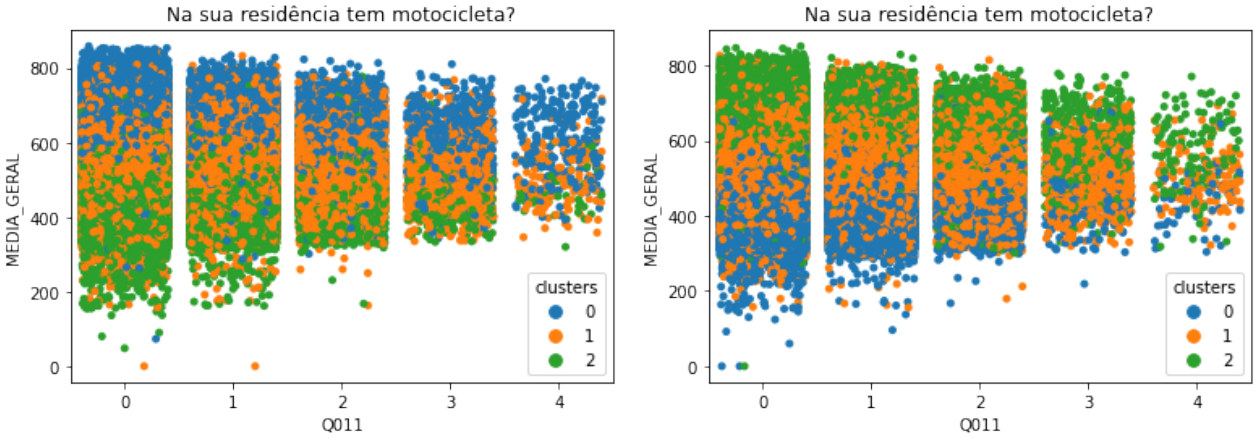
**Fonte: elaborado pelo autor.**

### 4.4.11 Q011

A Figura 29 apresenta os *clusters* em relação as respostas da questão socioeconômicas "Q011", equivalente a pergunta "Na sua residencia tem motocicleta?". As respostas, referente a essa questão podem ser visualizadas na Tabela 21.

Os agrupamentos, nessa visualização, estão divididos de forma mais proporcional que as outras visualizações apresentadas anteriormente. Com exceção da resposta 4, que

**Figura 29 – Visualização dos *clusters*, pela média geral e "Q011" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

aparenta ter a predominância do *cluster* 0 (azul), na edição de 2018, e do *cluster* 2 (verde), na edição de 2019, as outras respostas apresentam os 3 agrupamentos de forma similar.

**Tabela 21 – Questão "Q011" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

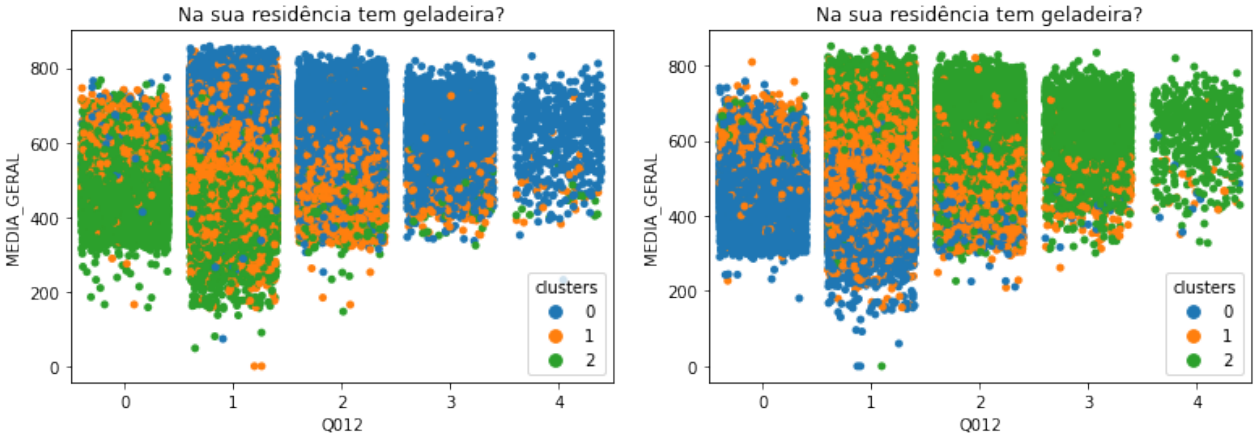
**Fonte: elaborado pelo autor.**

### 4.4.12 Q012

A Figura 30 é referente a questão socioeconômica "Q012", equivalente a "Na sua residencia tem geladeira?". As respostas, referente a essa questão podem ser visualizadas na Tabela 22.

Nessa visualização, os agrupamentos 2 (verde), do ENEM de 2018, e 0 (azul), do ENEM de 2019, aparecem em maior proporção nas respostas 0 e 1. O *cluster* 1 (laranja), de ambas edições, são mais evidentes nas respostas 1 e 2. Já os *clusters* 0 (azul) e 2 (verde), respectivamente das edições de 2018 e 2019, aparecem mais nas repostas 2, 3 e 4. É importante destacar que os participantes que responderam "Não.", não alcançaram médias de notas tao altas como os participantes que responderam a questão com outra resposta.

**Figura 30 – Visualização dos *clusters*, pela média geral e "Q012" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

**Tabela 22 – Questão "Q012" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

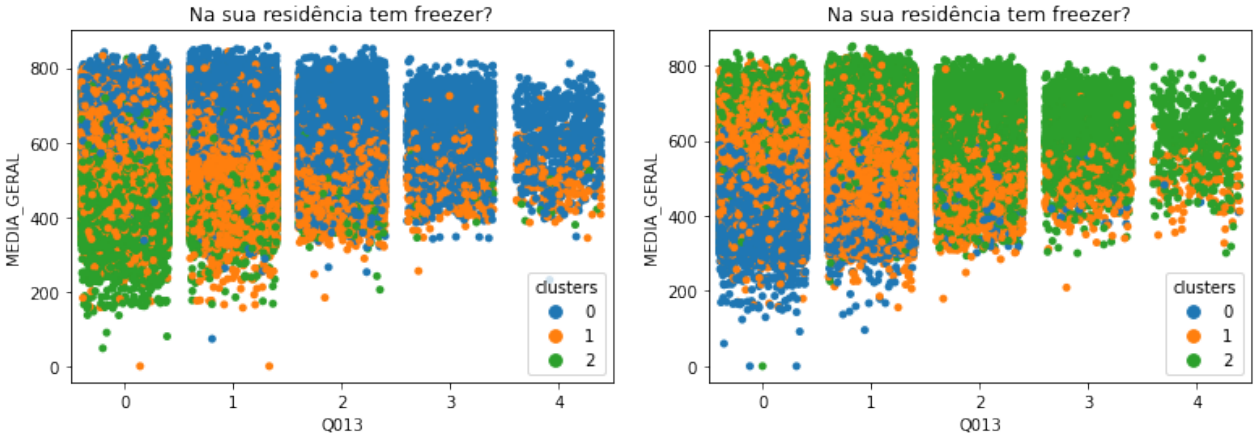
**Fonte: elaborado pelo autor.**

### 4.4.13 Q013

A Figura 31 é referente a questão socioeconômica "Q013", equivalente a "Na sua residencia tem freezer (independente ou segunda porta da geladeira)?". A questão "Q013" possui 5 respostas possíveis. A equivalência em formato de texto, para cada resposta, está representada na Tabela 23.

