Revista Eletrônica de Iniciação Científica em Computação, 2026, 24:1
ISSN: 3085-8461 • doi: 10.5753/reic.2026.7207
This work is licensed under a Creative Commons Attribution 4.0 International License.

Artigo de Pesquisa/Research Paper

Aplicação de Métodos de Aprendizado de Máquina no Contexto Nutricional na Região do Seridó/RN

Application of Machine Learning Methods for Nutritional Assessment in the Seridó/RN Region

Alec Can Yalçin [Universidade Federal do Rio Grande do Norte | *alec.yalcin.700@ufrn.edu.br*]
Karliane Medeiros Ovidio Vale [Universidade Federal do Rio Grande do Norte | *karliane.vale@ufrn.br*]
Flavius da Luz e Gorgônio [Universidade Federal do Rio Grande do Norte | *flavius.gorgonio@ufrn.br*]
Raí Nabichedí da Silva [Universidade Federal do Rio Grande do Norte | *rai.nabichedi@gmail.com*]

Bacharelado em Sistemas de Informação, Universidade Federal do Rio Grande do Norte, Campus Caicó, Rua Joaquim Gregório, 296, Penedo, Caicó–RN, 59300-000, Brasil.

Resumo. A identificação de padrões alimentares na nutrição tem incorporado a mineração de dados para analisar grandes volumes de informação. Este estudo exploratório analisou 407 habitantes da região do Seridó para buscar padrões de consumo de frutas e sua associação com variáveis sociodemográficas, antropométricas e de preferências pessoais. O método envolveu formulários estruturados, pré-processamento e os *k-means* e Apriori para identificar agrupamentos e regras de associação nos dados. Os resultados revelaram três padrões de consumo que seguem uma ordem hierárquica, influenciados por idade e horários de consumo, nos quais os grupos apresentaram avaliações progressivamente maiores para as frutas. O consumo de frutas com caroço (ameixa, cereja, pêssego e caqui) associou-se fortemente à aceitação de outros grupos de frutas. Conclui-se que idade, preferências pessoais e horários de consumo são preditores de dietas com maior variedade de frutas dentro da população documentada, fornecendo uma base teórica para análises nutricionais regionalizadas.

Abstract. The identification of dietary patterns in nutrition has incorporated data mining to analyze large volumes of information. This exploratory study analyzed 407 inhabitants of the Seridó region to investigate fruit consumption patterns and their association with sociodemographic, anthropometric, and personal preference variables. The method involved structured forms, data pre-processing, and the use of *k-means* and Apriori algorithms to identify clusters and association rules. The results revealed three hierarchical consumption patterns influenced by age and consumption times, in which the groups showed progressively higher fruit ratings. The consumption of stone fruits (plum, cherry, peach, and persimmon) was strongly associated with the acceptance of other fruit groups. It is concluded that age, personal preferences, and consumption times are predictors of diets with greater fruit variety within the studied population, providing a theoretical basis for regionalized nutritional analyses.

Palavras-chave: Mineração de Dados, Aprendizado de Máquina, Algoritmos de Agrupamento, Algoritmos de Regras de Associação, Nutrição, Padrão Alimentar

Keywords: Data Mining, Machine Learning, Algorithms, Clustering, Association Rules, Nutrition, Dietary Patterns

Recebido/Received: 16 December 2025 • **Aceito/Accepted:** 03 June 2026 • **Publicado/Published:** 12 June 2026

1 Introdução

A mineração de dados é o processo automático ou semiautomático de explorar analiticamente grandes bases de dados, com a finalidade de descobrir padrões relevantes que ocorrem nos dados [Silva *et al*., 2017]. No campo da nutrição, essa área tem sido aplicada para identificar comportamentos alimentares presentes em diferentes recortes populacionais, permitindo uma compreensão multifatorial do consumo alimentar [Mewes *et al*., 2021; Lazarou *et al*., 2012]. No cenário brasileiro, contudo, a utilização das técnicas de mineração de dados no contexto nutricional ainda é escassa [Silva *et al*., 2022]. Embora a mineração de dados seja eficaz em encontrar padrões abrangentes nesses cenários, tais análises negligenciam contextos regionais, fortemente ligados aos hábitos alimentares da população brasileira [Philippi, 2014]

Nesse sentido, os padrões alimentares variam entre os estados brasileiros [Alves *et al*., 2019; Souza *et al*., 2008], uma vez que o consumo é moldado pela interseção entre a disponibilidade local de alimentos e as particularidades socioeconômicas e culturais inerentes a cada território [Enthoven and Van den Broeck, 2021]. O interesse por entender precisamente como isso influencia o consumo de certos alimentos é frequente na nutrição, principalmente quando se trata do grupo de alimentos in natura ou minimamente processados, como o grupo das frutas [Martins *et al*., 2014, 2020; Donatti *et al*., 2023].

De acordo com Botelho [2006] regiões como o Seridó Potiguar apresentam uma diversidade de alimentos distinta da encontrada em ambientes metropolitanos, o que consolida uma individualidade dietética que diverge dos hábitos observados em outros centros urbanos. Diante dessa especificidade regional, a literatura vigente sobre mineração de dados, na condição apresentada, omite tais particularidades [Mewes *et al*., 2021; Silva *et al*., 2022], resultando em modelos descontextualizados ou incompletos para a análise da realidade local . Dessa forma, a mineração de dados ainda

não se configura como a ferramenta auxiliar ideal para a prática de profissionais da nutrição na região, haja vista a lacuna aqui explicitada. Sob essa ótica, há uma carência de análises baseadas em mineração de dados que mapeiem o consumo de frutas do Seridó Potiguar, moldado pela realidade climática e cultural local, em padrões adaptados ao contexto regional.

Diante do exposto, o presente estudo, de caráter exploratório, busca identificar padrões de consumo de frutas entre habitantes do Seridó Potiguar e sua associação com variáveis de escopo regional (sociodemográficas, antropométricas e preferências pessoais). Para tanto, foram utilizadas técnicas de mineração de dados com algoritmos de aprendizado de máquina não supervisionado, visando identificar perfis e extrair regras que traduzam as particularidades dietéticas da região. Com isso, espera-se que este trabalho forneça subsídios teóricos e práticos para os profissionais da área de alimentos da região, o que inclui nutricionistas e técnicos em nutrição, ao produzir informações que possam ser utilizadas como ferramenta auxiliar no exercício profissional.

Em síntese, o presente artigo está organizado da seguinte forma: a Seção 2 fundamenta teoricamente o estudo, revisando os principais conceitos de mineração de dados e de aprendizado de máquina; a Seção 3 apresenta trabalhos relacionados, situando esta pesquisa no contexto da literatura existente; a Seção 4 detalha a metodologia, explicando os procedimentos de coleta de dados, pré-processamento e aplicação dos algoritmos; a Seção 5 expõe os resultados obtidos e abre discussões, acerca das hipóteses levantadas na introdução; e, por fim, a Seção 6 discute as conclusões à luz das premissas estabelecidas, apresentando as limitações do estudo e sugestões para trabalhos futuros.

2 Fundamentação Teórica

Este capítulo apresenta a fundamentação teórica desta pesquisa com base em estudos da ciência da nutrição e análise de dados. Para compreender a complexidade do consumo alimentar no Seridó Potiguar, faz-se necessário estabelecer as bases da técnica dietética e dos sistemas alimentares regionais. Em seguida, discute-se a aplicação da mineração de dados como ferramenta capaz de decifrar os padrões de consumo propostos através do aprendizado de máquina com os algoritmos *k-means* e Apriori.

2.1 Sistemas Alimentares Regionais e o seu Impacto

O consumo alimentar é um fenômeno multifatorial que extrapola a necessidade biológica de ingestão de nutrientes. Segundo Enthoven and Van den Broeck [2021], os sistemas alimentares locais são caracterizados por uma forte dependência do contexto, na qual o território atua como o principal articulador entre a disponibilidade de alimentos e a identidade cultural de uma população. No cenário brasileiro, essa teoria é demonstrada por Alves *et al*. [2019], que desenvolveram uma pesquisa com a mesma base metodológica em diferentes estados brasileiros, demonstrando variações de uma mesma dieta com diferentes tipos de alimentos. Tal pressuposto é reforçado por Botelho [2006], ao estudar a culinária do Nordeste brasileiro, que destaca a singularidade da dieta regional, na qual até mesmo os mesmos pratos, quando em diferentes regiões, apresentam mudanças no modo de preparo e nos ingredientes. Esses achados reforçam que, em determinada localização, há muitos fatores que influenciam a escolha dos alimentos, inclusive a cultura desenvolvida pelos povos ali presentes. Para analisar e classificar esses costumes, é necessária uma organização precisa e bem fundamentada dessas informações.

