UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

Clarissa Cordeiro de Souza

Técnica de clusterização aplicada à análise de perfis socioeconômicos de estudantes concluintes de cursos de computação

Recife

Junho de 2022

Clarissa Cordeiro de Souza

Técnica de clusterização aplicada à análise de perfis socioeconômicos de estudantes concluintes de cursos de computação

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE
Departamento de Estatística e Informática
Curso de Bacharelado em Sistemas de Informação

Orientador: Roberta Macêdo Marques Gouveia

Recife
Junho de 2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

S729t Souza, Clarissa Cordeiro de
Técnica de clusterização aplicada à análise de perfis socioeconômicos de estudantes concluintes de cursos de computação / Clarissa Cordeiro de Souza. - 2022.
21 f. : il.

Orientadora: Roberta Macedo Marques Gouveia.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Recife, 2022.

1. clusterização. 2. ENADE. 3. perfis. 4. socioeconômico. 5. K-means. I. Gouveia, Roberta Macedo Marques, orient. II. Título

CDD 004

CLARISSA CORDEIRO DE SOUZA

TÉCNICA DE CLUSTERIZAÇÃO APLICADA À ANÁLISE DE PERFIS SOCIOECONÔMICOS DE ESTUDANTES CONCLUINTES DE CURSOS DE COMPUTAÇÃO

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada em: 03 de Junho de 2022.

BANCA EXAMINADORA

Roberta Macêdo Marques Gouveia (Orientadora)
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Maria da Conceição Moraes Batista
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Técnica de clusterização aplicada à análise de perfis socioeconômicos de estudantes concluintes de cursos de computação

[Clarissa Cordeiro de Souza] [1], **[Roberta Macêdo Marques Gouveia]** [1]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco
Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

clarissa.souza@ufrpe.br, roberta.gouveia@ufrpe.br

Resumo. *As diferentes classes sociais e econômicas de estudantes de cursos de graduação podem impactar no percurso de formação acadêmica e na permanência de tais alunos nas instituições de ensino superior brasileiras. Este trabalho de conclusão de curso aplicou uma técnica de mineração de dados chamada de clusterização K-means aos microdados do Exame Nacional de Desempenho dos Estudantes (ENADE) do ano de 2017, exame aplicado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), com o objetivo de analisar os contextos que separam os concluintes dos diversos cursos de computação, seja bacharelado ou licenciatura, utilizando os dados socioeconômicos. Os resultados apontaram para quatro grandes grupos de estudantes e, a partir das suas análises é possível elencar um perfil de estudante concluinte de computação no ano analisado, visto que os clusters apresentam várias características em comum, tais como: a maioria dos estudantes são do sexo masculino, solteiros, de cor branca, optaram pela modalidade presencial, cursaram o ensino médio em escolas públicas, entre outras. Contudo algumas características foram encontradas em grupos específicos, como por exemplo existe um grupo de concluintes que são de instituições públicas de turno integral.*

Abstract. *The different social and economic classes of undergraduate students can impact the course of academic training and the permanence of such students in Brazilian higher education institutions. This course conclusion work applied a data mining technique called K-means clustering to the microdata of the 2017 National Student Performance Exam (ENADE), an exam applied by the National Institute of Educational Studies and Research Anísio Teixeira (Inep), with the aim of analyzing the contexts that separate graduates from the various computer courses, whether bachelor's or licentiate, using socioeconomic data. The results pointed to four large groups of students and, based on their analysis, it is possible to list a profile of a graduate student of computing in the year analyzed, since the clusters have several characteristics in common, such as: most students are of the sex male, single, white, opted for the face-to-face modality, attended high school in public schools, among others. However, some characteristics were found in specific groups, for example there is a group of graduates who are from full-time public institutions.*

1. Introdução

A área da educação superior no Brasil vem gerando enormes quantidades de dados ao longo dos anos, seja no ingresso, durante a trajetória acadêmica ou no egresso dos estudantes. E quando tem-se dados em abudância é possível analisá-los para a obtenção de novos conhecimentos, que neste caso se originam de informações sobre os estudantes brasileiros. Um exemplo de como esses dados são gerados é a aplicação do Exame Nacional de Desempenho dos Estudantes (ENADE), onde é um requisito obrigatório que permite o estudante obter seu diploma de ensino superior. Para realizar a prova do ENADE, é necessário preencher um formulário com várias perguntas sobre o estudante. É nesse momento que são gerados dados que podem ser usados para futuras análises.

Analisar esses dados manualmente é inviável pois a quantidade de dados é enorme. Para tal, utiliza-se a Mineração de Dados, do inglês Data Mining (DM), que é o processo de explorar e analisar grande volumes de dados em busca de padrões, previsões, erros, associações entre outros (Amaral, 2016). No DM são aplicadas diferenciadas técnicas como análises estatísticas e aprendizagem de máquina para gerar conhecimento.

A análise de dados oriundos de educação chama-se Educational Data Mining (EDM), onde se concentra no desenvolvimento de métodos para explorar os tipos de dados que vêm de um contexto educacional (Romero et al., 2010). Assim, é possível compreender de forma mais eficaz e adequada os alunos, como eles aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem (Baker et al., 2011).

O objetivo desde trabalho foi formar clusters com os dados socioeconômicos dos estudantes concluintes dos cursos superiores da área da computação pelo Brasil no ano de 2017 e descrevê-los, utilizando o algoritmo de clusterização K-means. Esses dados foram obtidos pelos microdados do ENADE. Escolheu-se o ano de 2017 pois são os dados mais recentes, até o início da pesquisa, em que todos os 7 cursos de computação fizeram parte do ENADE, incluindo bacharelados e licenciaturas. A área de computação foi escolhida pois é a área em que a autora está inserida.

O Brasil é um país onde há diferentes classes sociais e econômicas e estas estão presentes na realidade dos estudantes de cursos de graduação. Como essas diferenças podem impactar no percurso de formação acadêmica e na permanência de tais alunos nas instituições de ensino superior brasileiras? Dessa forma, propõe-se encontrar possíveis similaridades e dissimilaridades nos referidos dados educacionais para melhor compreender a realidade dos estudantes que irão concluir um curso de ensino superior e irão ingressar no mercado de trabalho brasileiro.6

O ENADE é uma das avaliações do Sistema Nacional de Avaliação da Educação Superior (SINAES) que é realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) desde o ano de 2004. O ENADE tem como objetivo avaliar o rendimento dos alunos concluintes de cursos superiores no Brasil inteiro, considerando os conteúdos programáticos previstos nas diretrizes curriculares dos cursos, o desenvolvimento de competências e habilidades necessárias ao aprofundamento da formação geral e profissional, e o nível de atualização dos estudantes com relação à realidade brasileira e mundial.

A inscrição do estudante ENADE cabe exclusivamente à Instituição de Educação

Superior (IES). Todo estudante em posição de concluinte de um curso superior é obrigatória a inscrição e realização da avaliação, que para o estudante, começa com o preenchimento do formulário de inscrição juntamente com o questionário do estudante, este que é foco do estudo deste trabalho.

Este trabalho está organizado em 6 seções sendo elas: 1. Introdução, 2. Referencial teórico, 3. Trabalhos relacionados, 4. Metodologia, 5. Resultados, 6. Conclusões.

2. Referencial teórico

O Aprendizado de Máquina, do inglês Machine Learning (ML), é uma área de Inteligência Artificial cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática (Monard and Baranauskas, 2003). O ML é capaz de identificar padrões que dificilmente seriam encontrados manualmente ou mesmo utilizando técnicas triviais de análises de dados (Amaral, 2016). Ou seja, as máquinas são treinadas e aplicadas de forma automatizadas agilizando assim a obtenção de conhecimento.

Durante anos tem-se usado o ML por empresas de vários ramos para a descoberta de informações para a melhoria de seus serviços e produtos. Como por exemplo o sistema de recomendação de músicas e filmes baseados no que os usuários costumam ouvir e assistir em plataformas de streams. Na área de educação, o ML pode ser aplicado para analisar os dados provenientes de Ambientes Virtuais de Aprendizagem (AVAs) identificando as dificuldades de determinados alunos e oferecendo conteúdos de revisão aonde apresentam problemas. Esta identificação também possibilita que o professor saiba quais alunos ele pode atuar mais de perto até mesmo podendo ajudar a evitar evasões.