Os *clusters* 2 (verde), da edição de 2018, e 0 (azul), da edição de 2019, nessa visualização, são mais presentes na resposta 0, aparecendo em menor proporção nas outras. O agrupamento 1 (laranja), de ambas edições, aparecem mais nas respostas 0, 1 e 2. Os *clusters* 0 (azul), do ENEM de 2018, e 2 (verde), do ENEM de 2019, estão mais presentes nas repostas 1, 2 e 3. O número de participantes que responderam essa questão com a resposta "Sim, quatro ou mais." são menores em relação as outras respostas.

**Figura 31 – Visualização dos *clusters*, pela média geral e "Q013" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

**Tabela 23 – Questão "Q013" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

**Fonte: elaborado pelo autor.**

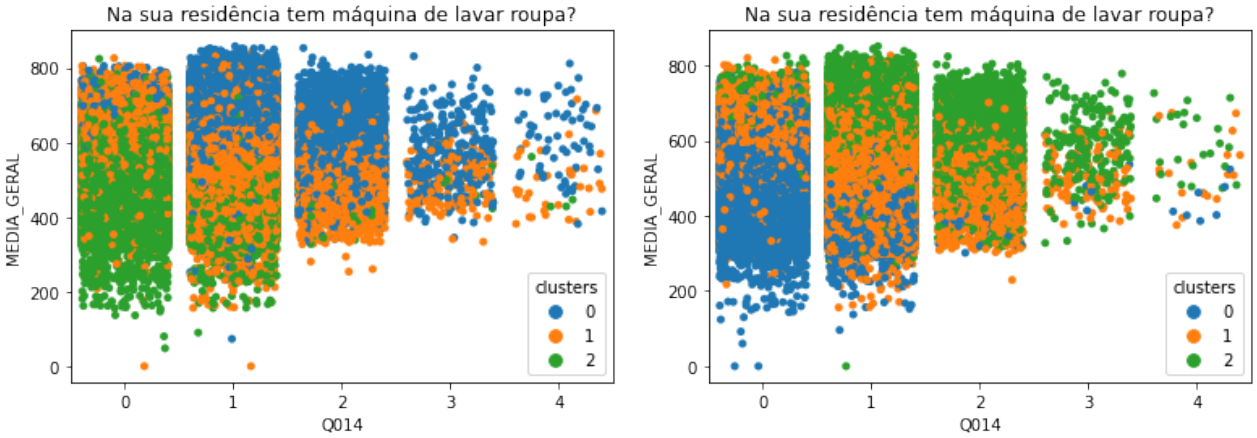
Os participantes que não tem freezer em casa pertencem ao *cluster* que possui a média de notas mais baixas em comparação com os agrupamentos restantes.

### 4.4.14 Q014

A Figura 32 é referente a questão socioeconômica "Q014", equivalente a "Na sua residencia tem maquina de lavar roupa? (o tanquinho NÃO deve ser considerado)". A questão "Q014" possui 5 respostas possíveis. A equivalência em formato de texto, para cada resposta, está representada na Tabela 24.

Nessa visualização, os agrupamentos estão segmentados de forma semelhante a Figura 31.

**Figura 32 – Visualização dos *clusters*, pela média geral e "Q014" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

**Tabela 24 – Questão "Q014" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

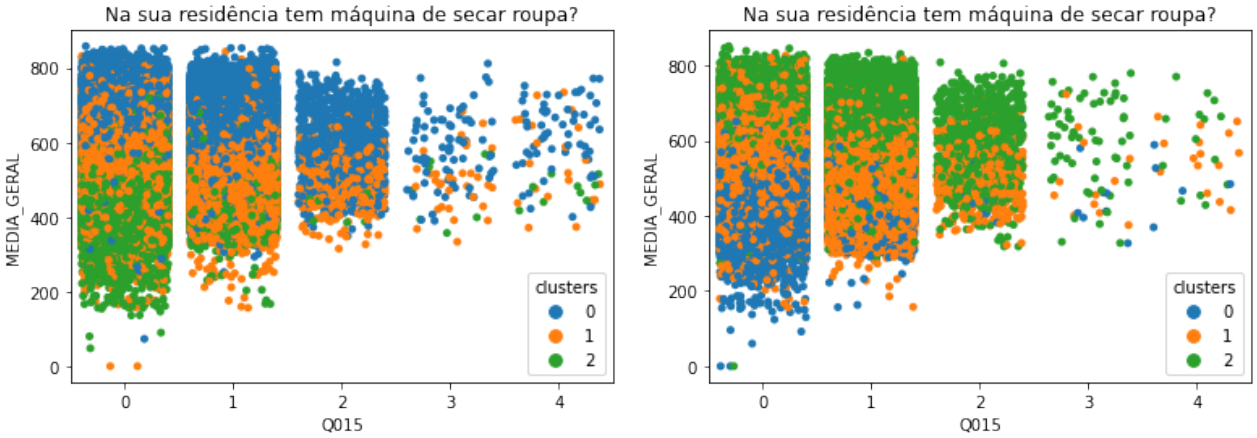
**Fonte: elaborado pelo autor.**

### 4.4.15 Q015

A Figura 33 é referente a questão socioeconômica "Q015", equivalente a "Na sua residencia tem maquina de secar roupa (independente ou em conjunto com a maquina de lavar roupa)?". A questão "Q015" possui 5 respostas possíveis. A equivalência em formato de texto, para cada resposta, está representada na Tabela 25.