2.2 Técnica Dietética e Categorização dos Alimentos

Nesse contexto, a dietética é a área que estuda e aplica os princípios e processos básicos da Ciência da Nutrição no organismo humano, permitindo o planejamento, a execução e a avaliação de dietas adequadas às características biopsicosociais dos indivíduos [Philippi, 2014]. Com isso, ela justifica a escolha dos recursos alimentares, sendo usada para identificar os padrões de consumo de alimentos presentes na população [Donatti *et al*., 2023].

Ornellas [2007], por sua vez, assinala a importância da categorização dos alimentos a partir de sua origem ou objetivo de sua produção. Dentro dessa classificação, um grupo de grande interesse é o de alimentos funcionais, responsáveis por: reduzir risco de doenças, fortificar o sistema imunológico e auxiliar no desempenho físico. Nessa categoria, destaca-se a presença de frutas, verduras e hortaliças, cujo consumo é fundamental para mensurar o estilo de vida de diferentes indivíduos, principalmente no que diz respeito as suas diferentes localizações geográficas [Martins *et al*., 2014, 2020; Donatti *et al*., 2023].

Conforme [Philippi, 2014], frutas são produto procedente da frutificação de uma planta sadia, destinada ao consumo *in natura*. Refere-se a parte polposa que envolve a semente de plantas. Possuem aroma característico, são ricas em suco, normalmente de sabor doce e podem, na maioria das vezes, serem consumidas cruas. Ornellas [2007] propõe uma organização desses alimentos baseada em características estruturais e botânicas em diferentes grupos. Essa taxonomia é de especial interesse para este estudo, pois permite a redução da granularidade das informações por meio de agrupamentos lógicos, mitigando a dispersão dos dados e facilitando a identificação de correlações diretamente ligadas às características biofísicas da região às quais essas frutas são nativas. Assim, o autor as classifica em:

- **Frutas com caroço:** como ameixa e pêssego;
- **Frutas duras:** como maçã e pera;
- **Frutas moles:** como morango e uva;
- **Frutas cítricas:** como laranja e limão;
- **Frutas mediterrâneas e tropicais:** como banana e melancia, dentre outras.

2.3 Big Data e Mineração de Dados

Apesar da classificação de frutas proposta por Ornellas [2007], a análise empírica desses dados enfrenta dificuldades relacionadas à quantidade de informações que podem ser associadas à dieta de uma dada população. Esse fenômeno é discutido em Shao *et al*. [2017], que descreve como o volume massivo de dados encontrados na *internet* (*Big Data*) pode ser utilizado em sistemas nutricionais de apoio à decisão. [Enthoven and Van den Broeck, 2021] continua a argumentação destacando as variáveis que influenciam a alimentação de um

indivíduo. Dentre elas, fatores sociais e econômicos são alguns dos principais demarcadores, assim como idade e cultura local, todos no escopo de variáveis sociodemográficas. Além disso, Shao *et al*. [2017] associa valores de IMC ao risco de problemas cardiovasculares e a fenótipos voltados ao desenvolvimento de doenças em uma população, evidenciando a participação de indicadores antropométricos, além dos anteriormente destacados, como componentes do perfil de saúde de uma população. Com isso, é possível entender que que essa ordem de dados (sociodemográficos e antropométricos) são importantes variáveis relacionadas a análise de padrões alimentares. Nesse sentido, ao integrar a preferência pessoal como componente cultural, também associado aos padrões alimentares, estabelece-se uma lista de variáveis que permite a realização de uma investigação nutricional mais abrangente.

Considerando a existência de uma abundância de informações associadas ao perfil de consumo alimentar populacional, uma das áreas mais recorridas para facilitar o entendimento desses padrões é a mineração de dados, capaz de automatizar esse processo, tornando a análise mais fluida [Silva *et al*., 2017]. Dentro desse campo, o aprendizado de máquina (*Machine Learning*) constitui uma área dedicada ao desenvolvimento de modelos computacionais capazes de identificar padrões a partir de experiências anteriores, conforme destaca Dangeti [2017]. Segundo os autores, a construção desses modelos envolve um conjunto estruturado de etapas: (i) coleta de dados, na qual são obtidas as informações necessárias para o treinamento; (ii) preparação dos dados, responsável pela padronização e pela adequação do formato dessas informações; (iii) análise exploratória, etapa voltada à extração de características determinantes ao entendimento preliminar do comportamento dos dados, permitindo definir o tipo de algoritmo a ser empregado; (iv) treinamento, fase em que diferentes configurações são avaliadas para compor o modelo inicial; (v) testagem, realizada antes da implantação, por meio da comparação entre valores esperados e o comportamento do modelo sobre um conjunto de dados previamente separado; e, por fim, (vi) implantação, na qual o modelo treinado passa a ser utilizado para reconhecer, continuamente, os padrões presentes nos dados coletados.

Nesse processo, a etapa de treinamento permite ao modelo estabelecer relações entre as instâncias de dados e os rótulos associados a cada uma delas. Esses rótulos são um agrupamento predeterminado pelo supervisor do algoritmo, adicionado às instâncias ou já presentes nelas. Esse tipo de característica identifica o que é conhecido como aprendizado de máquina supervisionado [Hackeling, 2014]. No entanto, existe uma abordagem oposta, o aprendizado de máquina não supervisionado, que não classifica os dados antes de enviá-los ao algoritmo. Nesta abordagem, o objetivo é encontrar regularidades no conjunto de dados analisado por meio de diferentes modelos matemáticos, que fazem a classificação sob a supervisão do supervisor [Alpaydin, 2020]. O presente artigo fundamenta-se nessa abordagem, utilizando métodos de aprendizado de máquina não supervisionados para identificar padrões ocultos nos dados de consumo de frutas na região do Seridó Potiguar.

O aprendizado de máquina não supervisionado se divide em diferentes categorias, algumas das quais se destacam por estarem dedicadas às tarefas de análise de agrupamentos, que consistem no estudo de um conjunto de dados com o intuito de descobrir relações presentes nessas informações que denotem similaridade ou diferenças entre si [Hackeling, 2014]; e os voltados à identificação de regras de associação , definidas como a busca por ocorrências frequentes e similares entre os elementos de um contexto, com o objetivo de descobrir um comportamento previsível e reincidente no conjunto de dados [Silva *et al*., 2017]. Para representar essas duas abordagens, o presente estudo fará uso de dois algoritmos de aprendizado de máquina, o *k-means*, para a tarefa de análise de agrupamentos, e o Apriori, para a tarefa de identificação de regras de associação.

2.3.1 Algoritmo de Análise de Clusters - k-Means

O *k-means* é um algoritmo cujo processo consiste em mover os centroides (pontos centrais de um agrupamento) para a posição média de seus constituintes e atribuir as entidades mais próximas a cada um deles, até que não haja alterações significativas nos elementos de cada agrupamento após iterações sucessivas. Essa estratégia se concentra na ideia de que os dados de cada instância, se colocados em um gráfico, apresentam maior proximidade entre si à medida que se encontram mais próximos [Dangeti, 2017]. O algoritmo começa por meio da escolha aleatória de k entidades, esse valor é definido pelo usuário e indica a quantidade de agrupamentos que devem ser encontrados durante a sua execução. Cada agrupamento possui um centroide e atribui novas instâncias conforme a proximidade da instância com o centroide mais próximo. Uma das dificuldades encontradas durante a aplicação desse algoritmo é a escolha do valor de k. Essa limitação levou à criação de diferentes métodos de identificação do número ideal, como, por exemplo, *Elbow Method* e *Sillhouette Score*.

2.3.2 Algoritmo de Regras de Associação - Apriori

O algoritmo Apriori opera sobre transações, conjunto de itens que aparecem em uma mesma ocorrência de dados de uma instância, identificando quais *itemsets*, agregado de um ou mais itens que aparecem em uma mesma transação, são mais frequentes conforme o número total de aparições entre as transações. O usuário escolhe um valor de suporte mínimo, definido em porcentagem, que indica ao algoritmo que as relações abaixo dele devem ser descartadas [Carvalho *et al*., 2011].

A partir desses dados, o próximo passo é gerar regras de associação com o padrão $X \rightarrow Y$, onde X é uma associação antecedente e Y é uma associação consequente. A força da regra é medida pela sua confiança (*confidence*), que expressa a probabilidade de Y ocorrer dado que X ocorreu. O Apriori filtra as regras cuja confiança seja superior a um limite mínimo estabelecido pelo usuário. Esse procedimento produz associações do tipo "se o antecedente (X) ocorre, existe uma determinada probabilidade (confiança) do consequente (Y) também ocorrer", inferindo padrões de co-ocorrência nos dados. A quantidade de vezes que esse padrão se repete na base dados é chamado de suporte (*support*). Castro and Ferrari [2016] discutem como a confiança pode não ser um critério confiável de análise, mesmo com o suporte. Ele não considera o suporte do conjunto de itens no consequente da regra, ou seja, caso esse consequente seja muito frequente em uma base

dados, ele irá criar regras com todos os outros elementos repetidamente. Para resolver isso, outro parâmetro é adicionado a equação: o chamado *lift*, calculado ao dividir a confiança de uma regra pelo suporte do consequente dessa regra, obtendo o quanto a ocorrência de X aumenta a probabilidade de Y aparecer no *itemset*.

