UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE QUIXADÁ

CURSO DE GRADUAÇÃO EM ENGENHARIA DE SOFTWARE

HIGOR DA SILVA CAMELO

CIÊNCIA DE DADOS NA ANÁLISE DE PERFIS E VARIÁVEIS DA INSEGURANÇA ALIMENTAR EM RESIDÊNCIAS DO CEARÁ

QUIXADÁ

2023

HIGOR DA SILVA CAMELO

CIÊNCIA DE DADOS NA ANÁLISE DE PERFIS E VARIÁVEIS DA INSEGURANÇA ALIMENTAR EM RESIDÊNCIAS DO CEARÁ

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus de Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Orientadora: Profa. Drª. Lívia Almada Cruz.

QUIXADÁ

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C189 Camelo, Higor da Silva.
Ciência de dados na análise de perfis e variáveis da insegurança alimentar em residências do Ceará / Higor da Silva Camelo. – 2023.
68 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Engenharia de Software, Quixadá, 2023.
Orientação: Profa. Dra. Lívia Almada Cruz.

1. Insegurança alimentar. 2. Aprendizagem não supervisionado. 3. Análise de Dados. I. Título.
CDD 005.1

HIGOR DA SILVA CAMELO

CIÊNCIA DE DADOS NA ANÁLISE DE PERFIS E VARIÁVEIS DA INSEGURANÇA ALIMENTAR EM RESIDÊNCIAS DO CEARÁ

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus de Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Aprovada em: _____/_____/_____

BANCA EXAMINADORA

Profa. Drª. Lívia Almada Cruz (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Regis Pires Magalhães
Universidade Federal do Ceará (UFC)

Prof. Me. Luís Gustavo Coutinho Rego
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

AGRADECIMENTOS

Primeiramente, quero de expressar meus sinceros agradecimentos a Deus, cuja ação, mesmo de formas misteriosas, me proporcionaram paz de espírito para me sentir melhor comigo mesmo, a fé n'Ele e a reflexão sempre como a última esperança quando o mundo e a vida pareciam confusos e desesperadores demais.

Aos meus pais, meu profundo agradecimento pelo apoio e dedicação, proporcionando os recursos necessários para minha educação. Sua generosidade e esforços têm sido fundamentais para meu crescimento como homem e cidadão.

Expresso minha sincera gratidão aos professores que, mesmo não construindo relações de amizade direta, sempre demonstraram cordialidade, respeito e uma disposição incansável para impulsionar meu aprendizado. Suas aulas e orientações foram essenciais para minha formação acadêmica, e cada contribuição foi valorizada.

Um agradecimento especial à minha orientadora, Prof.ª Dr.ª Lívia Almada Cruz, cuja dedicação ao longo deste ano foi notável. Sua orientação sábia, apoio constante e comprometimento ao longo deste ano de 2023 foram fundamentais para o desenvolvimento deste trabalho.

Gostaria de agradecer também ao meu gato Koichi e minha psicóloga, Dr.ª Adalgisa Bueno Guimarães, que nos piores momentos da minha vida, cada um de sua maneira, me deram o apoio necessário para que eu me recuperasse e tivesse a determinação e a esperança para seguir em frente e não desistir de tudo.

E, principalmente, gostaria de agradecer aos meus amigos, os quais eu considero meu principal pilar para o meu desenvolvimento pessoal e meu constante desejo de querer me tornar uma pessoa melhor. Nunca antes eu tive uma gama de amigos tão vasta e diversa, me expondo a diferentes pontos de vista e modos de viver, e acima de tudo, me dando o conforto e a segurança para eu compreender melhor minha identidade e estilo e poder expressá-los. Tenho um eterno sentimento de gratidão a todos, e, mesmo que daqui a uns anos cada um siga sua vida em sua respectiva cidade ou estado, sempre estarei disposto a fazer o que for possível para ajudar, apoiar e tornar a vida dessas pessoas melhor, assim como elas fizeram com a minha. Todos vocês são as minhas jóias da coroa e merecem ser tratados a peso de ouro, nada menos do que isso!

Por fim, agradecer a cidade de Quixadá, que me acolheu quando meu senso de responsabilidade e independência era muito pequeno, no fim da minha vivência aqui, me sinto pronto para prosseguir com a minha vida de forma mais independente e conhecedora da

realidade. Por extensão, a todos os funcionários e servidores do meu eternamente amado Campus de Quixadá, que sempre colaboraram para o bem-estar de todos, cada um de sua maneira.

Meu muito, muito obrigado a todos, vocês estarão sempre no meu coração.

RESUMO

O presente trabalho centra-se na análise da insegurança alimentar em famílias cearenses, empregando a ciência de dados como ferramenta principal. O contexto aborda não apenas a identificação do problema, mas também a busca por soluções eficazes e direcionadas por meio da aplicação de técnicas analíticas avançadas. Inicialmente, utilizou-se um extenso conjunto de dados por meio da Pesquisa CMIC (Caracterização das Famílias em Situação de Extrema Vulnerabilidade Social), feita pela Secretaria de Proteção Social do Ceará. Com cerca de 34.000 famílias e 183 perguntas, o banco de dados representativo foi filtrado para focar especificamente nas questões relacionadas à segurança alimentar e demografia, totalizando 29 perguntas essenciais. A metodologia adotada engloba o desenvolvimento e a avaliação de um modelo de aprendizado de máquina não supervisionado, com ênfase no algoritmo k-means. Esse algoritmo de clusterização é aplicado aos dados socioeconômicos, demográficos e comportamentais das famílias, identificando perfis semelhantes em relação à insegurança alimentar. A análise do modelo inclui métricas de desempenho como inércia, distância média dos centroides e silhueta, dentre outros, proporcionando uma avaliação abrangente da capacidade do modelo em realizar agrupamentos precisos. A interpretação dos resultados busca identificar variáveis-chave e padrões associados à ocorrência da insegurança alimentar, fornecendo *insights* valiosos para a alocação estratégica de recursos públicos. A relevância do estudo reside na contribuição para o avanço do conhecimento na área de segurança alimentar, apresentando uma abordagem inovadora e baseada em aprendizado de máquina para a identificação de perfis socioeconômicos em contextos de escassez de alimentos. Os resultados obtidos não apenas têm potencial para direcionar recursos e ações específicas, mas também oferecem um auxílio na tomada de decisões para a melhoria da qualidade de vida das famílias mais vulneráveis no Ceará. Este trabalho representa um esforço para integrar tecnologias modernas, como aprendizado de máquina, na abordagem de problemas sociais complexos, exemplificando o papel da ciência de dados na resolução de desafios reais.

Palavras-chave: insegurança alimentar; aprendizado não supervisionado; análise de dados

ABSTRACT

The present work focuses on the analysis of food insecurity in families from Ceará, employing data science as the primary tool. The context addresses not only the identification of the problem but also the search for effective and targeted solutions through the application of advanced analytical techniques. Initially, an extensive dataset from the CMIC Survey (Characterization of Families in Extremely Vulnerable Social Situation), conducted by the Social Protection Secretariat of Ceará, was used. With approximately 34,000 families and 183 questions, the representative database was filtered to specifically focus on issues related to food security and demographics, totaling 29 essential questions. The adopted methodology encompasses the development and evaluation of an unsupervised machine learning model, with emphasis on the k-means algorithm. This clustering algorithm is applied to the socioeconomic, demographic, and behavioral data of families, identifying similar profiles regarding food insecurity. The model analysis includes performance metrics such as inertia, average distance from centroids, and silhouette, among others, providing a comprehensive assessment of the model's ability to perform accurate groupings. The interpretation of the results seeks to identify key variables and patterns associated with the occurrence of food insecurity, providing valuable *insights* for the strategic allocation of public resources. The relevance of the study lies in contributing to the advancement of knowledge in the field of food security, presenting an innovative, machine-learning-based approach to identifying socioeconomic profiles in contexts of food scarcity. The obtained results not only have the potential to guide specific resources and actions but also offer assistance in decision-making to improve the quality of life for the most vulnerable families in Ceará. This work represents an effort to integrate modern technologies, such as machine learning, into the approach of complex social problems, exemplifying the role of data science in addressing real challenges.

Keywords: food insecurity; unsupervised learning; data analysis

LISTA DE ILUSTRAÇÕES

Figura 1 – Subdivisões do aprendizado de máquina e suas aplicações 17
Figura 2 – Comparação entre a Aprendizagem Supervisionada e a Não Supervisionada 19
Figura 3 – Sequência de atividades para execução do projeto 30
Figura 4 – Impacto do COVID-19 na disponibilidade de alimentos 37
Figura 5 – Preocupação acerca da falta de alimentos 38
Figura 6 – Porcentagem de famílias atendidas pelo CREAS 38
Figura 7 – Programas sociais estaduais mais frequentes 39
Figura 8 – Pontuação de silhueta simulada por número de *clusters*. 40
Figura 9 – Pontuação de inércia simulada por número de *clusters* — Método do Cotovelo. 41
Figura 10 – Visualização do *cluster* para $k = 5$ 41
Figura 11 – Visualização do *cluster* para $k = 6$ 42
Figura 12 – Visualização utilizando t-SNE 46
Figura 13 – Principais Fontes de Renda do *Cluster* 0 48
Figura 14 – Presença de água canalizada para famílias do *Cluster* 0 48
Figura 15 – Frequência de formação em cursos de qualificação e interesse no *Cluster* 1 . 50
Figura 16 – Situação da posse da residência das famílias do *Cluster* 1 51
Figura 17 – Presença de trabalhadores remunerados no *Cluster* 2 53
Figura 18 – Forma de acesso a água para uso doméstico *Cluster* 2 54
Figura 19 – Frequência da coleta de lixo das famílias do *Cluster* 3 56
Figura 20 – Presença da assistência do Centro de Referência de Assistência Social (CRAS) no *Cluster* 3 57
Figura 21 – Presença da assistência do Centro de Referência Especializado de Assistência Social (CREAS) no *Cluster* 4 59

LISTA DE QUADROS

Quadro 1 – Comparação dos Artigos 29
Quadro 2 – Perguntas e Respostas sobre Segurança Alimentar 34
Quadro 3 – Perguntas sobre Infraestrutura 34
Quadro 4 – Perguntas sobre Renda e Trabalho 35
Quadro 5 – Perguntas sobre Assistencialismo 35
Quadro 6 – Insegurança Alimentar por Cluster 43
Quadro 7 – Resultados das Métricas do Modelo K-Means 45
Quadro 8 – Comparação de Valores entre *Clusters* 59
Quadro 9 – Variáveis Mais Impactantes para Insegurança Alimentar 61

LISTA DE ABREVIATURAS E SIGLAS

CMIC	Pesquisa de Caracterização das Famílias em Situação de Extrema Vulnerabilidade Social
CRAS	Centro de Referência de Assistência Social
CREAS	Centro de Referência Especializado de Assistência Social
FAO	Food and Agriculture Organization
PCA	Principal Component Analysis
SVM	Support Vector Machine
WFP	World Food Programee

SUMÁRIO

1 INTRODUÇÃO 13
1.1 Justificativa 13
1.2 Objetivo Geral 14
1.2.1 Objetivos Específicos 14
2 FUNDAMENTAÇÃO TEÓRICA 15
2.1 Insegurança Alimentar no Ceará 15
2.2 Análise de Dados 16
2.3 Aprendizado de Máquina 16
2.3.1 Algoritmos de Aprendizado de Máquina 18
2.3.2 Clusterização de Dados 18
2.4 Pré-processamento de Dados 20
2.5 Avaliação de Modelos de Clusterização 21
3 TRABALHOS RELACIONADOS 23
3.1 ***The Use of Support Vector Machine to Analyze Food Security in a Region of Brazil*** 23
3.2 ***Machine learning can guide food security efforts when primary data are not available*** 24
3.3 ***A data-driven approach improves food insecurity crisis prediction*** 26
3.4 Análise Comparativa 27
3.4.1 Similaridades 27
3.4.2 Diferenças 28
3.4.3 Contribuição Única 28
4 METODOLOGIA 30
4.1 Natureza da Pesquisa 30
4.2 Abordagem da Pesquisa 30
4.3 Objetivos da Pesquisa 31
4.4 Definição de ferramentas e métodos 31
4.5 Conjunto de dados 32
4.6 Pré-processamento de dados 33
4.7 Análise Exploratória 36

4.8 Clusterização de Dados . . . 38
4.9 Aplicação do algoritmo e avaliação . . . 40
4.10 Análise dos *clusters* e formação dos perfis familiares . . . 43
4.11 Discussão de resultados e aplicações . . . 44
5 RESULTADOS . . . 45
5.1 Resultados das avaliações dos *clusters* . . . 45
5.1.0.1 Utilização de t-SNE Para Visualização de Clusters . . . *46*
5.2 Formação dos *Clusters* e suas Características . . . 47
5.2.1 Cluster 0 - Maior insegurança . . . 47
5.2.1.1 Análise da Renda e Trabalho . . . *47*
5.2.1.2 Análise da Infraestrutura . . . *47*
5.2.1.3 Análise do Acesso à Assistência Estatal . . . *49*
5.2.2 Cluster 1 - Baixa insegurança . . . 49
5.2.2.1 Análise da renda e trabalho . . . *49*
5.2.2.2 Análise da infraestrutura . . . *50*
5.2.2.3 Análise do acesso à assistência estatal . . . *51*
5.2.2.4 Correlação de Variáveis . . . *51*
5.2.3 Cluster 2 - Média insegurança . . . 52
5.2.3.1 Análise da renda e trabalho . . . *52*
5.2.3.2 Análise da infraestrutura . . . *53*
5.2.3.3 Análise do acesso à assistência estatal . . . *54*
5.2.3.4 Correlação de Variáveis . . . *54*
5.2.4 Cluster 3 - Insegurança média/baixa . . . 54
5.2.4.1 Análise da Renda e Trabalho . . . *55*
5.2.4.2 Análise da Infraestrutura . . . *55*
5.2.4.3 Análise do Acesso à Assistência Estatal . . . *57*
5.2.4.4 Correlação de Variáveis . . . *57*
5.2.5 Cluster 4 - Média/baixa insegurança . . . 57
5.2.5.1 Análise da Renda e Trabalho . . . *58*
5.2.5.2 Análise da Infraestrutura . . . *58*
5.2.5.3 Análise do Acesso à Assistência Estatal . . . *58*
5.2.5.4 Correlação de Variáveis — Cluster 4 . . . *58*

5.2.6 ***Perfil Socioeconômico: Clusters em Perspectiva*** 59
6 **CONCLUSÃO** 62
REFERÊNCIAS 66

1 INTRODUÇÃO

A insegurança alimentar é um problema persistente em muitas regiões do mundo, afetando milhões de pessoas e tendo sérias consequências para a saúde e o bem-estar das famílias (Food and Agriculture Organization of the United Nations, 2019; International Food Policy Research Institute, 2021). A compreensão dos fatores relacionados à insegurança alimentar é fundamental para o desenvolvimento de estratégias eficazes de mitigação e intervenção (Pinstrup-Andersen, 2009). Nesse contexto, o uso de técnicas de aprendizado de máquina tem se mostrado uma abordagem promissora para a análise da insegurança alimentar em famílias (Fernandes *et al.*, 2021; Poudel *et al.*, 2021).

O Ceará é um estado localizado na região Nordeste do Brasil, seus cenários de insegurança alimentar possuem realidade multifacetada moldada por uma interseção de fatores climáticos, sociais e históricos. Com um clima predominantemente semiárido, marcado por longos períodos de seca e chuvas irregulares, a região enfrenta desafios constantes na produção agrícola, resultando em escassez de alimentos e perda de safras. Segundo o Instituto Nacional de Meteorologia (INMET), a média pluviométrica anual tem se mantido abaixo do necessário para sustentar uma agricultura robusta (INMET, 2020). Além disso, a concentração de propriedades de terra e a distribuição desigual de recursos exacerbaram as desigualdades sociais, deixando muitas comunidades em situação de vulnerabilidade (IBGE, 2019). A histórica falta de investimentos em infraestrutura de irrigação e tecnologias agrícolas ressalta a fragilidade do sistema alimentar cearense (Silva, 2018). O estado do Ceará possui um histórico marcante de grandes secas, como a "Grande Seca de 1877 – 1879", que teve impactos profundos na população e na economia local (Silva, 2005); a literatura também retratou essas adversidades, como demonstrado por Rachel de Queiroz em sua obra "O Quinze", que ressaltou as duras condições enfrentadas pela população cearense durante a seca (Queiroz, 1930). O legado de décadas de seca e dificuldades socioeconômicas é profundamente enraizado na realidade do Ceará, impactando negativamente a segurança alimentar de sua população.

1.1 Justificativa

Espera-se que este trabalho contribua para o avanço do conhecimento na área de segurança alimentar, fornecendo uma abordagem baseada em aprendizado de máquina para a identificação dos perfis socioeconômicos da população cearense inserida em cenários de escassez

de alimentos. Os resultados obtidos poderão auxiliar no direcionamento de recursos e ações para mitigar a insegurança alimentar e melhorar a qualidade de vida das famílias mais vulneráveis.

1.2 Objetivo Geral

O objetivo principal deste trabalho é empregar técnicas de aprendizado de máquina, especificamente o modelo não-supervisionado *k-means*, para analisar e detectar perfis de famílias cearenses afetadas pela insegurança alimentar. A pesquisa utiliza dados abrangentes, englobando aspectos socioeconômicos, demográficos e comportamentais dessas famílias, proporcionando uma visão detalhada dos fatores relacionados à insegurança alimentar. Ao aplicar o algoritmo *k-means*, busca-se identificar e agrupar essas famílias em clusters com características semelhantes em relação à insegurança alimentar.

1.2.1 Objetivos Específicos

Os objetivos específicos deste trabalho estão listados a seguir:

- Avaliar e compreender fatores sócio-econômico-demográficos que causam insegurança alimentar;
- Identificar quais os perfis de famílias mais propensas a sofrerem de insegurança alimentar;
- Aplicar os conceitos da inteligência artificial para apontar quais as áreas mais deficitárias que estão relacionadas ao cenário de carência de alimentos;

2 FUNDAMENTAÇÃO TEÓRICA

A insegurança alimentar é um desafio global que afeta milhões de pessoas em todo o mundo. No contexto específico do estado do Ceará, no Nordeste do Brasil, a insegurança alimentar é uma preocupação significativa devido às condições socioeconômicas desafiadoras e à alta taxa de insegurança alimentar na região (IPECE, 2022). A compreensão dos fatores relacionados a este assunto e a busca por abordagens eficazes de análise são fundamentais para o desenvolvimento de políticas e intervenções que visem mitigar esse problema.

2.1 Insegurança Alimentar no Ceará

A insegurança alimentar é um desafio enfrentado por milhões de pessoas em todo o mundo, incluindo regiões específicas, como o estado do Ceará, no Brasil. Compreender a insegurança alimentar no contexto do Ceará requer uma análise abrangente dos fatores socioeconômicos, ambientais e políticos que afetam o acesso a alimentos adequados e nutritivos. O contexto socioeconômico do Ceará desempenha um papel fundamental na compreensão da insegurança alimentar. A análise de indicadores como pobreza, desigualdade de renda, acesso a serviços básicos e a estrutura econômica local fornece *insights* sobre as condições socioeconômicas que podem contribuir para a insegurança alimentar na região (IPECE, 2022). Além disso, é importante examinar a agricultura e a produção de alimentos no Ceará. Características agrícolas, como tipos de cultivos, sistemas de produção e disponibilidade de água para irrigação, têm impacto direto na oferta de alimentos na região. Considerar também os efeitos das mudanças climáticas na produção de alimentos é fundamental para entender os desafios enfrentados pelos agricultores e a consequente insegurança alimentar (Mendes *et al.*, 2020).