Essa visualização apresenta, na resposta 0, os 3 agrupamentos, em ambas edições. Na resposta 1 também, porém a proporção do *cluster* 2 (verde), da edição de 2018 e do *cluster* 0 (azul), da edição de 2019, muito menores. A resposta 2 apresenta, em maior proporção, 2 agrupamentos: o agrupamento 0 (azul) e 1 (laranja), do ENEM de 2018 e 2 (verde) e 1 (laranja), do ENEM de 2019. As respostas 3 e 4 apresentam um número muito menor de participantes.

**Figura 33 – Visualização dos *clusters*, pela média geral e "Q015" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

**Tabela 25 – Questão "Q015" e possíveis respostas**

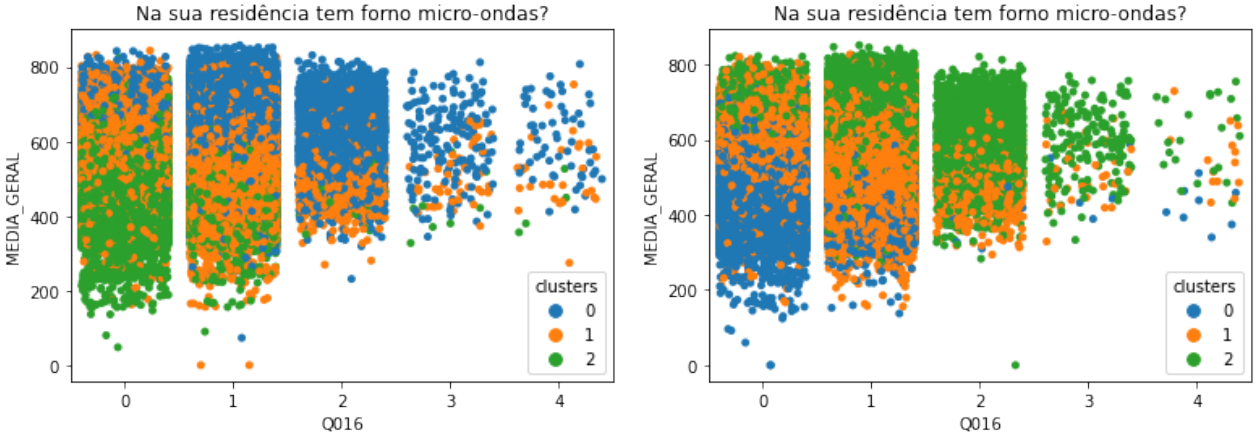
| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

**Fonte: elaborado pelo autor.**

### 4.4.16 Q016

A Figura 34 é referente a questão socioeconômica "Q016", equivalente a "Na sua residencia tem forno micro-ondas?". A questão "Q016" possui 5 respostas possíveis. A equivalência em formato de texto, para cada resposta, está representada na Tabela 26.

**Figura 34 – Visualização dos *clusters*, pela média geral e "Q016" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

Nessa visualização os agrupamentos estão sendo exibidos de forma semelhante a Figura 33.

**Tabela 26 – Questão "Q016" e possíveis respostas**

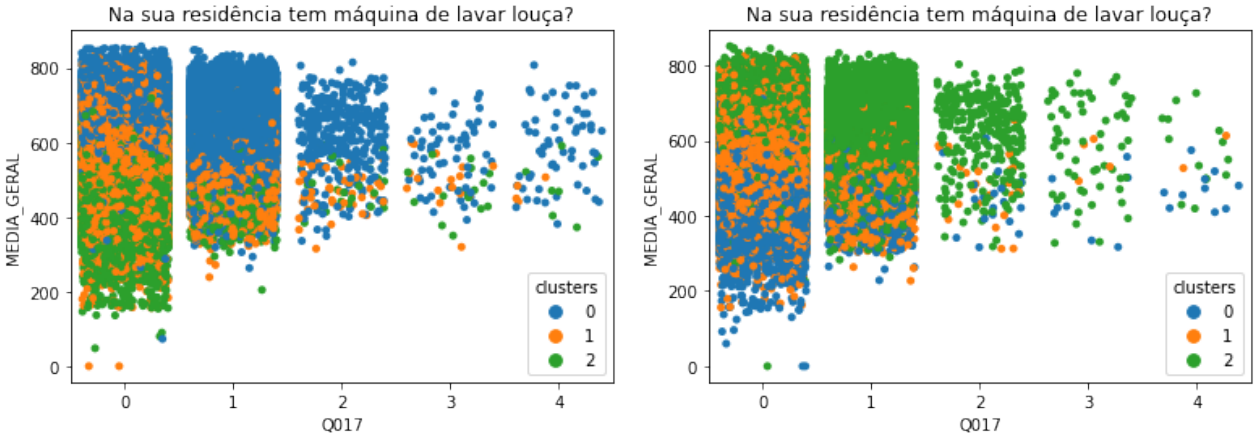
| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

**Fonte: elaborado pelo autor.**

### 4.4.17 Q017

A Figura 35 é referente a questão socioeconômica "Q017", equivalente a "Na sua residencia tem máquina de lavar louça?". A questão "Q017" possui 5 respostas possíveis. A equivalência em formato de texto, para cada resposta, está representada na Tabela 27.

**Figura 35 – Visualização dos *clusters*, pela média geral e "Q017" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

Essa figura apresenta gráficos semelhantes para as edições de 2018 e 2019. A resposta "Não." apresenta os 3 agrupamentos, em ambos *datasets*. Já a resposta 1, é composta, em grande parte, pelos *clusters* 0 (azul), na edição de 2018, e 2 (verde), na edição de 2019.

**Tabela 27 – Questão "Q017" e possíveis respostas**

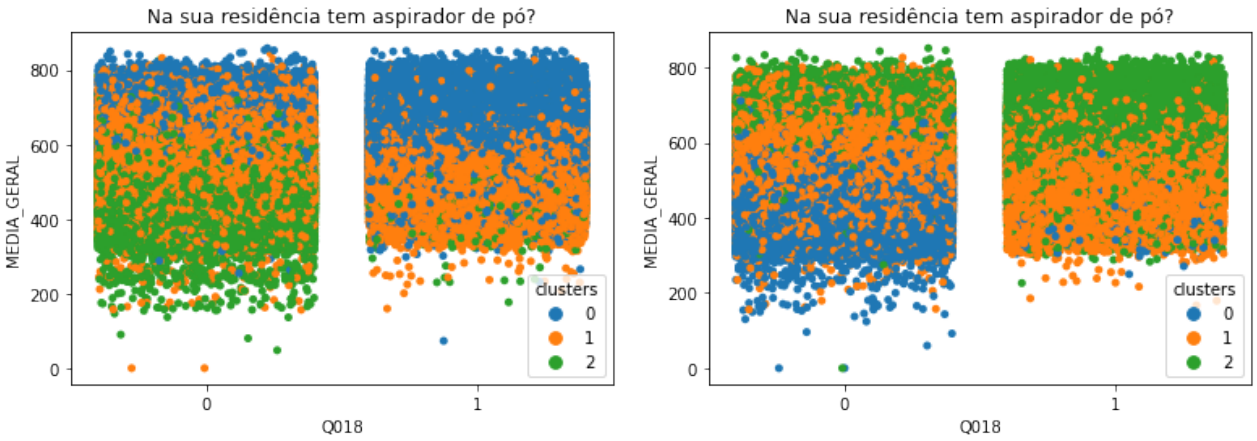
| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

**Fonte: elaborado pelo autor.**

### 4.4.18 Q018

A Figura 36 é referente a questão socioeconômica "Q018", equivalente a "Na sua residencia tem aspirador de pó?". A questão "Q018" possui 5 respostas possíveis. A equivalência em formato de texto, para cada resposta, está representada na Tabela 28.