Do ponto de vista das entradas e natureza de aprendizado, as tarefas e problemas relacionados a Machine Learning podem ser classificados em três principais categorias: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizado por reforço (Schneider, 2016).

A aprendizagem supervisionada exige a supervisão contínua de um analista de dados durante todo o processo de mineração. As máquinas nesta técnica necessitam dos dados de entrada e de saída para serem treinados. Aqui são encontradas as técnicas de classificação e regressão. Enquanto que o aprendizado não supervisionado pressupõe que apenas os dados de entrada são necessários para gerar um saída. Aqui encontram-se as técnicas de associação e de clusterização, técnica esta utilizada neste trabalho, pois tem-se um grande volume de dados e agrupá-los para observar as características que os unem é uma boa e eficiente opção para entender os dados.

O aprendizado por reforço, é um tipo de treinamento baseado na experiência, no qual a máquina deve lidar com o que deu errado anteriormente e encontrar a abordagem correta. Aqui usa-se a lógica de ”tentativa e erro”. Um exemplo bastante conhecido são as recomendações de produtos semelhantes ao carrinho de compras de um cliente. Se o cliente não clica nas sugestões, a máquina entende que os resultados não foram bons e tenta de novo com a exibição de outras sugestões.

2.1. Clusterização de dados

Os dados ao nosso redor foram divididos em grupos, como por exemplo no Português as palavras são divididas em substantivos, adjetivos, advérbios. Na matemática, temos os

grupos dos números decimais, números ordinais, números cardinais. Esses grupos são gerados a partir das características que esses dados possuem em comum, que os fazem ser do mesmo tipo.

Clusterização é uma classficação não-supervisionada de padrões em clusters (grupos) (Jain et al., 1999). Com essa técnica são gerados clusters automáticos dos dados levando em consideração as suas similaridades. O objetivo do agrupamento de dados, também conhecido como análise de agrupamento, é descobrir o(s) agrupamento(s) de um conjunto de padrões, pontos ou objetos (Jain, 2010). Segundo Jain (2010), o agrupamento de dados tem sido usado para os três propósitos principais a seguir.

- Estrutura subjacente: para obter informações sobre dados, gerar hipóteses, detectar anomalias e identificar características salientes.
- Classificação natural: para identificar o grau de semelhança entre formas ou organismos (relação filogenética).
- Compressão: como método para organizar os dados e resumi-los por meio de protótipos de cluster.

2.1.1. Algoritmo de Clusterização K-means

K-means é o algoritmo mais simples e mais famoso quando se trata de criar grupos de dados. Segundo Park et al. (2013), o K-means forma os clusters de objetos com base na distância Euclidiana entre eles. A Figura 1 apresenta uma ilustração do algoritmo K-means.

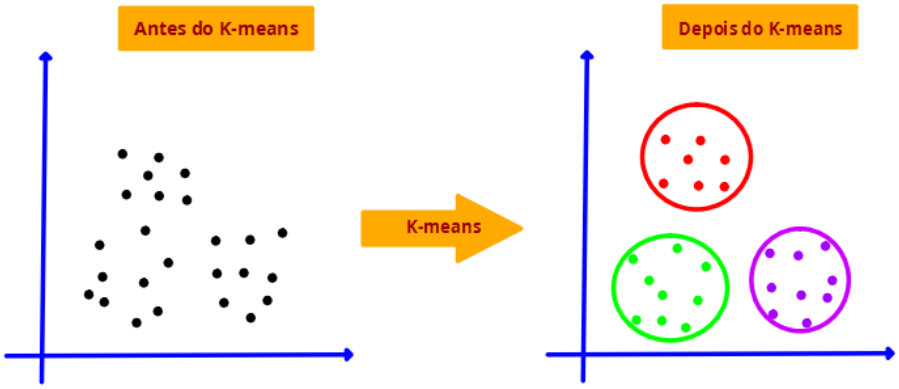

Figura 1. Representação do algoritmo K-means.
Fonte: A autora (2022)

O K-means é baseado no conceito de centroides como protótipos representativos de grupos, onde um centroide representa o centro do grupo e é calculado como a média de todos os objetos do grupo. Quando os centróides dos grupos param de mudar, ou quando um número predeterminado de iterações é concluído, o processo iterativo termina (Chicon and Telocken, 2021). O passo a passo mais básico para o algoritmo K-means é:

Algorithm 1: Funcionamento do K-means

Informar o valor de k (número de clusters que serão formados);
Inicializa k centroides aleatoriamente;
Repete a atribuição de cada ponto ao seu centroide mais próximo;
Repete o calculo do novo centroide (média) de cada cluster;
O algoritmo para suas repetições até que a posição dos centroides não mudem.

O algoritmo K-means depende do valor de k, onde k se refere ao número de clusters que serão formados, que sempre precisa ser especificado para realizar qualquer análise de agrupamento. Agrupar com diferentes valores de k eventualmente produzirá resultados diferentes (Ahmed et al., 2020).

Não é recomendado escolher o valor de k manualmente. Uma abordagem que é utilizada para a descoberta do melhor valor de k para o uso no K-means é o método Elbow (cotovelo, do inglês), onde no gráfico gerado é formado uma ponta parecida com um cotovelo é o melhor valor para k. Na Figura 2 é possível ver um exemplo gráfico do método Elbow, indicando k = 3 como o valor ideal, onde forma-se a ponta (cotovelo).

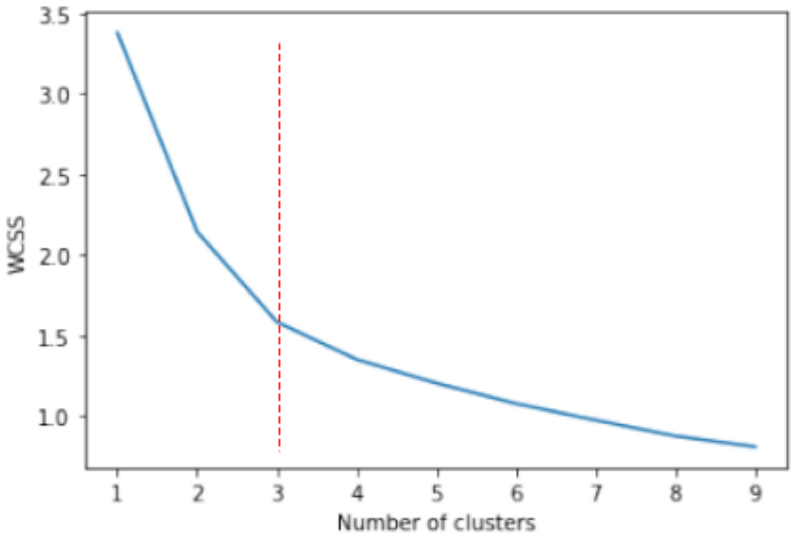

Figura 2. Representação gráfica do método Elbow.
Fonte: A autora (2022)

No método Elbow, o melhor valor de clusters é obtido a partir do valor de Sum of Square Error (SSE). Segundo Nainggolan et al. (2019), os estágios para determinar o melhor valor de k usando o método Elbow são:

1. Iniciar o valor de k (geralmente começa com k=2);
2. Aumentar o valor de k;
3. Calcular o resultado de SSE de cada valor de k;
4. Analisar o resultado do SSE do valor de k que diminuiu drasticamente;
5. Localizar e definir o valor de k que forma um cotovelo.

