INSTITUTO FEDERAL GOIANO
CAMPUS URUTAÍ
NÚCLEO DE INFORMÁTICA
CURSO DE SISTEMAS DE INFORMAÇÃO

JAQUELINE SOUZA MORENO

CARACTERIZAÇÃO DO PERFIL SENSORIAL DE CREAM CHEESE: UMA ABORDAGEM DA MINERAÇÃO DE DADOS

Urutaí, 22 de fevereiro de 2024

INSTITUTO FEDERAL GOIANO
NÚCLEO DE INFORMÁTICA
SISTEMAS DE INFORMAÇÃO

JAQUELINE SOUZA MORENO

CARACTERIZAÇÃO DO PERFIL SENSORIAL DE CREAM CHEESE: UMA ABORDAGEM DA MINERAÇÃO DE DADOS

Monografia apresentada ao Núcleo de Informática, curso de Sistemas de Informação, do Instituto Federal Goiano, como parte das exigências para obtenção do título de Bacharel em Sistemas de Informação.

Orientador(a):
Cristiane de Fátima dos Santos Cardoso

Urutaí, 22 de fevereiro de 2024

Sistema desenvolvido pelo ICMC/USP
Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas - Instituto Federal Goiano

MJ36c

Moreno, Jaqueline Souza
Caracterização do perfil sensorial de cream cheese: Uma abordagem da mineração de dados / Jaqueline Souza Moreno; orientadora Cristiane de Fátima dos Santos Cardoso. -- Urutaí, 2024.
43 p.

TCC (Graduação em Bacharelado em Sistemas de Informação) -- Instituto Federal Goiano, Campus Urutaí, 2024.

1. Clusterização. 2. Mineração de dados. 3. Análise sensorial. 4. Árvore de decisão. I. dos Santos Cardoso, Cristiane de Fátima, orient. II. Título.

Responsável: Johnathan Pereira Alves Diniz - Bibliotecário-Documentalista CRB-1 n°2376

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO

PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610, de 19 de fevereiro de 1998, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano a disponibilizar gratuitamente o documento em formato digital no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

IDENTIFICAÇÃO DA PRODUÇÃO TÉCNICO-CIENTÍFICA

- ☐ Tese (doutorado)
- ☐ Dissertação (mestrado)
- ☐ Monografia (especialização)
- ☑ TCC (graduação)
- ☐ Artigo científico
- ☐ Capítulo de livro
- ☐ Livro
- ☐ Trabalho apresentado em evento
- ☐ Produto técnico e educacional - Tipo:

Nome completo do autor:
Jaqueline Souza Moreno

Matrícula:
2019101202010272

Título do trabalho:
Caracterização do perfil sensorial de cream cheese: Uma abordagem da mineração de dados

RESTRIÇÕES DE ACESSO AO DOCUMENTO

Documento confidencial: ☑ Não ☐ Sim, justifique:

Informe a data que poderá ser disponibilizado no RIIF Goiano: 20 /02 /2024

O documento está sujeito a registro de patente? ☐ Sim ☑ Não

O documento pode vir a ser publicado como livro? ☐ Sim ☑ Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O(a) referido(a) autor(a) declara:

• Que o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;

• Que obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autoria, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;

• Que cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Urutaí — Local
20 /02 /2024 — Data

gov.br Documento assinado digitalmente
JAQUELINE SOUZA MORENO
Data: 20/02/2024 23:16:17-0300
Verifique em https://validar.iti.gov.br

Assinatura do autor e/ou detentor dos direitos autorais

Ciente e de acordo:

gov.br Documento assinado digitalmente
CRISTIANE DE FATIMA DOS SANTOS CARDOSO
Data: 21/02/2024 14:53:27-0300
Verifique em https://validar.iti.gov.br

orientador(a)

JAQUELINE SOUZA MORENO

CARACTERIZAÇÃO DO PERFIL SENSORIAL DE CREAM CHEESE: UMA ABORDAGEM DA MINERAÇÃO DE DADOS

Monografia defendida por Jaqueline Souza Moreno e aprovada em 05 de fevereiro de 2024, pela banca examinadora constituída pelos membros:

gov.br
Documento assinado digitalmente
CRISTIANE DE FATIMA DOS SANTOS CARDOSO
Data: 05/02/2024 18:19:55-0300
Verifique em https://validar.iti.gov.br

Dra. Cristiane de Fátima dos Santos Cardoso
Orientador

gov.br
Documento assinado digitalmente
JUCELINO CARDOSO MARCIANO DOS SANTOS
Data: 05/02/2024 18:59:36-0300
Verifique em https://validar.iti.gov.br

Dr. Jucelino Cardoso M. dos Santos
Avaliador

gov.br
Documento assinado digitalmente
JEAN TOMAZ DA SILVA
Data: 05/02/2024 18:53:01-0300
Verifique em https://validar.iti.gov.br

Me. Jean Tomaz da Silva
Avaliador

Urutaí, 05 de fevereiro de 2024

Dedico este trabalho aos meus familiares e professores

AGRADECIMENTOS

Agradeço a todos que contribuíram para a realização e conclusão desta monografia e que tornaram esta jornada possível:
Primeiramente, gostaria de expressar minha sincera gratidão a minha orientadora, Cristiane de Fátima Dos Santos Cardoso, por seu apoio, orientação e paciência ao longo deste processo. Todo o seu conhecimento foi essencial e fundamental para o desenvolvimento deste trabalho. E sou grata imensamente por sua disponibilidade em compartilhar todo seu conhecimento, tempo e *expertise* para a construção da monografia.
Agradeço também aos demais professores e membros do corpo docente do Núcleo de Informática que contribuíram com seus valores, conhecimentos e *insights* e sugestões durante a apresentação de seminários, discussões em sala de aula, e em todo o processo de aprendizado que obtive durante minha jornada ao longo dos anos. Seu comprometimento com o ensino e a pesquisa acadêmica me inspirou e moldou minha compreensão sobre o assunto abordado nesta monografia.
Agradeço imensamente também aos meus colegas de classe e amigos, que proporcionaram um ambiente acolhedor, de apoio e incentivo. Todos os debates e discussões moldaram minha compreensão e me ajudaram a compreender todas as minhas ideais e argumentos dentro do processo acadêmico.
Agradeço também aos meus familiares pelo constante apoio e encorajamento ao longo de toda a graduação, sua presença, amor e apoio em todos os momentos foram essenciais para que eu possa ter levado toda a formação em um ambiente acolhedor e ao mesmo tempo contribuíram para que eu tenha sempre perseverança em minhas metas.
A todos vocês que estiveram comigo, meus mais sinceros agradecimentos por serem fundamentais para a conclusão deste trabalho e por terem acreditado em mim ao longo de toda a minha formação acadêmica.

RESUMO

A mineração de dados é muito utilizada atualmente na separação e catalogação, auxiliando na compreensão das preferências dos consumidores. A análise de grupos com o uso de dados sensoriais resulta em informações que podem melhorar buscas, determinar a autenticidade de produtos, além de possibilitar avaliações para a geração de projetos ou relatórios para diferentes análises e melhorias. Assim, este trabalho apresenta um estudo que tem como objetivo caracterização do perfil sensorial do *cream cheese* por meio da clusterização de uma base dados contendo informações sensoriais tais como cremosidade, granulosidade, nível de manteiga etc. O objetivo principal desse estudo é criar *clusters* que coloquem esses dados em seus respectivos grupos, de forma que cada amostra seja parecida internamente com as demais em seu grupo e entre outros. Ao mesmo tempo deve haver diferenças entre os grupos (*clusters)*. Para isso, utilizou análise de grupos, árvores de decisão, juntamente com os algoritmos de clusterização da linguagem R, gerando uma acurácia de 94,48% no conjunto de treinamento e de 82,11% no conjunto de teste. O resultado final da clusterização é gerado por meio do PCA e da árvore de decisão.

.

Palavras-chave:

Análise sensorial. Mineração de dados. Clusterização. Árvore de decisão.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

AUC Área sob curva. Pag.28

DKM *Clustering K-means*, diferenciável camada para compreensão de redes neurais. Pag.27

ORT Indice de raio do ganho em transversal. Pag.28

PCA *Principal Component Analysis* - Análise de Componentes Principais. Pag.24

TMI Taxa de Mortalidade Infantil. Pag.21

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

A análise sensorial visa analisar e interpretar as reações dos sentidos humanos às características dos alimentos, é um campo importante uma vez que os dados obtidos a partir de uma análise sensorial podem ser utilizados no desenvolvimento de produtos, controle de qualidade, pesquisa de mercado, inteligência artificial, mapeamento de preferências, sendo que os resultados obtidos podem ser usados em outras áreas, como psicometria, biometria ou quimioterapia (QANNARI, 2017).

A análise de dados sensoriais por meio técnicas de mineração de dados é relativamente nova na literatura (SILVA et al., 2016), mas apresenta um grande potencial. A mineração de dados é uma importante ferramenta para o processo de tomada de decisão e é definida como uma tecnologia para descobrir padrões em bases de dados, assim, a partir de dados fornecidos, são gerados padrões de comportamento, que podem ser expressos de diversas formas, por exemplo, por meio de uma função de mapeamento (SILVA et al., 2016).

É bastante comum o uso de algoritmos de mineração com a finalidade de catalogação e separação, uma vez que ele faz separações e classifica cada elemento de acordo com o seu tipo e características similares (CAROLINA; DIAS, 2012). Nesse contexto um subgrupo de técnicas é relativo à análise de grupos, o uso de algoritmos de análise de grupos (*cluster*) é uma forma de apresentar uma visão mais conceitual e ao mesmo tempo abstrata da forma como grupos estão divididos entre si, dentro de uma escala ou de um monte de informações em uma área específica e ajuda na forma que elas serão visualizadas entre suas partes (CARUSO et al., 2017). O uso de árvores de decisão em mineração de dados serve para mapear os resultados e com isso elaborar as diretrizes e passos que serão tomados ao longo da pesquisa, utilizados em passos seriais ou

paralelos de acordo com a eficiência do algoritmo (TOURNIER et al., 2007).

O uso da análise de *clusters* para explorar dados sensoriais é frequentemente utilizado no âmbito da ciência da computação em conjunto com a ciência de alimentos. É importante essas contribuições em um estudo amplo em relação ao conceito de cremosidade em certos tipos de alimentos, para identificar usando o sensoriamento verbal juntamente com a análise de *cluster*, com o objetivo principal de identificar se o consumidor leva esses fatores em conta ou não na avaliação de produtos cremosos, e também, o estudo de componentes químicos em alimentos através da técnica de análise de *cluster* pode ser usado para a identificação de componentes reativos e bioquímicos (CARUSO et al., 2017). Observando também a importância dos laticínios na alimentação das pessoas, esse estudo apresenta uma aplicação de agrupamento de dados (clusterização), com a utilização de algoritmos de agrupamento e árvore de decisão, em análises sensoriais de amostras de *cream-cheese*.

1.1 Estrutura do trabalho

Este trabalho está organizado em 4 capítulos, sendo o primeiro a presente introdução. O capítulo 2 apresenta os fundamentos básicos em relação ao *cream cheese*, a clusterização, o agrupamento dos dados e, a definição de técnicas e as etapas da mineração de dados. O capítulo 3 apresenta os materiais e métodos usados para a realização da clusterização, desde a tabela de amostras sensoriais do *cream cheese* até e a explicação dos algoritmos usados. O capítulo 4 apresenta e discute os resultados obtidos através dos algoritmos. Por fim é apresentada a conclusão.