3 Trabalhos Relacionados

Diversos estudos têm explorado técnicas computacionais para analisar hábitos alimentares em populações específicas, aplicando aprendizado de máquina para relacionar esses hábitos aos riscos à saúde [Lazarou *et al*., 2012; Chika *et al*., 2024; Mewes *et al*., 2021; Olutunde *et al*., 2024; Silva *et al*., 2022]. No estudo de Lazarou *et al*. [2012], que investigou a relação da alimentação de crianças com o risco à obesidade, foram encontradas relações entre essa condição e o consumo de alimentos fritos, doces e ultraprocessados em suas dietas, isso através da implementação de algoritmos de aprendizado de máquina, se tornando um marco para análises futuras nessa área.

As tecnologias utilizadas para mineração de dados variam entre os trabalhos, buscando interpretar de maneiras diferentes informações semelhantes mediante diferentes algoritmos ou abordagens de organização entre eles [Chika *et al*., 2024; Mewes *et al*., 2021]. Em Chika *et al*. [2024], o uso do algoritmo *C4.5* como meio de identificação de regras de classificação resultou na associação entre hábitos alimentares "não-saudáveis"(refeições fora do horário, consumo de ultraprocessados e outros alimentos com alto valor calórico) com fatores de risco à saúde, como a obesidade, em pacientes com a Doença do Refluxo gastroesofágico. Já em Mewes *et al*. [2021], cujo estudo investigou como os hábitos da população adulta da suiça podem levar a problemas de saúde, o algoritmo utilizado para regras de associação foi o *Apriori*, foram encontradas regras que se associaram a problemas relacionados a hipertensão, hipercolesterolemia e diabetes como consequentes de uma alimentação pobre em vegetais, com ausência de nutrientes essenciais como vitaminas e minerais, e com maus hábitos de saúde, como o tabagismo.

Além disso, a mineração de dados também pode ser empregada na identificação de outros padrões de consumo, conforme destacado por Silva *et al*. [2022]. Estes autores propõem a existência de dois padrões de dieta na população brasileira conforme os resultados obtidos pelo algoritmo de agrupamento *k-means*: o primeiro padrão, nomeado de dieta ocidental, refere-se a ingestão elevada de alimentos calóricos (cereais refinados, feijões, carne vermelha, leites gordurosos, laticínios e bebidas açucaradas); e o segundo, nomeado de padrão prudente, caracteriza-se pelo predomínio da ingestão de alimentos considerados saudáveis (frutas, verduras, carnes brancas, cereais inteiros, leite com baixo teor de gordura). O estudo demonstrou que variáveis como idade, gênero e a frequência de atividades físicas são grandes preditores da qualidade das dietas entre indivíduos dessa população.

No estaudo de Olutunde *et al*. [2024], que investigou como técnicas de aprendizado de máquina podem identificar padrões alimentares, e, por conseguinte, relacioná-los a riscos de doenças crônicas e apoiar a criação de ferramentas voltadas a recomendações nutricionais personalizadas; foi empregada uma abordagem híbrida na construção de duas soluções principais: (1) um módulo capaz de agrupar hábitos de consumo e gerar sugestões individualizadas de alimentos, e (2) um sistema que estima indicadores de risco para problemas de saúde com base na dieta atual de cada indivíduo. Os resultados demonstraram elevado potencial de aplicação para profissionais de nutrição, servindo como ferramenta complementar na avaliação de padrões alimentares e na recomendação de alimentos para os pacientes.

De modo geral, os estudos apresentados ao longo desta seção relacionam algumas variáveis antropométricas, dados demográficos e avaliações individuais de preferência alimentar associados a fatores de risco à saúde, além de abordarem a criação de sistemas capazes de recomendar alternativas mais adequadas ou saudáveis na dieta de um indivíduo ou mais indivíduos. Esses avanços concentram-se em estudos voltados à análise dos padrões dietéticos e na avaliação de padrões gerais na alimentação de uma população, agrupando os alimentos em grandes conjuntos para produzir resultados mais abrangentes. No entanto, a eficácia desses padrões permanece uma lacuna em dados regionalmente distintos como os do Seridó Potiguar. Portanto, esse estudo aplica essas metodologias para identificar se a redução da escala geográficas é capaz de revelar padrões de associação que análises abrangentes não são capazes de capturar.

4 Metodologia

Trata-se de um estudo de caráter exploratório. A opção justifica-se pela necessidade de identificar padrões alimentares relacionados ao consumo de frutas na região do Seridó Potiguar, por meio da mineração de dados, o que oferece uma base teórica que fundamenta análises de consumo alimentar mais precisas da população dessa microrregião. Para isso, esta seção apresenta as características das variáveis coletadas, os critérios de seleção e o processo de construção da base de dados. Em seguida, descreve-se a etapa de pré-processamento, realizada com o objetivo de assegurar a consistência e a qualidade das informações analisadas. Na sequência, apresenta-se o *design* do experimento, no qual se aplica o aprendizado de máquina não supervisionado. O algoritmo de agrupamento *k-means* foi empregado para identificar padrões ocultos nos dados, acompanhado das etapas necessárias à sua parametrização e dos métodos utilizados para determinar o número ideal de *clusters*. Adicionalmente, o algoritmo Apriori foi utilizado para extrair associações e regras relevantes, filtrando-as com base nas métricas de suporte e confiança, com limiares de 10% e 80%, respectivamente. A análise das regras resultantes foi conduzida com base em métricas como *lift* e suporte, a fim de eliminar associações redundantes ou de baixa relevância.

4.1 Descrição dos Dados

As variáveis selecionadas abrangeram dados demográficos e antropométricos, bem como informações individuais sobre o consumo de frutas e as condições de saúde dos participantes. Entre os registros demográficos, foram coletados dados de idade (considerando os valores de ano a partir da data de nascimento até o instante em que o formulário foi respondido), gênero (masculino e feminino) e município de residência. Esses atributos permitiram a categorização dos participantes

em faixas etárias e grupos demográficos. Entre as variáveis antropométricas, altura (em metros) e peso (em quilogramas) foram utilizados para o cálculo do índice de massa corporal (IMC), conforme a fórmula $IMC = P(kg)/E^2(m)$ [WHO Working Group, 1986], o que permitiu examinar a relação entre a composição corporal e os padrões de consumo de frutas.

Com relação às variáveis individuais, foram incluídas informações sobre as comorbidades dos participantes, com categorias de resposta que abrangeram hipertensão, diabetes e doenças cardíacas. Também foram coletadas informações sobre o consumo de frutas e os horários em que essa prática ocorre. A variável de consumo foi mensurada por meio de uma escala *Likert* de seis pontos, adotada conforme Sullivan and Artino Jr [2013], com as seguintes opções: "muito bom", "bom", "neutro", "ruim', "muito ruim" e "nunca comi". O conjunto de frutas apresentado aos participantes, listado na Tabela 1, segue a classificação proposta por Philippi [2014] e age como uma maneira de filtrar dados em categorias próximas para facilitar a análise de padrões, diminuindo a granularização dos resultados. Além disso, os participantes indicaram horários de preferência para consumo de frutas, podendo selecionar múltiplas opções entre "manhã", "tarde", "noite" e "madrugada", informações utilizadas para verificar possíveis variações na alimentação dos participantes em períodos temporais.

Tabela 1. Lista de frutas selecionadas para o estudo segundo suas categorias

Categoria	Lista de Frutas
Caroço	Ameixa, Cereja, Pêssego, Caqui
Duras	Maçã, Pêra
Moles	Morango, Uva
Cítricas	Limão, Acerola, Laranja
Mediterrâneas	Kiwi, Cajú, Maracujá, Abacaxi, Goiaba, Manga, Jaca, Abacate, Banana, Figo, Melão, Melancia, Mamão, Pinha
Oleaginosos	Amêndoa, Avelã, Amendoim, Cacau, Castanha-do-pará, Coco, Noz

Fonte: adaptado de Ornellas [2007].

4.2 Coleta de Dados

A coleta de dados foi realizada por meio de um formulário de pesquisa no Google Forms [1], divulgado através de redes sociais, tais como: WhatsApp, Facebook e Instagram, para alcançar habitantes do Seridó Potiguar e região. O formulário foi aberto durante o período de 02/10/2024 a 11/12/2024 (70 dias) e coletou 407 respostas. A pesquisa obteve aprovação pelo Comitê Central de Ética em Pesquisa da UFRN, sob o CAEE: 78931924.5.0000.5537 e Número do Parecer: 7.083.409.