Segundo Kwedlo (2011), o SSE é definido como:

$$SSE(X, \prod) = \sum_{i=1}^{K} \sum_{x_j \epsilon C_i} \|x_j - m_i\|^2, \quad (1)$$

onde:
X = $\{x_1, ..., x_i, ..., x_N\}$, onde $x_i \epsilon \Re^M$, onde o N representa o números de pontos
$\prod = \{C_1, C_2, ..., C_K\}$, onde $\forall_{i \neq j} C_i \cap C_j = \varnothing, \cup_{i=1}^{K} C_i = X, \forall_i C_i \neq \varnothing$
$\|.\|$ = distância Euclidiana
m_i = centroide do cluster C_i

2.1.2. Conversão de variáveis categóricas em numéricas

Como a maioria dos algoritmos de ML entende inteiros "não texto", converter variáveis categóricas em numéricas é um passo necessário, para que um algoritmo de ML seja capaz de calcular a correlação entre eles e fazer as previsões corretas (Al-Shehari and Alsowail, 2021). O algoritmo K-means possui essa particularidade, ele é aplicado apenas em dados númericos. Para a conversão, é comumente usado 2 métodos: Label Encoder e One-hot Encoder.

O Label Enconder transforma os dados categóricos em inteiros exclusivos, exemplo: Azul = 0, Verde = 1, Preto = 2. Contudo, um algoritmo de ML pode entender que existe uma ordem nos dados e considerar que Preto = 2 tem peso maior do que Azul = 0 e isso causar influência errada nos resultados. Há situações em que é interessante a ordem, e é por isso existe o label encoder. Exemplo: atributo série do estudante. O One-Hot Enconder transforma os dados categóricos em valores binários (0 ou 1), onde o número 1 é atribuído quando há a presença de um valor e o número 0 quando não há. Exemplo: o atributo tipo de sexo do estudante com os valores M e F. Neste caso, cria-se uma nova variável especificando se o registro pertence ao sexo M (1) ou não (0). Dessa forma os valores binários não sofrem com ordenação, sendo uma boa opção para aplicar em dados categóricos e foi a abordagem usada neste estudo.

3. Trabalhos Relacionados

Esta seção apresenta estudos relacionados a este trabalho, observando os cenários de mineração de dados educacionais que abordam clusterização.

A pesquisa realizada por Vista et al. (2017), teve como objetivo aplicar a técnica de clusterização Hierarchical Clustering (Agrupamento Hierárquico) nos microdados do ENADE do ano de 2014 para extrair informações referentes ao desempenho dos acadêmicos do curso de Ciência da Computação no Rio Grande do Sul. Para a aplicação do algoritmo, utilizaram o software estatístico R. O objetivo da pesquisa foi poder descrever as instituições que apresentaram bons e ruins desempenhos acadêmicos, onde essas informações possam auxiliar na tomada de decisão que resultem numa melhoria no ensino de nível superior brasileiro. Esta pesquisa chegou a um resultado de formação de 4 grupos formados pelas instituições de ensino superior do Rio Grande do Sul com foco no conceito do ENADE.

A pesquisa realizada por de Almeida Lima et al. (2019), teve como objetivo desenvolver uma abordagem híbrida (HA) que usa clusterização e regressão para prever o

desempenho acadêmico dos alunos no contexto da mineração de dados. Para isso, utilizaram o conjunto de dados do ENADE dos anos de 2014 e 2017. As técnicas usadas foram o algoritmo de clusterização K-means e a regressão robusta. Conseguiram verificar que a presença de monitores e tutores para auxiliar os alunos no processo de aprendizagem é fundamental, influencia o desempenho dos alunos nos dois anos estudados. A implementação da HA mostrou-se satisfatória ao apresentar bons resultados em testes de hipóteses.

A pesquisa realizada por da Silva et al. (2019), teve como objetivo analisar os perfis dos docentes em instituições de ensino superior no estado do Espírito Santo, aplicando técnicas de mineração de dados para extrair informações relevantes da base de dados do censo superior de 2016, disponibilizados pelo INEP. Para o estudo, utilizou-se o software Weka 3.8.3. foi utilizado para este estudo a classificação dos dados com a clusterização do "Simple K-means"e a análise de árvore de decisão usando o algoritmo J48. A pesquisa resultou na criação de 7 clusters, além de detalhar algumas ramificações contidas na árvore de decisão criada. Uma das conclusões que pode ser citada como resultado deste estudo é que as instituições Públicas Federais, que em sua maioria localizam-se na capital do estado enquadram-se no perfil de índice Geral de Cursos nota D nota próxima da máxima aferida pelo censo INEP 2016, mostrando que há uma excelência no ensino superior público do estado do Espírito Santo.

Diferenciando dos trabalhos citados acima, este trabalho tem foco em analisar os dados socioeconômicos de concluintes de cursos de computação no Brasil inteiro, não apenas em estados específicos, aplicando a técnica de clusterização K-means nos dados socioeconômicos do ENADE de 2017, sejam eles numéricos ou categóricos, para compressão dos dados possibilitando assim a definição dos perfis desses concluintes.

4. Metodologia

A metodologia utilizada para o desenvolvimento deste trabalho foi dividida em 5 etapas: a primeira etapa foi a obtenção dos dados, a segunda etapa foi o pré-processamento dos dados com a finalidade de preparar, organizar e estruturar os dados para a implementação do algoritmo, a terceira etapa foi a aplicação de uma análise exploratória simples a fim de entender os dados, a quarta etapa foi a implementação do algoritmo de clusterização K-means e a última etapa foi direcionada a análise e descrição dos grupos gerados. O conjunto de atividades realizadas para o desenvolvimento deste estudo pode ser conferida na Figura 3.

Seguindo o diagrama descrito na Figura 3, o desenvolvimento deste trabalho se iniciou com a obtenção dos microdados referentes ao ano de 2017 no portal do Inep. Após sua obtenção, fez-se uma leitura de seu dicionário de variáveis (onde vem junto com os microdados do portal do Inep) para entender a semântica e de que tipo de dados este conjunto se tratava. A base de dados original possui 150 atributos com 537.437 registros com os dados dos estudantes. O foco desde trabalho foram os dados referentes ao questionário socioeconômico que os estudantes precisam preencher para se inscreverem no ENADE. As notas da prova e alguns dados referentes ao curso também foram usados, como região, nome e turno dos cursos.

A segunda etapa foi o pré-processamento dos dados, onde ocorreram a limpeza, transformação e adequação dos dados. Todas as análises e adequações feitas neste tra-

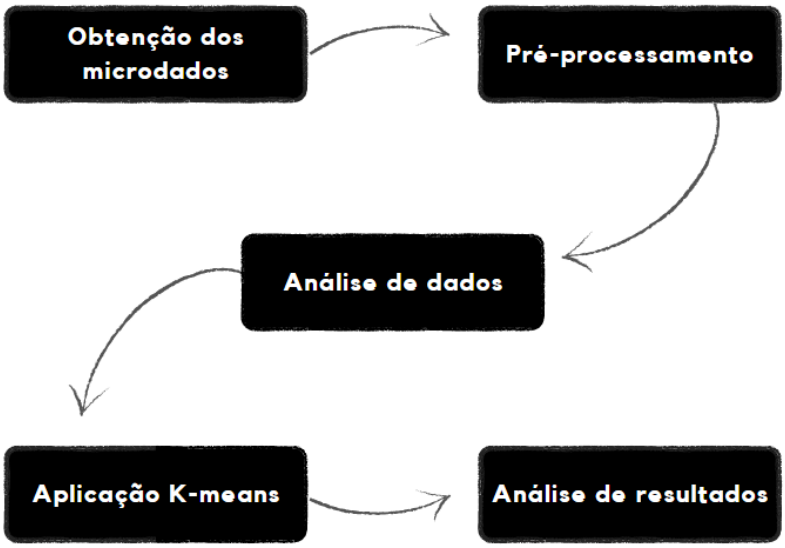

Figura 3. Diagrama das etapas realizadas nesta pesquisa.
Fonte: A autora (2022)

balho foram realizadas na linguagem de programação Python versão 3.7.13, utilizando a ferramenta gratuita de programação em nuvem Google Colaboratory. As bibliotecas utilizadas foram: Pandas, Pandas Profiling, Scikit-Learn, Numpy, Seaborn e Matplotlib.