As políticas públicas e os programas de segurança alimentar implementados no Ceará são elementos essenciais na abordagem da insegurança alimentar. Programas de transferência de renda, como o Programa Bolsa Família, têm o potencial de reduzir a vulnerabilidade das famílias em situação de insegurança alimentar. Além disso, iniciativas voltadas para o fortalecimento da agricultura familiar e o acesso a alimentos saudáveis desempenham um papel importante na promoção da segurança alimentar na região (Pinheiro *et al.*, 2016). A insegurança alimentar tem impactos significativos na saúde, educação e bem-estar das famílias e comunidades do Ceará. Estudos demonstram que a insegurança alimentar está associada a problemas de desnutrição, saúde mental, desempenho acadêmico e desenvolvimento infantil inadequado. Compreender

esses impactos é essencial para desenvolver estratégias eficazes de combate à insegurança alimentar e promover um ambiente mais saudável e sustentável (Monteiro *et al.*, 2018).

2.2 Análise de Dados

A análise de dados é uma área multidisciplinar que envolve a aplicação de técnicas e métodos estatísticos para explorar, compreender e interpretar conjuntos de dados (Jr *et al.*, 2019). Essa análise é fundamental para a obtenção de *insights*, tomada de decisões e geração de conhecimentos em diversas áreas, como ciência, negócios, saúde, entre outras.

A análise de dados pode ser dividida em duas abordagens principais: análise descritiva e análise inferencial. A análise descritiva envolve a utilização de medidas estatísticas descritivas, como média, mediana, desvio padrão e frequência, para resumir e descrever as características dos dados. Essa análise fornece uma visão geral do conjunto de dados, permitindo uma compreensão inicial dos padrões e tendências presentes.

A análise inferencial, por sua vez, busca fazer inferências e tirar conclusões sobre uma população maior com base em uma amostra dos dados. Essa abordagem utiliza técnicas estatísticas para estimar parâmetros, testar hipóteses e fazer generalizações sobre a população. A análise inferencial é especialmente importante quando se deseja tirar conclusões mais amplas a partir de um conjunto limitado de dados (Tabachnick; Fidell, 2013).

Além dessas abordagens, a análise de dados também envolve a utilização de técnicas de visualização, como gráficos e diagramas, para representar e comunicar as informações contidas nos dados de forma mais clara e acessível. A visualização dos dados facilita a identificação de padrões, tendências e *outliers*, auxiliando na interpretação e na comunicação dos resultados.

Existem diversas ferramentas e softwares disponíveis para realizar a análise de dados, desde planilhas eletrônicas, como o Microsoft Excel, até linguagens de programação especializadas, como R (Field *et al.*, 2012) e Python. Cada ferramenta oferece diferentes recursos e funcionalidades para manipular, explorar e analisar os dados.

2.3 Aprendizado de Máquina

O aprendizado de máquina, um subcampo da inteligência artificial, refere-se ao desenvolvimento de algoritmos e técnicas que permitem que um sistema computacional aprenda a partir de dados, sem ser explicitamente programado. Ele é baseado na ideia de que os

computadores podem aprender padrões e fazer previsões ou tomar decisões com base nesses padrões identificados.

Uma das principais abordagens do aprendizado de máquina é o aprendizado supervisionado (Figura 1), onde um algoritmo é treinado em um conjunto de dados rotulados, consistindo em entradas (características) e suas respectivas saídas desejadas. Durante o treinamento, o algoritmo ajusta seus parâmetros internos para encontrar um mapeamento que relacione as entradas às saídas esperadas. Exemplos de algoritmos amplamente utilizados nessa abordagem incluem as redes neurais artificiais (Rumelhart *et al.*, 1986) e as máquinas de vetor de suporte (Support Vector Machine (SVM)) (Cortes; Vapnik, 1995).

Figura 1 – Subdivisões do aprendizado de máquina e suas aplicações

Fonte: Adaptado de Kumar (2020)

Outra abordagem comum é o aprendizado não supervisionado, que lida com dados não rotulados, onde o objetivo é encontrar padrões ou estruturas intrínsecas nos dados. Algoritmos de agrupamento, como o *k-means*(Lloyd, 1982), são exemplos populares nessa categoria, onde os dados são divididos em grupos com base em sua similaridade. Tanto o aprendizado supervisionado como o não supervisionado são empregados em diversas aplicações, como classificação de imagens e previsão de mercado para o primeiro; e como sistemas de recomendação ou marketing focalizado para o segundo(Figura 1).

Além disso, existem técnicas de aprendizado de máquina que exploram a interação entre agentes em um ambiente, conhecidas como aprendizado por reforço. Nessa abordagem, um agente aprende a tomar ações em um ambiente para maximizar uma recompensa cumulativa.

O algoritmo de Q-Learning (Watkins; Dayan, 1992) é um exemplo clássico de aprendizado por reforço.

É importante mencionar também que a avaliação e seleção de modelos de aprendizado de máquina são etapas cruciais. Métodos como a validação cruzada (Kohavi, 1995) e a curva de aprendizado (Mitchell, 1997) são utilizados para estimar o desempenho do modelo em dados não vistos e evitar problemas de sobreajuste.

2.3.1 Algoritmos de Aprendizado de Máquina

Os algoritmos de clusterização são estruturas matemáticas ou estatísticas fundamentadas na identificação de padrões e estrutura nos dados, sem a necessidade de rótulos prévios. Diferentemente dos modelos preditivos supervisionados, que se baseiam em conjuntos de dados rotulados, onde a variável alvo é conhecida para cada exemplo de treinamento, os modelos de clusterização operam de maneira não supervisionada.

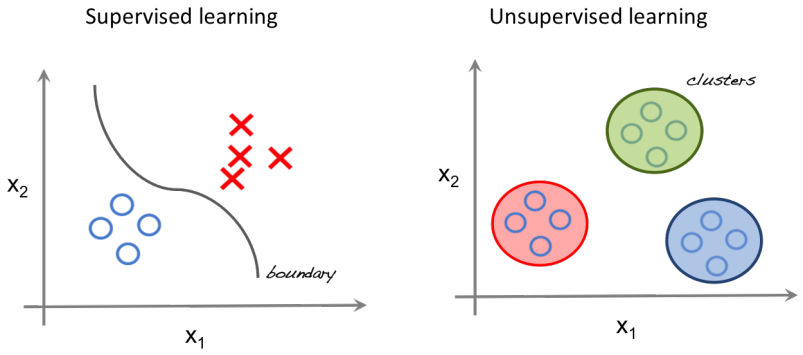
O principal propósito dos algoritmos de clusterização é agrupar observações semelhantes com base em características compartilhadas, visando identificar estruturas intrínsecas nos dados. Essa tarefa é realizada por meio de algoritmos de aprendizado não supervisionado, que exploram a similaridade entre as instâncias para formar grupos ou clusters. Exemplos de algoritmos de clusterização incluem o *K-Means*, a Propagação de Afinidade, o Agrupamento Hierárquico e o DBSCAN. Esses métodos são aplicados em diversas áreas, como ciência de dados e análise exploratória, para identificar padrões emergentes e estruturas latentes nos conjuntos de dados (Hastie *et al.*, 2009). Os algoritmos de aprendizado não supervisionados exploram a estrutura intrínseca dos dados, sem a necessidade de rótulos ou variáveis alvo conhecidas. Esses modelos são usados principalmente para tarefas de agrupamento (*clustering*), onde os dados são agrupados com base em sua similaridade ou propriedades comuns. Algoritmos populares de agrupamento incluem o *k-means*, agrupamento hierárquico e mistura de gaussianas (Hastie *et al.*, 2009).

2.3.2 Clusterização de Dados

A clusterização de dados, também conhecida como análise de agrupamento, é uma técnica de aprendizado não supervisionado amplamente utilizada para identificar grupos ou *clusters* naturais em um conjunto de dados. O objetivo principal da clusterização é agrupar objetos similares em um mesmo *cluster* e objetos distintos em *clusters* diferentes, com base

em suas características e propriedades (Figura 2). Esse processo permite explorar padrões e estruturas subjacentes nos dados, bem como para segmentar dados em grupos homogêneos.

Figura 2 – Comparação entre a Aprendizagem Supervisionada e a Não Supervisionada



Fonte: (Ribeiro, 2017).

Existem vários algoritmos de clusterização disponíveis, e um dos mais populares e amplamente utilizados é o algoritmo *k-means*(MacQueen, 1967). O *k-means* é um método iterativo que parte do pressuposto de que cada *cluster* é representado por seu centroide, calculado como a média dos objetos pertencentes ao *cluster*. O algoritmo inicia atribuindo aleatoriamente centroides iniciais e, em seguida, alterna entre duas etapas principais até convergir para uma solução: atribuição de objetos aos *clusters* com base na distância euclidiana em relação aos centroides e atualização dos centroides com base na média dos objetos atribuídos a cada *cluster*.

A distância *k-means* é uma medida utilizada para avaliar a qualidade dos *clusters* formados pelo algoritmo *k-means*(Jr *et al.*, 2019). Essa métrica calcula a média das distâncias euclidianas entre os objetos e o centroide de seus *clusters*. Quanto maior a distância *k-means*, maior é a separação média entre os *clusters*, indicando uma melhor separação e distinção entre os grupos. Dessa forma, a distância *k-means* fornece uma medida global da qualidade da clusterização.

É importante ressaltar que o algoritmo *k-means* tem algumas limitações, como a sensibilidade à inicialização dos centroides e à presença de *outliers*, além de ser aplicável apenas a dados numéricos. No entanto, ele continua sendo amplamente utilizado devido à sua simplicidade, eficiência computacional e interpretabilidade dos resultados.

Além do *k-means*, existem outros algoritmos de clusterização, como o DBSCAN (Ester *et al.*, 1996) e o *Hierarchical Agglomerative Clustering* (Everitt *et al.*, 2011). Cada algoritmo possui suas próprias características e critérios de formação de *clusters*, sendo importante selecionar o método mais adequado conforme a natureza dos dados e os objetivos da análise.

A clusterização de dados é amplamente utilizada em diversas áreas, incluindo ciência de dados, reconhecimento de padrões, bioinformática, *marketing* e análise de mercado. Na área da segurança alimentar, essa prática pode ser aplicada para identificar grupos de famílias com características semelhantes em termos de insegurança alimentar, permitindo uma compreensão mais aprofundada dos fatores que contribuem para essa condição e facilitando a elaboração de estratégias de intervenção mais eficazes.

2.4 Pré-processamento de Dados

O pré-processamento de dados é uma etapa fundamental no processo de análise de dados e construção de modelos de aprendizado não supervisionados. Essa etapa envolve uma série de técnicas e procedimentos para preparar e transformar os dados brutos em um formato adequado para análise e aplicação de algoritmos de aprendizado de máquina.

1. Limpeza de Dados: A limpeza de dados refere-se à identificação e tratamento de valores ausentes, inconsistentes ou errôneos nos dados. Isso pode incluir a remoção de registros duplicados, o preenchimento de valores ausentes por meio de técnicas como média, mediana ou imputação estatística, e a correção de erros de digitação ou formatos inconsistentes (Han *et al.*, 2011).
2. Transformação de Dados: A transformação de dados envolve a aplicação de técnicas para modificar a escala, distribuição ou formato dos dados. Isso pode incluir a normalização ou padronização dos valores das características, a aplicação de transformações logarítmicas ou exponenciais, e a redução da dimensionalidade por meio de técnicas como análise de componentes principais (PCA) ou seleção de características (Witten *et al.*, 2016).
3. Codificação de Variáveis Categóricas: Em muitos conjuntos de dados, as variáveis categóricas precisam ser convertidas em uma forma numérica para que os algoritmos de aprendizado de máquina possam processá-las adequadamente. Isso pode ser feito por meio de técnicas como codificação *one-hot*, codificação *LabelEncoder* ou codificação de frequência (Kuhn; Johnson, 2013).
4. Detecção e Tratamento de *Outliers*: *Outliers* são pontos de dados que diferem significativamente do restante dos dados. A detecção e o tratamento desses *outliers* são importantes para evitar que eles afetem negativamente a análise e os modelos de IA. Isso pode ser feito por meio de técnicas estatísticas, como o uso de intervalos interquartis (IQR) ou métodos robustos de detecção de *outliers* (Hawkins *et al.*, 1980).

2.5 Avaliação de Modelos de Clusterização

A avaliação de modelos não supervisionados, como os baseados em técnicas de clusterização, desempenha um papel crucial no entendimento da estrutura dos dados e na validação dos resultados obtidos. Diferentemente dos modelos supervisionados, onde a tarefa é prever rótulos conhecidos, os modelos não supervisionados exploram padrões subjacentes nos dados, agrupando-os de acordo com características semelhantes.

Durante o processo de clusterização, várias métricas podem ser utilizadas para avaliar a qualidade dos *clusters* formados e auxiliar na seleção do número adequado de *clusters*. Algumas das métricas mais comuns incluem inércia (Hastie *et al.*, 2009), silhueta (Rousseeuw, 1987), distância *k-means* (Tibshirani *et al.*, 2001), índice Davies-Bouldin (Davies; Bouldin, 1979) e índice Calinski-Harabasz(Calinski; Harabasz, 1974).

A inércia é uma métrica utilizada no algoritmo de clusterização *k-means*. Ela mede a soma das distâncias quadráticas dos objetos em relação ao centroide de cada *cluster*. Quanto menor a inércia, mais compactos e similares são os objetos dentro de cada *cluster* (JR. *et al.*, 2019). A fórmula matemática para a inércia, considerando n objetos (x_i) e k clusters (C_j) com centroides (μ_j), é dada pela equação 2.1:

$$\text{Inércia} = \sum_{i=1}^{n} \min_{j=1}^{k} \left\| x_i - \mu_j \right\|^2 \tag{2.1}$$

A métrica de silhueta (S) é uma medida de validação interna que avalia a qualidade da clusterização considerando a separação entre os *clusters* e a coesão dos objetos dentro de cada *cluster*. Ela varia de -1 a 1, onde valores próximos de 1 indicam que os objetos estão bem agrupados, enquanto valores próximos de -1 indicam que os objetos foram erroneamente atribuídos a *clusters* incorretos. Valores próximos de 0 indicam sobreposição entre os *clusters* (Field *et al.*, 2012). A fórmula matemática para a silhueta média ($S_{\text{média}}$) considerando n objetos, distâncias médias intra-cluster (a_i) e distâncias médias inter-cluster (b_i) é dada pela equação 2.2:

$$S_{\text{média}} = \frac{1}{n} \sum_{i=1}^{n} \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{2.2}$$

O índice Davies-Bouldin (DB) compara a dispersão dentro de cada *cluster* com a separação entre os clusters. Quanto menor o valor do índice, melhor é a qualidade da clusterização, indicando uma maior separação entre os *clusters* e uma menor dispersão dentro de cada

cluster (Davies; Bouldin, 1979). A fórmula matemática para o índice Davies-Bouldin é dada pela equação 2.3:

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{a_i + a_j}{d_{ij}}\right) \tag{2.3}$$

O Índice Calinski-Harabasz (CH) também é empregado na avaliação de algoritmos de clusterização, como o *k-means*. Esta métrica visa medir a qualidade da clusterização considerando tanto a coesão dos objetos dentro de cada *cluster* quanto a separação entre os agrupamentos. O cálculo do índice envolve a razão entre a dispersão média nos *clusters* (S_B) e a dispersão média dentro destes (S_W). Em termos simples, quanto maior o valor do Índice Calinski-Harabasz, melhor é a qualidade da clusterização, indicando uma maior separação entre os *clusters* e uma menor dispersão interna (Calinski; Harabasz, 1974). A fórmula matemática para o índice Calinski-Harabasz (equação 2.4), onde n é o número de objetos e k é o número de *clusters*, é dada por:

$$CH = \frac{S_B}{S_W} \times \frac{n-k}{k-1} \tag{2.4}$$

Em resumo, a avaliação de modelos de aprendizado não supervisionado e de clusterização é essencial para medir a qualidade dos resultados obtidos e tomar decisões informadas. As métricas específicas fornecem informações valiosos sobre o desempenho dos modelos e a qualidade dos *clusters* formados, permitindo ajustes e melhorias necessárias ao longo do processo.

3 TRABALHOS RELACIONADOS

Neste capítulo, serão dispostos alguns outros trabalhos que apresentam similaridades com o projeto aqui estabelecido, tão como em qual aspecto este projeto diverge desses.

3.1 *The Use of Support Vector Machine to Analyze Food Security in a Region of Brazil*

O artigo de Barbosa e Nelson (2016) apresenta uma análise detalhada da utilização da técnica de Máquina de Vetores de Suporte, do inglês SVM, para analisar a segurança alimentar em famílias no estado do Ceará. Sendo a desnutrição uma preocupação crucial em todo o mundo, e entender a situação em uma região específica pode fornecer inferências valiosas para o desenvolvimento de políticas e estratégias adequadas. Para conduzir a análise, foram coletados dados de diversas fontes, como instituições governamentais, órgãos de pesquisa e organizações não governamentais. Os dados abrangeram indicadores socioeconômicos, dados demográficos, informações sobre produção agrícola, disponibilidade de alimentos e dados nutricionais.

A metodologia adotada por Barbosa e Nelson (2016) nesse estudo envolveu várias etapas. Inicialmente, os dados coletados foram submetidos a um processo de pré-processamento, isso incluiu a limpeza dos dados para remover entradas duplicadas ou inconsistentes, bem como a transformação dos dados em um formato adequado para a análise. Em seguida, os dados foram divididos em conjuntos de treinamento e teste.

A técnica do SVM foi escolhida como método de análise devido à sua capacidade comprovada de lidar com conjuntos de dados complexos e multidimensionais (Cortes; Vapnik, 1995). O modelo SVM foi treinado usando os dados de treinamento, buscando estabelecer um modelo capaz de classificar corretamente as diferentes categorias relacionadas à segurança alimentar. Para isso, foram consideradas variáveis como renda per capita, acesso a serviços de saúde, educação, infraestrutura e características da produção agrícola local.

Após o treinamento do modelo SVM, foi realizada uma avaliação usando os dados de teste. Isso permitiu verificar a eficácia do modelo na classificação e previsão da segurança alimentar na região estudada. Os resultados foram analisados com o intuito de identificar áreas com maior risco de insegurança alimentar e compreender os fatores socioeconômicos e demográficos associados a essa situação.

Os resultados obtidos forneceram inferências valiosos sobre a segurança alimentar na região em estudo. Os mapas resultantes da análise do SVM permitiram identificar áreas com

maior vulnerabilidade, auxiliando na priorização da alocação de recursos e esforços para melhorar a situação. Além disso, a análise estatística dos fatores socioeconômicos e demográficos revelou quais características estão mais fortemente relacionadas à insegurança alimentar, possibilitando a identificação de pontos de intervenção estratégica.

Em conclusão, esse estudo demonstrou a eficácia do uso da Máquina de Vetores de Suporte como uma abordagem promissora para analisar a segurança alimentar em uma região específica do Brasil. A metodologia empregada permitiu obter informações detalhadas sobre a situação alimentar, fornecendo uma base sólida para a formulação de políticas e a tomada de decisões relacionadas à segurança alimentar, visando melhorar a qualidade de vida da população local.