CAPÍTULO 2

FUNDAMENTOS BÁSICOS

2.1 Cream cheese

O *cream cheese* é um produto alimentar fabricado da coagulação do leite com uma bactéria denominada *Streptococcus thermophilus*. Sua fabricação é realizada em aproximadamente 3 etapas, a primeira etapa é a separação do leite em soro e matéria sólida, por meio de um processo de fermentação no qual adiciona-se a bactéria e é realizado um descanso de 20 minutos, após o qual são acrescentados novos ingredientes para a finalização da sua composição. Nessa segunda etapa é feita a pasteurização que é o aquecimento do soro até a temperatura de 65ºC durante 30 minutos, pois a partir desse aquecimento que as enzimas trabalham na composição do creme. Por fim, é feito o seu repouso e consequente resfriamento, após esse processo, verifica-se a sua qualidade, fazendo ajustes, caso necessário e o produto está pronto para uso (SAINANI et al., 2004).

É importante observar que o queijo tradicional também é fabricado a partir da coagulação do leite juntamente com uma bactéria de acidificação, responsável pela estrutura rígida e acidez. No entanto há uma diferença importante, que é relativa às enzimas, além de quimosima, são usadas bactérias como a *Lactococci*, *Lactobacilli* ou *Streptococci*. Também são aplicados outros ingredientes como soro artificial, feito isso, o queijo fica no mínimo 24 horas em estado de descanso, em um ambiente fechado e sem umidade, após esse período checa-se o sabor e textura, realizando correções e um novo descanso, caso seja necessário. Portanto seu processo de fabricação é mais complexo e demorado que do *cream cheese* (JOHNSON, 2017).

Como resultado, o processo de fabricação do *cream cheese* resulta em um produto com uma textura mais leve e uma cremosidade mais acentuada, sendo que o seu pH é mais neutro,

enquanto um queijo comum possui um nível mais elevado de acidez, uma textura menos cremosa e uma composição mais rígida e estável. (BRIGHENTI et al., 2020).

O *cream cheese* possui teor de gordura que varia, podendo ter um alto teor de gordura ou não, esse último é usado na fabricação de *cream cheese light*. O *cream cheese* tradicional tem nível de gordura variando de 30% a 60% em relação a sua composição geral (SAINANI et al., 2004). Ele também pode variar em cremosidade e em outras características conforme a maneira como é fabricado e também conforme a matéria prima utilizada (QANNARI, 2017). **A**ssim, tais características podem ser utilizadas para a classificação de diferentes tipos de *cream cheese*.

O *cream cheese* pode ser consumido diretamente, como pasta alimentícia ou como ingrediente em receitas. Também é possível observar seu uso na fabricação de outros alimentos da área de cremes ou produtos à base de lactose.

2.2 Análises sensoriais em *cream cheese*

Os sensores referem-se ás aplicações e condições físicas do ser humano que podem ser usados como objetos de pesquisa. Essencialmente de um conjunto de operações humanas, do sentido humano, que juntamente com os demais órgãos do corpo trabalham para o processamento da informação sensorial, e isso inclui visão, olfato, paladar, tato e audição. Todos esses sensores do corpo humano são essenciais na compreensão de analises em pesquisas. O sistema sensorial é concebido através dos neurônios sensoriais, mais precisamente são classificadas como células receptoras do cérebro humano, responsáveis por captar os sinais que recebem das ações e estímulos externos. Em decorrência disso os sensores enviam sinais e respostas para o cérebro, em resumo os sentidos humanos são tradutores do mundo físico para o cérebro humano.

Dentro de uma análise sensorial são avaliados aspectos variados, como aparência (estado), cor, textura, odor e sabor. Estes aspectos são minuciosamente interpretados para compreender quais sensações e percepções o produto poderá despertar no consumidor ou se pode ser usado como objeto de pesquisa, e por meio desses aspectos podem proporcionar melhores experiências do produto em relação a sua qualidade (BRIGHENTI et al., 2008). Em seu trabalho, Brighenti et al. (2008), cita que a temática da análise sensorial em *cream cheese* se baseia principalmente na definição da cor exata de cada amostra (sensor da visão), no cheiro da amostra de *cream cheese* (sensor do olfato), e em aspectos relativos ao paladar, notadamente, a sua textura ao ser ingerido, sendo que o seu nível pode variar de uma textura cremosa para um ultra-cremoso, seu nível de acidez, sendo ele leve, moderado ou alto (BRIGHENTI et al., 2008)

Johnson (2017) trata do sentido paladar e destaca que além da cremosidade há também uma característica sensorial importante do *cream cheese* que é a sensação de arenosidade ou granulação, sendo que o nível de granulação, varia de 30% a 40% dependendo do tipo de *cream cheese*. O percentual da granulação indica um valor de 1 a 100 que reflete o quanto a amostra pode ser áspera (granulosa), sendo que os valores de 30% a 40% são margens consideradas medianas na comparação. O mesmo se aplica ao nível da arenosidade, quanto maior mais se assemelha a textura de areia (JOHNSON, 2017).

2.3 Mineração de dados

A mineração de dados envolve a aplicação de técnicas e algoritmos para mapear e buscar dados, organizando-os em grupos ou de padrões com base em suas características. O conhecimento produzido pela mineração facilita na hora de fazer agrupamentos, criar hipóteses, criar ou remodelar regras para a utilização dos dados, e permite a geração de árvores de decisão, grafos, matrizes e imagens formatadas em arquivos 3D que auxiliam a compreensão dos dados (LATOUCHE; RAMASWAMI, 1999).

É importante observar que as técnicas de mineração podem se diversificar conforme o contexto em que estão sendo aplicadas, como área de finanças, bancárias, vendas, marketing e muitos outras áreas. Um dos usos mais comuns de mineração é a busca por semelhanças entre variáveis ou de elementos, muito utilizada em varejos e vendas, a partir da informação gerada pela mineração é possível estabelecer padronizações de grupos. Á técnica de mineração é classificada como não supervisionada. A principal ideia da técnica é utilizar associações entre elementos para buscar resultados precisos para o usuário, vinculadas aos atributos ou a algo que esteja em concordância com a preferência do cliente. As principais etapas da mineração de dados são descritas nas seções seguintes.

2.3.1 Seleção

Nessa fase são selecionados os conjuntos de dados, gráficos, artigos textuais ou imagens, conforme o problema em questão e a solução a ser considerada. Geralmente os dados provenientes de bases de dados possuem mais informações que o necessário, então a seleção de dados e recursos faz essa divisão, separando o que é útil e o que não é (MAIONE et al., 2017), portanto, a seleção de recursos é fundamental para a criação de um bom modelo, pois corrobora com a redução de cardinalidade(SHERER; DUNCAN, 2022).

Um exemplo de seleção dos dados é apresentado por Braz et al. (2020) cujo objetivo é a realização da análise de *clusters* para tipologia de paisagens. Observou-se a existência de diferentes taxonomias e noções de zonalidade. Diante disso, foi escolhido um modelo taxonômico e um modelo teórico de geossistemas, também foram selecionadas apenas paisagens do município de Mineiros, em Goiás.

2.3.2 Pré-processamento

O pré-processamento trabalha com a organização e preparação dos dados que foram selecionados para os experimentos e observações, resultando em uma melhor estruturação do que vai ser utilizado na fase de processamento. Uma de suas principais tarefas é realizar uma limpeza de dados por meio da remoção de dados desnecessários, deixando apenas o que será utilizado. Também é importante que se tenha os dados bem definidos e estruturados, assim, essa etapa também é responsável por realizar uma melhor estruturação dos dados, possibilitando melhores resultados (FRØST, 2002). Como exemplo de pré-processamento, o trabalho de (BRAZ et al., 2020) destaca a importância da correção de incidências das unidades na base de dados. Isso inclui o arredondamento de valores, a agregação de valores reais a unidades que estejam em falta, a eliminação de redundâncias e a depuração de dados ruidosos da tipologia de paisagens. Além disso, o estudo enfatiza a eliminação de informações redundantes e valores inconsistentes, informações aleatórias.

Em outras palavras, o processo de pré-processamento visa remover dados estranhos que o sistema não conseguirá interpretar, bem como valores aleatórios ou conjuntos de características que não podem ser adequadamente interpretados pelo sistema. Esse cuidadoso tratamento dos dados contribui significativamente para a qualidade e confiabilidade das análises subsequentes.

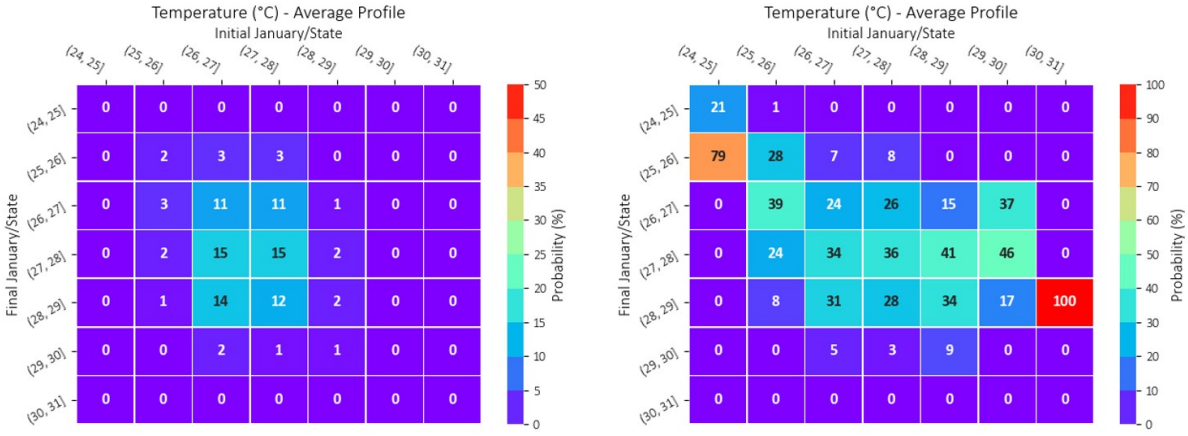
2.3.3 Mineração

Nessa etapa, técnicas de mineração de dados são empregadas com o propósito de identificar padrões, tendências e relações entre os conjunto de dados. Algoritmos e métodos estatísticos são aplicados para explorar os dados de forma a prever ou descrever fenômenos. Assim, dois formatos de saída comuns são as matrizes e a identificação de agrupamentos. Esses processos ajudam a extrair *insights* valiodos.Carvalho e Bleninger (2021) mostra o uso de técnicas de mineração de dados para identificar as variações da temperatura da água em um reservatório. A Figura 2.1 exibe uma matriz em que as colunas representam os possíveis estados iniciais e

as linhas os possíveis estados finais, sendo que a matriz da esquerda trata das probabilidades referentes a temperatura da água em um reservatório. A matriz da direita é a matriz de transição de estado da mesma coluna de água, ou seja reflete as possíveis mudanças de estado. A mineração de dados auxilia na análise das temperaturas que resulta na geração da matriz de estados.