Em seguida ocorreram exclusões de atributos que foram julgados desnecessários para este estudo, como por exemplo os atributos referentes a composição da nota, atributos referentes ao tipo de situação das questões da parte discursiva e o atributo que se refere ao ano de realização do ENADE. Onde apenas ficaram os que se julgou fundamentais para entender quais são os perfis dos concluintes de cursos de computação. Após a exclusão dos atributos, foi feito o preenchimento de dados faltantes (NaN) com moda para variáveis nominais (referente a dados tipo texto ou ordinal) e com mediana para dados numéricos. Dessa forma os registros com dados faltantes não são perdidos (por exclusão) e a base não sofre com alterações, já que a moda e mediana são os dados estatísticos mais prováveis de existirem na base.

Em seguida foi feita uma junção na coluna tipo de Instituição, onde juntou-se as Instituições de Ensino Superior (IES) estaduais e municipais em uma única categoria para diminuição da granularidade desta coluna. Também houve a exclusão dos registros referentes ao tipo de IES denominada Outras, pois só tinham apenas 2038 registros, equivalente a 0,38% dos dados, e por ser um número pequeno em relação ao total de registros da base de dados, foi considerado irrelevante para este estudo.

Após isso, realizou-se a renomeação dos atributos para nomes que se referem ao que armazenam. Exemplo: o atributo que se chamava Q1, referente a primeira pergunta do questionário do estudante Qual seu estado civil?, passou a ser chamar Estado_Civil. Esse processo ocorreu com todos os atributos restantes.

Continuando o pré-processamento, foi feita a renomeação das categorias dos atributos onde era necessário para entendimento dos dados sem a necessidade de consultar no dicionário dos microdados. Exemplo: o atributo Estado_Civil tinha como possibilidade

de resposta A, B, C e D, passou a ser Solteiro, Casado, Viuvo, Separado, respectivamente, conforme dicionário de dados disponibilizado pelo Inep.

Ainda na etapa de pré-processamento, também foi criado um novo atributo, Tempo_Transcorrido, expressado em anos, onde contém o resultado da subtração dos atributos "ano de ingresso na instituição superior"e "ano de conclusão no ensino médio". Após a criação, os atributos usados para a subtração foram excluídos. O objetivo da criação deste atributo foi diminuir o número de dimensões da base.

Após os passos informados anteriormente, foi feita a seleção dos dados de acordo apenas com os cursos referentes à computação. Foi uma decisão tomada para este estudo: avaliar somente os cursos da área de computação no Brasil, pois é a área do curso superior da autora em que este estudo foi elaborado. Os nomes dos 7 cursos podem ser conferidos na Tabela 1.

Tecnologia em Análise e Desenvolvimento de Sistemas
Tecnologia em Redes de Computadores
Engenharia da Computação
Ciência da Computação (Bacharelado)
Ciência da computação (Licenciatura)
Sistemas de Informação
Tecnologia em Gestão da Tecnologia da Informação

Tabela 1. Listas de cursos da área de computação.
Fonte: A autora (2022)

Antes que fosse feita qualquer implementação de algoritmo de clusterização ou conversões dos dados, fez-se uma análise exploratória simples na etapa 3 com o objetivo de ter os principais estimadores dos atributos (tais como moda, min, max, média, mediana, desvio padrão, etc.), sendo possível obter uma visão geral de suas distribuições e entender como os dados se comportam. É importante ressaltar que apenas alguns atributos foram escolhidos para exemplificar a análise neste documento. Para os atributos numéricos referentes a nota geral, nota de componente específico, nota de formação geral, idade e tempo transcorrido algumas informações podem ser conferidas nas Figuras 4 e 5. Já para os atributos categóricos, o resultado da análise pode ser conferido na Figura 6.

Como pode ser conferido na Figura 4 a base apresenta, para idade, valor mínimo de 17, máximo de 84 e mediana de 25. Os valores da média e desvio padrão são 27,16 e 6,33, respectivamente. Já para tempo transcorrido, os valores são: -1 para mínimo, 53 para máximo, 3 para mediana, 4,96 para a média e 5,39 para o desvio padrão. Como este atributo possui outliers que possivelmente são anomálias, este atributo foi desconsiderado para a formação dos clusters.

Na Figura 5 é possível observar que, para nota geral, os valores são: 0 para mínimo, 96,20 para máximo e 43,52 para a mediana. A média é 42,83 e o desvio padrão é 12,09. Enquanto que para a nota de formação geral, nota que se refere a parte da prova com questões de Língua Portuguesa, os valores são: 0 para mínimo, 98,80 para máximo e 51,42 para a mediana. Os valores achados para a média foi 50,78 e 15,81 para o desvio padrão. E, por fim, os valores para a nota do componente específico, nota que se re-

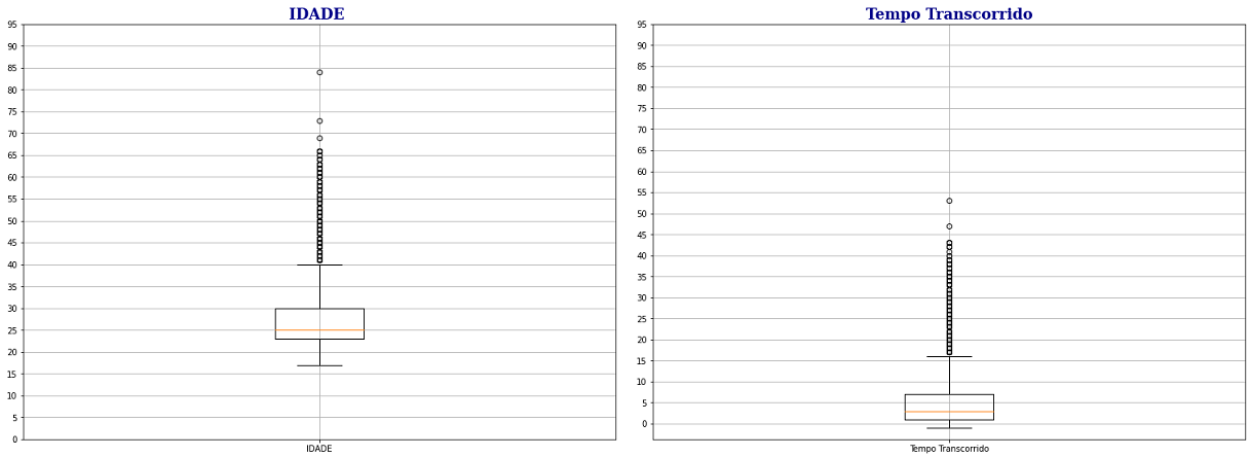

Figura 4. Gráficos de caixa dos atributos idade e tempo transcorrido antes da aplicação do K-means.

Fonte: A autora (2022)

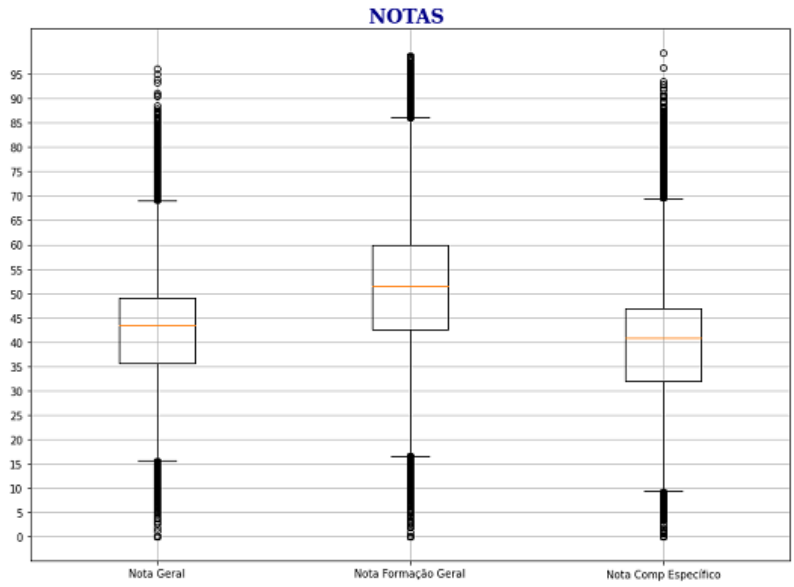

Figura 5. Gráficos de caixa dos atributos notas antes da aplicação do K-means.