Em comparação com o estudo conduzido por Barbosa e Nelson (2016), este trabalho adota uma abordagem distinta para analisar a insegurança alimentar. Enquanto Barbosa utilizou SVM, este trabalho emprega o algoritmo *k-means* de aprendizado de máquina não-supervisionado. Enquanto o SVM é notório por sua eficácia em classificação, o *k-means* destaca-se na identificação de padrões e grupos em conjuntos de dados complexos.

Assim, embora ambos os estudos busquem abordar a problemática da insegurança alimentar, as metodologias e variáveis consideradas neste trabalho oferecem uma perspectiva única e complementar, contribuindo para a ampliação do entendimento sobre essa questão complexa.

3.2 *Machine learning can guide food security efforts when primary data are not available*

A pesquisa de Martini *et al.* (2022) examina o uso de técnicas de aprendizado de máquina para orientar os esforços de segurança alimentar quando os dados primários não estão disponíveis. Sendo este um problema mais presentes em regiões menos desenvolvidas socioeconomicamente, a existência de dados demográficos e precisos pode demonstrar-se escassa ou mesmo inexistente, por falta de recursos, infraestrutura ou questões logísticas, dificultando assim a realização de pesquisas acerca do cenário e o possível planejamento de políticas de combate a insegurança alimentar. Deste modo, houve a oportunidade de utilizar conceitos de aprendizado de máquina para mensurar os esforços de segurança alimentar utilizando dados que não dizem a respeito propriamente às famílias pesquisadas, mas sim utilizando dados mais abrangentes naquela comunidade, como a inflação de alimentos, a precipitação das chuvas, a densidade populacional, dentre outros.

Foram aplicadas técnicas de aprendizado de máquina, como algoritmos de classificação, regressão e agrupamento, para analisar os dados secundários e realizar previsões e estimativas relacionadas à segurança alimentar, modelos foram treinados e ajustados utilizando-se algoritmos apropriados e validados usando técnicas adequadas de avaliação. Os resultados do estudo mostraram que as técnicas de aprendizado de máquina puderam fornecer *insights* valiosos sobre a segurança alimentar, mesmo quando os dados primários não estavam disponíveis.

Por meio da análise dos dados secundários e da aplicação de modelos de aprendizado de máquina, foram identificados padrões e correlações que permitiram uma compreensão mais profunda da situação alimentar nas regiões estudadas.

Os resultados destacaram áreas com maior risco de insegurança alimentar, identificando populações vulneráveis e indicando quais fatores socioeconômicos e demográficos estavam associados a essa situação. Além disso, os modelos de aprendizado de máquina permitiram fazer previsões sobre tendências futuras e avaliar o impacto de diferentes intervenções e políticas na segurança alimentar.

É importante ressaltar que, embora os resultados tenham sido promissores, existem limitações inerentes ao uso de dados secundários e modelos de aprendizado de máquina. A precisão e a confiabilidade dos resultados dependem da qualidade dos dados secundários e da correta implementação das técnicas de aprendizado de máquina. Portanto, é necessário interpretar os resultados com cautela e considerar a necessidade de validação por meio de dados primários sempre que possível.

Em relação à pesquisa conduzida por Martini *et al.* (2022), este trabalho apresenta diferenças substanciais em termos de abordagem e foco. Enquanto Martini se concentra na utilização de aprendizado de máquina quando dados primários sobre famílias específicas estão indisponíveis, este estudo aborda a insegurança alimentar por meio da análise direta de dados socioeconômicos, demográficos e comportamentais de famílias em uma região específica.

O autor propõe uma metodologia que utiliza dados secundários, como inflação de alimentos, precipitação das chuvas e densidade populacional, para fornecer apontamentos sobre a segurança alimentar em regiões com limitações na disponibilidade de dados primários. Em contraste, este trabalho emprega algoritmos de aprendizado não supervisionado, especificamente o *k-means*, para identificar padrões e agrupar famílias com características semelhantes em relação à insegurança alimentar.

Os resultados de Martini *et al.* (2022), embora promissores, ressaltam a importância

da interpretação cuidadosa de resultados baseados em dados secundários e da validação por meio de dados primários sempre que possível. Por outro lado, este estudo analisa dados mais diretos do modo de viver das famílias, como questões de renda, assistencialismo e infraestrutura, visando uma compreensão dos fatores que contribuem para a insegurança alimentar em uma comunidade específica.

3.3 *A data-driven approach improves food insecurity crisis prediction*

No trabalho de Lentz *et al.* (2019), o objetivo consiste em avaliar se a análise de dados pode melhorar a previsão de crises de insegurança alimentar e, em caso afirmativo, identificar quais variáveis são as mais importantes para prever essas crises. Além disso, os autores destacam a importância do estudo para desenvolver e padronizar metodologias que poderão ser utilizadas para complementar o Sistema de Classificação de Fase de Segurança Alimentar (IPC), o qual age como indicador para o nível de segurança alimentar de uma população.

Metodologicamente, o trabalho predisse três índices voltados para a segurança alimentar e utilizado por órgãos internacionais como a Agência dos Estados Unidos para o Desenvolvimento Internacional (USAID) e o Programa Alimentar Mundial (World Food Programee (WFP)), sendo estes índices o Índice Reduzido de Estratégias de Enfrentamento (rCSI), a Pontuação de Diversidade Alimentar Familiar (HDDS) e a Pontuação de Consumo de Alimentos (FCS). Para predizer os indicadores de segurança alimentar, os dados disponíveis foram agrupados em três classes com requerimentos de processamentos aumentativos e disponibilidade decrescente; os dados de Classe 0, considerada uma classe a parte das demais, consistem nos números do índice IPC; os de Classe 1 se referem aos dados relacionados a precipitação, preços dos alimentos, qualidade do solo e variáveis geográficas em conjunto com o padrão IPC da Classe 0; na Classe 2, foram classificados características como o tipo de revestimento do teto das residências e a posse de telefones celulares pelos habitantes; por fim, na Classe 3, foram considerados dados demográficos sobre as famílias, como gênero, idade, número de membros de uma família.

Os resultados mostraram que o modelo de aprendizado de máquina teve uma precisão significativamente maior na previsão de crises de insegurança alimentar do que os modelos tradicionais baseados em indicadores econômicos. Além disso, os autores identificaram as variáveis mais importantes para prever crises de insegurança alimentar, sendo a pobreza e as mudanças climáticas as mais relevantes. Destaca-se a importância de analisar cenários de

insegurança alimentar de forma integrada e multidisciplinar, não considerando apenas fatores diretamente referentes a nutrição de uma população, mas também aspectos sociais, econômicos e demográficos.

O estudo conduzido por Lentz *et al.* (2019) concentra-se na avaliação do potencial da análise de dados para aprimorar a previsão de crises de insegurança alimentar. Em contraste, este trabalho se volta para a análise direta de dados socioeconômicos, demográficos e comportamentais de famílias em uma região específica do Ceará, utilizando o algoritmo *k-means* como ferramenta principal.

Enquanto Lentz busca predizer índices amplamente utilizados, como o Índice Reduzido de Estratégias de Enfrentamento (rCSI), a Pontuação de Diversidade Alimentar Familiar (HDDS) e a Pontuação de Consumo de Alimentos (FCS), este estudo adota uma abordagem mais específica, buscando identificar perfis de famílias em relação à insegurança alimentar. A metodologia utilizada pelo autor envolve a categorização de dados em diferentes classes, considerando uma variedade de variáveis, desde indicadores climáticos até dados demográficos.

Os resultados de Lentz indicam que modelos de aprendizado de máquina superam significativamente os modelos tradicionais na previsão de crises de insegurança alimentar, destacando a importância de variáveis como pobreza e mudanças climáticas. Por outro lado, este trabalho, ao empregar o *k-means*, visa identificar grupos de famílias com características semelhantes, fornecendo uma compreensão mais detalhada dos fatores subjacentes à insegurança alimentar em uma região específica.

3.4 Análise Comparativa

Ao comparar o presente trabalho com as abordagens propostas nos estudos de Barbosa e Nelson (2016), Martini *et al.* (2022), e Lentz *et al.* (2019), destacam-se tanto semelhanças quanto diferenças significativas (Quadro 1).

3.4.1 Similaridades

O ponto de convergência entre os estudos reside no uso de técnicas de aprendizado de máquina para abordar a problemática da insegurança alimentar. Todos os trabalhos, incluindo este, reconhecem o potencial dessas técnicas na análise e previsão de cenários relacionados à segurança alimentar.

3.4.2 Diferenças

No entanto, as metodologias, fontes de dados e objetivos específicos divergem consideravelmente. O estudo de Barbosa e Nelson (2016) priorizou a coleta de dados primários por meio de questionários aplicados a famílias no Ceará, enquanto este trabalho concentrou-se na análise de dados secundários disponíveis para o mesmo contexto. Já Martini *et al.* (2022) visou prever a insegurança alimentar em regiões carentes de dados primários, empregando o algoritmo *Random Forest*.

A abordagem de Lentz *et al.* (2019), por sua vez, destaca-se por tentar aprimorar a previsão global de crises de insegurança alimentar, incorporando índices específicos, como rCSI, HDDS e FCS. Essa ênfase em indicadores globalmente reconhecidos diferencia-se da proposta deste trabalho, que visa identificar perfis específicos de famílias em uma região do Ceará.

3.4.3 Contribuição Única

Enquanto Barbosa e Nelson (2016) e Martini *et al.* (2022) exploraram aspectos preditivos, este trabalho, inspirado por Lentz *et al.* (2019), enfatizou a identificação de perfis e características intrínsecas às famílias. Essa abordagem única visa oferecer uma compreensão mais profunda da insegurança alimentar ao nível local, permitindo estratégias mais direcionadas e contextualizadas.

A comparação evidencia a complementaridade entre as abordagens, destacando a importância de considerar diferentes metodologias para uma compreensão abrangente da insegurança alimentar no Ceará.

Quadro 1 – Comparação dos Artigos

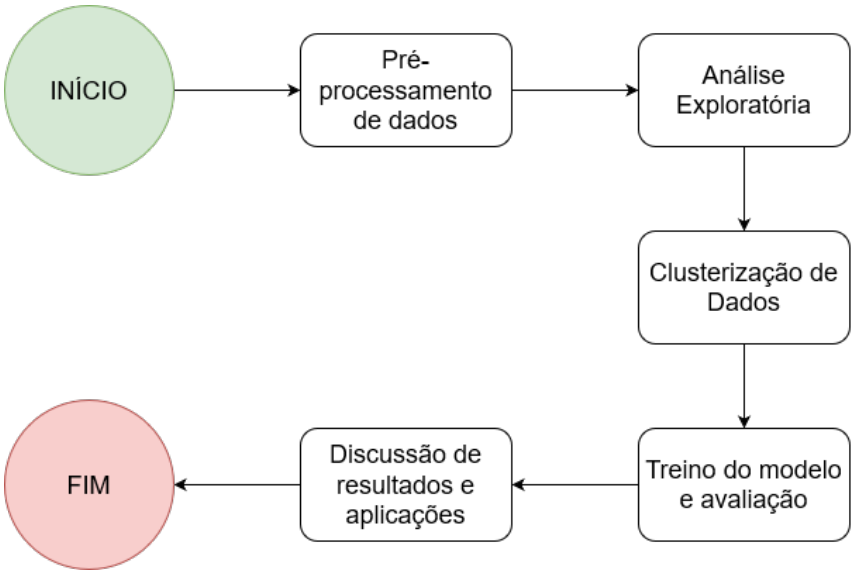
Artigo	**Algoritmo principal**	**Modo de aprendizado**	**Fonte de dados**	**Objetivo**
Barbosa e Nelson (2016)	SVM	Supervisionado	Questionários aplicados a famílias do Ceará	Identificação de fatores-chave que afetam a segurança alimentar e previsão futura da situação alimentar
Martini *et al.* (2022)	*Random Forest*	Supervisionado	*Food Security Information Network*	Prever insegurança alimentar em regiões sem dados primários
Lentz *et al.* (2019)	Árvore de decisão	Supervisionado	Banco Mundial, WFP e Food and Agriculture Organization (FAO)	Melhoria da previsão global de crises de insegurança alimentar através da análise de dados
Este trabalho	*K-means*	Não supervisionado	Pesquisa de Caracterização das Famílias em Situação de Extrema Vulnerabilidade Social (CMIC)	Caracterização e análise de famílias em diferentes níveis de insegurança alimentar

Fonte: Elaborado pelo autor.

4 METODOLOGIA

Este capítulo define e apresenta as etapas planejadas para a realização dos objetivos deste projeto, juntamente com as características do modelo de aprendizado não supervisionado e da análise de dados propostos. O diagrama presente na Figura 3 ilustra o fluxo do trabalho.

Figura 3 – Sequência de atividades para execução do projeto



Fonte: elaborado pelo autor.

4.1 Natureza da Pesquisa

A natureza da presente pesquisa é de cunho aplicado, uma vez que visa direcionar seus resultados para a contribuição ao combate de um problema prático e atual: a insegurança alimentar em famílias cearenses. A abordagem aplicada se alinha à necessidade de compreender os fatores socioeconômicos e demográficos que contribuem para essa mazela e, consequentemente, fornecer informações relevantes para a formulação de políticas públicas eficazes. A escolha da natureza aplicada foi fundamentada na urgência de fornecer *insights* práticos para a mitigação do deficit nutricional e no potencial impacto positivo que essa pesquisa pode ter na qualidade de vida das famílias mais vulneráveis.

4.2 Abordagem da Pesquisa

A abordagem de pesquisa adotada neste estudo é quantitativa, visto que visa quantificar e analisar dados socioeconômicos e demográficos dos núcleos familiares em carência

alimentar. A escolha pela abordagem quantitativa é respaldada pela necessidade de obter informações numéricas que permitam identificar tendências, grupos, padrões e relações entre as variáveis analisadas. A coleta, análise de dados quantitativos e o processo de clusterização são essenciais para a aplicação de técnicas estatísticas e algoritmos de aprendizado de máquina, que serão utilizados na construção do modelo analítico de insegurança alimentar.

4.3 Objetivos da Pesquisa

O objetivo exploratório deste estudo consiste em compreender os padrões e características associados à insegurança alimentar em famílias cearenses, por meio da análise de dados socioeconômicos e demográficos. A abordagem exploratória permite explorar a complexidade dessa problemática e identificar possíveis fatores de influência, sem impor hipóteses rígidas ou imutáveis. O objetivo é fornecer *insights* iniciais e aprofundar a compreensão sobre os determinantes da insegurança alimentar nesse contexto específico. Dessa forma, será possível direcionar futuras análises e desenvolver estratégias mais eficazes para combater o problema e melhorar a qualidade de vida dessas comunidades vulneráveis.

4.4 Definição de ferramentas e métodos

Este projeto utilizou diversas tecnologias e bibliotecas em Python, oferecendo uma abordagem completa para a análise exploratória de dados, modelagem e avaliação de resultados. A escolha da linguagem de programação Python foi motivada por sua flexibilidade e pela riqueza de ferramentas disponíveis para ciência de dados e aprendizado de máquina.

De acordo com Pedregosa *et al.* (2011), o scikit-learn é uma biblioteca em *Python* que fornece ferramentas eficientes para análise de dados e clusterização, incluindo algoritmos de aprendizado de máquina. A linguagem *Python*, por sua vez, é conhecida por sua facilidade de uso e rica biblioteca de suporte, tornando-a uma escolha ideal para o desenvolvimento do sistema.

Para visualização de dados, foram utilizadas as bibliotecas *matplotlib* (Hunter, 2007), *seaborn* (Waskom, 2021) e *yellowbrick* (Ince *et al.*, 2021). O *matplotlib* permitiu a criação de gráficos detalhados e personalizáveis, enquanto o *seaborn*, construído sobre o *matplotlib*, adicionou estilos atraentes e facilitou a criação de visualizações estatísticas complexas. A biblioteca *yellowbrick* desempenhou um papel crucial ao fornecer ferramentas específicas para a

avaliação de modelos de aprendizado de máquina e para a decisão final da quantidade de *clusters* utilizados, incluindo visualizações de curvas de aprendizado e gráficos de dispersão (Ince *et al.*, 2021).

A biblioteca Pandas (McKinney, 2021) foi a escolha natural para manipulação e análise de dados, com suas estruturas de dados poderosas, como o *DataFrame*, facilitando tarefas como limpeza, transformação e agregação de dados.

Para a redução de dimensionalidade e visualização de dados de alta dimensão, adotou-se a técnica t-SNE *(t-distributed Stochastic Neighbor Embedding)* (Maaten; Hinton, 2008). Essa abordagem permitiu representar dados complexos de forma mais acessível, facilitando a identificação de padrões e estruturas subjacentes (Maaten; Hinton, 2008).

Em resumo, a integração dessas tecnologias proporcionou uma base sólida para a condução da análise de dados, modelagem e interpretação de resultados. As referências apropriadas garantem a solidez metodológica do projeto, fundamentando as escolhas tecnológicas e destacando as melhores práticas em ciência de dados e aprendizado de máquina.

4.5 Conjunto de dados

Apesar dos desafios gerenciais, financeiros e técnicos envolvidos na realização de pesquisas em larga escala, a Secretaria de Proteção Social do Ceará (SPS) desenvolveu uma pesquisa para coletar informações sobre uma série de características sociais e econômicas de famílias em extrema vulnerabilidade social. Até agosto de 2022, a Pesquisa CMIC (Caracterização das Famílias em Situação de Extrema Vulnerabilidade Social), como ficou conhecida, foi coletada em todo o estado, abrangendo todos os 184 municípios e cerca de 50.000 famílias, para este projeto, foi disponibilizado um recorte até julho de 2022, contendo aproximadamente 34.000 registros.

Esse conjunto de dados é o objeto de análise deste trabalho, constituindo um esforço sistemático do governo do Ceará para avaliar as necessidades de famílias de baixa renda, especificamente aquelas com crianças e sem acesso adequado a serviços básicos. Vale ressaltar que o conjunto de dados completo é composto por 183 perguntas, algumas das quais focadas especificamente no desenvolvimento das crianças e outras na figura parental, como a saúde física e mental da mãe. Para este trabalho, foram filtradas as questões mais específicas para a segurança alimentar e para a demografia, reduzindo o escopo para 29 perguntas, os quais podem ser agrupados em 4 conjuntos, perguntas sobre nutrição (Quadro 2), infraestrutura e lazer

(Quadro 3), trabalho e renda (Quadro 4) e assistencialismo (Quadro 5).

Além disso, é importante destacar que os dados das famílias foram disponibilizados já anonimizados, garantindo assim a privacidade e confidencialidade das informações. Essa prática é fundamental para a ética em pesquisa, protegendo a identidade das famílias participantes. A anonimização é um procedimento padrão adotado para assegurar que nenhum dado pessoal identificável seja divulgado ou utilizado indevidamente.

Ressalta-se também que o acesso a esses dados foi obtido mediante um acordo de confidencialidade estabelecido com o Projeto *Big Data* e em parceria com o laboratório de pesquisa *Insight Data Science Lab*, da Universidade Federal do Ceará. Esse acordo reforça o compromisso com a segurança e o uso responsável das informações, garantindo que os dados sejam utilizados exclusivamente para os propósitos declarados na pesquisa, sem qualquer forma de divulgação não autorizada.

A análise desses dados baseada no aprendizado não supervisionado permite uma compreensão mais aprofundada da realidade dessas famílias, suas necessidades e os desafios enfrentados no acesso a serviços básicos. Além disso, são exploradas possíveis relações e correlações entre as variáveis presentes nas respostas obtidas na pesquisa, a fim de fornecer inferências relevantes para o contexto social.