4.3 Pré-processamento

No processo de aprendizado de máquina, é necessário que as informações enviadas aos algoritmos estejam padronizadas no formato exigido [Dangeti, 2017]. Conforme Castro and Ferrari [2016], faz-se necessária a etapa de pré-processamento para ingerir os dados coletados pela pesquisa e realizar o processo de limpeza — no qual se filtram os dados existentes em um padrão mínimo —, imputação — processo que visa preencher lacunas nas informações a partir de alguma inferência estatística —, e transformação — na qual os dados podem passar de um valor a outro conforme alguma função de conversão desejada.

4.3.1 Limpeza

Os dados obtidos via formulário do Google Forms, em planilha do Excel, originalmente com 43 colunas e 407 linhas, passaram por uma organização inicial. Colunas geradas automaticamente, como carimbo de data e termo de adesão, foram descartadas. Em seguida, as colunas foram padronizadas e renomeadas para facilitar a identificação. Por exemplo, o termo "Cidade de residência" foi alterado para "Cidade" e "Marque se você é portador de alguma doença" para "Condições Médicas". As colunas que continham as notas atribuídas às frutas também foram renomeadas com o nome de cada fruta. Após essa etapa, a tabela passou a ter 41 colunas. A partir dos dados estatísticos de cada coluna, concluiu-se que a coluna "Condições Médicas", preenchida somente em 5% da amostra, não era necessária. O mesmo aconteceu com a coluna "Cidade", na qual 86% dos participantes eram provenientes de um único município (Caicó-RN).

4.3.2 Imputação

Após a limpeza, o passo seguinte consistiu na correção de dados inconsistentes, como as datas de nascimento de alguns participantes ($\leq$ 5 anos ou $\geq$ 200 anos). Para tratar esses registros, que representam 2% (n=7) da amostra, foi utilizado o algoritmo *Hot Deck Imputation* [Castro and Ferrari, 2016], selecionado por preservar a variabilidade original da amostra e lidar adequadamente com variáveis categóricas. Essa técnica preenche valores ausentes ou inválidos com base em registros semelhantes no próprio conjunto de dados. Para sua aplicação, o conjunto de respostas foi ordenado por gênero, cidade e altura, de modo que as idades inconsistentes fossem substituídas por valores de participantes com características comparáveis.

4.3.3 Transformação

Por fim, muitas colunas tiveram seus dados transformados em outros valores para serem reconhecidos pelos algoritmos ou por serem de interesse para este estudo. A lista a seguir conta com a descrição dos processos realizados para cada coluna:

- **Altura e peso**: substituídos pelo IMC, motivado pelo interesse desta pesquisa em associar esse valor ao consumo.
- **Data de nascimento**: convertida em idade para trabalhar com faixas numéricas e diferenciar os resultados a partir desse valor.
- **Gênero e Horários de Consumo**: transformados em colunas binárias, com valores 0 ou 1. Essa escolha se deu pela representação exigida por ambos os algoritmos, que reconhecem informações textuais com maior precisão quando organizadas em colunas binárias.
- **Frutas**: agrupadas nas categorias apresentadas na Tabela 1 utilizando a média da avaliação de cada fruta presente na categoria.

[1] https://forms.gle/rKYHgPCdDxjzEjwD8

4.4 Design do Experimento

Dentre os algoritmos de agrupamento disponíveis, optou-se pelo *k-means* em razão de sua eficiência e de sua aplicação em estudos alinhados ao objetivo desta pesquisa [Silva *et al*., 2022]. Além deste, utilizou-se o algoritmo Apriori para regras de associação, também por sua implementação em trabalhos de mesmo foco [Chika *et al*., 2024]. Para construir os modelos de mineração de dados, foi utilizada a linguagem de programação *Python 3.11* com as bibliotecas *scikit-learn 1.5*, para a aplicação do *k-means*, e *mlxtend 0.23*, para a utilização do Apriori. A edição da base de dados foi realizada com o pacote *Pandas 2.2.3*, módulo esse utilizado para manipular a base de dados.

4.4.1 Aplicação de Agrupamentos

A execução do *k-means* foi estruturada nessa ordem de etapas: (i) seleção e normalização dos dados, (ii) escolha do número ideal de k grupos, (iii) diferenciação entre os grupos mediante métricas estabelecidas.

Primeira etapa. Foram utilizadas todas as variáveis tratadas ao longo do capítulo. Foi utilizada *Min-Max Normalization* para normalizar todos os valores das colunas entre 0 e 1. Estabelecendo que o maior valor da coluna será 1 e o menor 0, com o restante deles variando nessa faixa [Castro and Ferrari, 2016]. O algoritmo utiliza a distância euclidiana para definir os grupos ao longo da execução; se os valores das colunas estiverem em escalas diferentes, essa seleção pode ser incorreta [Niño-Adan *et al*., 2022]. A tabela foi criada com as seguintes informações:

- **Tabela de Clusterização:** Idade, Categoria de Frutas (Caroço, Duras, Moles, Cítricas, Mediterrâneas, Oleoginosos), Horários de Consumo (Manhã, Tarde, Noite, Madrugada) e IMC.

Segunda etapa. Foram utilizadas duas metodologias de identificação do número ideal de grupos em uma tabela de dados para o k-means. A primeira, o *Elbow Method*, avalia a compactação dos grupos resultantes do algoritmo k-means, buscando identificar o ponto onde o aumento no número de grupos (k) deixa de trazer ganhos significativos na redução das distâncias internas.

Complementarmente, aplicou-se o *Silhouette Score*, que é um indicador de coesão e separação dos pontos em relação aos *clusters* resultantes. Ele verifica quanto os elementos de um grupo estão próximos entre si e o quão distantes estão do grupo mais próximo. O resultado transita em uma faixa de -1 até $+1$, onde valores próximos a $+1$ indicam as melhores taxas nessas duas métricas.

Ambas as abordagens foram escolhidas devido às suas capacidades de definir os grupos com medidas de avaliação confiáveis e recomendação geral na literatura [Dangeti, 2017; Silva *et al*., 2017]. O pacote *mlxtend* conta com essas técnicas em algoritmos próprios. Conforme demonstrado na Figura 1, a interseção entre os melhores candidatos de ambos definiu o k final, sendo este $k = 2$ ou $k = 3$. Optou-se como k final o resultado $k = 3$ pois ainda há uma boa diminuição de inércia de $2 \rightarrow 3$ e a diferença de silhueta é pequena o bastante para não causar perdas significativas.

Terceira Etapa. Após executar o algoritmo com o número ideal de grupos, foram empregados métricas e parâmetros de avaliação para determinar as diferenças entre os resultados obtidos. Buscou-se destacar informações importantes para a pesquisa, sendo estas: (a) avaliação geral das notas, (b) avaliação individual de frutas, (c) avaliação por categoria de frutas e (d) padrões temporais de consumo. Todos os grupos passaram por esses indicadores e ainda foram subdivididos entre idade, gênero e IMC, por serem as variáveis presentes na tabela enviada ao algoritmo.

4.4.2 Aplicação das Regras de Associação

O processo de extração das regras de associação feito através do algoritmo Apriori aplicou uma segmentação de passos: (i) seleção e transformação dos dados, (ii) aplicação do algoritmo, (iii) filtragem e seleção das regras geradas.

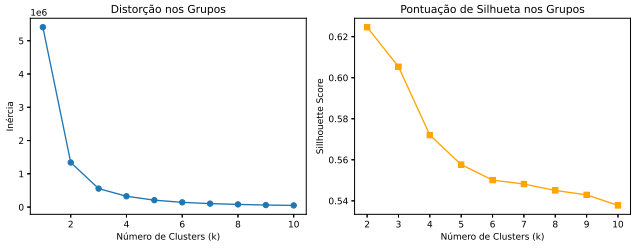
Primeira etapa. A base de dados foi adequada a demanda do algoritmo Apriori que aceita somente dados binários, passando por *One-hot Encoding*. Um processo de identificar os valores nas linhas de uma tabela de dados em colunas binárias, criando uma distinção numérica entre elas sem formar hierarquias. [Castro and Ferrari, 2016],

- **Tabela de Regras de Associação:** Idade foi transformada em Faixa Etária, seguindo classificações recomendadas em diretrizes da área da saúde. O IMC foi substituído por Classificação do IMC, com base em dados do Ministério da Saúde [BRASIL, 2020]. Categorias de Frutas foram classificadas em escala *likert* de três pontos, pelos resultados de categoria serem médias, foi aplicada uma aproximação a todos os valores: *gosta* $(4-5)$*, neutro* (3)*, detesta* $(1-2)$ e *desconhece* (0).

Segunda etapa. O algoritmo foi configurado com o suporte e confiança mínimos de 10% e 80% respectivamente. Valores estes empregados para se obter a maior quantidade de regras e a melhor qualidade substancial, pois certas variáveis em menor quantidade, como a coluna de homens (33,8% dos participantes) precisam de maior suporte para aparecer nas associações.