Fonte: A autora (2022)

fere a parte do contéudo específico do curso do concluinte, foram: 0 para mínimo, 99,30 máximo e 40,87 para a mediana. Apresentando 40,17 para a média e 13,41 para o desvio padrão.

Como mostra na Figura 6, a raça/cor que mais possui registros é a Branca, com 62,94%, seguida da raça/cor Parda com 24,93% dos registros. A categoria de instituição que possui mais registros é a Privada, com 66,28% dos registros. Os cursos que mais registraram concluintes foram os de Sistemas de Informação, Tecnologia em Análise e Desenvolvimento de Sistemas e Bacharelado em Ciência da Computação, com 29,46%, 24,64% e 19,87% respectivamente. As regiões que detiveram os maiores números de concluintes foram a região Sudeste com 52,95%, a Sul com 17,43% e a Nordeste com 15,84%. 73,56% dos concluintes do ano de 2017 afirmaram que alguém de suas famílias

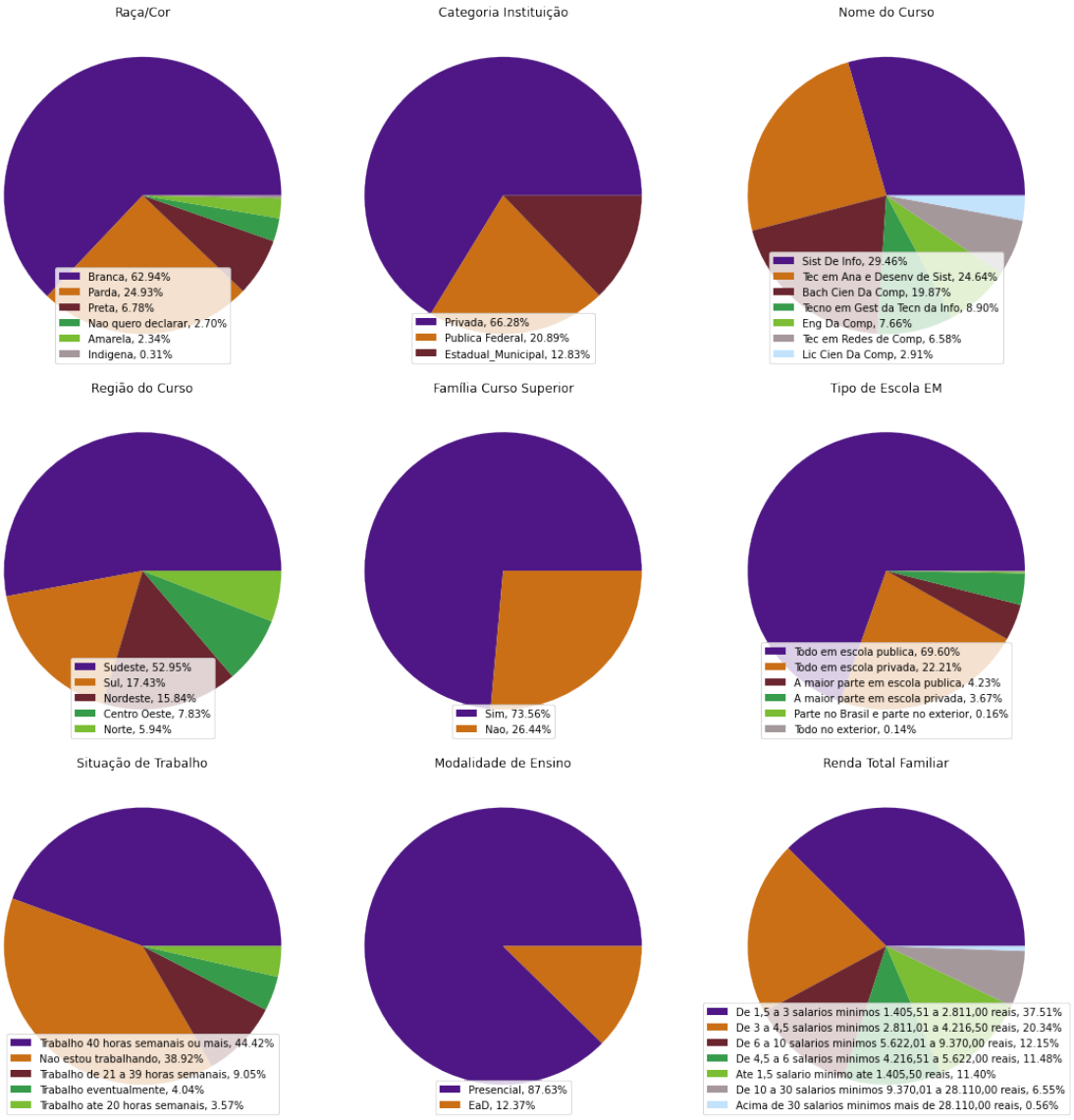

Figura 6. Resultado da análise exploratória simples nos atributos categóricos antes da aplicação do K-means.

Fonte: A autora (2022)

concluiu um curso de ensino superior. O tipo de escola no ensino médio que mais teve registro foi o tipo Pública, com 69,60% contra 22,21% do tipo Privada. 44,42% dos concluintes afirmaram estarem trabalhando 40 horas semanais ou mais enquanto 38,92% afirmaram não estarem trabalhando. A modalidade de ensino que mais obteve registro foi a Presencial com 87,63% contra 12,37% da modalidade de ensino a distância. 37,51% dos concluintes afirmaram ter renda familiar de 1,5 a 3 salários mínimos (R$ 1.405,51 a 2.811,00 reais) e 20,34% afimaram ter renda familiar de 3 a 4,5 salários mínimos (R$ 2.811,00 a 4.216,50 reais.

Após isso feito na etapa de pré-processamento, com os dados devidamente tratados prosseguiu-se com a implementação do algoritmo de clusterização K-means. A base de dados resultante contém 38 atributos e 50.802 registros. Pode-se conferir a seleção dos

atributos na Tabela 2. As descrições e categorias de cada atributo podem ser conferidas no dicionário de dados disponível no link: https://bityli.com/aDXAVM.

categoria_instituicao	turno_graduacao	escolarizacao_mae
nome_curso	nt_geral	renda_total_familiar
modalidade_ensino	nt_formacao_geral	situacao_financeira
uf_curso	nt_comp_espec	situacao_trabalho
regiao_curso	estado_civil	bolsa_estudos
nu_idade	cor_raca	auxilio_permanencia
tp_sexo	escolarizacao_pai	bolsa_academica
organizacao_dedicacao	atividades_praticas	conhecimentos_atualizados
infraestrutura_salas	ambientes_equipamentos	biblioteca_fisica
infraestrutura_instituicao	temp_transcorrido	biblioteca_virtual
politica_acao_afirmativa	tipo_escola_em	familia_curso_superior
horas_semanas_estudos	motivo_escolha_curso	motivo_escolha_instituicao
relacao_professor_aluno	dominio_professores	

Tabela 2. Atributos Finais.
Fonte: A autora (2022)

A próxima etapa foi a implementação do K-means. Como o algoritmo K-means só é implementado para dados numéricos, foi escolhido o método de conversão One-Hot Encoding para os dados categóricos, visando assim que as tranformações não tenham ordenação como caso fosse usado o método Label Encoding (exemplo: transformar Solteiro = 1 e Casado = 2 e o algoritmo considerar que Casado tem maior peso que Solteiro e agrupar erroneamente). Para tal conversão foi utilizada a função get_dummies(), função esta encontrada na biblioteca Pandas versão 1.3.5 da linguagem de programação Python.

O algoritmo K-means tem uma sensibilidade em relação a escala que os dados se encontram pois utiliza cálculos de distâncias. Para sua aplicação é ideal que os dados se encontrem na mesma escala. Para resolver tal problema e evitar que haja interferência no agrupamento dos dados, utilizou-se a função MinMaxScaler() da biblioteca Scikit-Learn versão 1.0.2 nos dados numéricos nt_geral, nt_comp_espec, nt_formacao_geral, nu_idade e temp_transcorrido, de forma que a escala deles fossem entre 0 e 1.