4.6 Pré-processamento de dados

O pré-processamento dos dados é uma etapa fundamental para garantir a qualidade e adequação dos dados ao modelo de aprendizado de máquina. Os dados foram submetidos a um processo de limpeza para eliminar registros inconsistentes, duplicados ou faltantes. De acordo com Jr *et al.* (2019), a limpeza dos dados é essencial para evitar distorções e garantir a confiabilidade dos resultados. Técnicas como remoção de *outliers* e eliminação de valores faltantes foram aplicadas para tratar essas questões.

De início, em um *dataframe* contendo 34.616 registros, notou-se 809 linhas em que apenas o identificador familiar estava presente, sem qualquer outro valor, sendo assim, a exclusão dessas linhas foi a medida necessária.

Posteriormente, com o banco de dados com 33.807 valores, durante a avaliação dos dados de renda, percebe-se que existem valores de renda mensal muito superiores ao normal para uma família em estado de insegurança alimentar, como quantias acima de R$10.000,00. Deste modo, procurou-se uma “renda de corte” que respeitasse a proporcionalidade e a demografia do

Quadro 2 – Perguntas e Respostas sobre Segurança Alimentar

Pergunta	**Respostas**
Nos últimos 3 meses, houve preocupação com a falta de comida?	Sim: alguns dias; Sim: quase todo dia; Não; Sim: 1 ou 2 dias
Nos últimos 3 meses, houve falta de comida antes de ter dinheiro?	Sim: alguns dias; Sim: quase todo dia; Não; Sim: 1 ou 2 dias
Nos últimos 3 meses, você comeu menos por falta de dinheiro?	Sim: alguns dias; Sim: quase todo dia; Não; Sim: 1 ou 2 dias
Nos últimos 3 meses, houve falta de dinheiro para alimentação saudável?	Sim: alguns dias; Sim: quase todo dia; Não; Sim: 1 ou 2 dias
Nos últimos 3 meses, houve redução de alimentos por falta de dinheiro?	Sim: alguns dias; Sim: quase todo dia; Não; Sim: 1 ou 2 dias
Mudança na disponibilidade de alimentos após a COVID-19?	Sim: diminuiu; Sim: aumentou; Não mudou
Cria animais para consumo da família?	Sim; Não
Planta alimentos para consumo da família?	Sim; Não

Fonte: Elaborado pelo autor.

Quadro 3 – Perguntas sobre Infraestrutura

Pergunta	**Respostas**
A casa onde a família mora é:	Própria; Emprestada/cedida; Alugada
Material predominante nas paredes externas:	Alvenaria sem revestimento; Alvenaria com revestimento; Taipa não revestida; Taipa revestida; Madeira aproveitada; Palha; Outro
Forma de abastecimento de água:	Rede geral de distribuição; Poço ou nascente; Outra forma; Cisterna; Não sei
Água canalizada em pelo menos um cômodo?	Sim; Não
Água para beber é:	Sem tratamento; Filtrada; Tratada de outra forma; Água mineral; Fervida
Banheiro ou sanitário no domicílio?	Sim; Não
Coleta de lixo pela Prefeitura:	Mais de duas vezes; Nenhuma vez; De uma a duas vezes
Forma de iluminação mais utilizada:	Elétrica; Óleo, querosene ou gás; Outra; Vela
Existem locais para atividades próximos à casa?	Crianças brincarem; Atividades culturais; Atividades esportivas
Domicílio em área de conflito/violência?	Sim; Não; Não sei

Fonte: Elaborado pelo autor.

Quadro 4 – Perguntas sobre Renda e Trabalho

Pergunta	**Respostas**
Recebem algum benefício do governo do Estado do Ceará?	Pagamento do Cartão Mais Infância; Vale gás; Cesta básica; Isenção da tarifa de energia; Isenção da tarifa de água; Alimentos in natura; Virando o jogo
Alguém no domicílio tem trabalho remunerado atualmente?	Não; Sim; Não sabe
Se sim, quantos trabalham?	Valor numérico condicionado à pergunta anterior
Fontes de renda da família:	Bolsa Família; Cartão Mais Infância; Agricultura, Pecuária, Pesca ou Aquicultura; Pensão; Trabalho como assalariado; Outros (valor aberto)
Total de ganhos no último mês:	Valor numérico aberto
Nos últimos 12 meses, alguém no domicílio fez algum curso de qualificação?	Não sabe; Não; Sim
Gostaria de fazer algum curso de qualificação?	Não sabe; Não; Sim

Fonte: Elaborado pelo autor.

Quadro 5 – Perguntas sobre Assistencialismo

Pergunta	**Respostas**
É atendido pelo CRAS?	Não; Sim; Não sei
É atendido pelo CREAS?	Não; O município não possui CREAS; Não sei; Sim
Recebem algum benefício do governo do Estado do Ceará?	Pagamento do Cartão Mais Infância; Vale gás; Cesta básica; Isenção da tarifa de energia; Isenção da tarifa de água; Alimentos in natura; Virando o jogo

Fonte: Elaborado pelo autor.

dataframe, chegando em um valor de R$3.000,00, assim, a quantidade de valores considerados inválidos foram de 39. Em complemento, também detectou-se 2 valores vazios e 2 valores não numéricos, totalizando 43 registros acerca de renda mensal inválidos.

Em sequência, os valores acerca do número de trabalhadores por família também possuíam valores incorretos, 9 eram números decimais e também havia valores muito altos e ilógicos, para isso, foi definido um máximo de 12 pessoas, logo, 36 registros foram desconsiderados, por fim, não havia valores negativos ou não numéricos.

Por fim, somando todos os valores inválidos, chegou-se a quantidade de 888 padrões desconsiderados para este projeto, totalizando aproximadamente 0,025% do banco de dados original, assim, optou-se pela exclusão, logo, o tamanho final do *dataframe* para este projeto foi

de 33.728 famílias. Essa abordagem justifica-se principalmente pelo baixíssimo impacto que esses valores incorretos para o algoritmo de clusterização, julgando-se mais eficiente a remoção do que a imputação de novos valores baseado em métricas como média ou mediana dos demais registros.

Após a limpeza, foi realizada a codificação de variáveis categóricas em valores numéricos, a fim de torná-las adequadas aos algoritmos de aprendizado de máquina, utilizando-se da função *LabelEncoder*, do *Scikit-learn*. Após isso, foi aplicado a técnica de normalização ou padronização dos dados via *MinMaxScaler*, também do *Scikit-learn*, visando colocar todas as variáveis na mesma escala.

No contexto desta aplicação, a inversão e a ordinalidade da codificação foi adotada, de modo que valores mais altos representassem uma situação mais desfavorável, indicando maior insegurança alimentar. Esse processo teve como objetivo garantir que as variáveis codificadas refletissem de maneira apropriada o grau de desfavorabilidade das respostas. A adequação dessas bibliotecas ao contexto deste trabalho é crucial para a interpretação e validação dos resultados obtidos.

Somado a esses processos, a redução de dimensionalidade é útil quando se lida com conjuntos de dados com muitas variáveis, o que pode dificultar a análise e o desempenho dos algoritmos de aprendizado de máquina. Essa técnica permite simplificar e compactar os dados, preservando as informações mais relevantes (Hastie *et al.*, 2009). Métodos como Análise de Componentes Principais (PCA) ou Seleção de Características podem ser utilizados para reduzir o número de variáveis, mantendo a maioria da variabilidade original dos dados.

4.7 Análise Exploratória

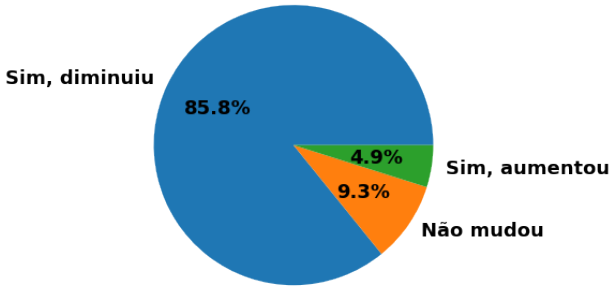
A análise exploratória dos dados é uma etapa crucial no processo de modelagem e compreensão dos dados coletados. Essa etapa explorou e examinou os dados de forma sistemática, identificando padrões, relações e características relevantes que possam fornecer *insights* para o desenvolvimento do modelo de aprendizado de máquina.

Inicialmente, realiza-se uma análise descritiva das variáveis socioeconômicas e demográficas coletadas das famílias do Ceará. Essa análise incluiu medidas de tendência central, como média, mediana e moda, além de medidas de dispersão, como desvio padrão e intervalo interquartil. Essas medidas forneceram uma visão geral das características dos dados e ajudarão a identificar possíveis valores discrepantes (*outliers*) (Hastie *et al.*, 2009).

Em seguida, são utilizadas técnicas de visualização de dados, como gráficos de barras, histogramas, gráfico de setores, e gráficos de dispersão, para explorar as relações entre as variáveis e identificar possíveis correlações. Por exemplo, segundo Tukey (1977), pode-se atestar se existe uma relação entre o nível de renda familiar e o nível de insegurança alimentar.

Em uma pequena visualização prática, aplicando estas metodologias (Figura 4), pode-se perceber o impacto negativo da pandemia do COVID-19 na disponibilidade de alimentos um grupo de famílias cearenses selecionadas para a coleta de dados.

Figura 4 – Impacto do COVID-19 na disponibilidade de alimentos



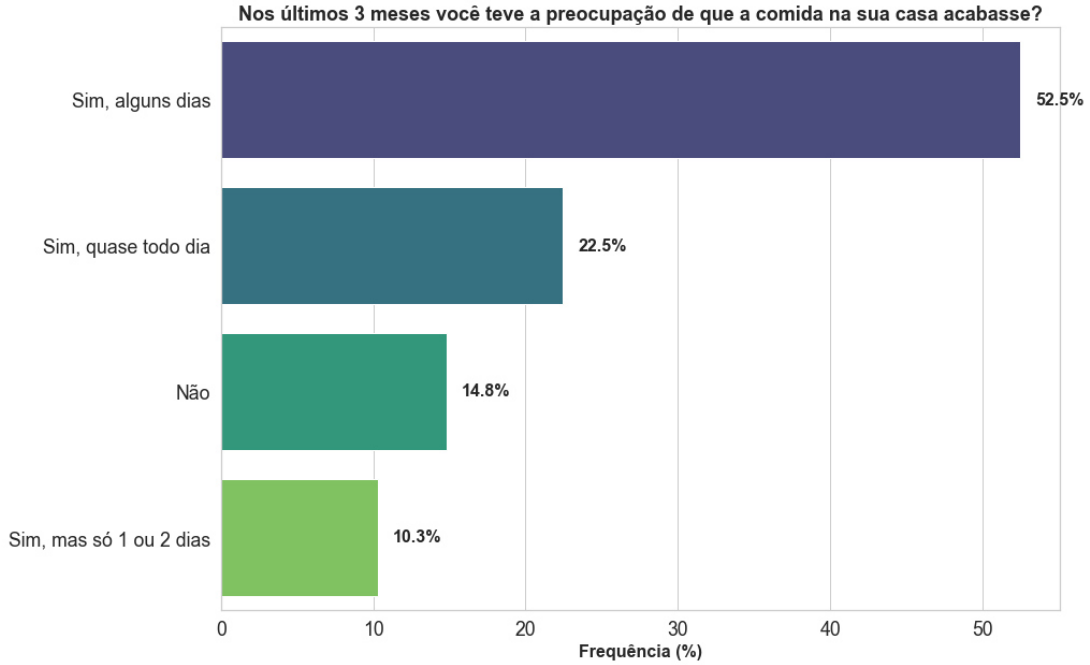
Fonte: elaborado pelo autor.

Utilizando outra métrica e estilo de visualização (Figura 5), pode-se perceber que 89,7% das pessoas entrevistadas possuem preocupações frequentes sobre a sua segurança alimentar.

Já no contexto dos programas sociais, é perceptível que uma minoria extremamente reduzida possui acesso ao CREAS (Centro de Referência Especializado de Assistência Social), dificultando o acesso a políticas públicas (Figura 6). Além disso, também é destacável a grande presença do Cartão Mais Infância e do vale-gás nas famílias beneficiada por programas assistencialistas estaduais (Figura 7).

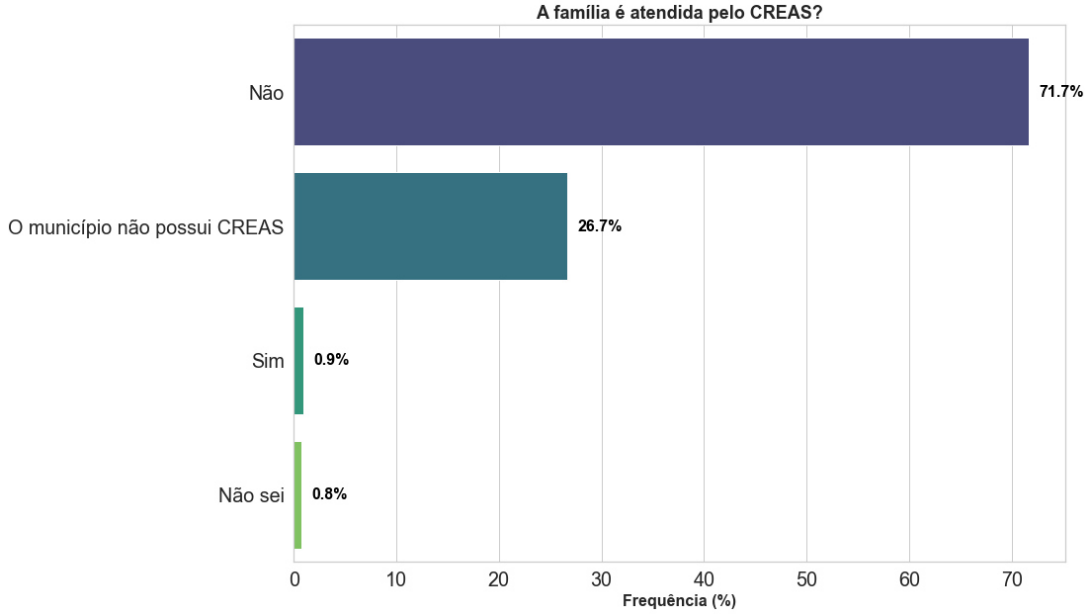
É importante ressaltar que as técnicas utilizadas na análise exploratória podem variar conforme a natureza dos dados e as questões de pesquisa. Portanto, métodos estatísticos adequados são aplicados, seguindo as melhores práticas e considerando a literatura especializada na área de análise exploratória de dados.

Figura 5 – Preocupação acerca da falta de alimentos



Fonte: elaborado pelo autor.

Figura 6 – Porcentagem de famílias atendidas pelo CREAS

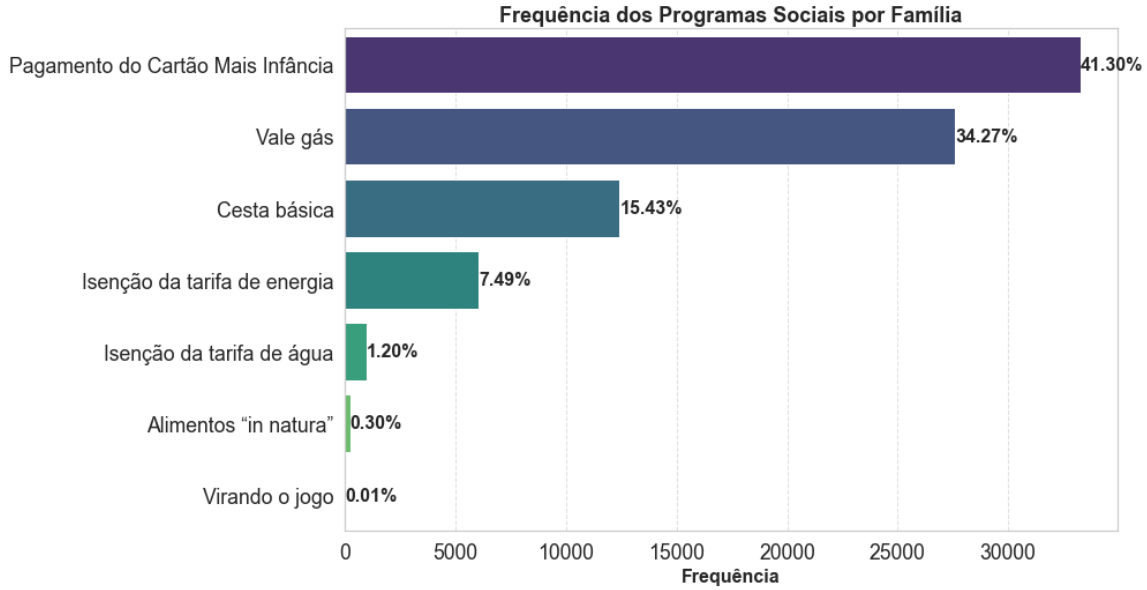


Fonte: elaborado pelo autor.

4.8 Clusterização de Dados

A clusterização representa uma técnica essencial para organizar dados em conjuntos ou *clusters*, com base em similaridades intrínsecas. Esta abordagem facilita a identificação de padrões ou grupos naturais presentes nos dados, conforme ressaltado por Hastie *et al.* (2009). Algoritmos de clusterização, notavelmente o *k-means* ou o DBSCAN, são empregados para categorizar famílias em diferentes grupos, considerando características socioeconômicas e

Figura 7 – Programas sociais estaduais mais frequentes



Fonte: elaborado pelo autor.

demográficas similares.

A aplicação do conjunto de dados em um modelo não supervisionado é uma etapa crucial para permitir que o algoritmo *k-means* analise e agrupe os dados de maneira apropriada. Conforme destacado por Müller e Guido (2017), no pré-processamento dos dados, são necessárias etapas como tratamento de valores ausentes, normalização de variáveis numéricas e codificação de variáveis categóricas. Essas medidas asseguram a consistência e uniformidade dos dados, possibilitando a correta aplicação do algoritmo *k-means*.

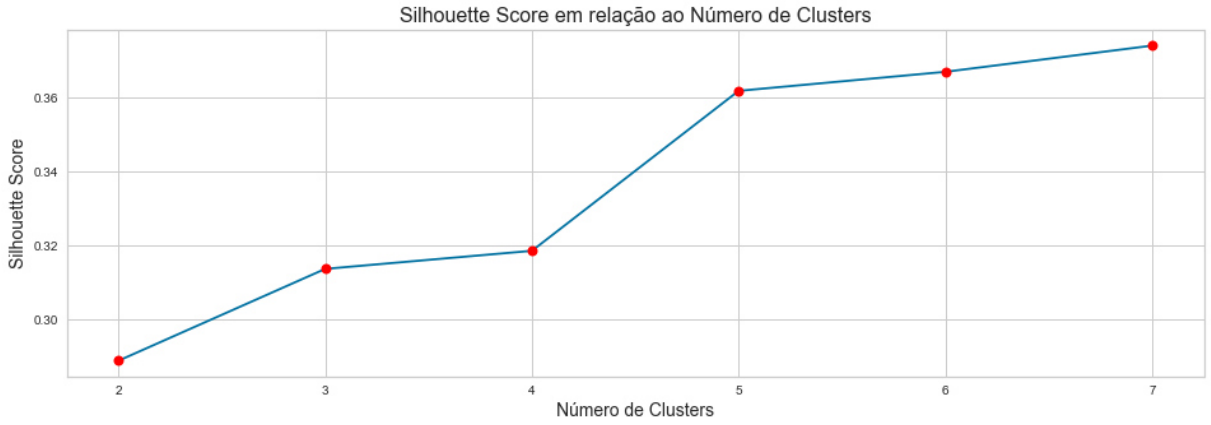
Após o pré-processamento, o conjunto de dados é introduzido no algoritmo de clusterização. Seguindo a abordagem de Raschka e Mirjalili (2017), o algoritmo *k-means* é aplicado aos dados, considerando as variáveis selecionadas como entrada. O objetivo principal é agrupar famílias em *clusters*, destacando suas características socioeconômicas e demográficas semelhantes.