Terceira etapa. Por fim, com o desfecho da execução, aplicou-se um filtro superficial e remoção de redundâncias. Para o filtro, a biblioteca utilizada retornou métricas além de suporte e confiança, o *lift* foi limitado a valores acima ou iguais a 1.2 por indicar maior independência da regra com relação ao seu consequente. Para reduzir redundâncias, foram removidas regras que possuíam uma dependência direta de sua confiança

Figura 1. *Elbow Method* e *Sillhouette Score*



Fonte: autoria própria, 2025.

com relação a outras regras geradas anteriormente. Esse tipo de redundância é comum nos resultados gerados pelo algoritmo, regras como $A \rightarrow B$ podem aparecer em outros locais, como $A + C \rightarrow B$, sendo C um conjunto de associações adicionais. No entanto, a gravidade ocorre quando C não aumenta a confiança dessa regra, e sim a diminui, ou seja, $A \rightarrow B$ é a única parte dessa regra que explica a confiança alta.

Inicialmente, foram encontradas 3098 regras de associação, após a primeira filtragem, restaram 2121 (-31%) e, com a redução de redundâncias, o valor baixou ainda mais, chegando a 846 (-60%), totalizando queda de 72% no número de regras. A seleção destas regras foi realizada a partir da análise individual com um profissional da área, para validação e acompanhamento dos resultados.

5 Resultados

Nesta seção, são apresentadas as informações extraídas da análise da base de dados, assim como os resultados da execução dos algoritmos. Primeiramente, são exibidas as estatísticas descritivas da amostra. Logo em seguida, os grupos retornados pelo algoritmo *k-means* e, por fim, as regras de associação identificadas pelo Apriori.

5.1 Estatísticas Descritivas dos Dados

A amostra é composta por 407 indivíduos, sendo $66,2\%$ mulheres e $33,8\%$ homens. As idades variam entre 8 e 77 anos; entretanto, a faixa dos 10 aos 50 anos concentra 93% da população total, com destaque para o intervalo de 20 a 25 anos, que representa aproximadamente um quarto da amostra.

Quanto às variáveis antropométricas, o peso médio foi de $69,9 \pm 17,94$ kg e a altura média de 164 ± 10 cm, resultando em um IMC médio de $25,76 \pm 5,4$ kg/m². Transformando esse dado em classificação, com base na tabela presente em [BRASIL, 2020], 10% dos indivíduos encontram-se abaixo do peso, 33% eutróficos, 35% em sobrepeso e 22% em situação de obesidade.

Os dados relativos à nota atribuída ao quesito "preferência de consumo" das frutas estão descritos na Tabela 2, enquanto a classificação geral por categoria de fruta foi apresentada na Figura 2. Convém observar que, em relação aos horários de consumo, a maioria dos indivíduos relatou ingerir frutas à tarde (73%), seguida pela manhã (63%), noite (44%) e madrugada (4%).

No que tange às avaliações por categorias, observou-se uma disparidade entre o consumo de itens locais e variedades específicas. Tomando como referência a classificação de frutas regionais por Ornellas [2007], as médias obtidas foram: manga (4.314), goiaba $(4,088)$, acerola $(3,955)$ e pinha $(3,008)$. Em contrapartida, os menores índices de aceitação concentraram-se nas categorias de frutas com caroço e oleaginosas, que registraram as médias mais baixas da amostra, com $2,32$ e $2,75$, respectivamente.

5.2 Resultado do Agrupamento (*k-means*)

Após o tratamento e análise inicial dos dados, foi executado o algoritmo k-means, o qual retornou três *clusters*. Na Tabela 3 encontram-se informações relacionadas à caracterização da amostra, em que se listam dados demográficos, antropométricos e sobre as condições médicas gerais dos indivíduos presentes em cada grupo. Já na Tabela 4 é possível observar dados das frutas com melhores e piores avaliações, assim como a média de notas atribuídas a cada categoria de fruta. Por fim, na Tabela 5 estão apresentadas as preferências de horário de consumo e a distribuição das notas.

Tabela 2. Apresentação das frutas segundo o grau de preferência, a partir da média de pontuação atribuída.

Fruta	Média de Nota
Banana	4.646
Laranja	4.476
Uva	4.449
Melancia	4.412
Maçã	4.327
Manga	4.314
Abacaxi	4.295
Maracujá	4.255
Goiaba	4.088
Morango	4.059
Acerola	3.955
Amendoim	3.875
Coco	3.864
Melão	3.809
Limão	3.809
Mamão	3.715
Cajú	3.492
Abacate	3.449
Kiwi	3.383
Castanha-do-pará	3.130
Ameixa	3.130
Pinha	3.008
Pêra	2.973
Cereja	2.702
Cacau	2.418
Jaca	2.189
Avelã	2.093
Amêndoa	1.928
Noz	1.926
Pêssego	1.872
Caqui	1.566
Figo	1.551

Fonte: autoria própria, 2025.

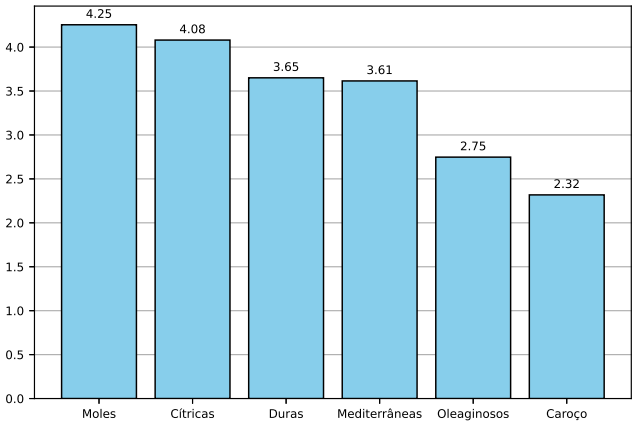
Tabela 3. Caracterização da amostra por clusters do algoritmo k-means

Variável	Grupo 1	Grupo 2	Grupo 3
Indivíduos (n, %)	136 (36.2%)	134 (35.6%)	106 (28.2%)
Idade (anos)	25.80 ± 11.25	29.88 ± 12.99	30.67 ± 15.12
IMC (kg/m²)	25.49 ± 5.64	25.96 ± 5.26	25.85 ± 5.44
Gênero (n, % relativo ao grupo)			
Mulheres	83 (61.02%)	98 (73.13%)	68 (64.15%)
Homens	53 (38.97%)	36 (26.86%)	38 (34.84%)
Condições Médicas (n, % relativo ao grupo)			
Diabetes	6 (4.41%)	4 (2.98%)	7 (6.60%)
Hipertensão	9 (6.61%)	8 (5.97%)	12 (11.32%)
Doenças cardíacas	3 (2.20%)	5 (3.73%)	3 (2.83%)

Fonte: autoria própria, 2025.

O Grupo 1 apresenta a população mais jovem dentre os demais, com a média de 25.80 anos, e também a maior quantidade de indivíduos $(36.2\%$ de toda a população da pesquisa). Sua peculiaridade é destacada nos 0% de consumo pela manhã, indicando uma forte preferência pela tarde e noite. Dentre os três apresentados, é o grupo que possui maior

Figura 2. Média de avaliações por categoria de fruta



Fonte: autoria própria, 2025.

Tabela 4. Ranqueamento das médias de avaliação de cada fruta e por categoria de fruta de cada grupo

Variável	Grupo 1	Grupo 2	Grupo 3
Melhores avaliações de frutas (média de notas)			
Fruta N°1	Banana (4.55)	Banana (4.65)	Banana (4.75)
Fruta N°2	Laranja (4.41)	Laranja (4.52)	Uva (4.62)
Fruta N°3	Melancia (4.40)	Maracujá (4.44)	Melancia (4.53)
Fruta N°4	Uva (4.32)	Uva (4.40)	Maçã (4.52)
Fruta N°5	Manga (4.26)	Abacaxi (4.34)	Laranja (4.5)
Piores avaliações de frutas (média de notas)			
Fruta N°32	Figo (1.19)	Figo (1.77)	Figo (1.72)
Fruta N°31	Caqui (1.20)	Caqui (1.78)	Caqui (1.76)
Fruta N°30	Pêssego (1.45)	Pêssego (2.05)	Amêndoa (1.94)
Fruta N°29	Noz (1.5)	Amêndoa (2.20)	Noz (1.98)
Fruta N°28	Amêndoa (1.64)	Jaca (2.23)	Avelã (1.99)
Avaliações de categoria de fruta (média de notas)			
Caroço	1.93	2.44	2.63
Oleoginosas	2.49	3.00	2.73
Mediterrâneas	3.40	3.64	3.83
Duras	3.50	3.59	3.90
Cítricas	3.90	4.14	4.22
Moles	4.09	4.24	4.46

Fonte: autoria própria, 2025.

distribuição de notas nas classificações “Ruins” (14.20%) e “Desconhece” (15.80%). É o grupo que menos se alimentou de frutas com caroço e o que apresenta melhor aceitação de frutas mediterrâneas, conforme a Figura 3.