**Figura 36 – Visualização dos *clusters*, pela média geral e "Q018" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

A partir dessa visualização, é possível identificar que a resposta 0 é composta pelos 3 *clusters*, porém, em menor proporção para o *cluster* 0 (azul), na edição de 2018, e 2 (verde), na edição de 2019. Já, na resposta 1, o agrupamento 1 (laranja), de ambas edições, aparece em grande proporção, além do *cluster* 0 (azul), do ENEM de 2018, e *cluster* 2 (verde), do ENEM de 2019.

**Tabela 28 – Questão "Q018" e possíveis respostas**

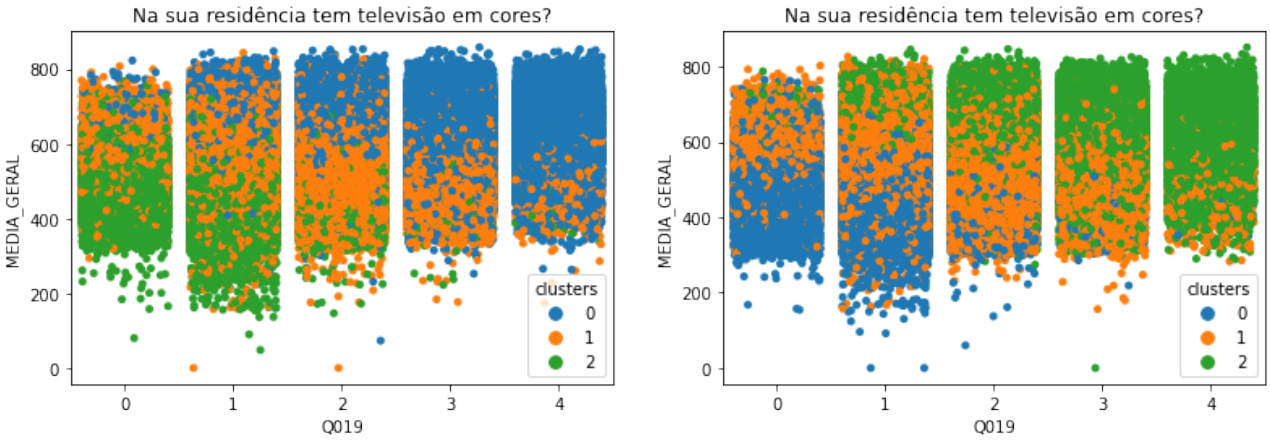
| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim. |

**Fonte: elaborado pelo autor.**

### 4.4.19 Q019

A Figura 37 é referente a questão socioeconômica “Q019”, equivalente a “Na sua residencia tem televisão em cores?”. A questão “Q019” possui 5 respostas possíveis. A equivalência em formato de texto, para cada resposta, está representada na Tabela 29.

**Figura 37 – Visualização dos *clusters*, pela média geral e “Q019” dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

Nessa visualização, os agrupamentos estão sendo exibidos de forma semelhante a Figura 28. Destaca-se o fato de que os participantes que responderam a questão socioeconômica “Q019” com as respostas “Não.” e “Sim, um.”, tendem a pertencer ao *cluster* com média de nota mais baixa que os demais agrupamentos.

**Tabela 29 – Questão “Q019” e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

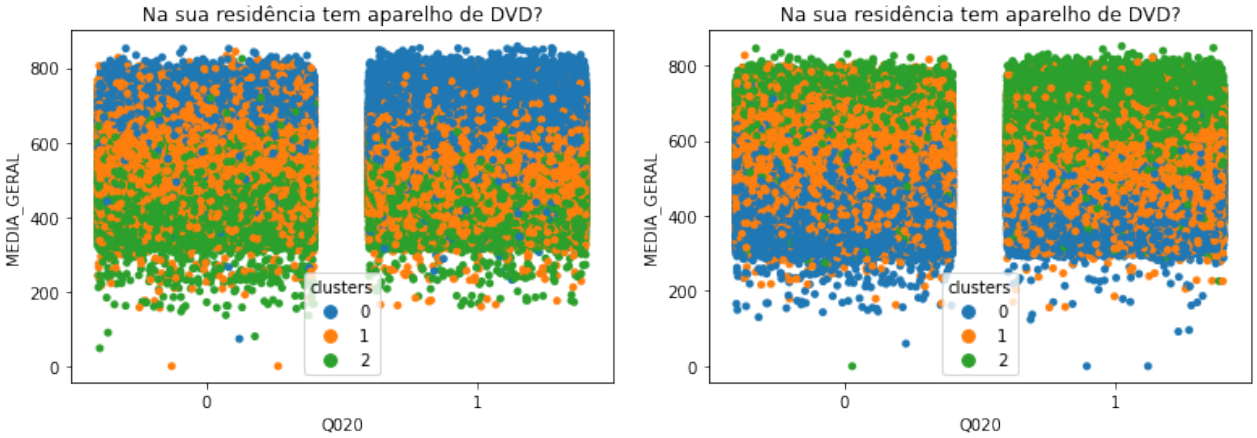
**Fonte: elaborado pelo autor.**

### 4.4.20 Q020

A Figura 38 é referente a pergunta “Em sua casa tem aparelho de DVD?”, equivalente a questão socioeconômica “Q020”. Os agrupamentos, nessa visualização, estão em proporção similar, em ambos *datasets*.

A questão “Q020” possui 2 respostas possíveis. A equivalência em formato de texto, para cada resposta, está representada na Tabela 30.