Já a tabela da Figura 2.7 reflete uma análise de *clusters* feita para uma pesquisa de tipologia de paisagens. Essa tipologia demonstra aplicabilidade na cartografia de imagens e no mapeamento de relevos. No contexto desse estudo, diversas paisagens geográficas foram agrupadas para análise de suas unidades ou classificações. Utilizando técnicas de mineração de dados em conjunto com análises de *clusters*, determinou-se a quantidade de *clusters*, e em qual *cluster* cada tipo de paisagem (unidade) vai ser encaixada, a tabela mostra o valor de definição de cada unidade e sua classificação no grupo (BRAZ et al., 2020).

Figura 2.1: Matriz de probabilidade (esquerda) e matriz de transição de estado (direita) – Perfil de temperatura média de profundidade da coluna de água.



Fonte: (CARVALHO; BLENINGER, 2021)

2.3.4 Análise e avaliação dos resultados

A análise dos resultados desempenha um papel crucial na garantia da utilidade dos dados para resolver o problema em questão. Nessa fase, métricas fundamentais como precisão, revocação, e acurácia são empregadas para mensurar o desempenho do modelo. Além disso, a análise de gráficos, como dispersão, linhas, barras, setores, entre outros, é aplicada para proporcionar uma compreensão visual dos resultados (SILVA et al., 2016). A avaliação abrangente dos resultados é seguida por uma seleção criteriosa dos melhores visando obter dados equilibrados em conformidade com os modelos propostos e os objetivos pretendidos (QURESHI et al., 2012). Um método de análise e avaliação de resultados foi usado na pesquisa de matrizes

Figura 2.2: Matriz resultante da análise de *cluster* por *k-médias* para tipologia de paisagens

Unidade	*Cluster* (Grupo)	Unidade	*Cluster* (Grupo)
1	25	45	2
2	25	**46**	2
3	25	[...]	[...]
[...]	[...]	270	24
43	17	**271**	**24**
44	17	272	24

Fonte: (BRAZ et al., 2020)

de estado da água como ferramenta de análise e previsão. Foi usado o método de clusterização que permitiu que os resultados fossem gerados em gráficos de reposição de linhas e colocados os resultados em notas de texto e valores absolutos ou fracionários. A partir das informações obtidas foi criada uma matriz de transição com intervalos discretos. Nesse contexto, é plaúsivel argumentar que os gráficos desempenham um papel crucial como ferramentas de representação no estudo para avaliar os critérios de condição do estado da água. Conforme evidenciado por esta pesquisa, os resultados foram apresentados de maneira elucidativa através de gráficos de linha. Essa escolha visual permite uma interpretação mais clara das tedências e variações ao longo do tempo, facilitando a análise dos resultados finais.

2.4 Clusterização

Categorizar e segmentar dados é sempre muito útil em diversos contextos, como no *marketing*, para estudos de mercado, ou até mesmo para pesquisas internas de instituições e de laboratórios. A clusterização é uma ferramenta muito importante pois visa facilitar o estudo sobre grupos de dados, permitindo uma análise abrangente, e até a exclusão de parâmetros irrelevantes para o problema em questão (MAIONE et al., 2017).

No processo de clusterização, uma técnica é utilizada para guiar a maneira como os elementos serão separados. Tais técnicas são concebidas seguindo diferentes métodos, tais como métodos hierárquicos, métodos por particionamento, métodos baseados em *grid* e métodos utilizando aprendizado de máquina, a Figura 2.3 mostra essa classificação. Métodos hierárquicos são subdivididos em métodos aglomerativos e divisores. O método de aglomeração (agrupamento)

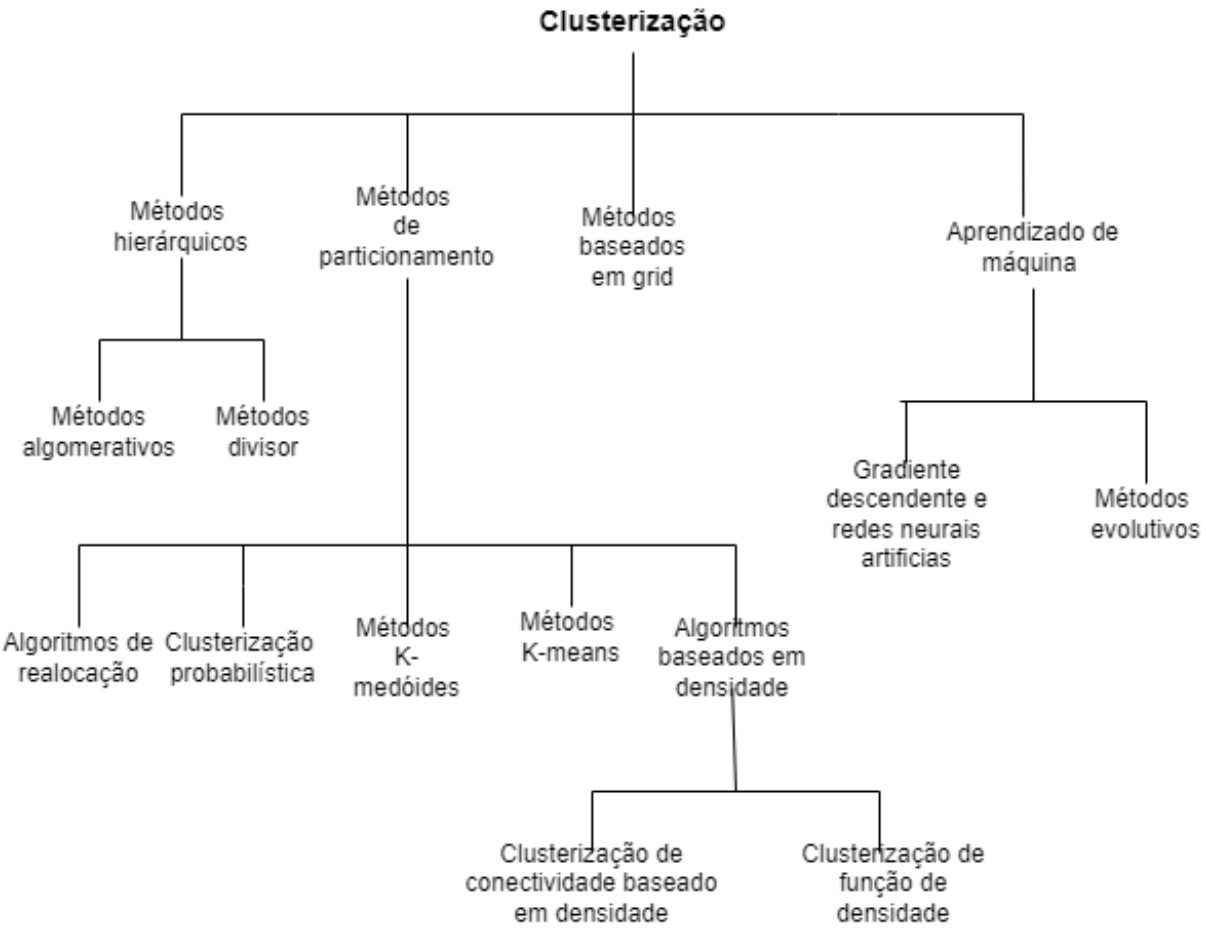
segue uma abordagem *bottom-up*, uma vez que esse método começa com cada objeto formando um *cluster* separado e após isso ele mescla os *clusters* próximos em ordem sucessiva até que todos os *clusters* sejam mesclados em um único grupo maior de hierarquia, ou até que uma condição final de encerramento seja válida.

O *clustering* hierárquico busca construir uma hierarquia de clusters, sendo que as estratégias utilizadas foram apresentadas anteriormente, na seção 2.3. Na abordagem aglomerativa ou “de baixo para cima” cada observação começa em seu próprio *cluster*, e pares de *clusters* são mesclados à medida que alguém sobe na hierarquia O método divisor começa com todos os objetos no mesmo *cluster*, em cada iteração um *cluster* é dividido em *clusters* menores, até que cada objeto esteja em um *cluster* ou uma condição de término seja alcançada. As observações começam em um *cluster* e se dividem, de maneira recursiva a medida que desce na hierarquia. O agrupamento hierárquico requer uma medida de dissimilaridade (ou distância) e um critério de aglomeração, esse método divisivo também é chamado de "de cima para baixo".

O método divisor tem início com todos os objetos atribuídos a um único *cluster*. A cada iteração, um *cluster* é subdividido em *clusters* menores, continuando esse processo até que cada objeto esteja alocado em um *cluster* individual ou até que uma condição de término predefinida seja satisfeita. Essa abordagem hierárquica implica que as observações começam em um único cluster e se subdividem entre si. Algoritmos típicos de particionamento incluem relocação, clusterização probabilística, métodos *k-medoids*, métodos *k-means* e algoritmos baseados em densidade. Uma outra classe muito importante de métodos de clusterização é relativa aos métodos de clusterização baseados em aprendizado de máquina, sendo os principais baseados no gradiente descendente e redes neurais.

Um exemplo de aplicação de método de clusterização é a análise de dados de *recall* automotivo realizado por Maione et al. (2023). O foco desse trabalho é investigar a tendência mundial de crescimento do número de *recalls*, e também o número de produtos envolvidos em cada campanha. Foram usados algoritmos de aprendizado de máquina não supervisionado para obter e agrupar os dados de 2010 a 2019, realizando-se em seguida a análise dos gráficos do Brasil e do EUA. A Figura 2.4 apresenta o gráfico dos dados agrupados de *recalls* no Brasil, e a Figura 2.5 apresenta o gráfico dos dados agrupados de *recalls* dos EUA, o primeiro gráfico representa os *clusters* com base no número de *recalls* por mês, o segundo, os *cluster* com base no número de veículos afetados por problemas relacionados a manutenção e o último é uma junção dos dois critérios, ou seja, o número de problemas de manutenção juntamente com o número de

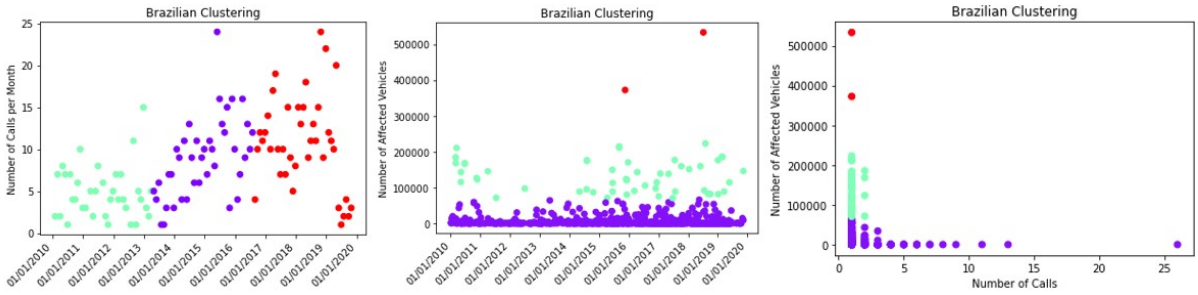
Figura 2.3: Tipos de algoritmos para *clusterização*



Fonte: Adaptado de <https://medium.com/@pytyagi/clustering-83f210299e47>

recalls relacionados simultaneamente. Os autores do trabalho determinaram que 3 clusters é o ideal para este problema.