No contexto específico do projeto, o processo para determinar o valor ótimo de k (número de *clusters*) envolve a análise do cotovelo e da silhueta. A análise do cotovelo identifica o ponto no qual o acréscimo no número de *clusters* não resulta em melhoria significativa na variação intracluster. A análise da silhueta avalia coesão e separação entre *clusters*, auxiliando na escolha de k que maximize esses valores. Essas técnicas proporcionam uma abordagem robusta para a determinação do valor adequado de k no contexto da clusterização com *k-means*.

4.9 Aplicação do algoritmo e avaliação

A aplicação do algoritmo *k-means* envolveu a definição do número de *clusters*, *k*, sendo uma decisão importante a ser tomada. De acordo com MacQueen (1967), o valor de *k* pode ser definido com base em conhecimentos prévios do domínio ou utilizando técnicas de validação cruzada, como a validação do cotovelo, que visa identificar o ponto de inflexão na curva de variância explicada pelos *clusters*. Além disso, foi buscado um equilíbrio entre o número de clusters e as métricas de avaliação utilizadas, como silhueta e o método do cotovelo, por exemplo.

Figura 8 – Pontuação de silhueta simulada por número de *clusters*.

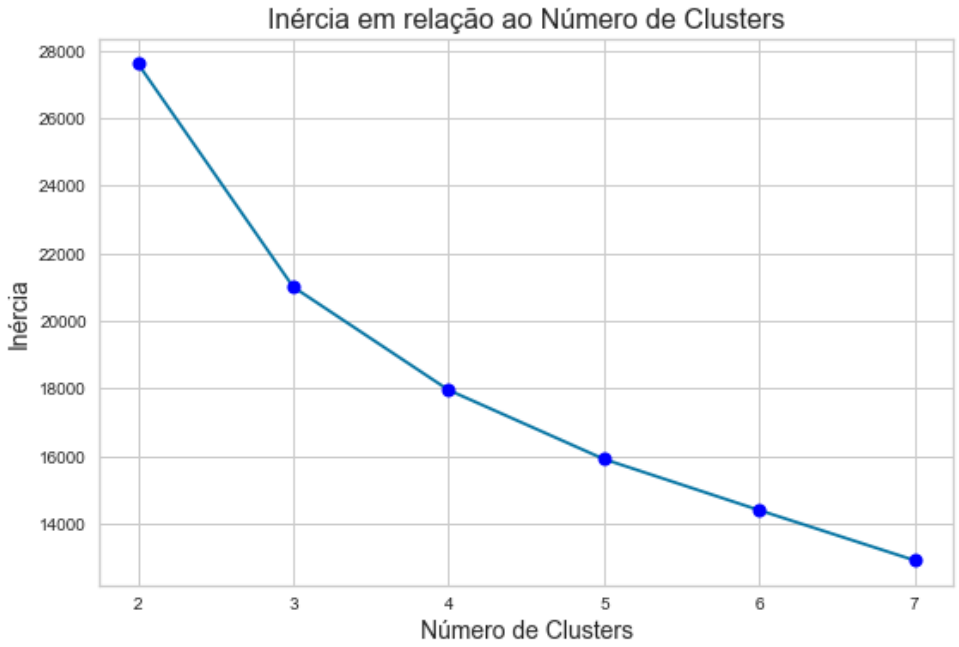


Fonte: Elaborado pelo autor.

Deste modo, seguindo a metodologia de usar gráficos e aplicando-a no *dataframe* utilizado neste projeto, pode-se perceber que na Figura 8 há um perceptível crescimento da pontuação de silhueta entre 2 e 5 *clusters*. A partir deste último valor, há a tendência a um platô, sem aumento considerável. Em sequência, utilizando-se do método do cotovelo (curva de *Elbow*) na Figura 9, observa-se que a curva não apresenta um "cotovelo" distintivo, mas exibe uma transição mais suave. Essa característica pode implicar na complexidade do processo de decisão, sendo necessárias mais métricas.

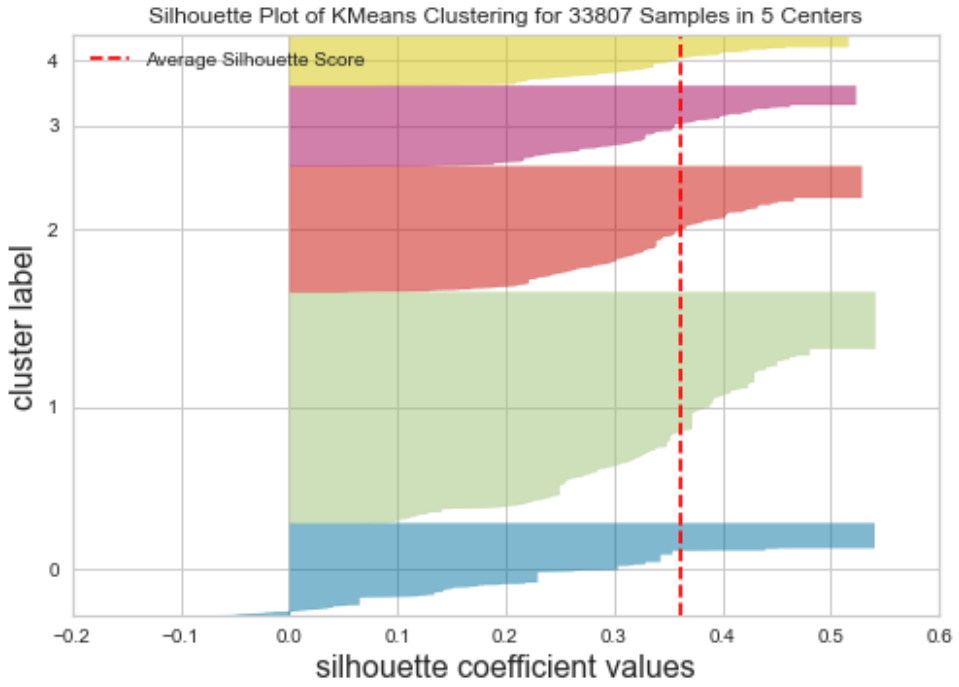
Assim, utilizou-se de outro meio de visualização de silhueta, onde não só o valor numérico é exibido, como também a representação visual dos clusters. Foi possível verificar sua compacticidade. Com base nos gráficos anteriores, utilizou-se de exemplo o *cluster* para $k = 5$ e $k = 6$. Nota-se que na configuração de 5 *clusters*, estes estão adequadamente compactados e com poucos valores de silhueta negativos. Já no contexto com 6 *clusters*, a compacticidade também é apropriada, mas é perceptível a maior quantidade de pontuações de silhueta negativas, padrão

Figura 9 – Pontuação de inércia simulada por número de *clusters* — Método do Cotovelo.



Fonte: Elaborado pelo autor.

Figura 10 – Visualização do *cluster* para $k = 5$

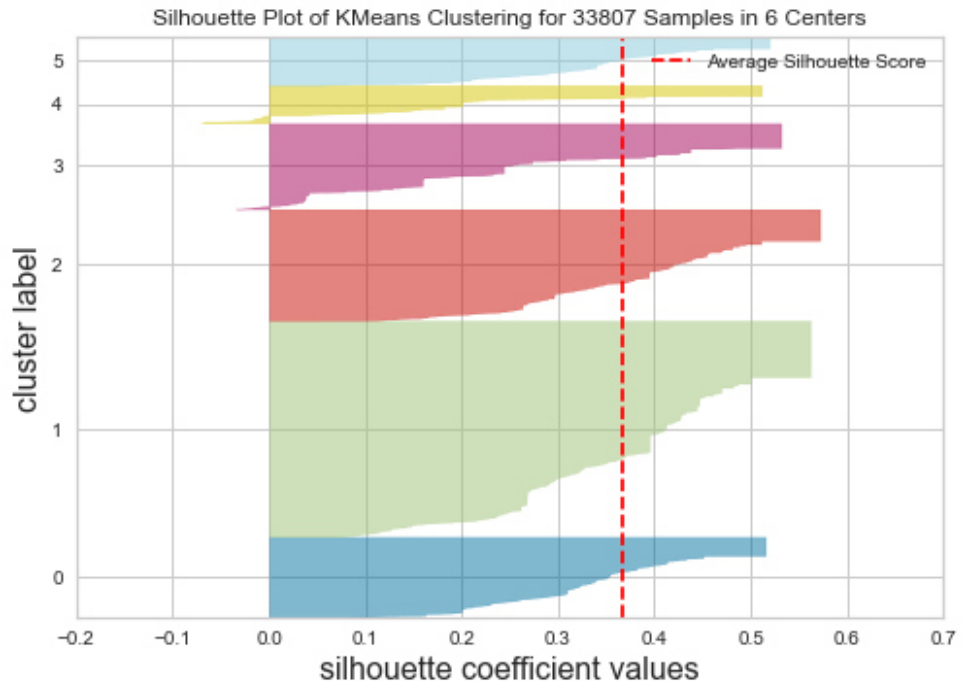


Fonte: elaborado pelo autor.

este também perceptível no intervalo abaixo de 5 e acima deste. Dessa forma, o valor de $k = 5$ apresenta um melhor equilíbrio de densidade de *clusters* e pontuação em métricas, sendo assim escolhido para este projeto.

Após esse período de experimentos, e com o valor 5 para k definido, o algoritmo *k-means* foi aplicado ao conjunto de dados. Durante a execução, os centroides dos *clusters* são

Figura 11 – Visualização do *cluster* para $k = 6$



Fonte: elaborado pelo autor.

atualizados iterativamente até que a convergência seja alcançada. Conforme destacado por Jain (2010), a convergência ocorre quando a posição dos centroides não muda significativamente entre as iterações.

Uma vez que o algoritmo tenha sido executado com o *k-means* e os *clusters*s tenham sido formados, é importante avaliar a qualidade do agrupamento. Para isso, métricas de avaliação de *clusters*, como o coeficiente de silhueta e o índice Davies-Bouldin, podem ser utilizadas. Conforme mencionado por Kaufman e Rousseeuw (1990), o coeficiente de silhueta mede a coesão interna e a separação entre os *clusters*. Já o índice de Davies e Bouldin (1979) avalia a compacidade e a separação dos *clusters*.

Uma vez formados os cinco clusters por meio do algoritmo *k-means*, procedeu-se à avaliação do nível de insegurança alimentar em cada grupo. Esse processo envolveu a consideração dos *loadings* obtidos a partir da Análise de Componentes Principais Principal Component Analysis (PCA). Os *loadings* representam as contribuições de cada variável original para os componentes principais, indicando a importância relativa na explicação da variabilidade dos dados. A fim de quantificar o nível de insegurança alimentar em cada família, foi realizado um processo de cálculo utilizando PCA. O primeiro passo consistiu na aplicação do procedimento matemático às variáveis categóricas codificadas, buscando identificar os pesos relativos de cada variável no primeiro componente principal. Os *loadings* resultantes do PCA indicaram a

importância de cada variável na variabilidade total dos dados. Posteriormente, esses *loadings* foram normalizados para garantir que somassem 1, transformando-os em pesos proporcionais à sua contribuição para o componente principal.

A insegurança alimentar agregada para cada família foi então calculada multiplicando os valores das variáveis codificadas pelos pesos normalizados e somando esses produtos. Essa abordagem, representada pela equação 4.1,

$$\text{Insegurança Alimentar} = \sum_{i=1}^{n} (\text{Variável}_i \times \text{Peso}_i) \tag{4.1}$$

proporcionou um índice que reflete o impacto conjunto das variáveis na insegurança alimentar, considerando seus pesos relativos identificados pelo método. A normalização dos pesos e a ponderação por PCA visam garantir que o índice capture de maneira robusta a complexidade das respostas, atribuindo maior peso às variáveis mais significativas na análise de componentes principais.

Quadro 6 – Insegurança Alimentar por Cluster

Cluster	**Insegurança Média**	**Classificação**
0	0,778	Alta
1	0,273	Baixa
2	0,476	Média
3	0,383	Média/Baixa
4	0,377	Média/Baixa

Fonte: Elaborado pelo autor.

4.10 Análise dos *clusters* e formação dos perfis familiares

Durante a análise detalhada de cada *cluster*, o processo consistiu em cruzar um novo conjunto de dados contendo os rótulos dos *clusters*, os identificadores únicos e anônimos de cada família (*family_id*) e os índices de insegurança alimentar de cada família. Essa abordagem permitiu uma investigação mais profunda e organizada das características específicas de cada grupo, identificando padrões distintos de insegurança alimentar.

Adicionalmente, foram integrados outros *dataframes* relacionados a programas sociais, infraestrutura e trabalho e renda. Esses conjuntos de dados foram previamente processados, incluindo a codificação de variáveis categóricas com o *LabelEncoder* e a normalização com o *MinMaxScaler*, em processo semelhante ao executado no banco de dados onde o *k-means* foi

realizado, garantindo consistência nas análises. Ao correlacionar as variáveis desses *datasets* com o índice de insegurança alimentar, pode-se identificar quais aspectos estavam mais fortemente associados ao aumento ou redução dessa pontuação.

O processo de análise envolveu a formação de gráficos para visualização intuitiva das relações entre as variáveis. Nos *dataframes* de infraestrutura, assistência e renda e trabalho, as respostas estavam disponíveis tanto em seus valores "originais", de forma categórica, quanto codificados. Essa abordagem permitiu uma análise abrangente, utilizando tanto representações visuais quanto métricas matemáticas, na busca por padrões e correlações relevantes. Os resultados dessas análises colaborou para a concepção dos perfis das famílias em situação de carência de alimentos.

4.11 Discussão de resultados e aplicações

A discussão dos resultados obtidos a partir da aplicação do modelo de clusterização de insegurança alimentar envolve uma análise detalhada das métricas de desempenho, bem como da influência das variáveis socioeconômicas e demográficas no agrupamento. Foram identificados os fatores mais relevantes para a ocorrência de insegurança alimentar, fornecendo *insights* sobre os principais determinantes desse problema no contexto específico das famílias do Ceará.

Além disso, foram exploradas as possíveis aplicações práticas dos resultados. O sistema de clusterização desenvolvido pode auxiliar na tomada de decisões relacionadas à segurança alimentar, direcionando políticas públicas e programas de assistência alimentar para as populações mais vulneráveis. A alocação eficiente de recursos e a identificação de famílias em maior risco de insegurança alimentar podem contribuir significativamente para a redução desse problema social.

5 RESULTADOS

Neste capítulo, serão mostrados os resultados dos experimentos de clusterização, destacando a análise socio-demográfica de cada cluster acerca de infraestrutura, programas sociais e renda. Além disso, métricas de avaliação também são discutidas.

5.1 Resultados das avaliações dos *clusters*

A avaliação do modelo *K-Means* revela resultados valiosos sobre a qualidade do agrupamento realizado. A inércia, medida em 15911,28, indica a compactação dos clusters, sendo desejável um valor menor (JR. *et al.*, 2019). Comparativamente, a inércia deve ser considerada em relação à modelos com diferentes números de clusters para identificar possíveis subestruturas nos dados.

A distância média aos centroides, avaliada em 1,27, fornece uma medida da dispersão média dos pontos dentro de cada cluster (Jr *et al.*, 2019). Valores menores indicam que os pontos estão mais próximos dos centroides, o que é um resultado desejável.

A silhueta média, com um valor de 0,36, varia de -1 a 1. Valores mais próximos de 1 indicam que os pontos estão bem agrupados, enquanto valores próximos de 0 indicam sobreposição entre clusters (Kaufman; Rousseeuw, 1990). Assim, 0,36 sugere uma clusterização razoavelmente boa, com pontos bem definidos em seus clusters.

O índice Davies-Bouldin, com um valor de 1,14, mede a "compacidade" e "separação" dos clusters. Quanto menor o índice, melhor, indicando clusters compactos e bem separados. Neste caso, 1,14 é considerado um resultado positivo (Davies; Bouldin, 1979).

Por fim, o índice *Calinski-Harabasz*, com um valor de 11330,95, compara a dispersão entre clusters com a dispersão nos clusters. Valores maiores indicam uma melhor separação entre clusters, indicando uma boa qualidade na formação dos grupos (Calinski; Harabasz, 1974).

Quadro 7 – Resultados das Métricas do Modelo K-Means

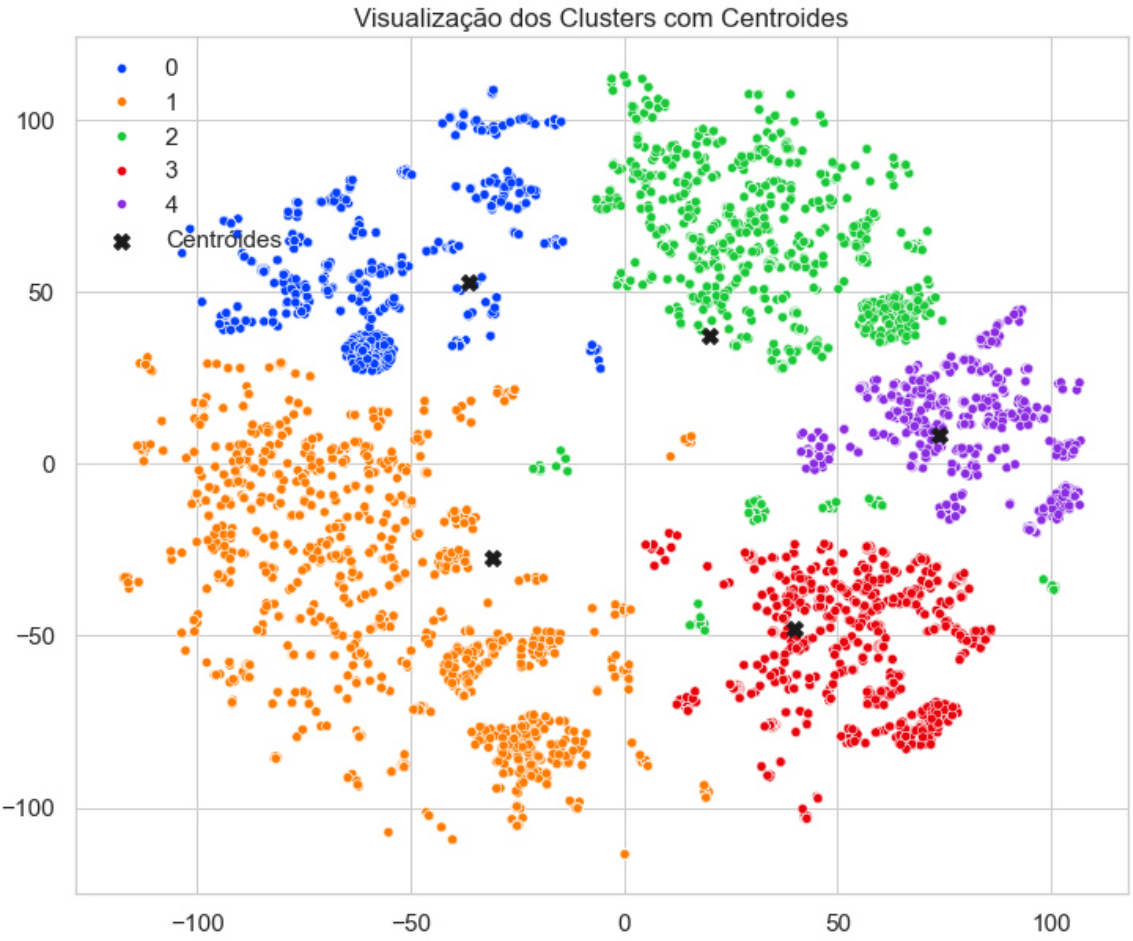
Métrica	**Valor**	**Qualidade**
Inércia	15911,28	Razoável
Distância Média aos Centroides	1,27	Alta
Silhueta Média	0,36	Razoável
Índice Davies-Bouldin	1,14	Boa
Índice Calinski-Harabasz	11330,95	Boa

Fonte: Elaborado pelo autor.