**Figura 38 – Visualização dos *clusters*, pela média geral e "Q020" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

**Tabela 30 – Questão "Q020" e possíveis respostas**

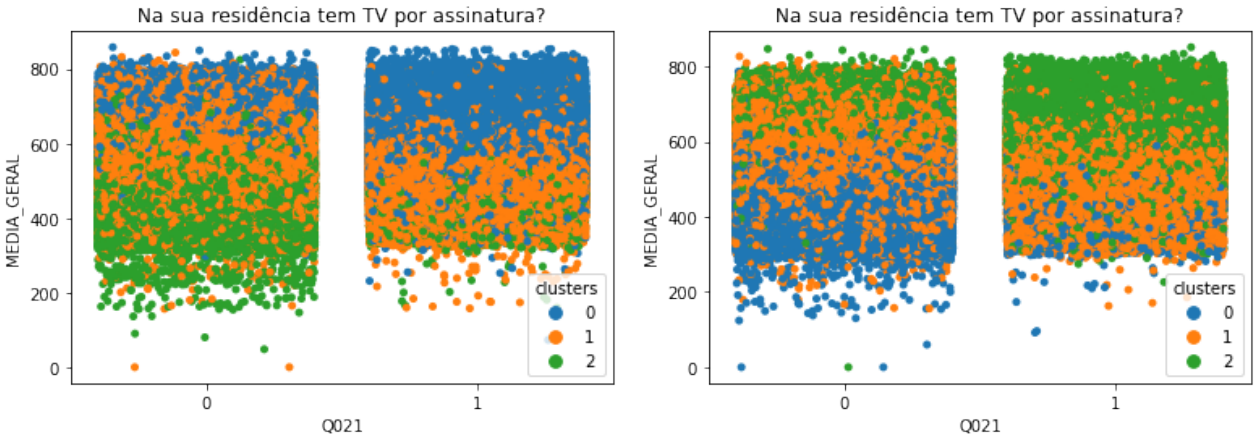
| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim. |

**Fonte: elaborado pelo autor.**

### 4.4.21 Q021

A visualização dos agrupamentos gerados, em relação a questão socioeconômica "Q021", equivalente a pergunta "Na sua residencia tem TV por assinatura?", está na Figura 39. Essa questão possui 2 respostas possíveis. A equivalência em formato de texto, para cada resposta, está representada na Tabela 31.

**Figura 39 – Visualização dos *clusters*, pela média geral e "Q021" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

Nessas visualizações, os 3 agrupamentos, em ambas bases de dados, aparecem nas duas respostas possíveis. Porém, os *clusters* 2 (verde), do ENEM de 2018, e 0 (azul), do

ENEM de 2019, em proporção muito menor na resposta “Sim.”. Destaca-se que, os alunos que pertencem a esse agrupamento, tendem a ter uma média da nota menor que os demais participantes.

**Tabela 31 – Questão “Q021” e possíveis respostas**

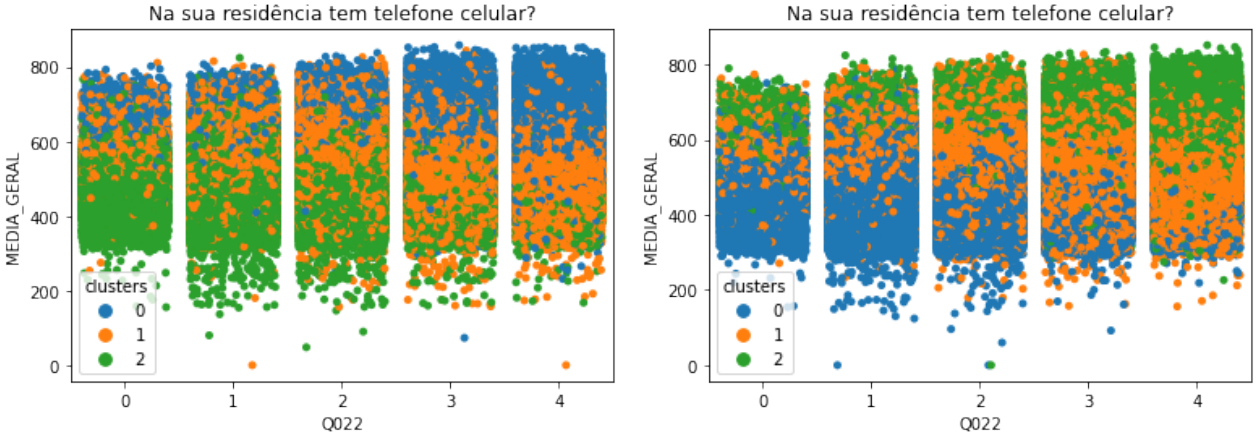
| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim. |

**Fonte: elaborado pelo autor.**

### 4.4.22 Q022

A Figura 40 é referente a pergunta “Em sua residencia tem telefone celular?”, equivalente a questão socioeconômica “Q022”.

**Figura 40 – Visualização dos *clusters*, pela média geral e “Q022” dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

A questão “Q022” possui 5 respostas possíveis. Os valores em formato de texto, para cada resposta, estão representados na Tabela 32.

**Tabela 32 – Questão “Q022” e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim, um. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

**Fonte: elaborado pelo autor.**

Os *clusters* 2 (verde), do ENEM de 2018, e 0 (azul), do ENEM de 2019, aparecem mais, em comparação ao outros *clusters*, nas respostas 0, 1 e 2. O *cluster* 1 (laranja), de

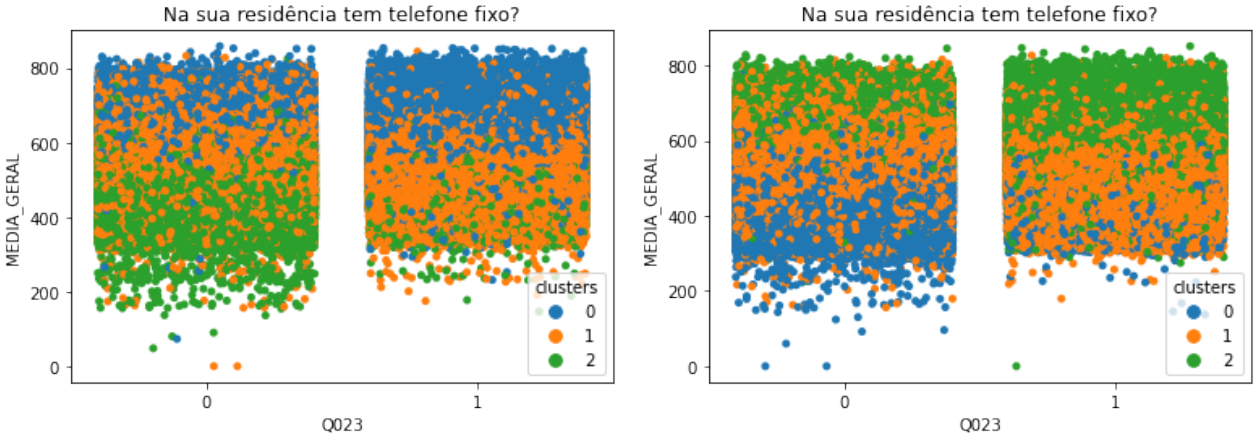
ambos *datasets*, aparecem em maior proporção nas respostas 2, 3 e 4. Já os agrupamentos 0 (azul), da edição de 2018, e 2, da edição de 2019, são mais predominantes nas respostas 3 e 4.