Com as transformações realizadas, a base que continha 38 atributos passou a ter 225, pois com a aplicação do One-Hot Encoder, cada categoria dos atributos se transformam em um atributo novo para receber os valores 0 ou 1. E assim, os dados estavam prontos para serem usados na implementação do algoritmo de clusterização K-means. A inicialização escolhida para este projeto foi a k-means++ que é a melhor abordagem, pois seleciona os centroides iniciais de forma inteligente para acelerar a convergência.

Como dito na seção de Referencial Teórico, é necessário informar qual o valor de k para que o algoritmo K-means seja implementado. Para isso, foi utilizado o método Elbow. O melhor número de k para este estudo foi k=4. A Figura 7 mostra o resultado. Com o valor ideal de k definido, foi possível seguir com a implementação do K-means.

Após a implementação do algoritmo K-means, foi adicionado a base de dados uma nova coluna chamada de Clusters_Previstos, onde foi extraído de kmeans.labels_, que resgata a informação onde em qual clusters cada registro foi agrupado. A partir

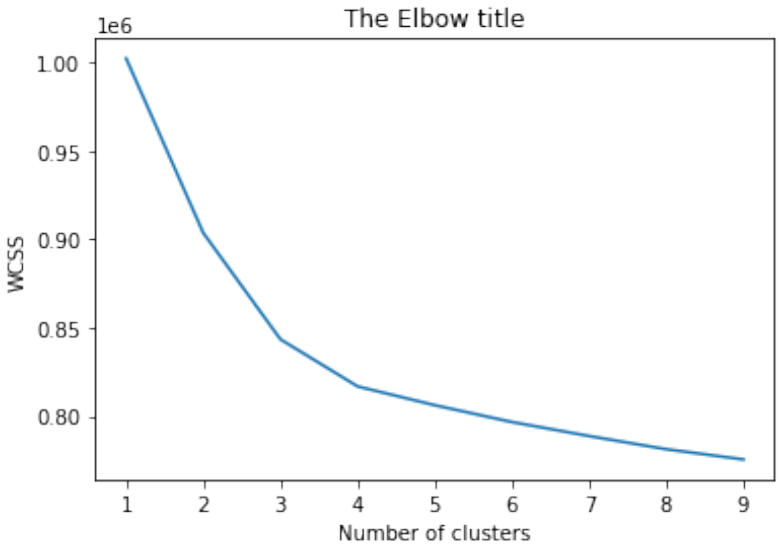

Figura 7. Método Elbow.
Fonte: A autora (2022)

disso, a base de dados estava pronta para a última etapa do estudo: receber análises com o objetivo de descrever os grupos gerados pelo K-means e encontrar possíveis perfis de estudantes de computação, ou seja, obter as características similares dos estudantes dentro dos grupos e as características distintas entre os grupos.

Para verificar o resultado da clusterização, foi implementado o Silhouette Analysis (SA), pois é usado para determinar o grau de separação entre os clusters. O coeficiente do SA pode assumir valores entre [-1,1]. Onde:

- SA = 0: a amostra está muito próxima dos clusters vizinhos.
- SA = 1: a amostra está distante dos clusters vizinhos.
- SA = -1: a amostra é atribuída aos clusters errados.

Segundo Wang et al. (2017), o SA analisa as distâncias de cada ponto de dados para seu próprio cluster e seu cluster vizinho mais próximo. O SA para um único ponto é calculado como:

$$sil(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2}$$

onde $a(i)$ é a distância de um ponto para seu próprio cluster e o $b(i)$ é a distância de um ponto para o seu cluster vizinho mais próximo. O SA para uma clusterização é definido como:

$$Sil = \frac{1}{n} \sum_{i=1}^{n} sil(i) \tag{3}$$

Neste estudo, o coeficiente do SA para k = 4 foi igual a 0,098, como pode ser conferido na Figura 8, indicando que os clusters estão separados pois estão entre o intervalo

[0,1] mas são clusters bem próximos, pois o coeficiente é mais próximo de 0, indicando que os clusters são parecidos em suas caracteristicas.

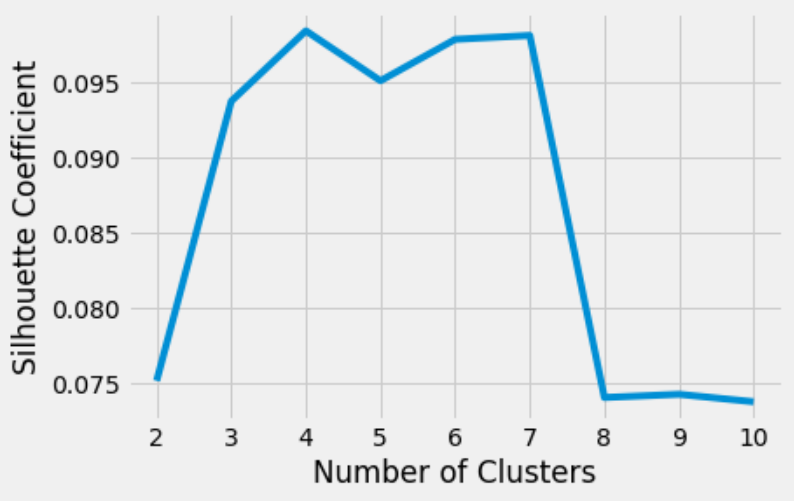

Figura 8. Resultado do Silhouette Analysis aplicado no estudo.
Fonte: A autora (2022)

5. Resultados

Os 50.802 registros foram distribuídos em 4 grandes grupos, chamados de Cluster 0 (C0), Cluster 1 (C1), Cluster 2 (C2) e Cluster 3 (C3). A Figura 9 exibe a visão geral dos 4 grupos formados neste estudo sobre os dados socioeconômicos dos concluintes de cursos de computação no ano de 2017 após a aplicação do algoritmo K-means.

Na Figura 9, é possível observar que 24,80% dos concluintes foram agrupados no C0. O C1 possui 28,90% concluintes, enquanto o C2 possui 16,10%. Por último, o C3 detém 30,20% dos concluintes. Perante o exposto, é conveniente afirmar que os grupos são equilibrados. Em relação ao tipo de instituição de ensino superior, o C0 conta com 66% de IES públicas federais. os grupos C1, C2 e C3 possuem 85%, 75% e 98% respectivamente de IES privadas.

Em relação a modalidade do curso todos os grupos apresentam a modalidade de ensino presencial como a mais cursada, registrando altos valores iguais a 96% para o C0, 91% para o C1, 84% para o C2 e, por fim, 80% para o C3. As regiões que se destacam no C0 são as sudeste e nordeste, com 47% e 23% respectivamente. No C1, as regiões são sudeste com 51% e sul com 18%. Já no C2, 55% região sudeste e 16% região nordeste. o C3 conta com 59% para a região sudeste e 22% para a região sul.

No que diz respeito a escolha do curso, o C0 mostra que 27% dos concluintes escolheram o curso de bacharelado em ciência da computação. 34% dos concluintes do C1 escolheram o curso de sistemas de informação. Os concluintes de C2 e C3 também escolheram o curso de sistemas de informação com os valores de 29% e 32% respectivamente.