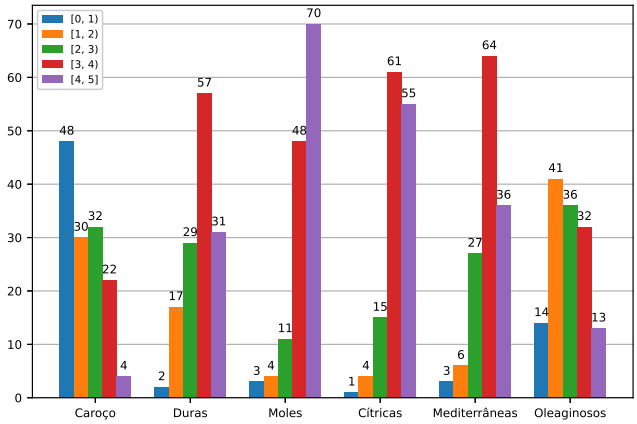
Diferentemente do Grupo 1, o Grupo 2 apresenta preferências de consumo concentradas no período da manhã, sem registros de consumo noturno. Reúne a maior proporção de mulheres (73.13%) e apresenta as maiores médias de avaliação nas categorias de fruta com pior aceitação geral do público nessa pesquisa: “Caroço” (2.44) e “Oleaginosas (3.00)”. Nesse grupo, observa-se a menor frequência da categoria “Desconhece” (9.77%) e a maior proporção de avaliações “Neutras” (20.49%).

Tabela 5. Distribuição de horários e notas

Variável	Grupo 1	Grupo 2	Grupo 3
Preferências de horário (n, % relativa ao grupo)			
Manhã	0 (0%)	134 (100%)	106 (100%)
Tarde	112 (82.35%)	87 (64.92%)	78 (73.54%)
Noite	62 (45.58%)	0 (0%)	106 (100%)
Madrugada	7 (5.14%)	1 (0.74%)	8 (7.54%)
Distribuição de notas (% relativa ao grupo)			
Boas	51.72%	56.71%	61.99%
Neutras	18.26%	20.49%	15.80%
Ruins	14.20%	13.01%	10.22%
Desconhece	15.80%	9.77%	11.96%

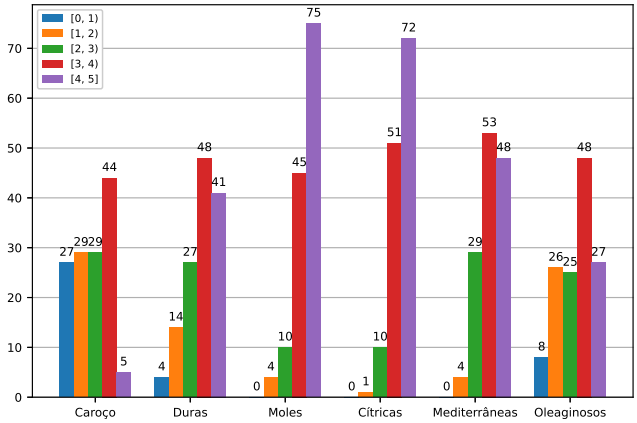
Fonte: autoria própria, 2025.

Figura 3. Distribuição de Notas em Porcentagem por Categoria de Fruta do Grupo 1



Fonte: autoria própria, 2025.

Figura 4. Distribuição de Notas em Porcentagem por Categoria de Fruta do Grupo 2

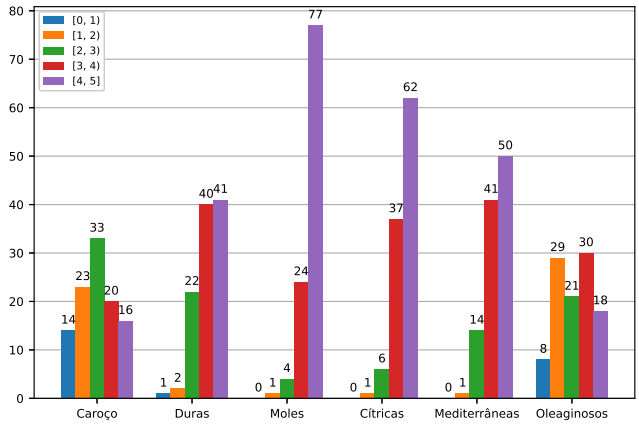


Fonte: autoria própria, 2025.

Por fim, o Grupo 3 é marcado por preferências de horário pela manhã e noite, a maior média de notas “Boas” (61.99%) e a menor média de notas “Ruins” (10.22%). As médias de avaliação por categoria demonstram avaliações muito positivas

para todas as categorias, exceto "Caroço" (2.63) e "Oleoginosas" (2.73). Através da Figura 5, é possível observar que essas categorias não possuem um padrão de notas definido no grupo, sendo bastante variáveis e diferentes das demais.

Figura 5. Distribuição de Notas em Porcentagem por Categoria de Fruta do Grupo 3



Fonte: autoria própria, 2025.

Quanto às avaliações, observa-se a presença constante de banana, laranja e uva nos três grupos (1, 2 e 3). Embora apresentem diferentes médias de pontuação, essas frutas ocupam as primeiras posições da Tabela 4. Nessa mesma tabela, a diferença entre as médias obtidas por cada fruta entre os grupos é similar, há variações somente nas combinações de frutas presentes.

Entre as frutas com menores médias, figo, caqui e pêssego aparecem de forma recorrente em todos os grupos. O Grupo 2, com essas frutas, possui médias próximas à faixa "Ruim" (2), enquanto os demais grupos (1 e 3) se aproximam do intervalo "Muito Ruim" (1).

Em todos os grupos, a hierarquia das categorias segue o mesmo ordenamento, do melhor ao pior: moles, cítricas, duras, mediterrâneas, oleaginosas e caroço; padrão identificado na análise descritiva geral apresentada na Figura 2.

As distribuições de notas, presentes nas Tabelas 3, 4 e 5, atribuídas às categorias moles, cítricas e duras, são semelhantes entre os grupos; contudo, as categorias mediterrâneas, oleaginosas e de caroço apresentam diferenças entre si.

5.3 Resultado das Regras de Associação (Apriori)

De maneira simultânea à análise de agrupamento, o algoritmo Apriori retornou 846 regras de associação durante sua execução, estas foram analisadas através das métricas suporte, confiança e *lift*. As Tabelas 6, 7 e 8 são um compilado de regras com antecedentes ou consequentes em comum, suporte mínimo de 10% e confiança acima de 80%. A Tabela 9, no entanto, é um comparativo entre diferentes regras de associação que possuem os mesmos antecedentes e consequentes, mas algumas dessas regras citam o IMC.

Na Tabela 6, o conjunto de regras destacado associa avaliações positivas da categoria de frutas "Caroço" a avaliações positivas para as demais categorias. Esse recorte apresentou o maior percentual de *lift* dentre os resultados obtidos pelo algoritmo Apriori.

Tabela 6. Regras de associação com "Caroço"

Regra	Suporte	Confiança	Lift
Gostar de frutas com caroço leva a gostar de frutas cítricas, moles, mediterrâneas e oleaginosas	14%	85%	318%
Gostar de frutas duras e com caroço leva a gostar de frutas cítricas, moles, mediterrâneas e oleaginosas	14%	91%	340%
Gostar de frutas com caroço leva a gostar de frutas cítricas, mediterrâneas e oleaginosas	14%	85%	309%
Gostar de frutas duras e com caroço leva a gostar de frutas cítricas, mediterrâneas e oleaginosas	14%	91%	330%
Gostar de frutas com caroço leva a gostar de frutas moles, mediterrâneas e oleaginosas	14%	87%	315%
Gostar de frutas duras e com caroço leva a gostar de frutas moles, mediterrâneas e oleaginosas	14%	93%	336%
Gostar de frutas com caroço leva a gostar de frutas moles, mediterrâneas, cítricas, duras e oleaginosas	14%	84%	354%

Fonte: autoria própria, 2025.

Tabela 7. Regras de associação "Detestar frutas com caroço"

Regra	Suporte	Confiança	Lift
Detestar frutas oleaginosas e ser neutro com frutas duras leva a detestar frutas com caroço	14%	80%	172%
Detestar frutas oleaginosas e ser jovem leva a detestar frutas com caroço	18%	82%	175%
Detestar frutas oleaginosas e ser neutro com frutas mediterrâneas leva a detestar frutas com caroço	16%	90%	192%
Ser neutro com frutas mediterrâneas e ser neutro com frutas duras leva a detestar frutas com caroço	11%	82%	174%

Fonte: autoria própria, 2025.