5.1.0.1 Utilização de t-SNE Para Visualização de Clusters

Para uma compreensão mais aprofundada (Figura 12), uma visualização em t-SNE *(t-distributed Stochastic Neighbor Embedding)* pode fornecer percepções adicionais. Esta técnica permite representar dados multidimensionais de forma mais clara e identificar possíveis sobreposições entre os clusters (Maaten; Hinton, 2008).

Figura 12 – Visualização utilizando t-SNE



Fonte: elaborado pelo autor.

É importante observar áreas na visualização onde as instâncias de diferentes clusters se aproximam, indicando similaridades que podem não ter sido completamente capturadas pelo algoritmo de clusterização. Essas sobreposições podem revelar nuances nas características das famílias que não foram totalmente consideradas durante a construção dos clusters.

5.2 Formação dos *Clusters* e suas Características

A aplicação do algoritmo *k-means* resultou na formação de cinco clusters distintos, cada um caracterizado por um perfil socioeconômico específico. Vale destacar que o valor utilizado como referência para o salário mínimo considerado nessa seção é relativo à data da realização da pesquisa CMIC (agosto de 2022), o qual à época era de R$1212,00. A seguir, uma análise detalhada de cada cluster é apresentada com base nos índices de insegurança alimentar:

5.2.1 Cluster 0 - Maior insegurança

Este cluster, composto por 5.390 famílias, exibe uma média de insegurança alimentar de 0,778, com um desvio padrão de 0,138. Observa-se que a média é relativamente alta, indicando um nível significativo de vulnerabilidade. As famílias neste cluster apresentam uma dispersão substancial em seus níveis de insegurança alimentar, conforme evidenciado pelo desvio padrão.

5.2.1.1 Análise da Renda e Trabalho

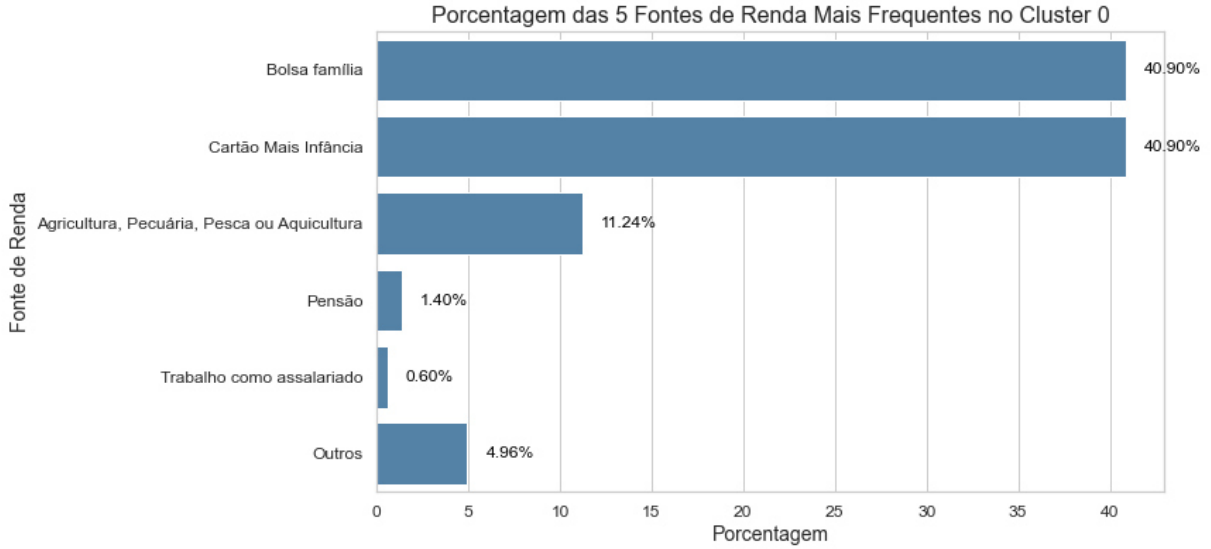
No que diz respeito à renda e ao trabalho, o *Cluster* 0 apresenta características distintivas. A maioria das famílias (86,9%) não possui trabalho remunerado, indicando uma alta dependência de programas sociais. A renda média é de menos de 0,5 salário mínimo (R$ 484,80), destacando uma situação econômica extremamente desafiadora. A diversificação das fontes de renda é limitada, com Bolsa Família e Cartão Mais Infância representando 40,90% cada (Figura 13). A participação em cursos de qualificação é mínima, indicando uma necessidade de ampliação de oportunidades de capacitação.

5.2.1.2 Análise da Infraestrutura

No contexto da infraestrutura, o *Cluster* 0 reflete condições de vida desafiadoras. Embora a maioria das famílias resida em propriedade própria (57,4%), uma parcela significativa mora em residências emprestadas ou cedidas (24%). O tipo de moradia e o material das paredes apresentam correlações positivas fracas com a insegurança alimentar, sugerindo que certos tipos de moradia podem estar associados a níveis mais altos de insegurança.

Quanto ao acesso à água, a maioria das famílias (64,3%) possui água canalizada, indicando um nível mediano de infraestrutura básica (Figura 14). No entanto, a correlação negativa moderada entre o abastecimento de água e a insegurança alimentar sugere que a falta

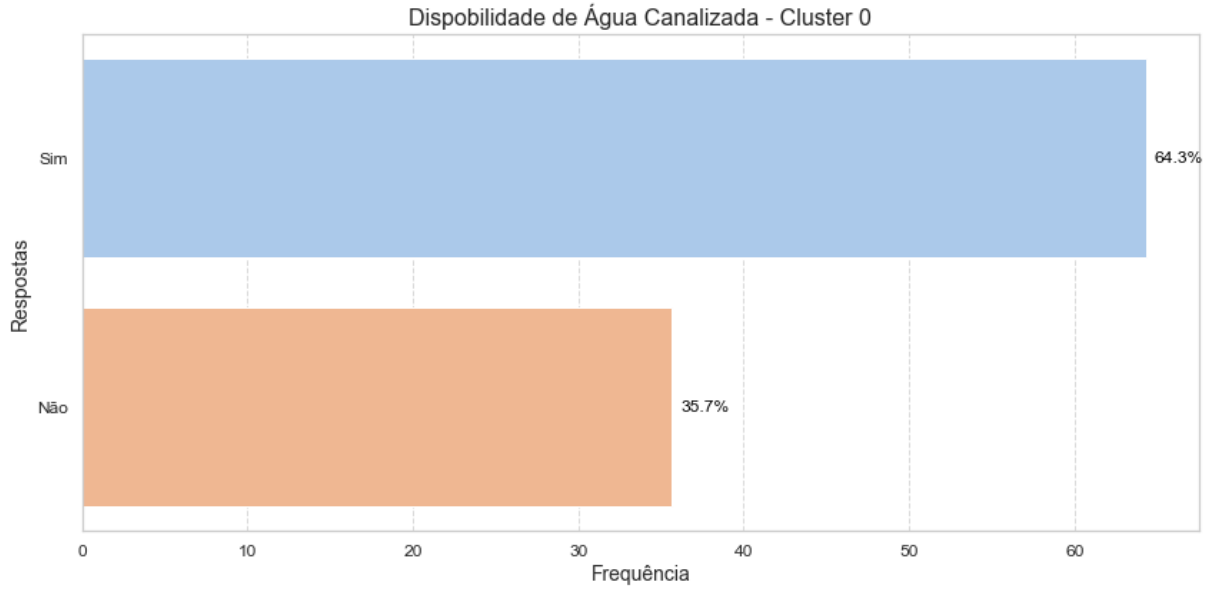
Figura 13 – Principais Fontes de Renda do *Cluster* 0



Fonte: elaborado pelo autor.

de acesso a fontes seguras de água pode agravar a vulnerabilidade alimentar. A diversidade nos métodos de tratamento de água para consumo indica uma conscientização sobre a importância da qualidade da água.

Figura 14 – Presença de água canalizada para famílias do *Cluster* 0



Fonte: elaborado pelo autor.

A presença de banheiro sanitário em casa (82,6%) é positiva para a saúde e o saneamento básico. No entanto, a frequência variada da coleta de lixo, com 32,5% das famílias mencionando nenhuma coleta, é preocupante e pode ter impactos negativos no ambiente e na saúde pública. A iluminação elétrica universal (97,1%) é um aspecto positivo para atividades

diárias e qualidade de vida.

A disponibilidade de lugares públicos para brincadeiras (21,3%), eventos culturais (2,9%), locais para práticas esportivas (17,4%), e a baixa incidência de domicílios em áreas de conflito (12%) são elementos adicionais a serem considerados. A limitada presença de espaços de lazer e entretenimento pode afetar o desenvolvimento social e cultural das famílias neste cluster, enquanto a presença de áreas de conflito pode contribuir para um ambiente mais inseguro.

5.2.1.3 Análise do Acesso à Assistência Estatal

No que diz respeito ao acesso à assistência estatal, o *Cluster* 0 apresenta uma significativa dependência de programas sociais, com destaque para o Cartão Mais Infância, Vale Gás e Cesta Básica. A presença de famílias assistidas pelo CRAS é expressiva (47,3%), enquanto a assistência pelo CREAS é baixa (1,1%). Vale ressaltar que 27,4% das famílias indicam que o município não possui CREAS. Essa alta dependência de programas sociais sugere uma necessidade de fortalecimento das políticas de assistência social. A correlação positiva entre a assistência pelo CRAS e a insegurança alimentar pode indicar a necessidade de revisão e reforço dos programas existentes para melhor atender às demandas específicas desse *cluster*.

5.2.2 Cluster 1 - Baixa insegurança

O segundo cluster, com 13.474 famílias, mostra uma média de insegurança alimentar mais baixa, alcançando 0,273, com um desvio padrão de 0,131. Este cluster apresenta uma menor dispersão em relação ao índice de insegurança alimentar, sugerindo uma maior homogeneidade nas condições socioeconômicas das famílias.

5.2.2.1 Análise da renda e trabalho

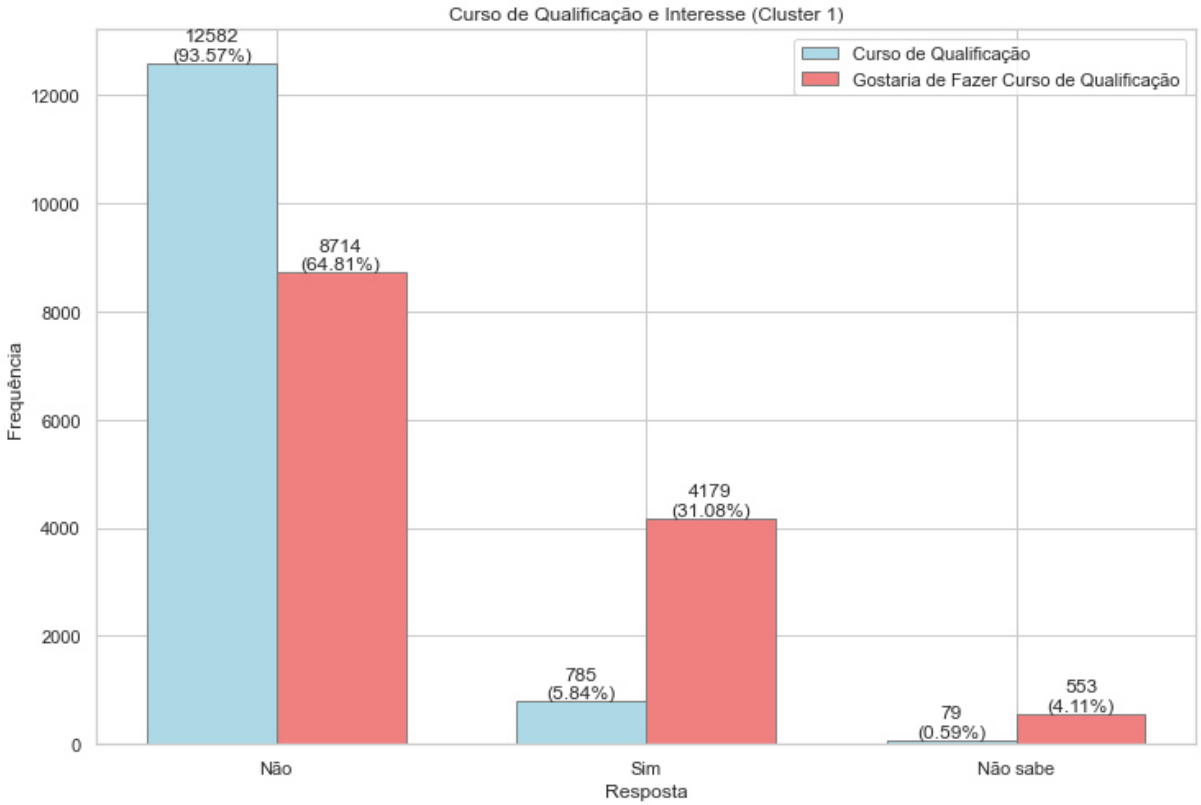
No âmbito da renda e do trabalho, o *Cluster* 1 destaca-se por uma renda média de menos de 0,5 salário mínimo (R$ 521,16). A maioria das famílias (81,3%) não está envolvida em trabalho remunerado, e a correlação entre trabalho remunerado e insegurança alimentar é negativa, indicando que a participação em trabalho remunerado está associada a níveis mais baixos de insegurança alimentar neste cluster.

O número de trabalhadores por família também mostra uma correlação negativa leve com a insegurança alimentar. Famílias com mais membros empregados tendem a apresentar

uma redução na insegurança alimentar. Quanto às fontes de renda, o *Cluster* 1 é caracterizado por uma distribuição diversificada, com destaque para o Bolsa Família e o Cartão Mais Infância como principais fontes.

A participação em cursos de qualificação é baixa (6,43% fizeram cursos), mas a expressiva parcela (31,08%), com interesse em fazê-los, pode indicar uma disposição para aprimoramento profissional, o que pode impactar positivamente na situação financeira das famílias (Figura 15).

Figura 15 – Frequência de formação em cursos de qualificação e interesse no *Cluster* 1



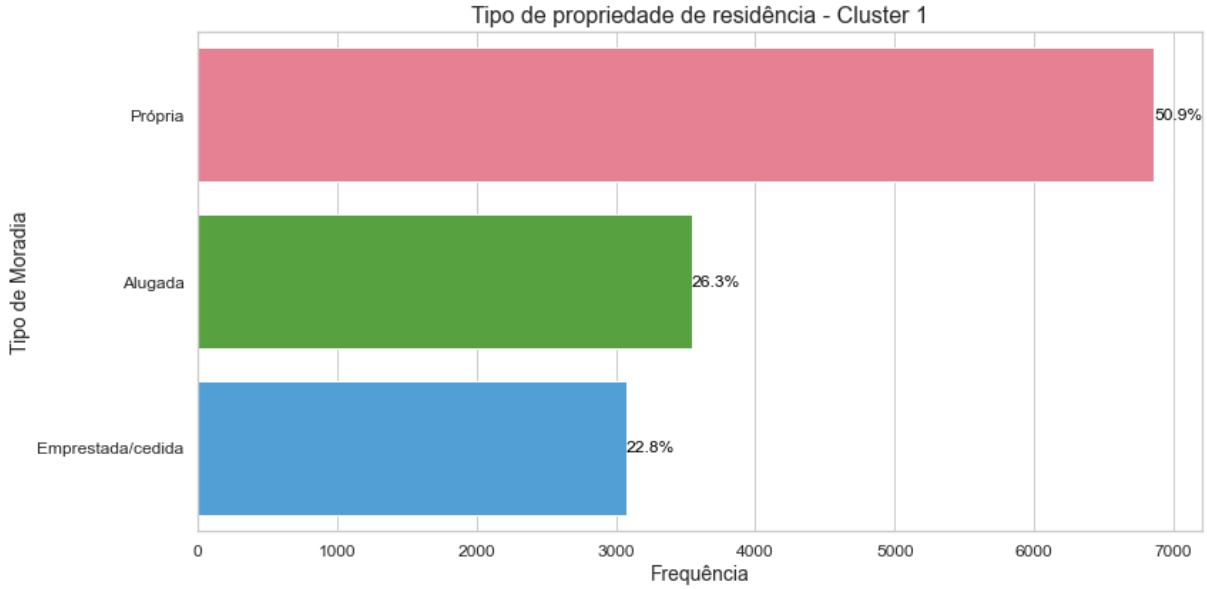
Fonte: elaborado pelo autor.

5.2.2.2 Análise da infraestrutura

O *Cluster* 1 apresenta condições de infraestrutura relativamente melhores em comparação com o *Cluster* 0. A maioria das famílias reside em propriedade própria (50,9%) (Figura 16), e o material das paredes mostra uma correlação positiva fraca com a insegurança alimentar, sugerindo que certos tipos de material de parede, de melhor qualidade, podem estar correlacionados, mas não determinantemente, a níveis mais baixos de insegurança.

O acesso à água é mais amplo, com 74,1% das famílias utilizando a rede geral de

Figura 16 – Situação da posse da residência das famílias do *Cluster* 1



Fonte: elaborado pelo autor.

distribuição. A presença de água canalizada (82,6%) e banheiro sanitário em casa (93%) é significativamente alta. A coleta de lixo é realizada com maior frequência, sendo que 51,4% das famílias relatam que o lixo é coletado de uma a duas vezes por semana.

A existência de lugares públicos para brincadeiras (31,2%), eventos culturais (4,6%), locais para práticas esportivas (26,9%), e a baixa incidência de domicílios em áreas de conflito (11,9%) também são aspectos adicionais a serem contabilizados. A presença de espaços de lazer e entretenimento é mais expressiva neste cluster, o que pode contribuir para um ambiente mais enriquecedor em termos sociais e culturais.

5.2.2.3 Análise do acesso à assistência estatal

A análise do acesso à assistência estatal mostra que a maioria das famílias não recebe assistência do CRAS (56,4%). A presença do CREAS é limitada (0,8%), e uma parcela considerável (22,3%) indica que o município não possui CREAS. Isso pode impactar a oferta de serviços sociais e de apoio às famílias deste cluster.

5.2.2.4 Correlação de Variáveis

A análise das correlações destaca que fatores como trabalho remunerado, número de trabalhadores, fontes de renda, participação em cursos de qualificação, renda mensal, e a própria insegurança alimentar apresentam associações que, em sua maioria, corroboram com o senso

comum. O *Cluster* 1 exibe um padrão onde a participação em atividades remuneradas, maior renda e diversificação nas fontes de renda estão associadas a níveis mais baixos de insegurança alimentar.

No que se refere à infraestrutura, a presença de água canalizada, banheiro sanitário em casa, melhores condições de moradia e frequência regular na coleta de lixo indicam um ambiente mais favorável para o *Cluster* 1 em comparação com o *Cluster* 0.