Todos os grupos apontam que a maioria de seus concluintes são do sexo masculino onde os valores são: 83% para o C0, 86% para o C1, 88% para o C2 e C3. 40% dos concluintes do C0 tiveram um curso de turno integral, 78% dos concluintes do C1 tiveram

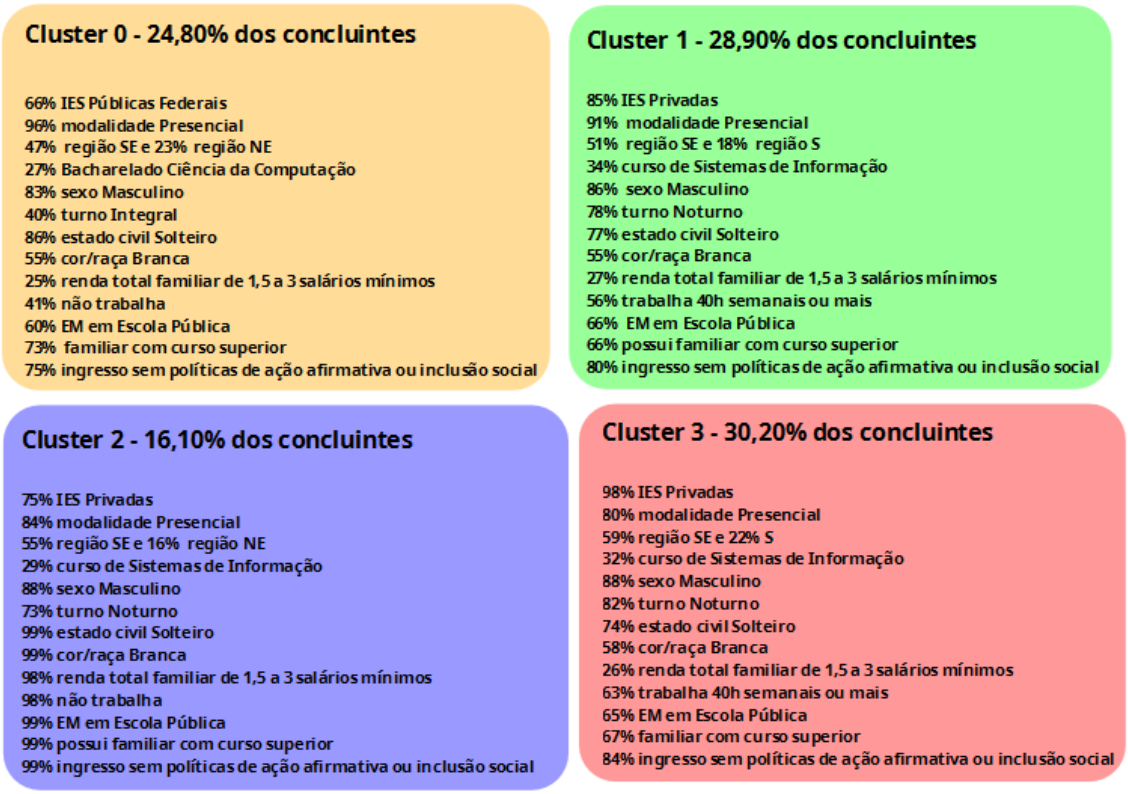

Figura 9. Visal geral dos clusters.
Fonte: A autora (2022)

um curso de turno noturno, 73% dos concluintes de C2 também foram de cursos à noite e, para terminar, 82% dos concluintes de C3 também cursaram cursos no período da noite.

Todos os grupos apresentam concluintes de estado civil solteiro e cor/raça branca. Os números são: 86% solteiro e 55% cor/raça branca no C0, 77% solteiro e 55% cor/raça branca no C1, 99% solteiro e 99% cor/raça branca e, por fim, 74% solteiro e 58% cor/raça branca no C3.

Quando analisamos a renda total familiar dos grupos também encontramos similaridade nesse quesito, onde todos os grupos apresentam que seus concluintes possuem renda total familiar de 1,5 a 3 salários mínimos (R$ 1.405,51 a 2.811,00 reais). Os valores encontrados foram: 25% no C0, 27% no C1, 98% no C2 e 26% no C3.

41% dos concluintes no C0 afirmaram que não trabalham. Já no C1, 56% afirmaram que trabalham 40 horas semanais ou mais. No C2, 98% declararam não trabalhar e 63% no C3 expuseram que trabalham 40 horas semanais ou mais. Ao que se refere ao tipo de escola no ensino médio, todos os grupos apontam para a escola pública: 60% no C0, 66% no C1, 99% no C2 e, por último, 65% no C3.

No quesito de curso superior na família, em todos os grupos os concluintes afirmaram que alguém de sua família cursou um curso de ensino superior. 73% para o C0, 66% para o C1, 99% para o C2 e 67% para o C3. Sobre o tipo de ingresso na instituição de ensino superior, 75% dos concluintes no C0, 80% no C1, 99% no C2 e 84% no C3 afirmaram que não tiveram nenhum tipo de ação afirmativa ou inclusão social.

O resultado da análise dos dados numéricos dos grupos formados pode ser conferido na Figura 10 para idade e nota geral e na Figura 11 para nota de formação geral e do componente específico. O C0 apresenta uma média de 25,89 para a idade, com idade mínima de 18 e máxima de 69. A mediana é de 24 anos. Para o C1 a média de idade é 26,94, com mínima de 17 e máxima de 73. A mediana registra 25 anos. Já para a idade no C2, a média é 28,97 com mínima de 18 e máxima de 84 e mediana 27. Por fim, no C3 a média é 27,44 com mínima de 18, máxima de 66 e mediana de 26.

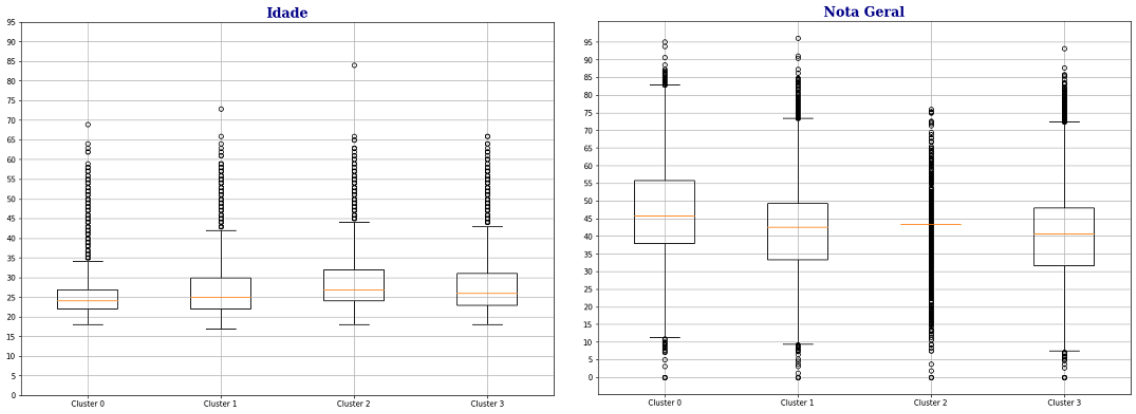

Figura 10. Gráficos de caixa dos atributos idade e nota geral dos Clusters.
Fonte: A autora (2022)

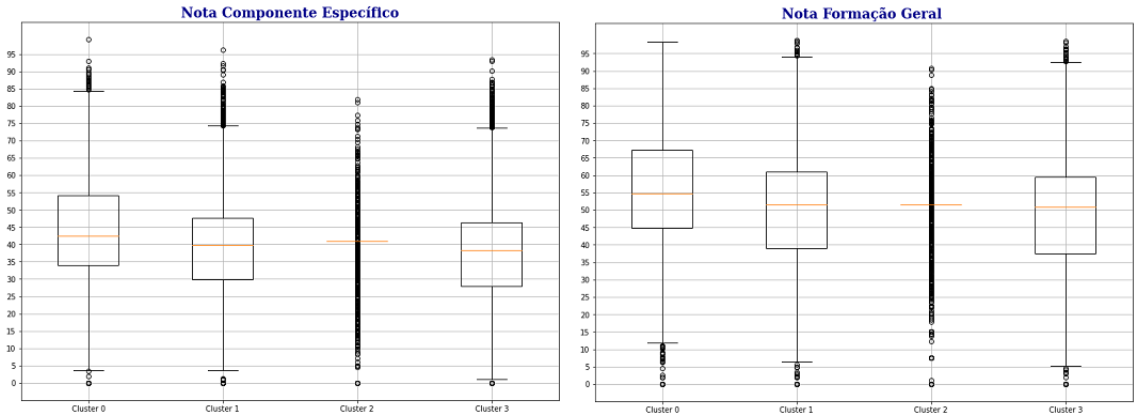

Figura 11. Gráficos de caixa dos atributos nota de formação geral e do componente específico dos Clusters.
Fonte: A autora (2022)

A nota geral, que é composta pela nota de formação geral (25%) e nota componente específico (75%), do C0 possui média de 46,72, mínima de 0, máxima de 95,10 e mediana de 45,85. Já no C1, a média é 41,84, com mínima de 0 e máxina de 96,20 e mediana de 42,70. No C2, a média apresentada é de 43,07, mínima de 0, máxima de 76 e mediana de 43,52. E, finalmente, no C3 a média da nota geral é de 40,47, com mínima de 0 e máxima de 93,30 e mediana de 40,80.