Os participantes que possuem 3 ou mais celulares em suas residências, tendem a pertencer ao *cluster* que possui maior média de nota.

### 4.4.23 Q023

A Figura 41 é referente a pergunta "Em sua residencia tem telefone fixo?", equivalente a questão socioeconômica "Q023".

**Figura 41 – Visualização dos *clusters*, pela média geral e "Q023" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

A questão "Q023" possui 2 respostas possíveis. O valor em formato de texto, para cada resposta, está representado na Tabela 33.

**Tabela 33 – Questão "Q023" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim |

**Fonte: elaborado pelo autor.**

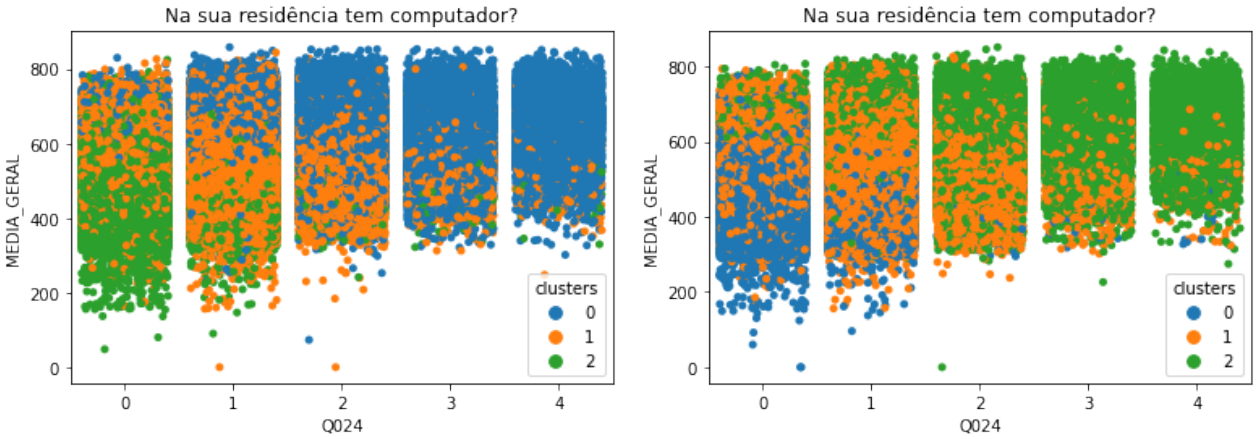
Na resposta 0 ("Não."), é composta pelos 3 *clusters*, porém, tem como maioria o agrupamento 2 (verde), na edição de 2018 e 0 (azul), na edição de 2019. Já a resposta 1 ("Sim."), é composta, na maior parte, pelo *cluster* 1, em ambos *datasets*, além do agrupamento 0 (azul), na edição de 2018, e 2 (verde), na edição de 2019.

### 4.4.24 Q024

A Figura 42 é referente a pergunta "Em sua residencia tem computador?", equivalente a questão socioeconômica "Q024". A questão "Q024" possui 5 respostas possíveis.

O valor em formato de texto, para cada resposta, está representado na Tabela 34.

**Figura 42 – Visualização dos *clusters*, pela média geral e "Q024" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

**Tabela 34 – Questão "Q024" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim. |
| 2 | Sim, dois. |
| 3 | Sim, três. |
| 4 | Sim, quatro ou mais. |

**Fonte: elaborado pelo autor.**

Nessa visualização, os *clusters* 2 (verde), do ENEM de 2018, e 0 (azul), do ENEM de 2019, aparecem mais nas respostas 0 e 1. O agrupamento 1 (laranja), de ambas edições, aparece mais nas respostas 0, 1 e 2. Os *clusters* restantes, 0 (azul), da edição de 2018, e 2, da edição de 2019, são mais presentes nas respostas 2, 3 e 4.

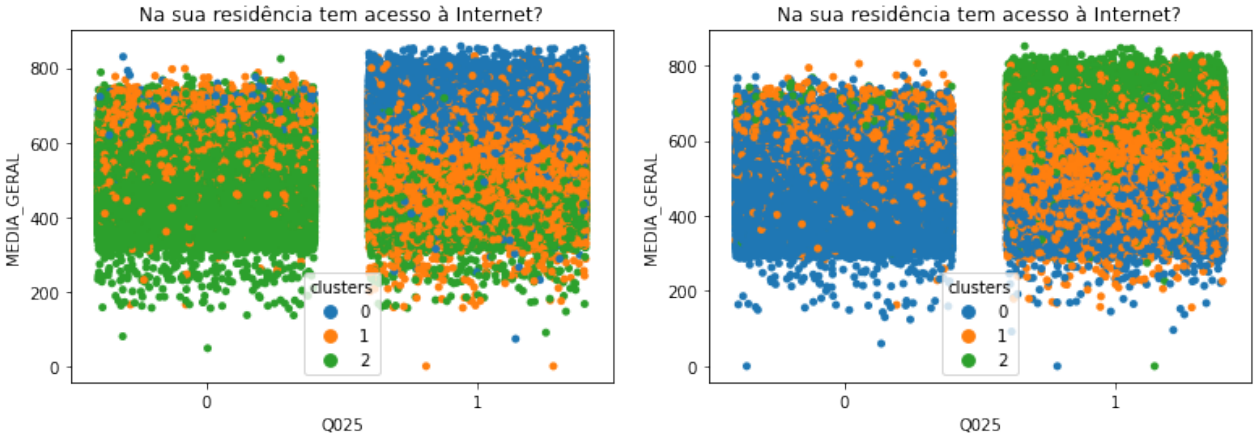
Os participantes que responderam a questão "Q024" com as respostas 3 e 4, tendem a pertencer ao grupo que tem a média de nota mais alta entre os *clusters*.