A assistência estatal, no entanto, é um ponto de atenção, pois a maioria das famílias não recebe assistência do CRAS, e a presença do CREAS é limitada. Isso sugere que, embora as condições de vida sejam melhores em termos de infraestrutura, o suporte social e assistencial pode ser aprimorado para melhorar ainda mais a situação dessas famílias.

5.2.3 Cluster 2 - Média insegurança

Composto por 7.330 famílias, o terceiro cluster exibe uma média de insegurança alimentar de 0,476, e um desvio padrão de 0,144. Este cluster representa uma categoria intermediária em termos de vulnerabilidade, com uma dispersão moderada nos níveis de insegurança alimentar.

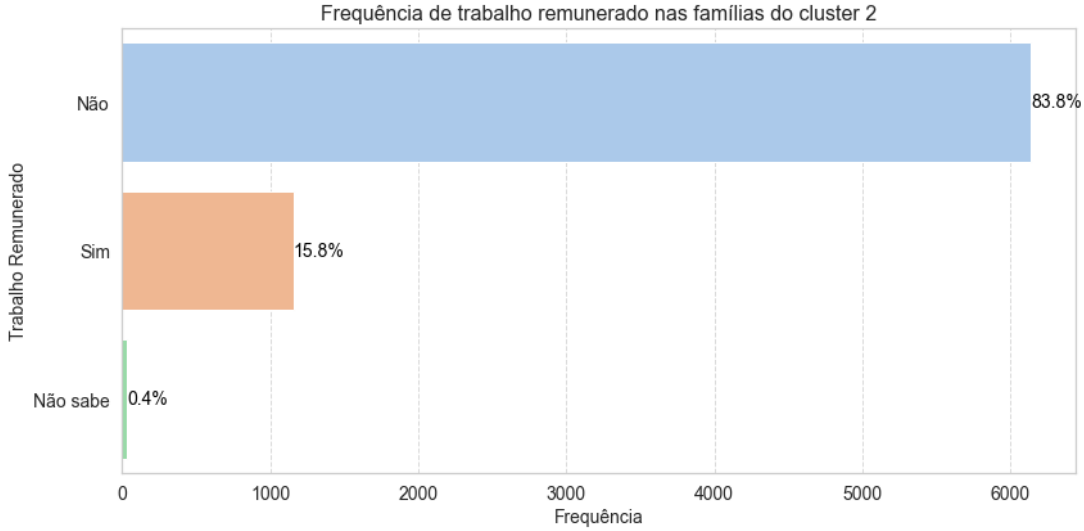
5.2.3.1 Análise da renda e trabalho

No cenário da renda e do trabalho, o *Cluster* 2 destaca-se pela presença de uma renda média inferior a 0,5 salário mínimo (R$ 521,16). A grande maioria das famílias (83,8%) não está envolvida em trabalho remunerado. A correlação negativa entre trabalho remunerado e insegurança alimentar sugere que a participação em atividades remuneradas está associada a níveis mais baixos de insegurança alimentar neste *cluster* (Figura 17).

A quantidade de trabalhadores por família apresenta uma correlação negativa leve com a insegurança alimentar. Nesse sentido, famílias com mais membros empregados tendem a experimentar uma redução na insegurança alimentar. Quanto às fontes de renda, o *Cluster* 2 se caracteriza por uma distribuição diversificada, sendo o Bolsa Família e o Cartão Mais Infância as principais fontes.

Apesar da baixa participação em cursos de qualificação (5,09%), uma parcela considerável (33,5%) expressa interesse em realizá-los, indicando uma disposição para o aprimoramento profissional.

Figura 17 – Presença de trabalhadores remunerados no *Cluster* 2



Fonte: elaborado pelo autor.

5.2.3.2 Análise da infraestrutura

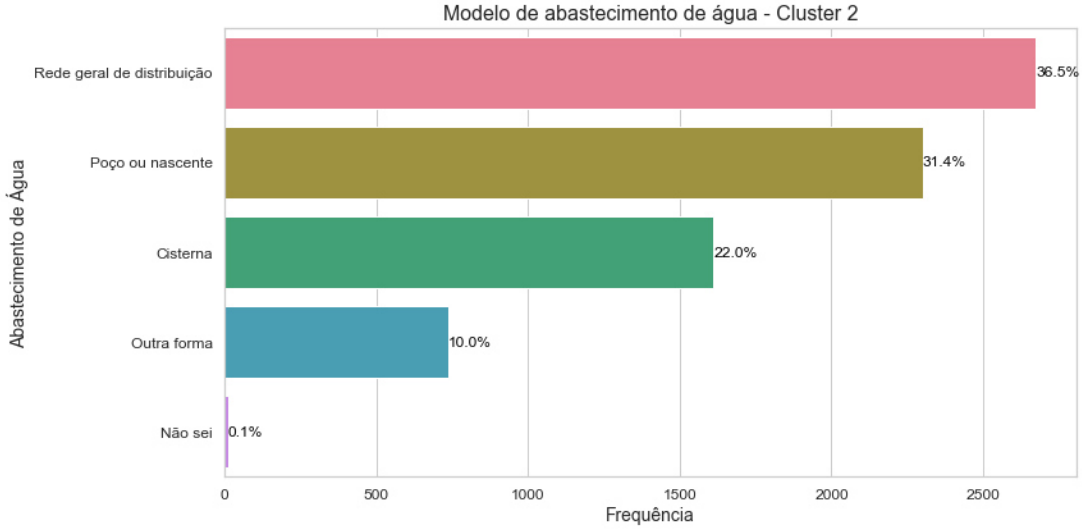
O *Cluster* 2 apresenta condições de infraestrutura intermediárias entre os *Clusters* 0 e 1. No que diz respeito à propriedade das residências, a maioria é própria (67,9%), com uma parcela significativa sendo emprestada/cedida (28,2%). O material das paredes exibe uma correlação positiva fraca com a insegurança alimentar, sugerindo que determinados tipos de material de parede podem estar associados a níveis mais baixos de insegurança.

O abastecimento de água é diversificado, com 36,5% utilizando a rede geral de distribuição, 31,4% por meio de poço ou nascente, 10% de outra forma, e 22% por meio de cisterna. A presença de água canalizada é de 57,4%, e a água para beber é tratada de várias maneiras (Figura 18).

A infraestrutura sanitária é relativamente boa, com 75,3% das famílias possuindo banheiro sanitário em casa. A coleta de lixo é realizada com frequência variada, com 6,1% das famílias indicando mais de duas vezes por semana, 63,8% nenhuma vez, e 30,1% de uma a duas vezes por semana.

O acesso a locais públicos para brincadeiras (9,4%), eventos culturais (2,2%), e locais para práticas esportivas (12,3%) é limitado, e a incidência de domicílios em áreas de conflito é de 3,9%.

Figura 18 – Forma de acesso a água para uso doméstico *Cluster* 2



Fonte: elaborado pelo autor.

5.2.3.3 Análise do acesso à assistência estatal

A análise do acesso à assistência estatal revela que 45,6% das famílias são assistidas pelo CRAS, enquanto apenas 1,1% recebem assistência pelo CREAS. Uma parcela significativa (30,4%) indica que o município não possui CREAS.

5.2.3.4 Correlação de Variáveis

No *Cluster* 2, variáveis como trabalho remunerado, diversificação nas fontes de renda e acesso à infraestrutura básica, especialmente água canalizada, emergem como mais impactantes na redução da insegurança alimentar. Enquanto o interesse em cursos de qualificação é expresso, a baixa participação destaca uma oportunidade para fortalecer programas de capacitação profissional.

Em comparação com os *Clusters* 0 e 1, o *Cluster* 2 representa um grupo intermediário, apresentando aspectos positivos e colaborando para o entendimento mais amplo dos cenários de insegurança alimentar. O entendimento dessas nuances é crucial para direcionar políticas e programas específicos que atendam às necessidades específicas desse grupo, promovendo uma melhoria contínua em sua qualidade de vida.

5.2.4 Cluster 3 - Insegurança média/baixa

O quarto cluster, com 4.682 famílias, apresenta uma média de insegurança alimentar de 0,383 e um desvio padrão de 0,136. Este cluster demonstra uma tendência de menor

vulnerabilidade, com uma dispersão moderada nos níveis de insegurança alimentar.

5.2.4.1 Análise da Renda e Trabalho

As fontes de renda no *Cluster* 3 estão mais diversificadas em comparação com o *Cluster* 0. Bolsa Família e Cartão Mais Infância representam 37,11% cada, enquanto a agricultura, pecuária, pesca ou aquicultura contribuem com significativos 18,65%. A participação em cursos de qualificação é mais expressiva, com 31,98% indicando interesse em fazer cursos. A renda média é de menos de 0,5 salário mínimo (R$ 521,16), sugerindo uma situação econômica ainda desafiadora, praticamente igual do que em clusters de maior insegurança alimentar, demonstrando a estagnação na renda média entre esses agrupamentos.

5.2.4.2 Análise da Infraestrutura

O *Cluster* 3 revela um perfil de famílias com condições de infraestrutura que, em geral, indicam níveis médios a satisfatórios de acesso a recursos básicos. Essas características podem estar associadas a uma maior estabilidade e qualidade de vida para as famílias neste grupo.

No que diz respeito à propriedade de residência, a maioria das famílias (63,4%) possui residência própria. Esse dado sugere uma certa estabilidade habitacional, o que pode contribuir para a sensação de segurança e pertencimento. Em contrapartida, 11,2% das famílias do cluster vivem em residências alugadas, indicando uma diversidade de situações habitacionais.

Ao analisar o material das paredes das residências, a predominância é de alvenaria com revestimento (67,6%). Esse é um indicativo de construção mais sólida e, possivelmente, melhor isolamento térmico, contribuindo para condições de moradia mais confortáveis. Cerca de 22,7% das residências têm paredes de alvenaria sem revestimento, o que, embora represente uma parcela menor, ainda é uma presença significativa.

Quanto ao abastecimento de água, a maioria das famílias (56,5%) tem acesso à rede geral de distribuição. Esse dado sugere um razoável acesso a um serviço essencial. A presença de água canalizada em 68,5% das residências é um indicativo adicional de infraestrutura básica, contribuindo para a comodidade no dia a dia.

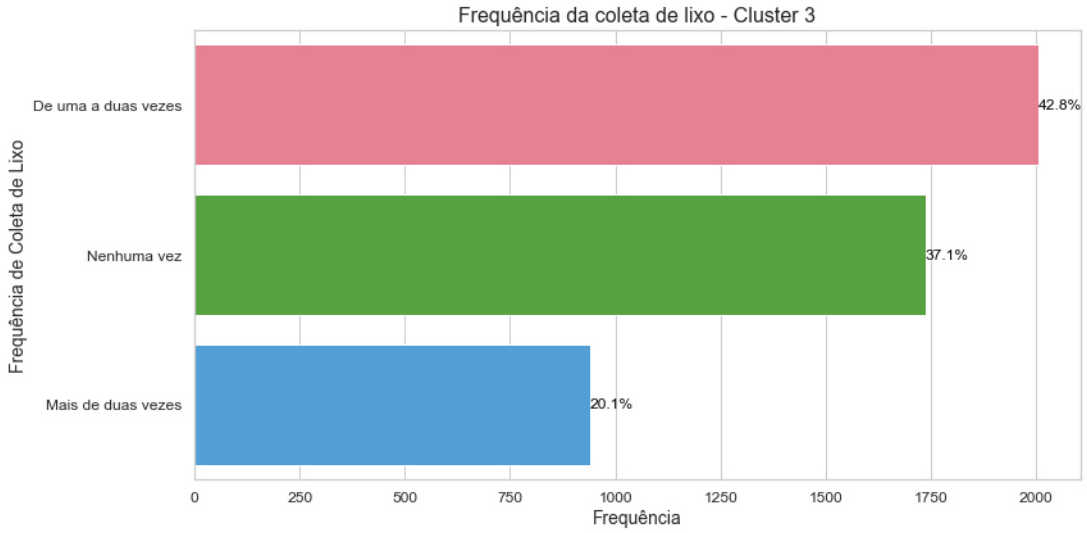
Em relação à água para beber, observa-se que a maioria das famílias (30,2%) consome água sem tratamento, enquanto 25,6% utilizam água filtrada. Esse aspecto pode indicar diferentes padrões de acesso à água potável e destaca a importância de políticas públicas para garantir água

de qualidade a todas as famílias.

A presença de banheiro sanitário em 84,5% das residências é um indicativo positivo de saneamento básico. No entanto, é importante destacar que 15,5% das famílias ainda não contam com esse serviço, o que pode afetar a qualidade de vida e a saúde.

Quanto à coleta de lixo, 42,8% das famílias têm o lixo coletado de uma a duas vezes por semana, indicando uma certa regularidade nesse serviço. No entanto, 37,1% relatam que o lixo não é coletado, o que pode impactar o ambiente local e a saúde das famílias (Figura 19).

Figura 19 – Frequência da coleta de lixo das famílias do *Cluster* 3



Fonte: elaborado pelo autor.

Em termos de iluminação, a quase totalidade das residências (98,7%) possui iluminação elétrica, indicando um bom acesso a essa forma de energia. No entanto, 0,6% ainda utilizam óleo, querosene ou gás para iluminação, ressaltando que alguns lares podem enfrentar desafios no acesso à energia elétrica.

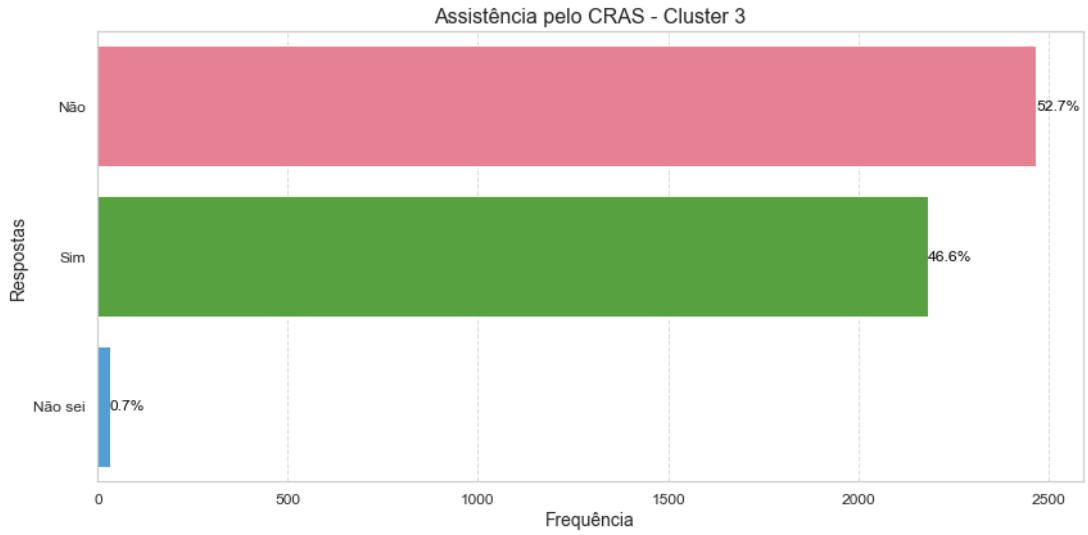
A presença de lugares públicos para brincadeiras (20%), realização de eventos culturais (2,9%) e locais para práticas esportivas (21,9%) sugere que, embora exista acesso a essas atividades, a participação não é generalizada. Isso pode indicar desafios adicionais, como falta de tempo ou recursos.

Em resumo, o *Cluster* 3 apresenta um perfil de infraestrutura que reflete condições melhores em comparação aos *clustrs* anteriores em diversos aspectos, incluindo moradia, abastecimento de água e acesso a serviços essenciais. No entanto, há nuances e desafios específicos, como a coleta de lixo e a participação em atividades culturais e esportivas, que podem ser áreas de foco para melhorias e intervenções sociais.

5.2.4.3 Análise do Acesso à Assistência Estatal

No *Cluster* 3, a assistência estatal é menos proeminente em comparação com o *Cluster* 0. A presença de famílias assistidas pelo CRAS é de 46,6% (Figura 20), enquanto o CREAS assiste apenas 0,9% das famílias. Além disso, 29,3% indicam que o município não possui CREAS. Esses números indicam uma dependência relativamente menor da assistência estatal em comparação com clusters de maior insegurança alimentar.

Figura 20 – Presença da assistência do CRAS no *Cluster* 3



Fonte: elaborado pelo autor.

5.2.4.4 Correlação de Variáveis

A análise de correlações indica que, no *Cluster* 3, a insegurança alimentar está mais levemente associada a fatores como o tipo de moradia, material das paredes, abastecimento de água e a presença de água canalizada. A participação em cursos de qualificação, interesse em cursos e a renda mensal mostram correlações muito fracas com a insegurança alimentar. Destaca-se a correlação positiva muito fraca entre a insegurança alimentar e a existência de áreas com conflitos (91,6% das famílias).

5.2.5 Cluster 4 - Média/baixa insegurança

O último cluster, composto por 2.931 famílias, revela uma média de insegurança alimentar de 0,377 e um desvio padrão de 0,144. Similar ao *Cluster 3*, este cluster sugere uma tendência de menor vulnerabilidade, com uma dispersão moderada nos níveis de insegurança

alimentar.

5.2.5.1 Análise da Renda e Trabalho

No *Cluster* 4, a dinâmica de renda e trabalho revela uma diversidade de fontes de sustento. A presença significativa de famílias sem trabalhadores adicionais (80,83%) destaca uma dependência considerável da renda principal. A participação ativa em programas sociais, como o Bolsa-Família e o Cartão Mais Infância, representa uma parcela substancial da renda média do cluster, indicando a importância desses benefícios como suporte financeiro. A baixa participação em cursos de qualificação (93,51% não fez) pode ser um ponto de intervenção para promover o desenvolvimento profissional, considerando o interesse expresso por uma parcela significativa em futuros cursos (32,59%).

5.2.5.2 Análise da Infraestrutura

O perfil de infraestrutura no *Cluster* 4 destaca condições habitacionais relativamente estáveis, com a maioria das famílias residindo em propriedade própria (64,2%) e utilizando paredes de alvenaria com revestimento (62,3%). A presença de água canalizada em 66,8% das residências é um indicativo positivo de acesso a serviços essenciais. No entanto, a porcentagem significativa de famílias sem banheiro sanitário (14,7%) ressalta desafios persistentes em saneamento básico que requerem atenção.

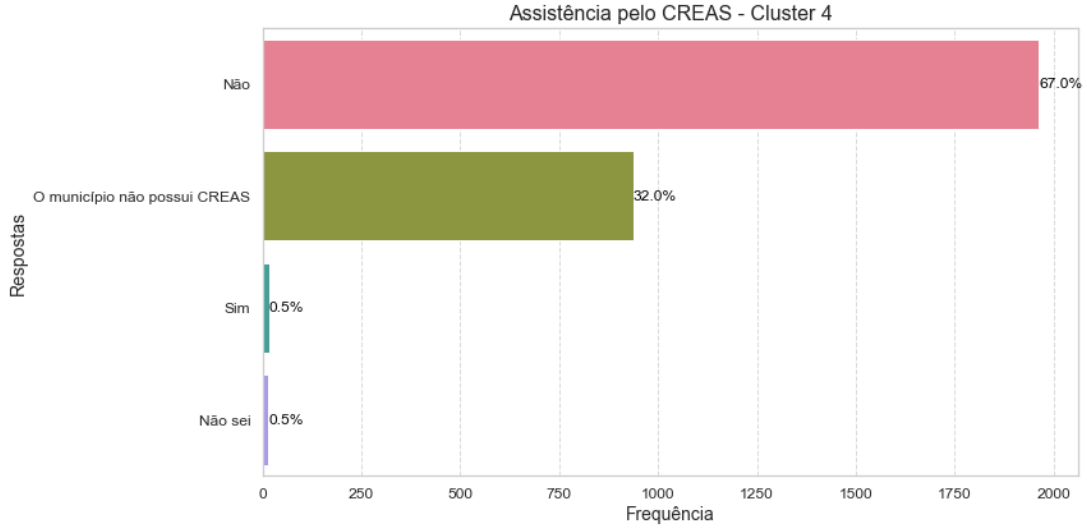
5.2.5.3 Análise do Acesso à Assistência Estatal

A assistência estatal, representada pelo CRAS, alcança 39,6% das famílias no *Cluster* 4. Isso evidencia uma demanda considerável por suporte social adicional. No entanto, a presença limitada do CREAS, com 32% das famílias residindo em municípios sem a instalação, sugerem áreas de melhoria na acessibilidade a esses serviços (Figura 21). Esses dados indicam a necessidade de esforços adicionais para garantir que as famílias estejam cientes e possam acessar integralmente os recursos disponíveis.