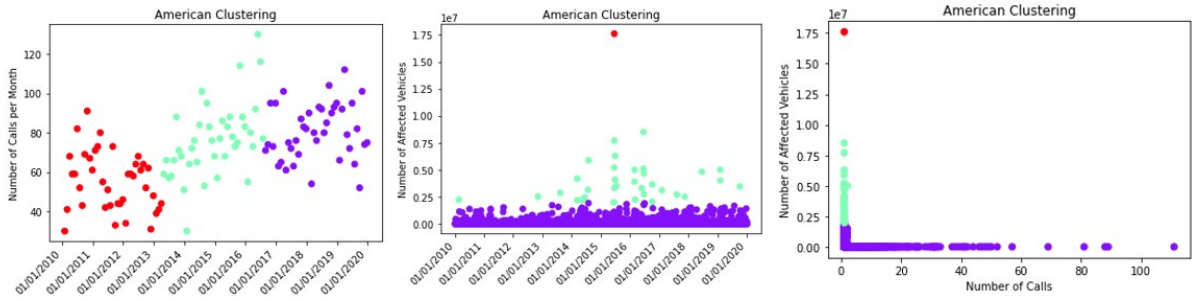
Figura 2.4: Dados agrupados de *recalls* brasileiros



Fonte: (MAIONE et al., 2023)

Considerando os modelos evolucionários pode-se citar a pesquisa realizada por (ALLEM et al., 2022) em que o objetivo é analisar a evolução da pandemia de Covid-19 no estado do Rio Grande do Sul aplicando ferramentas teóricas de grafo, juntamente com análises de *clusters*

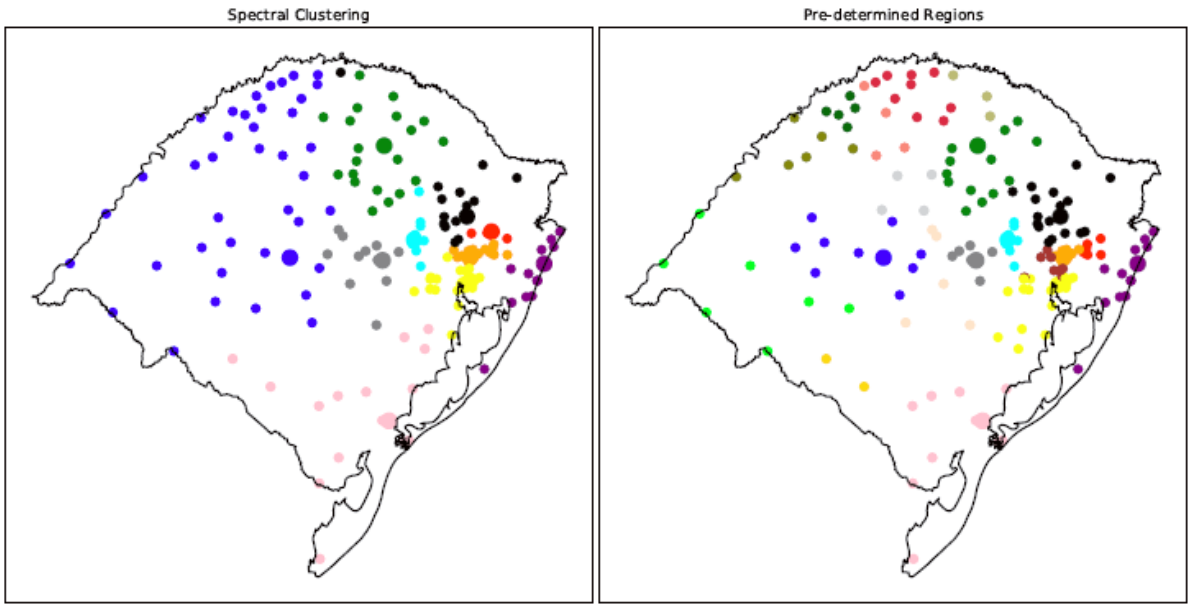
Figura 2.5: Dados agrupados de *recalls* dos EUA.



Fonte: (MAIONE et al., 2023)

espaciais evolutivos. Foram analisados os casos em 167 municípios do estado do Rio Grande do Sul, em que foram usadas técnicas de agrupamento espectral, baseada na teoria de grafos espectrais, a Figura 2.6 apresenta os resultados após o particionamento espectral para k=10 agrupamentos, sendo que a maior cidade em cada *cluster* está marcada com um círculo maior e na subfigura da direita é possível verificar a semelhança com as regiões definidas pelo governo.

Figura 2.6: Agrupamento obtido por agrupamento espectral em relação à medida A_0 para k=10 *clusters* em comparação com as regiões pré-determinadas

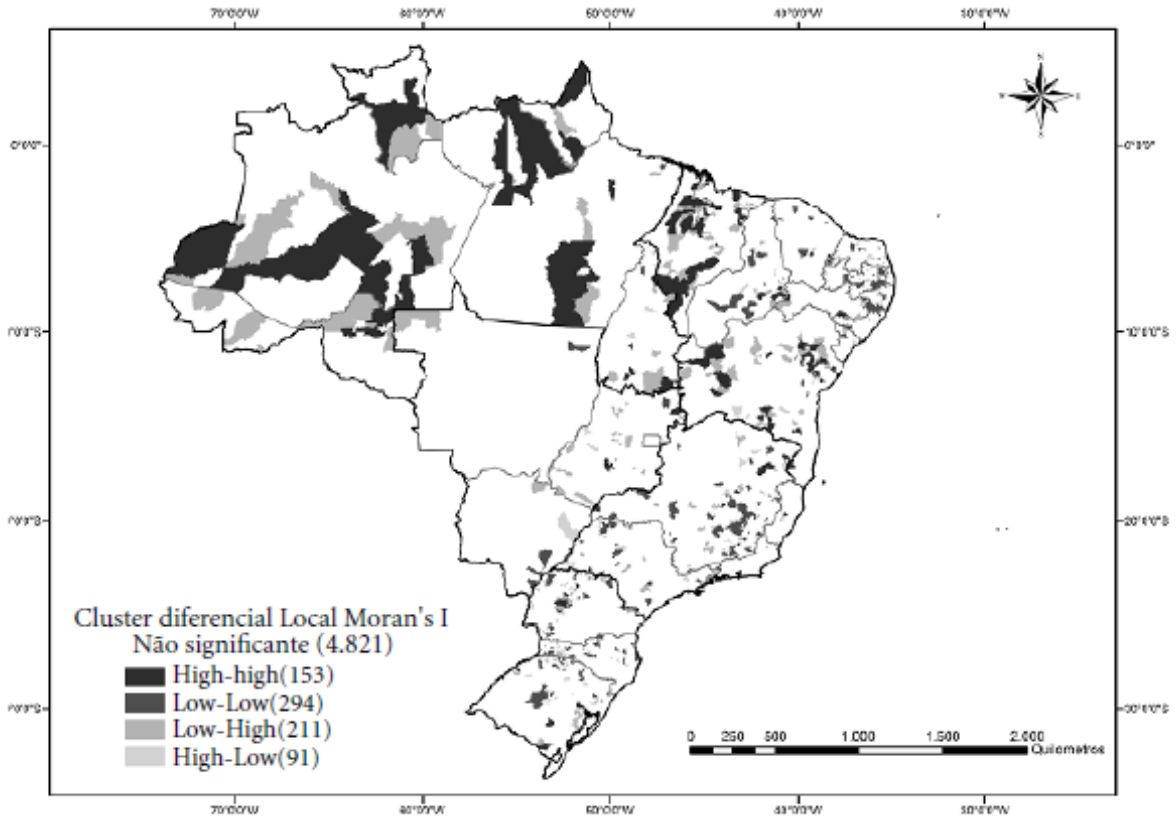


Fonte: (ALLEM et al., 2022)

Um exemplo de clusterização é o uso do k-means feito por Pasklan et al. (2021) em que o foco é realizar uma análise espacial da qualidade dos serviços de atenção primários a saúde objetivando a redução da mortalidade infantil. A coleta dos dados ocorreu entre julho de 2016 até maio de 2018, sendo que foi aplicada análise de clusterização espacial por meio de abordagem diferencial da estatística local de I de Moran, com o uso dessa técnica foi possível analisar de

forma diferencial a presença de *clusters* espaciais quanto a Taxa de Mortalidade Infantil (TMI) nos municípios brasileiros entre os anos de 2000 a 2015, (PASKLAN et al., 2021).

Figura 2.7: Cluster diferencial local Moran‘s I das taxas de mortalidade infantil entre 2000 e 2015



Fonte: (PASKLAN et al., 2021)

A Figura 2.7 apresenta a relação da taxa de mortalidade diferencial entre os municípios do Brasil. Nesta análise, um total de 749 cidades foram investigadas, e cada município foi categorizado com base no cluster ao qual foi atribuído, fornecendo uma visão abrangente das disparidades de mortalidade em diferentes regiões do país. O *High-High* referenciado na legenda e demarcado nos municípios apresenta uma classificação do TMI como muito alta nessas cidades, ou seja, não houve uma melhora da qualidade dos serviços para combater a TMI. Na demarcação identificada como "*Low-Low*", os municípios exibiram uma taxa de Mortalidade Infantil (TMI) notavelmente baixa. Isso sugere que essas localidades apresentam serviços de qualidade que desempenham um papel eficaz na prevenção de mortes infantis, ressaltando a eficácia de suas iniciativas de sáude e bem-estar. A legenda que indica os municípios com o *cluster Low-High*, apresenta as cidades onde teve uma equilibrada de TMI de baixo para alto, ou seja, essas cidades estavam com uma TMI baixa e por questões internas houve um pequeno aumento, ficando na margem de baixo para alto nessa classificação. Por fim o *cluster High-Low* apresenta uma

margem diferencial de municípios onde teve um aumento significativo de mortes na TMI, e gradualmente ela foi abaixando até entrar na categoria Low, ficando com um equilíbrio gradual de alto para baixo na TMI.

Portanto, cada problema exige uma abordagem diferente. Quando se tem uma pesquisa por exemplo para encontrar valores de *clusters* em uma proposta geral, é necessário acrescentar os parâmetros usados na pesquisa, depois aplicar esses valores no *k-means* ou outro algoritmo de clusterização, para ter os dados agrupados. Algoritmos como o *k-means* são responsáveis por separar os padrões formando grupos que ao mesmo tempo que são diferentes uns dos outros, mas que os dados de cada grupo devem possuir o máximo possível de características semelhantes entre si, assim, o objetivo é maximizar a similaridade intragrupo e ao mesmo tempo minimizar a similaridade intergrupo (SILVA et al., 2016). São feitas equiparações para se tentar chegar ao equilíbrio de um conjunto de valores que possam ser de 80% ou mais de semelhanças entre os dados usados e os resultados finais obtidos (CHARRAD et al., 2014).