Em relação aos resultados obtidos para o C2 quanto as notas, as plotagens apresentam uma nítida diferença em relação aos outros clusters. Esta particularidade ocorreu

pois o algoritmo agrupou os valores faltantes (NaNs) preenchidos pela mediana, na etapa de pré-processamento, neste cluster. Exemplo: para a nota geral, 7545 notas do C2 (equivalente a 93% do cluster) são equilaventes a 43,52. Com isso, as demais notas foram consideradas outliers pelo algoritmo de plotagem do gráfico de caixa. É importante ressaltar que os outliers neste estudo não foram tratados.

A partir dos resultados expostos, o C0, de forma geral, é composto por IES públicas federais, da região sudeste e nordeste, do curso de bacharelado em ciência da computação, na modalidade presencial e turno integral. Os concluintes são solteiros, do sexo masculino, brancos, não trabalham e possuem renda familiar entre 1,5 a 3 salários mínimos. As medianas de idade e nota geral são 24 e 45,85, respectivamente.

O C1 e C3, em sua maioria, são formados por IES privadas, da região sudeste e sul, do curso de sistemas de informação, na modalidde presencial e turno noturno. E seus concluintes são solteiros, do sexo masculino, brancos, trabalham 40 horas semanais ou mais e possuem renda familiar entre 1,5 a 3 salários mínimos. Em relação a idade e nota geral, C1 registra mediana de 25 para idade e 42,70 para nota geral. O C3 registra mediana de 26 para idade e 40,80 para nota geral.

E o C2, em geral, é formado por IES privadas, da região sudeste e nordeste, do curso de sistemas de informação, na modalidade presencial e turno noturno. Os concluintes são solteiros, do sexo masculino, brancos, trabalham 40 horas semanais ou mais e possuem renda familiar entre 1,5 a 3 salários mínimos. E, para concluir, As medianas de idade e nota geral são de 27 e 43,52, respectivamente.

6. Conclusões

Este trabalho objetivou a obtenção do perfil socieconômico de estudantes brasileiros concluintes de cursos superiores da área de computação no ano de 2017, para isso utilizou a técnica de clusterização K-means. Os dados analisados foram obtidos a partir dos microdados do ENADE 2017 no portal do Inep, de acesso público.

Os resultados apontaram para quatro grandes grupos de estudantes que não apresentaram grandes dissimilaridades entre eles. De acordo com os resultados obtidos pelos clusters é possível elencar um perfil de estudante concluinte de computação no ano analisado, visto que os clusters apresentam várias características em comum, tais como: a maioria dos estudantes são do sexo masculino, solteiros, de cor branca, com ingresso nas IES sem políticas de ações afirmativas ou inclusão social, optaram pela modalidade presencial, cursaram o ensino médio em escolas públicas, possuem renda total familiar de 1,5 a 3 salários mínimos e alguem familiar cursou um curso superior. Contudo algumas características foram encontradas em grupos específicos, por exemplo, o cluster 0 integra, em sua maioria, os estudantes de Ciência da Computação de turno integral de IES públicas federais, enquanto que os demais clusters correspondem, em sua maioria, a estudantes de Sistemas de Informação de turno noturno de IES privadas. Os clusters 0 e 2 englobam estudantes concluintes que não trabalham, em contrapartida, os clusters 1 e 3 possuem estudantes concluintes que trabalham 40 horas semanais ou mais.

Este trabalho ainda abre a possibilidade de realização de várias outras pesquisas nesta área, como a aplicação de diferentes técnicas de clusterização e associação na base usada neste estudo do ano de 2017, como também nas bases de dados de anos anteriores ou posteriores a 2017 quando estiverem disponíveis pois até o fim deste trabalho, a

base de dados de 2017 é a mais atual de concluintes de computação. Possibilitando assim observar se há mudança nas caracteristicas dos grupos ao longo dos anos. Pode-se também expandir a pesquisa para outras áreas, não somente a área de computação que é o foco do presente trabalho. Outro ponto importante a considerar em trabalhos futuros é a identificação e tratamento de outliers, que podem mudar os resultados obtidos.

Referências

Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8): 1295, 2020.

Taher Al-Shehari and Rakan A Alsowail. An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques. *Entropy*, 23(10):1258, 2021.

Fernando Amaral. *Aprenda mineração de dados: teoria e prática*, volume 1. Alta Books Editora, 2016.

Ryan Baker, Seiji Isotani, and Adriana Carvalho. Mineraçao de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, 19(02): 03, 2011.

Patricia Mariotto Mozzaquatro Chicon and Alex Vinicius Telocken. Otimização aplicada a técnica de clusterização. *REVISTA INTERDISCIPLINAR DE ENSINO, PESQUISA E EXTENSÃO*, 9(1):13–27, 2021.

André Bessa da Silva, Denilton Macário de Paula, and Geórgia Regina Rodrigues Gomes. Mineração de dados: Um estudo para identificação do perfil docente das ies com conceito 3 ou superior no igc avaliado em 2016 no estado do espírito santo. 2019.

Marilia Nayara Clemente de Almeida Lima, Geovanne Oliveira Alves, Wedson Lino Soares, and Roberta Andrade de Araújo Fagundes. Educational data mining: A hybrid approach to predicting academic performance of students. In *MLDM (2)*, pages 500–514, 2019.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, sep 1999. ISSN 0360-0300. doi: 10.1145/331499.331504. URL `https://doi.org/10.1145/331499.331504`.

Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, jun 2010. doi: 10.1016/j.patrec.2009.09.011. URL `https://doi.org/10.1016%2Fj.patrec.2009.09.011`.

Wojciech Kwedlo. A clustering method combining differential evolution with the k-means algorithm. *Pattern Recognition Letters*, 32(12):1613–1621, 2011.

Maria Carolina Monard and José Augusto Baranauskas. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, 1(1):32, 2003.

Rena Nainggolan, Resianta Perangin-angin, Emma Simarmata, and Astuti Feriani Tarigan. Improved the performance of the k-means cluster using the sum of squared error (sse) optimized by using the elbow method. In *Journal of Physics: Conference Series*, volume 1361, page 012015. IOP Publishing, 2019.

Geon Yong Park, Heeseong Kim, Hwi Woon Jeong, and Hee Yong Youn. A novel cluster head selection method based on k-means algorithm for energy efficient wireless sensor network. In *2013 27th international conference on advanced information networking and applications workshops*, pages 910–915. IEEE, 2013.

Cristobal Romero, Sebastian Ventura, Mykola Pechenizkiy, and Ryan SJd Baker. *Handbook of educational data mining*. CRC press, 2010.

Pedro Henrique Schneider. *Análise preditiva de Churn com ênfase em técnicas de Machine Learning: uma revisão*. PhD thesis, 2016.

Nicolas Pastorio Boa Vista, Michele Ferraz Figueiró, and Patricia Mariotto Mozzaquatro Chicon. Técnicas de mineração de dados aplicadas aos microdados do enade para avaliar o desempenho dos acadêmicos do curso de ciencia da computação no rio grande do sul utilizando o software r. *I Seminário de Pesquisa Científica e Tecnológica*, 1(1), 2017.

Fei Wang, Hector-Hugo Franco-Penya, John D Kelleher, John Pugh, and Robert Ross. An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 291–305. Springer, 2017.