A Tabela 7 engloba regras de associação com consequente de detestar frutas com caroço. Os dados indicam que as seguintes características são os principais fatores desse resultado: detestar frutas oleaginosas, ser neutro a frutas mediterrâneas ou duras e ser jovem.

Outro padrão encontrado na base de dados é o aumento de confiança e *lift* com a adição de "consumir de manhã" em regras pré-existentes com antecedentes "ser adulto" e "gostar de frutas oleaginosas", conforme a Tabela 8.

Regras relacionadas a alguma classificação de IMC não obtiveram resultados muito diferentes em comparação com as mesmas regras sem esses valores. Na Tabela 9 é possível observar um comparativo entre esses resultados.

Além dessas regras, houve uma análise da quantidade de regras associadas em diferentes faixas etárias. Jovens e adultos possuem 30 (3.5%) e 147 (17.4%) regras agrupadas, respectivamente. Os jovens, na faixa de 19 a 30 anos, corres-

pondem a 47% da população, enquanto os adultos, na faixa de 31 a 59 anos, representam 34% da população.

Tabela 8. Regras de associação com Idade e Gênero

Regra	Suporte	Confiança	Lift
Ser adulto e gostar de frutas oleaginosas leva a gostar de frutas cítricas, duras e mediterrâneas	14%	85%	195%
Consumir de manhã, ser adulto e gostar de frutas oleaginosas leva a gostar de frutas cítricas, duras e mediterrâneas	12%	94%	215%
Ser adulto e gostar de frutas oleaginosas leva a gostar de frutas moles, duras e mediterrâneas	14%	87%	197%
Consumir de manhã, ser adulto e gostar de frutas oleaginosas leva a gostar de frutas moles, duras e mediterrâneas	12%	94%	213%
Ser adulto e gostar de frutas oleaginosas leva a gostar de frutas cítricas, moles, duras e mediterrâneas	14%	85%	204%
Consumir de manhã, ser adulto e gostar de frutas oleaginosas leva a gostar de frutas cítricas, moles, duras e mediterrâneas	12%	94%	224%

Fonte: autoria própria, 2025.

Tabela 9. Comparação entre regras com e sem IMC

Regra	Suporte	Confiança	Lift
Gostar de frutas cítricas e duras leva a gostar de frutas mediterrâneas	43%	86%	139%
Gostar de frutas cítricas e duras tendo um IMC normal leva a gostar de frutas mediterrâneas	11%	81%	131%
Gostar de frutas cítricas, duras e moles leva a gostar de frutas mediterrâneas	42%	86%	139%
Gostar de frutas cítricas, duras e moles tendo um IMC normal leva a gostar de frutas mediterrâneas	11%	80%	129%
Ser neutro com frutas mediterrâneas leva a detestar frutas com caroço	23%	83%	178%
Ser neutro com frutas mediterrâneas e ter IMC normal leva a detestar frutas com caroço	10%	87%	185%

Fonte: autoria própria, 2025.

6 Discussão

A convergência entre os centróides do algoritmo *k-means* e as regras de associação obtidas pelo algoritmo Apriori fornece a base empírica necessária para identificar padrões alimentares na população do Seridó Potiguar. A discussão a seguir integra os resultados obtidos, com implicações culturais e geográficas, a partir da comparação entre os estudos da área apresentados na Seção 3. Primeiramente, serão discutidas as características dos grupos encontrados pelo *k-means*, em seguida, será feita uma análise dos recortes de regras encontradas pelo Apriori, e, por fim, uma síntese geral com os resultados dos dois algoritmos.

6.1 Discussão do Agrupamento (*k-means*)

Na comparação entre os três grupos, identificados cada um por um padrão de consumo de frutas distinto, observou-se que havia poucas diferenças demográficas e antropométricas entre si, considerando os seus membros. Ademais, fatores como IMC, gênero e condições médicas não se destacaram como variáveis que diferenciam os grupos, enquanto dados como preferência de horário, avaliação da categoria de fruta e média das notas atribuídas tiveram maior peso nesse quesito. A escolha de múltiplos períodos de consumo parece estar diretamente relacionada a maiores médias de avaliação, no entanto, é possível correlacionar esse fato com uma frequência maior de consumo, dado esse que não foi incluído na pesquisa.

Outro fator observado é a pouca presença de frutas regionais dentre as melhores avaliadas pela população do estudo. Com exceção de um índice no Grupo 1, com Manga ficando em 5º lugar, nenhuma outra fruta da região (goiaba, acerola e pinha, conforme Ornellas [2007]) aparece. Essa avaliação sugere que a variabilidade no consumo de frutas não depende da regionalidade da fruta.

Comparações com a literatura revelam grande proximidade de características com os padrões dietéticos identificados no estudo de Silva *et al*. [2022], visto que o grupo 3 apresentou dados demográficos e padrões alimentares de consumo de frutas que coincidem com o padrão da "Dieta Prudente" desenvolvida no artigo, onde a dieta é encontrada em uma população mais velha e consome uma maior quantidade de frutas, verduras e hortaliças. Nesse mesmo campo comparativo, em Martins *et al*. [2014] o percentual de consumo de frutas pelos adolescentes do estado do Maranhão coincide com os resultados obtidos até a segunda colocação nos grupos 1 e 2 (Banana e Laranja), mas difere no restante das colocações. Essas similaridades observadas entre os padrões de consumo das ietas indicam um nível de proximidade entre as dietas de diferentes populações, embora estas sofram pequenas alterações regionais.

6.2 Discussão das Regras de Associação (Apriori)

As associações encontradas destacaram três principais variáveis que levaram a maiores médias de avaliação das frutas. São elas: (i) Gostar de frutas com caroço e/ou oleaginosas, (ii) Faixa etária adulta e (iii) Consumo pela manhã. A primeira afirmação contrasta com a Tabela 4, por serem grupos com frutas de menor avaliação em todo o estudo. Devido a esse desse detalhe, avaliar positivamente o grupo de alimentos com caroço e/ou oleaginosas aumenta a média de avaliação de todas as outras frutas. Os outros dados contrastam com os grupos encontrados pelo *k-means*, reforçando a ideia de que idade é uma variável determinante nesse contexto de frutas e que consumir pela manhã tende a aumentar o consumo no resto do dia.

Além destas, a quantidade de regras associadas às faixas etárias de jovens e adultos levanta as seguintes hipóteses: (i) Os hábitos alimentares dos jovens são menos homogêneos e passam por transformações significativas ao longo dessa fase da vida, resultando em padrões avaliativos mais baixos; (ii) As

regras envolvendo a população adulta surgiram em maior número devido ao tamanho reduzido do grupo, o que facilitou a aparição de padrões com suporte a partir de 10%. A primeira hipótese apresenta fundamentos consistentes, uma vez que as regras associadas a adultos geralmente reforçam preferências por frutas, enquanto, entre os jovens, foram comuns regras que confirmavam aversões a determinadas categorias ou reiteravam tendências óbvias, como a preferência por frutas moles e cítricas, o que é comum a toda a população.

Pesquisas anteriores não utilizaram regras de associação nos mesmos moldes adotados neste estudo. Portanto, não há parâmetros comparativos com relação às associações que levam ao consumo de uma fruta a outra, ou ao de não gostar de certos tipos de fruta.

6.3 Síntese Integrada

Os resultados obtidos pelo *k-means* e Apriori reforçam que os hábitos alimentares relacionados ao consumo de frutas dependem de fatores como idade, horário de consumo e preferências por frutas de categorias como "com caroço e oleaginosas".

Apesar dos avanços, algumas limitações foram identificadas: a análise restringiu-se a categorias de frutas, o que reduziu a complexidade das associações geradas pelo algoritmo Apriori. Além disso, o instrumento de coleta de dados poderia ser mais abrangente, incluindo variáveis como a frequência de consumo e uma maior diversidade etária, especialmente em faixas etárias pouco representadas, como as de crianças e idosos.

Essas limitações abrem caminhos para pesquisas futuras, que podem incorporar questionários mais detalhados, aplicar algoritmos alternativos e expandir a amostra para outras populações. Nesse sentido, este estudo não apenas descreve padrões de comportamento relacionados ao consumo de frutas da região analisada, mas também demonstra o potencial do aprendizado de máquina como ferramenta de apoio à ciência da nutrição, oferecendo subsídios tanto para profissionais da saúde quanto para o avanço da pesquisa em ciência de dados aplicada à alimentação.

7 Conclusões

Este estudo apresentou a aplicação de técnicas de mineração de dados para identificar padrões de consumo de frutas entre os habitantes do Seridó Potiguar e suas associações com variáveis regionais. Os resultados revelaram que o consumo é fortemente mediado por fatores como idade, preferências pessoais e janelas de consumo. A baixa frequência de citação de frutas regionais nos grupos resultantes do *k-means* sugere que as preferências transcendem a disponibilidade local, sinalizando uma possível padronização dos hábitos alimentares.