Ao agrupar é necessário fazer a validação, que é a avaliação do resultado do agrupamento, para garantir que o agrupamento realizado condiz com a realidade. Essa validação utiliza cálculo da distância entre as amostras dentro dos grupos para verificar a separabilidade entre os grupos e a compacidade (compactação) do grupo, sendo que podem ser usadas diversas medidas de distância: máxima, mínima, distância média etc (SILVA et al., 2016). A avaliação é feita por meio dos chamados “índices”, que podem ser baseados em critérios internos ou externos. Índices externos necessitam de conhecimento sobre a estrutura do conjunto de dados, enquanto que índices internos não necessitam de qualquer conhecimento. Como exemplo de índice externo, pode-se citar: índice de *Rand*, índice *Jaccard*, índice de *Folkes* e *Mallows*, já índices internos tem se o índice *Dunn*, índice *Davies-Bouldin* dentre outros.

CAPÍTULO 3

MATERIAIS E MÉTODOS

3.1 Dataset

A base de dados utilizada nos experimentos foi proposta por Frøst (2002) e consiste de uma tabela de características do *cream-cheese*, em que os valores para cada atributo são obtidos por meio da classificação realizada por especialistas em análise sensorial. A tabela abrange aspectos como acidez, cremosidade, cores (branco, amarelo, cinza), resistência, níveis de sal e açucar, dentre outras características. A Figura 3.1 mostra a base em questão, sendo no total 240 amostras e 23 características.

Figura 3.1: Perfil sensorial de dados de *cream cheese*

	Product_name	N_Cream	N_Acidic	N_Butter	N_OldMilk	E_White	E_Grey	E_Yellow	E_Green	H_Resistance	E_Grainy	E_Shiny	M_Firm	M_Melt down	M_Resistance	M_Creamin
1	16%'	5.55	5.40	5.40	5.10	9.45	2.40	4.80	0.75	5.40	1.50	12.75	7.35	8.40	7.35	
2	16%'	8.55	6.75	7.80	3.00	9.90	3.60	3.75	3.00	4.20	4.05	12.90	2.70	4.80	5.40	
3	16%'	7.05	9.00	9.60	1.95	10.80	1.65	2.25	0.75	6.60	1.65	9.00	5.55	10.95	6.45	
4	16%'	8.55	6.00	10.80	1.95	7.95	4.05	4.65	1.80	4.50	7.95	10.35	4.20	13.20	1.65	
5	16%'	8.85	8.70	9.75	4.95	9.75	2.85	1.95	1.80	5.25	8.10	11.10	6.60	10.35	4.80	
6	16%'	10.65	3.90	9.60	3.00	10.20	2.10	3.30	2.55	4.20	3.30	12.30	7.35	10.05	5.40	
7	16%'	7.95	10.80	10.35	1.80	12.15	3.30	1.20	0.45	5.85	1.65	9.15	3.30	9.15	4.35	
8	16%'	9.00	6.75	6.45	3.15	6.75	5.55	4.20	3.30	6.00	5.25	5.85	6.00	4.95	6.30	
9	16%'	6.00	8.40	7.20	1.65	10.80	2.70	1.35	0.90	6.75	0.45	13.65	7.35	9.45	7.05	
10	16%'	7.65	6.60	6.90	2.70	9.15	2.70	4.20	2.70	5.55	3.90	10.50	4.20	7.20	4.35	
11	16%'	7.50	10.80	10.35	1.95	10.80	2.40	1.95	1.35	5.70	1.50	11.40	6.15	8.25	6.75	
12	16%'	5.10	12.00	4.05	7.05	9.45	4.20	4.95	1.80	2.85	2.70	10.95	5.70	10.35	4.65	
13	16%'	7.65	6.90	9.45	5.55	10.20	2.40	3.75	1.95	4.05	6.00	10.50	6.45	10.35	6.15	
14	16%'	7.80	7.95	9.30	1.50	10.35	1.95	1.95	1.50	3.00	7.20	10.95	3.60	10.65	4.35	
15	16%'	7.80	8.85	7.50	1.05	11.70	1.05	2.70	1.05	5.10	1.65	9.90	2.40	9.90	3.45	
16	16%'	7.20	5.55	6.60	1.95	8.10	2.55	5.55	1.65	6.30	3.60	6.00	5.25	6.75	4.95	
17	16%'	6.00	7.20	6.15	1.80	10.50	2.40	0.90	0.00	5.40	1.20	13.05	7.65	7.35	7.05	
18	16%'	8.55	7.95	7.65	3.00	9.90	2.40	3.00	2.10	2.85	3.45	12.30	5.10	6.60	10.20	
19	16%'	8.70	10.95	10.35	1.65	9.75	3.30	2.55	1.50	3.75	1.50	11.40	6.90	7.65	7.65	
20	16%'	2.40	8.40	7.80	5.40	9.00	2.40	3.30	1.95	3.90	5.25	12.60	4.65	12.00	2.55	
21	16%'	8.10	9.45	8.55	4.80	10.20	2.40	2.85	2.10	4.05	3.00	10.95	3.00	11.85	2.70	
22	16%'	7.20	7.35	10.50	2.25	10.20	1.65	3.30	1.65	4.20	4.50	11.85	4.35	12.15	2.85	

Fonte:(FRØST, 2002)

A Tabela 3.1 apresenta uma breve explicação dos dados usados para formar os clusters com base nos *cream cheese* apresentados, na coluna da direita tem-se a definição de cada

atributo de cada amostra do *cream cheese* e na coluna da esquerda uma breve explicação sobre a característica analisada na pesquisa sensorial.

Tabela 3.1: Características do *cream cheese*

Atributo	Característica do atributo
N-Cream	Nível de cremosidade do produto
N-Acidic	Nível de acidez do produto
N-Butter	Nível de manteiga/gordura em produto
N-OldMilk	Quantidade e nível de leite velho na composição do produto
E-White	Nível do quanto o produto tem de brancura na sua aparência
E-Grey	Nível do quanto o produto tem de cinza na sua aparência
E-Yellow	Nível do quanto o produto tem de amarelo na sua aparência e composição
E-Green	Nível do quanto o produto tem de verde na sua aparência e composição
H-Resistance	Nível da resistência da amostra do *cream cheese*
H-Grainy	Nível de granulosidade da amostra do *cream cheese*
E-Shiny	Nível do quanto a amostra tem de aparência brilhante
M-Firm	Nível do quanto a amostra tem de firmeza em sua textura
M-Meltdown	Nível de fusão da amostra em relação aos componentes gerais do *cream cheese*
M-Resistance	Nível da resistência em relação a proximidade de firmeza dos demais níveis
M-Creaminess	Nível da cremosidade da amostra em textura
M-Grainy	Nível da granulosidade da amostra em textura com relação aos demais níveis de comparação.
M-Chalky	Nível do quanto a amostra tem de duro e sólido na amostra
M-Cream	Nível do quanto a amostra tem de creme em sua composição
M-Fat	Nível de gordura da amostra
M-Butter	Nível de manteiga na composição da amostra de *cream cheese*
M-Salt	Nível de manteiga na composição da amostra de *cream cheese*
M-Sour	Nível do quanto de azedura existe em sua composição da amostra
M-Sweet	Nível de açúcar na amostra do *cream cheese*

3.2 PCA

O PCA (*Principal Component Analysis* - Análise de Componentes Principais) é uma técnica que tem a função de derivar um conjunto de dados de baixa dimensão a partir de um grande conjunto de objetos. Essa técnica está classificada como aprendizado de máquina não supervisionado e pode ser usado na limpeza e pré-processamento dos dados, o PCA também contribui para compactar as informações e transmiti-las.

Uma vantagem muito importante do PCA em conjunto com a aprendizado de máquina

ajuda a simplificar algoritmos mais complexos de negócios, minimiza a variação mais significativa de dimensões, facilita a eliminação de informações irrelevantes, tais como ruídos ou fatores externos. Ele também pode ser usado para redimensionar imagens, para analisar dados de estoque ou dados que possam ser previstos (GARG; CHADHA, 2020). Conforme (SILVA et al., 2016), O algoritmo para obter o PCA é dado pelos seguintes passos: inicialmente considere a projeção em M quando $dim(M) = 1$. Seja um vetor unitário $u_1 \in D$ a direção de M. Então a projeção de uma observação $x_n \in X$ em M é $u_1^T x_n$ e a variância dos dados projetados é

$$\frac{1}{N}\sum_{n=1}^{N}\{u_1^T x_n - u_1^T \bar{x}\} = u_1^T S u_1 \tag{3.1}$$

onde $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_i$ é a média do conjunto amostral e S é a matriz de covariância do conjunto de dados X:

$$S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T \tag{3.2}$$

A maximização de (3.1) é mantida em um círculo unitário, pois escolhemos u_1 s.t. $||u_1|| = u_1^T u_1 = 1$. Portanto, precisamos encontrar o máximo da próxima função de Lagrange:

$$L(X, \lambda_1) = u_1^T S u_1 + \lambda_1(1 - u_1^T u_1) \tag{3.3}$$

Ao definir a derivada em relação a u_1 igual a zero, descobrimos que no ponto estacionário u_1 precisa ser um autovetor de S:

$$S u_1 = \lambda_1 u_1 \tag{3.4}$$

Agora, quando multiplicamos à esquerda por u_1^T e usamos $u_1^T u_1 = 1$, descobrimos que a variância é dada por

$$u_1^T S u_1 = \lambda_1 \tag{3.5}$$

e então a variância será máxima quando definirmos u_1 igual ao autovetor tendo o maior autovalor λ_1. Este ID de autovetor é chamado de primeiro principal componente (SILVA et al., 2016). Os próximos componentes principais podem ser encontrados seguindo o mesmo procedimento e escolhendo cada nova direção de forma que maximize a variância projetada entre todas as direções possíveis ortogonais àquelas já consideradas.

3.3 Kmeans

O algoritmo *K-means* faz parte da família de algoritmos de agrupamento com centróides, otimizando a organização de dados em *clusters*, distintos com características bem definidas.

O processo inicia com a definição prévia do número de *clusters* (K), seguido pela criação de centróides inicias para cada grupo. A etapa subsequente envolve a associação iterativa de cada objeto ou dado ao seu *cluster* mais próximo, assim o algoritmo do k-means trabalha em iterações, até que cada objeto esteja em seu *cluster* e não tenha mais trocas ou iterações (MIGUEL, 2023).

Algoritmo 4-3: Algoritmo para agrupamento por partição k-médias

Parâmetros de entrada:

X_{tr}: conjunto de exemplares não rotulado de treinamento,ou seja, $X_{tr} = \left\{ \overrightarrow{X_i} \right\}, i = 1, ..., n$;

K: o número de partições (ou grupos) a serem descobertos;

$dist$: medida de distância;

Parâmetros de saída:

k: vetores que representam centroides de grupos, representando as partições descobertas;

Passo 1: escolha aleatoriamente um conjunto de vetores distintos para representar os centroides, ou seja, $C = \left\{ \left(\overrightarrow{c_p} \right) \right\}, p = 1, ..., K)$

Passo 2: **enquanto** houver alterações nas associações dos exemplares aos grupos representados por cada centroide **faça**

Passo 2.1: verifique a distância $dist(\overrightarrow{X_i}, \overrightarrow{c_p})$ para cada exemplar em i em n e cada centroide p em k;

Passo 2.2: associe cada exemplar $\overrightarrow{x_i}$ ao vetor $\overrightarrow{c_p}$ que minimiza $dist(\overrightarrow{X_i}, \overrightarrow{c_p})$, formando cada uma das k partições do conjunto $\overrightarrow{x_{tr}}$;

Passo 2.3: atualize o conjunto de centroides C, de forma que cada vetor $\overrightarrow{c}$ seja a média dos vetores de $\overrightarrow{x}$ associados a ele.