### 4.4.25 Q025

A Figura 43 é referente a pergunta "Em sua residencia tem acesso a internet?", equivalente a questão socioeconômica "Q025". A questão "Q025" possui 2 respostas possíveis. O valor em formato de texto, para cada resposta, está representado na Tabela 35.

Essa visualização apresenta uma grande diferença de proporção dos *clusters* em cada resposta. A resposta 0 é constituída, em maior parte, pelos agrupamentos 2 (verde), na edição de 2018, e 0 (azul), na edição de 2019. Já, na resposta 1, é possível visualizar os 3 agrupamentos. Destaca-se o fato de que os participantes que responderam a

**Figura 43 – Visualização dos *clusters*, pela média geral e "Q025" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

questão "Q025" com "Não.", tendem a não alcançar notas mais altas como nos demais agrupamentos .

**Tabela 35 – Questão "Q025" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 0 | Não. |
| 1 | Sim. |

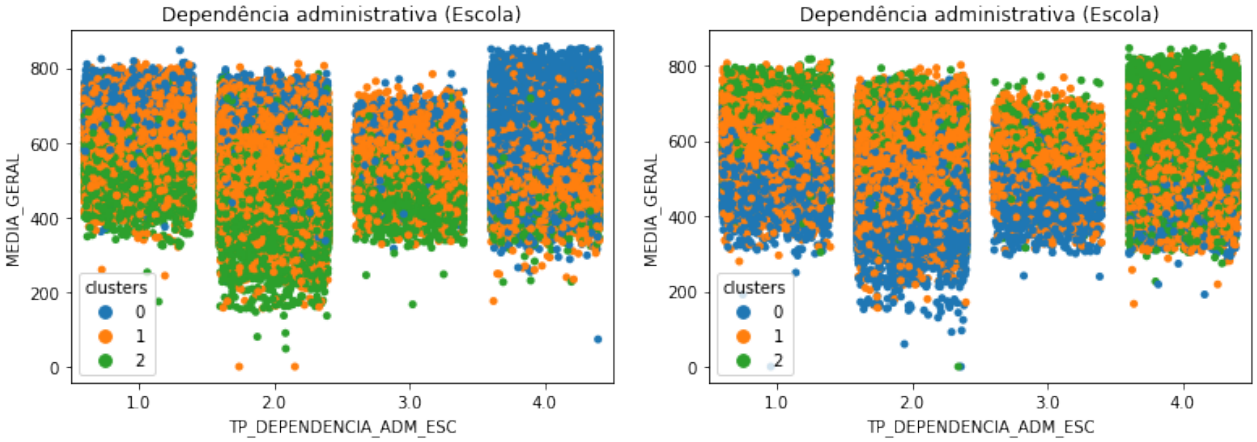
**Fonte: elaborado pelo autor.**

### 4.4.26 Dependência administrativa da escola em relação a nota média

A Figura 44 é referente ao tipo de administração da escola, equivalente a questão "TP_DEPENDENCIA_ADM_ESC". A questão "TP_DEPENDENCIA_ADM_ESC" possui 4 respostas possíveis. O valor em formato de texto, para cada resposta, está representado na Tabela 36.

Nessa visualização, os *clusters* 2 (verde), do ENEM de 2018, e 0 (azul), do ENEM de 2019, aparecem mais nas respostas 1, 2 e 3. O agrupamento 1 (laranja), de ambas edições, aparece em todas as respostas. Os *clusters* restantes, 0 (azul), da edição de 2018, e 2, da edição de 2019, são mais presentes na resposta 4. A partir dessa visualização é possível afirmar que os participantes de escolas municipais, tendem a não alcançar notas médias mais altas, em comparação as escolas federais, estaduais e privadas.

**Figura 44 – Visualização dos *clusters*, pela média geral e "Q025" dos *datasets* de 2018 e 2019.**




**Fonte: elaborado pelo autor.**

**Tabela 36 – Questão "TP_DEPENDENCIA_ADM_ESC" e possíveis respostas**

| Resposta em formato numérico | Significado da resposta |
|---|---|
| 1 | Federal. |
| 2 | Estadual. |
| 3 | Municipal. |
| 4 | Privada. |

**Fonte: elaborado pelo autor.**

Após a análise de cada visualização, as principais características de cada *cluster* ficam evidentes. O agrupamento 2 (verde), do ENEM de 2018, é composto, em geral, por participantes menos privilegiados em questões socioeconômicas. Destaca-se também que, em geral, a média desse agrupamento é menor que nos demais *clusters*. O mesmo acontece com o agrupamento 0 (azul), do ENEM de 2019.

O *cluster* 1 (laranja), é o mesmo em ambas edições do ENEM. As características desse agrupamento variam, pois, diferentes dos demais, está presente, em geral, em todas as respostas. Consequentemente, as notas dos participantes que participam desse agrupamento variam mais que nos outros *clusters*.

Já os agrupamentos 0 (azul), de 2018, e 2 (verde), de 2019, são compostos, em geral, por participantes mais privilegiados em questões socioeconômicas. Os alunos que pertencem a esses *clusters*, tendem a alcançar médias de notas mais altas que os demais agrupamentos.

# 5 CONCLUSÃO

Neste trabalho foi feita uma revisão sistemática sobre aplicação de clusterização nos dados do ENEM. A revisão sistemática teve como objetivo a identificação de algoritmos, técnicas, métodos de validação, linguagens de programação, ferramentas e atributos da base do ENEM, utilizados para clusterização. Após a análise dos resultados, identificou-se que o algoritmo de clusterização mais utilizado pelos trabalhos selecionados, é o *k-means*. Além disso, foi feita uma revisão bibliográfica sobre *machine learning*.

Para a aplicação da clusterização foi utilizado o método de validação WSS, a fim de utilizar o melhor valor possível, dentro dos cenários do trabalho, como parâmetro (número de agrupamentos). O número de *clusters* ideal encontrado, a partir dos dados selecionados, foi 3. Após a aplicação do algoritmo de agrupamento, foram geradas 52 gráficos, metade sobre o ENEM de 2018 e a outra metade sobre o ENEM de 2019.

As características que variam entre os 3 *clusters* gerados, estão relacionadas às condições socioeconômicas dos participantes. Os agrupamentos 0 (azul), do ENEM de 2018, e 2 (verde), do ENEM de 2019, são compostos, em grande maioria, por participantes que possuem mais bens materiais, que possuem maior renda familiar, que estudaram em escolas privadas e que os pais têm maior grau de estudo. Já os participantes dos *clusters* 2 (verde), da edição de 2018, e 0 (azul), da edição de 2019, tendem a ter poucos bens materiais, baixa renda familiar, em comparação aos demais agrupamentos, e pais com menor grau de estudo. O agrupamento 1 (laranja), de ambos *datasets*, é composto por uma maior variedade de participantes, onde as condições financeiras estão mais distribuídas entre as respostas possíveis. Com base na análise dos *clusters*, participantes que pertencem a classes sociais com mais recursos, em média, tendem a alcançar notas mais altas no ENEM. Além disso, participantes que pertencem a classes sociais com menos recursos, em média, tendem a alcançar notas mais baixas no ENEM.