5.2.5.4 Correlação de Variáveis — Cluster 4

Ao explorar as correlações, observamos uma relação inversa fraca, mas significativa, entre a insegurança alimentar e variáveis como trabalho remunerado (-0,07), número de trabalha-

Figura 21 – Presença da assistência do CREAS no *Cluster* 4



Fonte: elaborado pelo autor.

dores (-0,07), fontes de renda (-0,08), e renda mensal (-0,11). Esses resultados sugerem que, à medida que essas variáveis aumentam, a insegurança alimentar tende a diminuir. A presença de correlações negativas destaca a importância de abordagens multifacetadas, integrando suporte financeiro e oportunidades de emprego para mitigar a insegurança alimentar no *Cluster* 4.

5.2.6 Perfil Socioeconômico: Clusters em Perspectiva

A análise dos clusters revela diferentes perfis socioeconômicos e demográficos, refletindo nuances nas condições de vida das famílias (Tabela 8).

Quadro 8 – Comparação de Valores entre *Clusters*

Variável	***Cluster 0***	***Cluster 1***	***Cluster 2***	***Cluster 3***	***Cluster 4***
Insegurança Alimentar (Índice)	0,778	0,273	0,476	0,383	0,377
Trabalho Remunerado	13,1%	18,7%	16,2%	16,2%	19,2%
Famílias sem Renda	86,9%	81,3%	83,8%	74,4%	80,8%
Acesso à Rede de Água	64,3%	74,1%	36,5%	56,5%	66,8%
Banheiro em Casa	82,6%	93,0%	75,3%	84,5%	85,3%
Assistência pelo CRAS	47,3%	56,4%	45,6%	46,6%	39,6%
Assistência pelo CREAS	1,1%	0,8%	1,1%	0,9%	0,4%

Fonte: Elaborado pelo autor.

Cluster 0 - Extrema Insegurança:

Este cluster destaca-se por um cenário de extrema insegurança alimentar. As condições precárias de habitação, com moradias improvisadas e a falta de acesso a serviços básicos,

indicam uma vulnerabilidade financeira significativa. A dependência de programas sociais, como o Cartão Mais Infância e Vale Gás, evidencia a necessidade urgente de suporte econômico para essas famílias.

Cluster 1 - Menor Insegurança:

Embora ainda enfrente insegurança alimentar, o Cluster 1 mostra sinais de melhoria em comparação com o Cluster 0. Residências mais estáveis e maior participação em atividades sociais e culturais sugerem um maior investimento público em proporcionar estas experiências. A presença de membros engajados em trabalhos remunerados indica uma busca ativa por estabilidade financeira, apesar da dependência contínua de programas sociais.

Cluster 2 - Média Insegurança:

O Cluster 2 representa um grupo com nível intermediário de insegurança alimentar. Condições de moradia melhores, com aumento na propriedade de residências, indicam uma melhoria gradual. No entanto, a participação moderada em atividades sociais e culturais, juntamente com a persistente dependência de programas sociais, destaca desafios econômicos persistentes que precisam ser abordados.

Cluster 3 - Média/Baixa Insegurança:

Este cluster evidencia um nível de insegurança alimentar intermediário com tendências para um quartil mais baixo. A diversidade nos materiais de construção das moradias, embora indicativa de algumas melhorias, ainda carece de condições básicas. A pronunciada dependência do CRAS destaca a necessidade de intervenções sociais. A limitada participação em atividades sociais e culturais e em trabalhos remunerados reflete desafios substanciais que precisam ser enfrentados.

Cluster 4 - Média/Baixa Insegurança (Complexidade Única):

O Cluster 4 se destaca pela variação na insegurança alimentar, indicando contextos socioeconômicos diversos. A diversidade nas condições habitacionais e na participação em programas sociais sugere uma complexidade única neste grupo. É crucial explorar mais profundamente essas nuances para entender as dinâmicas específicas que contribuem para a insegurança

alimentar variada neste cluster.

Relações entre Variáveis e Insegurança Alimentar:

Material das Paredes e Insegurança Alimentar:

A presença significativa de taipa sem revestimento no Cluster 3 está correlacionada com maior insegurança alimentar, destacando a influência direta das condições habitacionais na vulnerabilidade alimentar. Estratégias de melhoria nas condições de moradia podem ser chave para reduzir a insegurança alimentar nesse grupo.

Participação em Atividades Sociais e Culturais e Insegurança Alimentar:

Clusters com maior participação nessas atividades, como o Cluster 1, tendem a apresentar menor insegurança alimentar. Isso sugere que o engajamento social desempenha um papel crucial na melhoria das condições de vida, enfatizando a importância de iniciativas comunitárias.

Dependência de Programas Sociais e Insegurança Alimentar:

A forte correlação entre a dependência de programas sociais, como o Cartão Mais Infância e Vale Gás, e a insegurança alimentar destaca a importância desses programas na mitigação dos desafios econômicos. Reforçar e expandir esses programas pode ser vital para apoiar os clusters mais vulneráveis.

Quadro 9 – Variáveis Mais Impactantes para Insegurança Alimentar

Variável	**Impacto na Insegurança Alimentar**
Material das Paredes	Cluster 3: Taipa sem revestimento
Participação em Atividades Sociais e Culturais	Cluster 1: Maior participação, menor insegurança
Dependência de Programas Sociais	Todos os Clusters: Relação significativa

Fonte: Elaborado pelo autor.

6 CONCLUSÃO

Recapitulação dos Principais Achados

Ao longo desta pesquisa, mergulhou-se na análise dos diferentes *clusters* que emergiram dos dados, proporcionando uma compreensão abrangente das variáveis socioeconômicas e demográficas associadas à insegurança alimentar. Cada *cluster* apresentou nuances específicas, destacando disparidades marcantes nas condições de vida das famílias investigadas.

Cluster 4: Uma Complexidade a Ser Desvendada

Dentre os *clusters* identificados, o *Cluster* 4 se destaca como uma entidade única e complexa. Sua heterogeneidade aponta para realidades socioeconômicas diversas, exigindo uma análise mais aprofundada para desvendar as variáveis-chave que contribuem para essa complexidade.

Variáveis Determinantes na Insegurança Alimentar

A análise revelou que as condições habitacionais desempenham um papel crucial na vulnerabilidade alimentar, sendo o material das paredes uma variável significativa. Moradias precárias, como taipa sem revestimento, foram associadas a níveis mais elevados de insegurança alimentar.

Paralelamente, observa-se que a participação em atividades sociais e culturais surge como um fator positivo, indicando que o engajamento comunitário pode desempenhar um papel na melhoria das condições de vida. A dependência de programas sociais, como o Cartão Mais Infância e Vale Gás, destaca-se como um elemento crucial na mitigação dos desafios econômicos enfrentados por essas famílias.

Necessidade de Abordagens Específicas diante da Diversidade

A conclusão extraída das análises realizadas reforça a imperatividade de estratégias distintas e sensíveis que considerem as particularidades de cada agrupamento identificado. É patente que não há uma panaceia aplicável universalmente, tornando essencial a formulação de políticas públicas que sejam eficazes ao incorporar a diversidade inerente aos contextos delineados nos clusters.

Nesse sentido, a complexidade revelada pelos dados demanda uma abordagem cuidadosa e embasada. As soluções não podem ser generalizadas, visto que cada agrupamento reflete dinâmicas socioeconômicas e demográficas singulares. Assim, a eficácia de políticas públicas na mitigação da insegurança alimentar está diretamente vinculada à sua capacidade de adaptação a diferentes realidades.

Ausência de Solução Padrão e a Relevância da Sensibilidade Contextual

No âmbito das políticas públicas, torna-se perceptível que não existe uma abordagem única que possa resolver de maneira abrangente os desafios apresentados pelos diversos *clusters* identificados. A heterogeneidade das condições de vida e a multiplicidade de fatores associados à insegurança alimentar requerem uma análise detalhada de cada contexto específico.

A necessidade de políticas públicas sensíveis contextualmente é incontestável. A uniformidade de intervenções pode resultar em equívocos e na subutilização de recursos, enfraquecendo a efetividade das ações governamentais. Portanto, para que se alcance êxito na promoção da segurança alimentar, é fundamental que as políticas sejam moldadas conforme as particularidades de cada agrupamento.

Trabalhos Futuros

Embora esta pesquisa forneça uma visão aprofundada das dinâmicas da insegurança alimentar na região estudada, há espaço para investigações futuras que possam expandir e aprimorar o conhecimento nesta área. Alguns caminhos para trabalhos futuros incluem:

- **Análise Longitudinal:** Realizar estudos longitudinais para compreender as mudanças ao longo do tempo nas condições socioeconômicas e nos níveis de insegurança alimentar. Isso permitiria uma análise mais dinâmica e a identificação de tendências.
- **Intervenções Eficazes:** Investigar a eficácia de intervenções específicas, sejam elas programas sociais, iniciativas comunitárias ou políticas governamentais, no combate à insegurança alimentar em diferentes *clusters*.
- **Impacto das Mudanças Climáticas:** Explorar como as mudanças climáticas podem impactar a segurança alimentar na região, considerando fenômenos climáticos extremos e variações sazonais.
- **Abordagem Multidisciplinar:** Promover estudos multidisciplinares que integrem diferentes campos, como economia, nutrição, saúde pública e ciências sociais, para obter uma

compreensão mais abrangente e holística da insegurança alimentar.

– **Envolvimento Comunitário:** Avaliar o papel do envolvimento comunitário no desenvolvimento de soluções sustentáveis, incentivando a participação ativa das comunidades na busca por estratégias eficazes.

Limitações da Pesquisa

Este estudo, apesar de contribuir para a compreensão da insegurança alimentar na região, apresenta algumas limitações que devem ser consideradas:

– **Generalização Restrita:** Os resultados são específicos para a população estudada e podem não ser generalizáveis para outras regiões com características socioeconômicas diferentes.

– **Dependência de Dados Autodeclarados:** A pesquisa dependeu, em grande parte, de dados autodeclarados pelos participantes, o que pode introduzir vieses ou imprecisões.

– **Contexto Temporal e Geográfico:** O estudo está situado em um contexto temporal e geográfico específico, e as condições podem ter evoluído desde então. Outras regiões podem ter desafios distintos.

– **Complexidade Multifatorial:** A insegurança alimentar é um fenômeno complexo influenciado por várias variáveis, e a análise pode não ter abrangido todas as nuances envolvidas.

Ao considerar essas limitações, é essencial que pesquisas futuras abordem essas lacunas e refinem ainda mais o entendimento sobre a insegurança alimentar, contribuindo para a formulação de estratégias mais eficazes e contextuais.

Consideração Final

Em síntese, esta pesquisa não apenas auxiliou no entendimento sobre as interseções entre variáveis socioeconômicas e insegurança alimentar, mas também destacou a necessidade contínua de estudos mais aprofundados e políticas específicas para enfrentar os desafios complexos que permeiam essa questão. Ao enfrentar esses desafios com uma abordagem informada e sensível, é possível contribuir para a construção de comunidades mais dignas e humanizadas. A história nos ensina que catástrofes como A Grande Fome não devem ser esquecidas, mas sim utilizadas como impulsores para ações presentes e futuras. O povo cearense, e, por extensão, o povo nordestino, merece uma realidade onde a prosperidade não é apenas um sonho, mas uma certeza. Ao compreender as nuances da insegurança alimentar e agir de acordo, pode-se

pavimentar o caminho para uma região mais resiliente e próspera, onde cada indivíduo tem garantido o direito fundamental à alimentação e dignidade.

REFERÊNCIAS

BARBOSA, R. M.; NELSON, D. R. The use of support vector machine to analyze food security in a region of brazil. **Applied Artificial Intelligence**, v. 30, n. 4, p. 318–330, 2016.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics - Theory and Methods**, v. 3, n. 1, p. 1–27, 1974.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, n. 2, p. 224–227, 1979.

ESTER, M.; KRIEGEL, H.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)**. [*S. l.*: *s. n.*], 1996. p. 226–231.

EVERITT, B.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster Analysis**. 5th. ed. [*S. l.*]: John Wiley & Sons, 2011.

FERNANDES, A. F.; FIEDLER, R.; SILVEIRA, A. L. da. Application of machine learning techniques for food insecurity analysis: A systematic review. **Agronomy**, v. 11, n. 7, p. 1433, 2021.

FIELD, A.; MILES, J.; FIELD, Z. **Discovering Statistics Using R**. [*S. l.*]: SAGE Publications Ltd., 2012.

Food and Agriculture Organization of the United Nations. **The State of Food Security and Nutrition in the World 2019**. 2019.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining**: Concepts and techniques. [*S. l.*]: Morgan Kaufmann, 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning:**: Data mining, inference, and prediction. [*S. l.*]: Springer, 2009.

HAWKINS, D. M.; BRADU, D.; KASS, G. V. Location of several outliers in multiple-regression data using elemental sets. **Technometrics**, v. 22, n. 4, p. 427–431, 1980.

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing in Science & Engineering**, v. 9, n. 3, p. 90–95, 2007.

IBGE, I. B. de Geografia e E. **Censo Demográfico 2010**: Características gerais dos domicílios e dos moradores - brasil. 2019. Acesso em: 12 abr. 2023. Disponível em: https://censo2010.ibge.gov.br/resultados.html.

INCE, D. C. *et al.* Yellowbrick v1.3—visualizer. **Journal of Open Source Software**, v. 6, n. 60, p. 3131, 2021.

INMET, I. N. de M. **Normais Climatológicas do Brasil 1981-2010**. 2020. Acesso em: 12 abr. 2023. Disponível em: https://portal.inmet.gov.br/normais.

International Food Policy Research Institute. **Global Hunger Index 2021**. 2021.

IPECE, I. de Pesquisa e Estratégia Econômica do C. Perfil básico municipal do ceará 2022. 2022.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651–666, 2010.

JR, J. F. H.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E. **Multivariate Data Analysis**. 7. ed. [*S. l.*]: Bookman, 2019.

JR., W. C. *et al.* Machine learning for clustering: A review. **Journal of Big Data**, Springer, v. 6, n. 1, p. 1–38, 2019.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data**: An introduction to cluster analysis. [*S. l.*]: Wiley-Interscience, 1990.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: **Proceedings of the 14th International Joint Conference on Artificial Intelligence**. [*S. l.*: *s. n.*], 1995. v. 2, p. 1137–1143.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. [*S. l.*]: Springer, 2013.

KUMAR, A. **Machine learning algorithm types vis-a-vis real-world applications**. 2020. Acesso em: 10 jul. 2023. Disponível em: https://vitalflux.com/great-mind-maps-for-learning-machine-learning/.

LENTZ, E. C.; MICHELSON, H.; BAYLIS, K.; ZHOU, Y. A data-driven approach improves food insecurity crisis prediction. **World Development**, v. 122, 2019.

LLOYD, S. Least squares quantization in pcm. **IEEE Transactions on Information Theory**, v. 28, n. 2, p. 129–137, 1982.

MAATEN, L. v. d.; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, p. 2579–2605, 2008.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics**. [*S. l.*]: University of California Press, 1967. p. 281–297.

MARTINI, G.; BRACCI, A.; RICHES, L. *et al.* Machine learning can guide food security efforts when primary data are not available. **Nat Food**, v. 3, p. 716–728, 2022.

MCKINNEY, W. pandas: Powerful data structures for data analysis, time series, and statistics. **Journal of Open Source Software**, v. 6, n. 60, p. 2302, 2021.

MENDES, L. M.; LIMA, A. S. de; GUERRA, G. A situação da agricultura no semiárido cearense: uma análise dos sistemas agrícolas. **Revista de Política Agrícola**, v. 29, n. 3, p. 61–76, 2020.

MITCHELL, T. M. **Machine Learning**. [*S. l.*]: McGraw Hill, 1997.

MONTEIRO, J. S.; ASSIS, M. M. A. de; KONSTANTYNER, T.; TADDEI, J. A. A. C.; SILVA, L. A. da. Segurança alimentar e nutricional, saúde e desenvolvimento infantil no brasil. **Revista de Nutrição**, v. 31, n. 6, p. e180027, 2018.

MüLLER, A. C.; GUIDO, S. **Introduction to Machine Learning with Python**: A guide for data scientists. [*S. l.*]: O'Reilly Media, 2017.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PINHEIRO, A. R. R.; LUIZ, R. R.; COSTA, T. H. M. da. Políticas públicas de segurança alimentar e nutricional no brasil. **Revista de Nutrição**, v. 29, n. 1, p. 165–176, 2016.

PINSTRUP-ANDERSEN, P. **Food Security**: A comprehensive framework for action. [*S. l.*]: Princeton University Press, 2009.

POUDEL, S.; TSUKAMOTO, H.; KOYAMA, T. Application of machine learning for food security assessment: A systematic review. **Computers and Electronics in Agriculture**, v. 182, p. 105974, 2021.

QUEIROZ, R. d. **O Quinze**. [*S. l.*]: José Olympio, 1930.

RASCHKA, S.; MIRJALILI, V. **Python Machine Learning**. [*S. l.*]: Packt Publishing, 2017.

RIBEIRO, L. P. M. **Clusterização K-Means Paralelo Aplicado na Classificação de Alvos em Imagens de Alta Resolução**. 2017. Acesso em: 05 jul. 2023. Disponível em: http://wiki.dpi.inpe.br/lib/exe/fetch.php?media=cap-378-topicos:relatorio_luis_2017.pdf.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, v. 20, p. 53–65, 1987.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, v. 323, n. 6088, p. 533–536, 1986.

SILVA, A. H. A seca de 1877-1879 e suas representações. **História: Questões & Debates**, v. 42, n. 1, p. 11–41, 2005.

SILVA, J. S. Políticas públicas de convivência com o semiárido: Uma análise da experiência do ceará. **Revista Econômica do Nordeste**, v. 49, n. 1, p. 27–45, 2018.

TABACHNICK, B. G.; FIDELL, L. S. **Using Multivariate Statistics**. 6. ed. [*S. l.*]: Pearson Education, 2013.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 63, n. 2, p. 411–423, 2001.

TUKEY, J. W. **Exploratory Data Analysis**. [*S. l.*]: Addison-Wesley, 1977.

WASKOM, M. Seaborn: Statistical data visualization. **Journal of Open Source Software**, v. 6, n. 60, p. 3021, 2021.

WATKINS, C. J. C. H.; DAYAN, P. Q-learning. **Machine Learning**, v. 8, n. 3-4, p. 279–292, 1992.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining**: Practical machine learning tools and techniques. [*S. l.*]: Morgan Kaufmann, 2016.