3.4 nbCluster

O pacote *nbClust* se destaca como uma ferramenta abrangente, oferecendo 30 índices para a determinação eficiente do número de *clusters*. Sua abordagem inovadora propõe ao usuário o melhor esquema de *clusterização*, "explorando uma ampla gama de resultados obtidos através da variação sistemática de números de *clusters*, medidas de distância e métodos de clusterização" (CHARRAD et al., 2014). São utilizados o k-means e o agrupamento hierárquico com diferentes medidas de distância e métodos de agregação como métodos de clusterização e em função da grande variabilidade utilizada, torna-se possível a avaliação de vários esquemas

de agrupamento simultaneamente. As distâncias utilizadas são: distância euclidiana, distância máxima, *manhatan*, *canberra*, binária e *minkowski*, enquanto que os métodos de aglomeração são *Ward*, *ward.d2*, *single*, *complete*, *average*, *mcquitty*, *media-n*, *centroid* e outras.

Em muitas situações que envolvem clusterização, o usuário enfrenta uma dificuldade que é escolher o melhor número de *clusters* ou partições nos dados subjacentes, assim, o **NbCLuster** é um pacote da linguagem R que pode ser usado no geral para selecionar o número ideal de *clusters* para cada valor de índice, n= número de observações, p = número de variáveis, q = número de clusters, e o resultado do *nbCluster* é o possível número de *clusters* que devem ser formados com os dados em questão. O *nbCluster* também pode colaborar em tarefas com o foco de encontrar semelhanças entre os coeficientes de valores em uma pesquisa (CORNELISSEN, 2021).

3.5 Árvore de decisão

Árvore de decisão é um modelo hierárquico de suporte a decisões, usado principalmente na classificação de dados, no qual as consequências de um fluxo guia a tarefa de classificação, é uma maneira de apresentar um algoritmo com condicionais aninhadas. A árvore de decisão é composta por uma série de nós e ramos, sendo que cada ramo representa um curso alternativo de ação ou decisão. No final de cada ramo ou curso alternativo há outro nó que representa um evento de condição verdadeira, caso contrário uma condição falsa será redirecionada (MAGEE, 1964). Assim, a estrutura de uma árvore é muito semelhante a um fluxograma, em que cada nó interno representa um "teste"em um atributo, ou seja uma decisão tomada para dar continuidade aquele determinado fluxo em segmento.

Árvores de decisão são comumente usadas em pesquisa operacional ou gerenciamento de operações, e também para a classificação ou para regressão, uma vez que apresentam uma melhor visualização do esquema de classificação das amostras, tornando mais fácil a justificativa na hora de tomar decisões, pois permitem ao leitor uma visão ampliada do passo a passo da quebra de um problema (BREIMAN et al., 2013), (SILVA et al., 2016). Esse esquema é feito por meio de uma ordem de avaliação das características de tais amostras e de como seria o seu resultado em uma árvore organizada escalonável, para isso, critérios de seleção de atributos são adotados, como exemplo de tais critérios tem-se: ganho de informação com base na entropia, índice Gini com base na impureza, *Likelihood Ration*, DKM (*Clustering K-means*, diferenciável camada para compreensão de redes neurais) , raio do ganho, *twoing*, ORT (Indice de raio do ganho em

transversal) , *Kolmogorov_Smirnov*, AUC (Área sob curva) , *splitting*, assim, tais critérios são aplicados para dizer qual atributo é o mais adequado em um dado momento, para ser associado ao nó da árvore.

3.5.1 Árvore de decisão C5.0

Considerando a existência de diversos critérios de seleção, e diversas maneiras de construir e manter uma árvore, existe vários modelos de árvore de decisão. Um modelo que tem sido bastante utilizado, é a árvore de decisão do tipo c5.0, esse algoritmo que se tornou padrão pelas suas vantagens, e possui desempenho quase tão bom quanto modelos de aprendizado de máquina mais sofisticados e avançados, como redes neurais e máquinas de vetor suporte. As árvores de decisão no algoritmo *c5.0* são mais fáceis de entender e de implementar são bastante robustas na presença de problemas com dados incompletos e grande número de campos de entrada (SMITH; LEE, 2023).

Um modelo *c5.0* divide a amostra com base no campo que fornece o máximo de ganho de informações e cada subamostra definida pela primeira divisão é dividida novamente, geralmente com base em um campo diferente, repetindo-se esse processo até que não possa ser realizada nenhuma subdivisão. As divisões de nível inferior que não contribuem com o valor do modelo são podadas (SMITH; LEE, 2023).

Portanto, a definição desse tipo de árvore se baseia na entropia de valores dos dados para assim ser possível medir a pureza. A entropia de uma amostra de dados indica quão mistos são os valores das classes, a abordagem das árvores de decisão do tipo C5.0 para predição consiste em dividir os dados em grupos cada vez menores, ou seja, transformando cada grupo dessa partição em um ramo separado de conjunto para que se crie uma condição ou evento a ser seguido. Em cada observação de cada nó da árvore apresenta uma maioria de elementos da mesma classe (MAGEE, 1964).

3.6 Método proposto

Com o propósito de agrupar o c*cream cheese*, com base em suas características provenientes da análise sensorial, a metodologia incia-se pela obtenção de uma base de dados pública contendo tais informações. Em seguida, realiza-se uma minuciosa limpeza manual dos dados, excluindo informações que não contribuem para a análise sensorial. Após a limpeza, o *nbCluster* é aplicado para obter a quantidade ideal de *clusters* que as amostras podem gerar. Por fim, o

algoritmo *k-means* é usado para obter a clusterização.

1. **obter base de dados** utilização de base pública sobre análise sensorial do *cream cheese*;

2. **remover dados desnecessários** as colunas *panellist, Replicate, Session, Serving order e Productnames*, são variáveis desnecessárias para a obtenção dos clusters e por isso são removidas;

3. **criar PCA de visualização inicial** da tabela de amostras original, esse primeiro PCA mostra como são os dados sem a clusterização;

```
res.\sig{PCA} <- \sig{PCA}(dados[, c(2:24)], graph = FALSE)
factoextra::fviz_\sig{PCA}_ind(res.\sig{PCA},
         geom.ind = "point", # show points only (nbut not "text")
         col.ind = dados$Product_name, # clor by groups
         #palette = c("#00AFBB", "#E7B800", "#FC4E07"),
         addEllipses = TRUE, # Concentration ellipses
         legend.title = "Groups"
)
```

4. **aplicar nbcluster** com distância de *Manhattan*, o número mínimo de sementes igual a 2, o número máximo de sementes igual a 10, método de agregação *"complete"* e o *index* sendo em *"all"*, o nbcluster resulta nos possíveis números de clusters. *Complete* é um parâmetro dentro do algoritmo de formação do PCA para a criação do numero de clusters, simboliza a distancia D_{ij} entre dois *clusters* C_i e C_j, ou seja é a distância máxima entre dois pontos x_{C_i}, y_{C_j}. *All* simboliza o valor do *index* geral em que todos os índices de cada amostra deve ser calculado;

```
numG <-NbClust(dados[0,-1], distance="manhattan",
            min.nc=2, max.nc=10, method="complete",index="all")
```

5. **criar gráfico de barra** para visualizar o melhor número de *clusters*;

6. **fazer partição dos dados em clusters**, o *BestPartition* é usado para calcular e fazer partição dos *clusters*, trata-se da partição que corresponde ao melhor número de *clusters*, dado pelo NbCluster;

```
dados$Cluster <- numG$Best.partition
```

7. **particionar em conjunto de teste e treino** aplicar o algoritmo do createDataPartition para particionar em teste e treino, 60% das amostras clusterizadas vão para o conjunto de treinamento e 40% para o conjunto de teste;

```
#########################################################
 #Fazer particionamento dos dados na coluna Cluster
train <- createDataPartition(dados$Cluster, p=0.60, list=FALSE)

#train recebe apenas 60% dos dados em treino
df_train <- dados[train,]

#test recebe os outros 40% que falta
df_test <- dados[-train,]
```

8. **criar árvore do decisão do tipo c5.0** utilizando os valores do conjunto de treinamento;

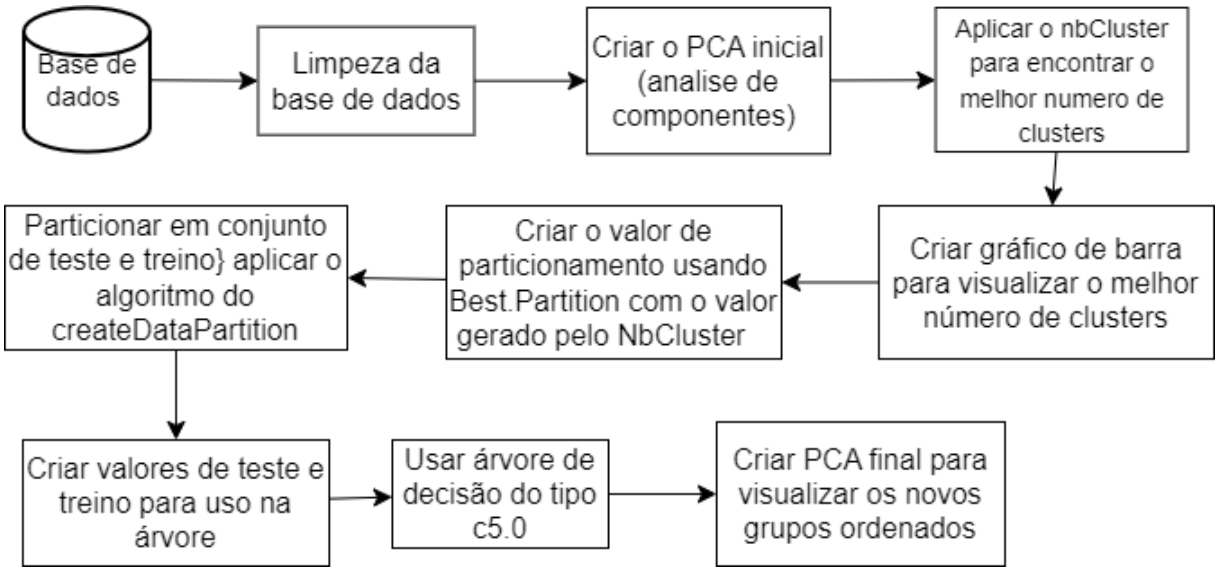
```
#semente de apontamento
    set.seed(20210309)
    #Fazer particionamento dos dados na coluna Cluster
    train_3 <- createDataPartition(dados$Cluster, p=0.60, list=FALSE)
    #train3 recebe apenas 60% dos dados em treino na variavel df_train3
    df_train3 <- dados[train_3,]
    #train3 recebe os outros 40% que falta dos dados
    df_test3 <- dados[-train_3,]
    str(df_train3)
    #treinar a arvore do tipo C5.0
    tree_3<-C5.0(Cluster ~., data=df_train3)
    #tree_3<- C5.0(x =df_train3[0,-24], y = dados$Cluster)
    plot(tree_3)
```

9. **aplicar PCA final** por fim aplicar o ultimo PCA para gerar a visualização de como ficou a organização dos grupos após a clusterização das amostras;

```
res.\sig{PCA} <- \sig{PCA}(dados[, c(2:24)], graph = FALSE)
predicoesTOTAL <- predict(tree_3, dados[,-25])
factoextra::fviz_\sig{PCA}_ind(res.\sig{PCA},
             geom.ind = "point", # show points only (nbut not "text")
             col.ind = dados$Cluster, # clor by groups
             #palette = c("#00AFBB", "#E7B800", "#FC4E07"),
             addEllipses = TRUE, # Concentration ellipses
             legend.title = "Groups"
```

A Figura 3.2 mostra a ordem em que estes métodos e algoritmos são aplicados.