Neste trabalho, foram analisados e comparados os *clusters* das edições do ENEM de 2018 e 2019. Como evolução deste trabalho, podem ser feitas comparações com os agrupamentos de anos anteriores e posteriores, além da possibilidade de geração de *clusters* unificando as edições do ENEM em uma única base de dados. Também é possível agrupar as informações referentes aos eletrodomésticos. Além disso, podem ser aplicados diferentes algoritmos de *clusterização*, a fim de comparar os resultados obtidos. O mesmo algoritmo usado, *k-means*, também pode ser aplicado, utilizando uma *Graphics Processing Unit* (GPU), para comparação do tempo necessário para a *clusterização*, em comparação a uma *Central Processing Unit* (CPU).

# REFERÊNCIAS

BURKOV, A. *The Hundred-Page Machine Learning Book*. [S.l.]: Andriy Burkov, 2019. ISBN 978-1999579500. Citado 2 vezes nas páginas 32 e 34.

CARMO, R. V. do; HECKLER, W. F.; CARVALHO, J. V. de. Uma análise do desempenho dos estudantes do rio grande do sul no enem 2019. *RENOTE*, v. 18, n. 2, p. 378–387, 2020. Citado 2 vezes nas páginas 40 e 42.

CARREIRA, S. da S. et al. Aplicação de data mining na base de dados do processo seletivo do exame nacional do ensino médio-enem 2010. 2012. Citado 7 vezes nas páginas 22, 23, 24, 25, 28, 29 e 30.

GOOGLE RESEARCH. *Colaboratory*. 2021. Disponível em: <https://research.google-.com/colaboratory/intl/pt-PT/faq.html>. Acesso em: 13 julho 2021. Citado na página 39.

HAN, J.; PEI, J.; KAMBER, M. *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*. ebook kindle. Morgan Kaufmann, 2011. 626 p. Disponível em: <https://lead.to/amazon/com-/?op=btla=ptcu=brlkey=B0058NBJ2M>. Citado 7 vezes nas páginas 14, 33, 34, 35, 37, 38 e 47.

HECKLER, W. F. AnÁlise preditiva sobre pacientes do "projeto de extensÃo reabilitaÇÃo pulmonar" da universidade feevale. 2018. Disponível em: <https://tconline-.feevale.br/NOVO/tc/files/0001_4637.pdfpage=30zoom=100,0,665>. Citado na página 19.

HURWITZ, J.; KIRSCH, D. Machine learning for dummies. *IBM Limited Edition*, John Wiley & Sons, Inc, v. 75, 2018. Citado 4 vezes nas páginas 14, 32, 34 e 35.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATíSTICA. *ENEM - Exame Nacional do Ensino Médio*. 2021. Disponível em: <https://ces.ibge.gov.br/base-de-dados/metadados/inep/exame-nacional-do-ensino-medio-enem.html>. Acesso em: 10 março 2021. Citado na página 13.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANíSIO TEIXEIRA. *INEP*. 2021. Disponível em: <http://portal.inep.gov.br/conheca-o-inep>. Acesso em: 10 março 2021. Citado na página 13.

JAIN, A. *Algorithms for clustering data*. Englewood Cliffs, N.J: Prentice Hall, 1988. ISBN 0-13-022278-x. Citado na página 14.

JUNIOR, R. d. A. L. et al. Mineração de dados abertos. 2018. Citado 8 vezes nas páginas 22, 23, 24, 25, 26, 28, 29 e 31.

JUPYTER. *Project Jupyter*. 2021. Disponível em: <https://jupyter.org/>. Acesso em: 13 julho 2021. Citado na página 40.

KITCHENHAM, B. A.; CHARTERS, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. [S.l.], 2007. Disponível em: <https://www.elsevier-.com/_data/promis_misc/525444systematicreviewsguide.pdf>. Citado na página 16.

LAPES. *StArt*. 2021. Disponível em: <http://lapes.dc.ufscar.br/tools/start_tool>. Acesso em: 15 abril 2021. Citado na página 18.

LEONI, R. C.; SAMPAIO, N. A. de S. 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. Citado 9 vezes nas páginas 22, 23, 24, 25, 26, 28, 29, 30 e 31.

LIMA, A. et al. Analysis of enem's attendants between 2012 and 2017 using a clustering approach. *Journal of Information and Data Management*, v. 11, n. 2, 2020. Citado 7 vezes nas páginas 22, 23, 24, 25, 28, 29 e 39.

MOREIRA, N. L. Detecção de atributos que melhor caracterizam perfis de inscritos do enem utilizando redução de dimensionalidade. 2016. Citado 10 vezes nas páginas 14, 22, 23, 24, 25, 27, 28, 29, 30 e 35.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 46.

SANTOS, D. A. et al. Pedagogical recommendation to improve the quality of writing: a case study in a public school. In: IEEE COMPUTER SOCIETY. *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*. [S.l.], 2018. p. 75–76. Citado 8 vezes nas páginas 22, 23, 24, 25, 27, 28, 29 e 30.

SHAI, B.-D. S. S.-S. *Understanding Machine Learning: From Theory to Algorithms*. draft. [S.l.]: CUP, 2014. ISBN 9781107057135. Citado na página 35.

SILVA, V. A. A. da et al. Identificação de desigualdades sociais a partir do desempenho dos alunos do ensino médio no ENEM 2019 utilizando mineração de dados. In: *Anais do XXXI Simpósio Brasileiro de Informática na Educação (SBIE 2020)*. Sociedade Brasileira de Computação, 2020. Disponível em: <https://doi.org/10.5753/cbie.sbie.2020.72>. Citado na página 14.

TAN, P.-N. et al. *Introduction to Data Mining*. 2. ed. [S.l.]: Pearson, 2018. (What's New in Computer Science). ISBN 2017048641,9780133128901,0133128903. Citado 3 vezes nas páginas 33, 36 e 37.

THEOBALD, O. *Machine Learning For Absolute Beginners: A Plain English Introduction*. 2 edition. ed. [S.l.]: Scatterplot Press, 2017. Citado 4 vezes nas páginas 32, 34, 35 e 36.