Do ponto de vista prático, esses achados permitem que profissionais de saúde desenvolvam orientações dietéticas segmentadas, utilizando as associações identificadas para promover uma alimentação mais diversificada e culturalmente sensível. A descoberta, via algoritmo Apriori, de que a preferência por frutas com caroço e oleaginosas atua como um preditor de uma dieta plural, constitui uma ferramenta útil para que nutricionistas antecipem a aceitação de grupos alimentares específicos em intervenções clínicas.

Por fim, embora o foco recaia sobre a realidade do Seridó Potiguar, o método estabelecido possibilita que outras regiões diagnostiquem suas próprias lacunas nutricionais e formulem intervenções baseadas em dados locais, superando as limitações de diretrizes nacionais genéricas. Trabalhos futuros devem aplicar estas técnicas em novos contextos geográficos, expandindo as variáveis coletadas para aprofundar a compreensão das dinâmicas alimentares regionais.

Declarações complementares

Contribuições dos autores

Alec Can Yalçin contribuiu com a extração de dados, mineração de dados, implementação dos algoritmos, escrita e revisão deste estudo. Karlianne Medeiros Ovidio Vale e Flavius da Luz e Gorgônio foram responsáveis pela concepção deste estudo, referencial teórico, revisão da literatura e avaliação geral dos softwares utilizados na metodologia. Raí Nabichedí da Silva apoiou a pesquisa com conhecimentos da área da nutrição através da revisão, refinamento de linguagem, comentários e sugestões. Alec Can Yalçin é o principal contribuidor e escritor deste manuscrito. Todos os autores leram e aprovam o manuscrito final.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

Disponibilidade de dados e materiais

O conjunto de dados produzido e utilizado durante o estudo se encontra disponível no Kaggle[2], todos os algoritmos e *scripts* utilizados na metodologia se encontram no Github[3].

Outras informações relevantes

A pesquisa obteve aprovação pelo Comitê Central de Ética em Pesquisa da UFRN, sob o CAEE: 78931924.5.0000.5537 e Número do Parecer: 7.083.409.

Referências

Alpaydin, E. (2020). *Introduction to machine learning. Fourth*. MIT Press.

Alves, M. d. A., Souza, A. d. M., Barufaldi, L. A., Tavares, B. M., Bloch, K. V., and de Vasconcelos, F. d. A. G. d. (2019). Dietary patterns of brazilian adolescents according to geographic region: an analysis of the study of cardiovascular risk in adolescents (erica). *´Cadernos de Saúde Pública*. DOI: 10.1590/0102-311X00153818.

Botelho, R. B. A. (2006). Culinária regional: o nordeste e a alimentação saudável. *Brasília: Universidade de Brasília*. Disponível em: `https://repositorio.unb.br/handle/10482/5236`.

BRASIL, M. d. S. (2020). Protocolo clínico e diretrizes terapêuticas de sobrepeso e obesidade em adultos. *DF: Ministério da Saúde/Conitec*. Disponível em: `https://www.gov.br/conitec/pt-br/midias/protocolos/20201113_pcdt_sobrepeso_e_obesidade_em_adultos_29_10_2020_final.pdf`. Acesso em: 9 out. 2025.

Carvalho, A., Faceli, K., Lorena, A., and Gama, J. (2011). *Inteligência Artificial–uma abordagem de aprendizado de*

[2] `https://www.kaggle.com/datasets/alecyalcin/dados-de-consumo-de-frutas-do-serid-potiguar-2025`

[3] `https://github.com/AlecYalcin/FruitLearning`

máquina, volume 2. Livros Técnicos e Científicos Editora Ltda.

Castro, L. N. d. and Ferrari, D. G. (2016). *Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações*, volume 5. Saraiva Uni.

Chika, P., Relita, B., and Sitompul, M. P. (2024). Application of apriori algorithm in determining behavioral patterns and lifestyle of gerd patients. *International Journal of Informatics, Economics, Management and Science*, 3(2):146–160. DOI: 10.52362/ijiems.v3i2.1593.

Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd, Birmingham, U.K.

Donatti, T., Henn, R. L., and Cremonese, C. (2023). Dietary patterns of elementary school students in southern brazil and associated factors: a cross-sectional school-based study. *Braz J Devel*. DOI: 10.34117/bjdv9n1-052.

Enthoven, L. and Van den Broeck, G. (2021). Local food systems: Reviewing two decades of research. *Agricultural systems*, 193:103226. DOI: 10.1016/j.agsy.2021.103226.

Hackeling, G. (2014). *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd, Birmingham, U.K.

Lazarou, C., Karaolis, M., Matalas, A.-L., and Panagiotakos, D. B. (2012). Dietary patterns analysis using data mining method. an application to data from the cykids study. *Computer Methods and Programs in Biomedicine*, 108(2):706–714. DOI: 10.1016/j.cmpb.2011.12.011.

Martins, I. C. V., Hardman, C. M., de Souza Andrade, M. L. S., dos Santos, A. R. M., Brito, A. L. S., Soares, F. C., and de Barros, M. V. G. (2020). Diferenças regionais do consumo de frutas, verduras e hortaliças em adolescentes. *RBONE-Revista Brasileira de Obesidade, Nutrição e Emagrecimento*, 14(88):906–913. Disponível em: `https://scholar.google.com/scholar?cluster=11126726254805246195`.

Martins, M. L. B., Tonial, S. R., Gama, M. E. A., Ribeiro, T. H., Ribeiro, J. M., Barbosa, J. M. A., *et al*. (2014). Consumo de alimentos entre adolescentes de um estado do nordeste brasileiro. *DEMETRA: Alimentação, Nutrição & Saúde*, 9(2):577–594. DOI: 10.12957/demetra.2014.9693.

Mewes, I. R., Jenzer, H., and Einsele, F. (2021). A study about discovery of critical food consumption patterns linked with lifestyle diseases for swiss population using data mining methods. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) - HEALTHINF*, pages 30–38. INSTICC, SciTePress. DOI: 10.5220/0010160200300038.

Niño-Adan, I., Landa-Torres, I., Portillo, E., and Manjarres, D. (2022). Influence of statistical feature normalisation methods on k-nearest neighbours and k-means in the context of industry 4.0. *Engineering Applications of Artificial Intelligence*, 111:104807. DOI: 10.1016/j.engappai.2022.104807.

Olutunde, T., Ani, C. L., and Adesue, G. A. (2024). Leveraging machine learning for personalized dietary recommendations, nutritional patterns, and health outcome predictions. *Journal of Science Research and Reviews*, 1(2):43–56. DOI: 10.70882/josrar.2024.v1i2.40.

Ornellas, L. H. (2007). *Técnica dietética: seleção e preparo de alimentos*. Atheneu Editora São Paulo Ltda., 8th edition.

Philippi, S. T. (2014). *Nutrição e técnica dietética*. Editora Manole Ltda., 3th edition.

Shao, A., Drewnowski, A., Willcox, D., Krämer, L., Lausted, C., Eggersdorfer, M., Mathers, J., Bell, J., Randolph, R., Witkamp, R., *et al*. (2017). Optimal nutrition and the ever-changing dietary landscape: a conference report. *European journal of nutrition*, 56(1):1–21. DOI: 10.1007/s00394-017-1460-9.

Silva, L. A. d., Peres, S. M., and Boscarioli, C. (2017). *Introdução à mineração de dados: com aplicações em R*. Elsevier Brasil.

Silva, V. C., Gorgulho, B., Marchioni, D. M., Araujo, T. A. d., Santos, I. d. S., Lotufo, P. A., and Benseñor, I. M. (2022). Clustering analysis and machine learning algorithms in the prediction of dietary patterns: Cross-sectional results of the brazilian longitudinal study of adult health (elsa-brasil). *Journal of Human Nutrition and Dietetics*, 35(5):883–894. DOI: 10.1111/jhn.12992.

Souza, R. S. d., Arbage, A. P., Neumann, P. S., Froehlich, J. M., Diesel, V., Silveira, P. R., Silva, A. d., Corazza, C., Baumhardt, E., and Lisboa, R. d. S. (2008). Comportamento de compra dos consumidores de frutas, legumes e verduras na região central do rio grande do sul. *Ciência Rural*, 38(2):511–517. DOI: 10.1590/S0103-84782008000200034.

Sullivan, G. M. and Artino Jr, A. R. (2013). Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, 5(4):541. DOI: 10.4300/JGME-5-4-18.

WHO Working Group (1986). Use and interpretation of anthropometric indicators of nutritional status. *Bulletin of the World Health Organization*, 64(6):929–941. Disponível em: `https://pmc.ncbi.nlm.nih.gov/articles/PMC2490974/`.