Figura 3.2: Fluxograma da *clusterização* do *Cream Cheese*



CAPÍTULO 4

RESULTADOS E DISCUSSÃO

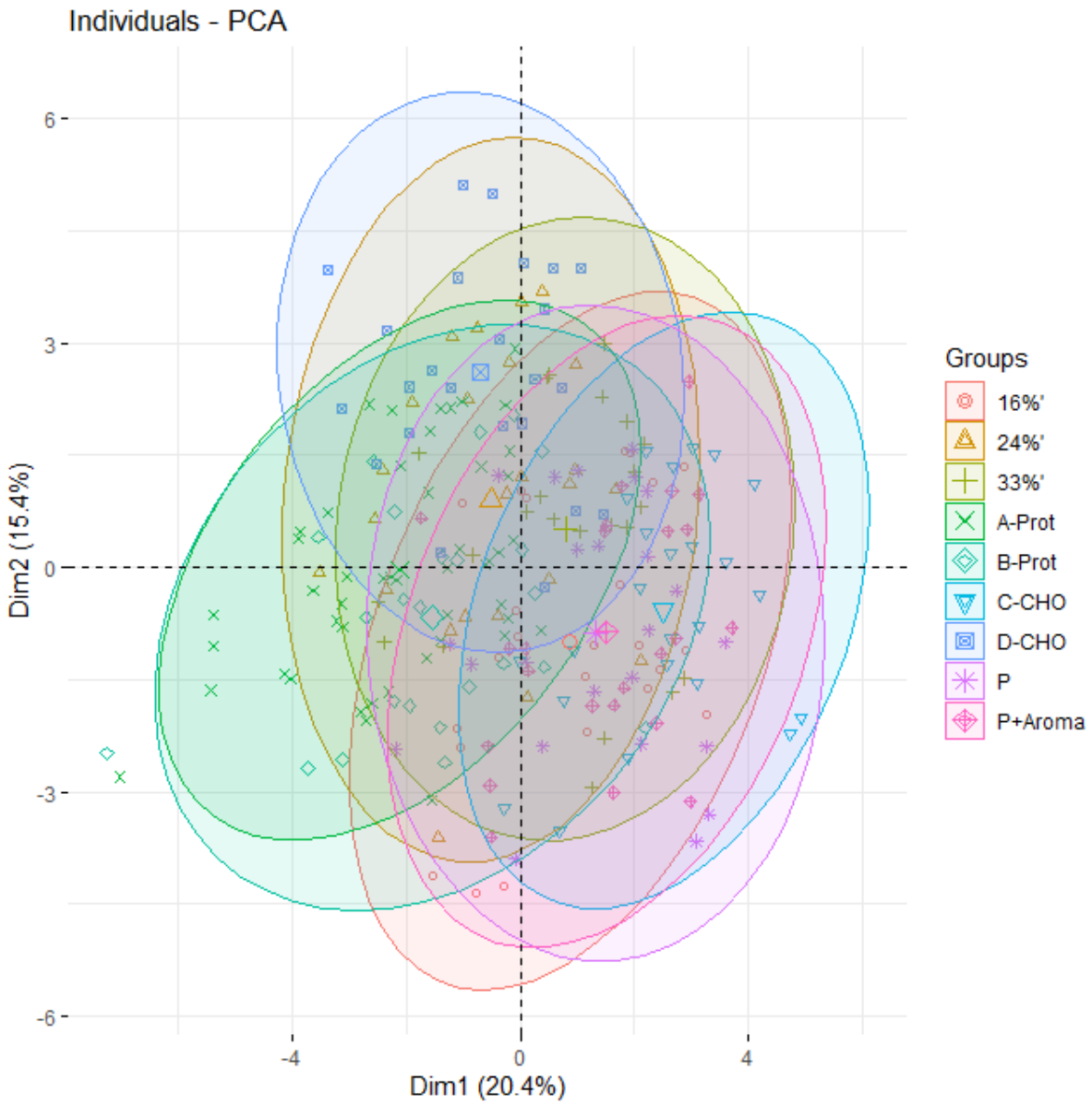
4.1 Etapas do processo de formação dos clusters

Para o processo de formação dos *clusters* em amostras do *cream cheese*, foi usada uma abordagem metodológica de pesquisa experimental com o uso de ferramentas definidas, principalmente a linguagem R. Primeiramente, foi obtida uma base pública de dados sensoriais em amostras de *cream cheese* na seção 3.1. No entanto, foi observada a presença de informações irrelevantes, sendo necessário fazer a limpeza da tabela das amostras, e para isso foi necessário retirar certas colunas, como a coluna de *Panellist*, *Replicate*, a coluna de *Session*, *Serving_Order*, e por fim a coluna de *Product_name* que não acrescentavam nenhuma informação útil sobre o problema.

Após a limpeza, foi criado o PCA dos *clusters* para se ter uma visualização de como é o estado natural dos dados antes do processo clusterização e organização dos dados. A figura 4.1, possibilita observar que as amostras não podem ser convenientemente agrupadas, uma vez que são gerados vários grupos, contendo muitas interseções entre eles e até sobreposições. Concluindo se, portanto, que as amostras produzem um padrão de desclassificação apresentando uma visualização de categorias imediata, problema este que pode ser resolvido com a adição de outras técnicas e métodos no processo de clusterização.

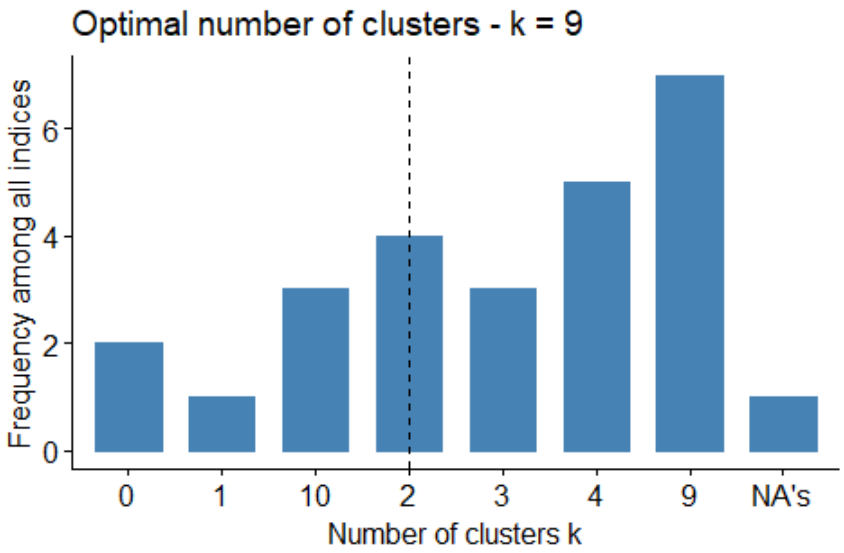
Diante disso, é feito o cálculo para determinar o número ideal de possíveis *clusters*, a Figura 4.2 apresenta o resultado do algoritmo *nbClust*, em que se observa que o melhor número de *clusters* possíveis para a quantidade de amostras da tabela é 9, uma vez que este valor apresenta a maior frequência entre todos os índices, assim, este número estabelece a quantidade máxima de *clusters* indicada para este trabalho.

Figura 4.1: PCA Inicial



Em seguida, o resultado da clusterização do NbClust é obtido por meio do *Best.Partition*. Observa-se que o NbClust utiliza o algoritmo do k-means nos dados limpos para fazer a separação das amostras em seus respectivos *clusters*, portanto, é criada a separação dos dados de acordo suas características próprias feitas nas amostras do *cream cheese*. Após a separação das amostras, é feita a separação dos *clusters* em treino e em teste, isso é feito por meio da função do *createDataPartition* que separa as amostras em dois conjuntos, que serão utilizados na classificação feita pela árvore de decisão, observando-se que 60% das amostras vão para o conjunto de treinamento e para teste os outros 40% são colocados na variável teste. A razão de se usar essa separação é que a maioria dos algoritmos de clusterização não pode "prever"

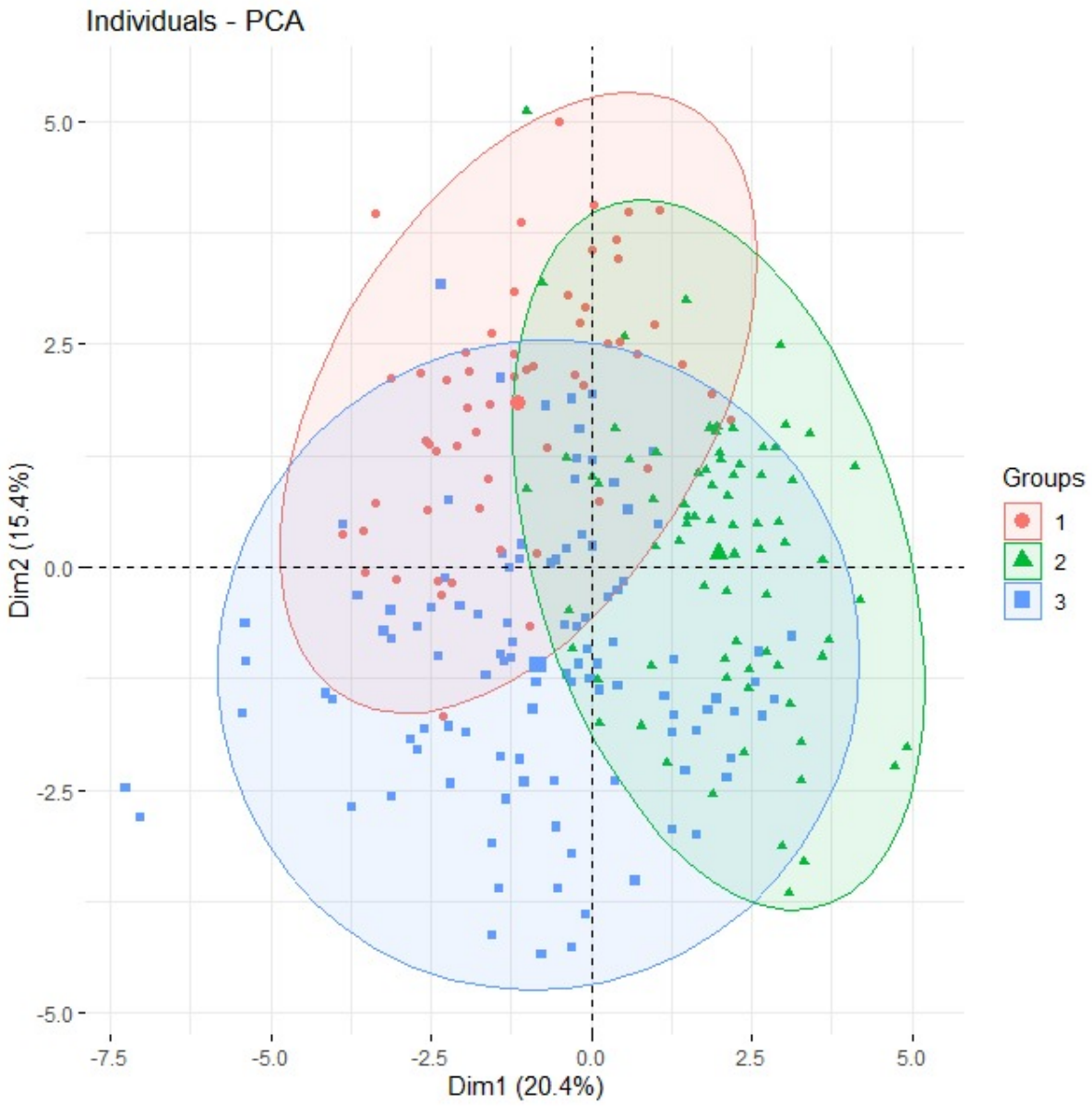
Figura 4.2: Gráfico mostrando o número de *clusters*



novos dados, o K-means é uma rara exceção, porque pode-se fazer a classificação do vizinho mais próximo nos centróides, prevendo assim a qual *cluster* o novo dado pertence. Mas para qualquer método que não use centróides, não está claro como se aplicaria aos dados de "teste". Essas abordagem também é interessante para evitar problemas de ajuste excessivo, no qual o modelo funciona bem apenas para o conjunto de dados que o originou. Por exemplo, aumentar o número de clusters sempre “aumentará o desempenho” e com isso o treinamento representa o valor de sucessos de execução do algoritmo na clusterização, um valor ideal e perfeito, enquanto que teste representa o valor possível de erros, um valor inferior mas que é usado para balancear a árvore de decisão.

Pode ser observado na Figura 4.3 uma visualização completa da classificação das amostras em seus respectivos *clusters*, observa-se a existência de 3 *clusters*. Assim, na Figura 4.3 as amostras do *cream cheese* estão inseridas no grupo que melhor as representa, de forma que todas as amostras foram organizadas em 3 grupos (*clusters*), em cada círculo com uma cor definida na legenda é a representação de um grupo, e os pontos dentro de cada círculo representam as amostras de *cream cheese*. É importante observar, que há sobreposição entre os *clusters* em função da própria indecidibilidade do problema em questão, que é uma situação muito comum em problemas de classificação.

Figura 4.3: PCA final de *clusters*



Após obter os conjuntos de treinamento e de teste, a árvore de decisão do tipo c.5 é construída usando os algoritmos da linguagem de programação R, tendo como entrada o conjunto de amostras de treinamento. A árvore de decisão da Figura 4.4 apresenta o resultado da classificação em *clusters* com base nas características sensoriais de cada amostra e na clusterização obtida anteriormente. Cada círculo representa um atributo de uma amostra, o valor que está na sua ligação com outra amostra (a linha) representa o seu valor de nivelamento em relação a sua característica principal, cada node (folha da árvore) representa em qual *cluster* a amostra vai ficar inserida, os valores 1,2 e 3 na horizontal de cada node representa o *cluster* em que a amostra foi colocada, e os valores de 0 a 1 na vertical representa o percentual de amostras

que ficou em cada *cluster*, a variável n representa quantas amostras no total entraram naquele node (ex: n=14, então 14 amostras no total entraram nesse node). Por exemplo, as amostras de *cream cheese* com NCream (cremosidade) cujo valor é maior que 8.1 estão no Nodo 9 da árvore (*cluster* 3), totalizando 3 amostras (n=3), e as amostras com valores menor ou igual a 8.1 estão classificados no *cluster* 2, totalizando 2 amostras (n=2).

4.2 Resultado

É possível observar na Figura 4.4 que as características mais relevantes para o contexto são: *N_Oldmilk* (nível de leite velho do produto), *M_Firm* (firmeza da amostra), *M_Butter* (manteiga da amostra), *M_Creaminess* (cremosidade da amostra), *E_Grainy* (granulosidade da amostra), *N_Cream* (cremosidade do produto), *M_Chalky* (quantidade de duro/sólido da amostra) e *E_Grainy* (granulosidade na aparência). Portanto, as características individuais das amostras sobressaem às características do produto, enquanto que as características de aparência possuem pouca relevância na clusterização. O atributo *N_OldMilk* é responsável por classificar mais de 20% das amostras, uma vez que um valor superior a 3,75 tende a classificar a amostra no *cluster* 3, o restante das amostras precisa ter outros atributos avaliados para serem inseridas em um *cluster*. Ao associar *M_Firm* igual ou inferior a 7,95 e *M_Butter* inferior ou igual a 6,6, também tem se uma boa caracterização do *cluster* 3. Por outro lado, valores de M_Butter superior a 6,6 demandam a inclusão do atributo E_Grainy para classificar, menor ou igual a 5,1 aponta para o *cluster* 2, enquanto que valores maiores que 5,1 demandam o uso do atributo *N_Cream*, que se for inferior ou igual a 8,1 indicam *cluster* 2, do contrário, *cluster* 3. O *Cluster* 1 é caracterizado pela Tabela 4.1.

Atributo	Valores 1	Valores 2
N_Oldmilk	$\leq$ 3,75	$\leq$ 3,75
M_Firm	$>$ 7,95	$>$ 7,95
M_Creaminess	$\leq$ 10,35	$\leq$ 10,35
M_Chalky	$\leq$ 6,3	>6,3
E_Grainy		$\leq$ 4,2

Tabela 4.1: Caracterização do cluster 1

Em classificação é comum o uso de medidas para quantificar o quão bem sucedida foi a tarefa, uma dessas medidas é a acurácia, ela se refere a taxa de classificações corretas realizadas pelo método, e é dada por (SILVA et al., 2016):

$$Acurácia = |y - f(\omega) = 0| \tag{4.1}$$

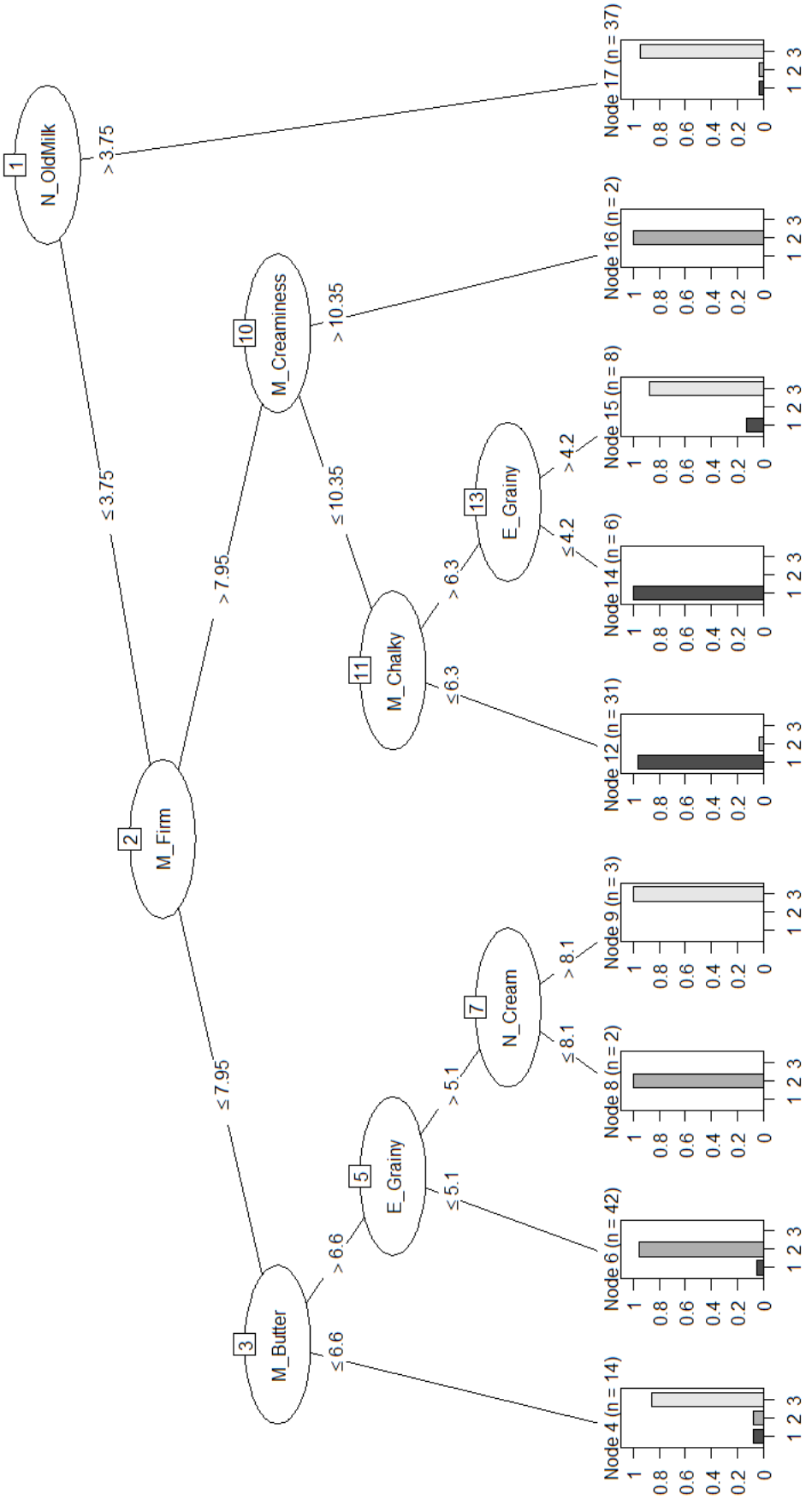
sendo que $|x|$ é a quantidade de vezes que a expressão x é verdadeira, f é o modelo preditivo, ω é o subconjunto de dados para o qual o modelo está sendo avaliado, $f(.)$ é a classificação dada pelo modelo para cada um dos exemplares, e y e a classe esperada como resposta.

A tabela 4.2 apresenta o resultado da classificação em função da acurácia no conjunto de testes e de treinamento, em que se observa uma acurácia de 94,48% no conjunto de treinamento e de 82,11% no conjunto de testes. O valor da 94,48% citado está se referindo a quantos acertos o método teve em relação às amostras do conjunto de treinamento, resultando portanto em um treinamento bastante favorável em relação a formação dos *clusters*, e para o conjunto das amostras de teste, o valor da acurácia é de 82,11% o que também demonstra a eficiência do método em predizer. A acurácia é a medida mais comum usada para avaliar o desempenho de um modelo preditivo de classificação (ANTONIO et al., 2018). Como a precisão de um modelo preditivo é normalmente alta (mais de 90%), é comum resumir o desempenho de um modelo em termos da taxa de erro do modo.

Tabela 4.2: Resultado

Conjunto	**Acurácia** (%)
Treinamento	94,48%
Teste	82,11%

Figura 4.4: Árvore de decisão c5.0



CONCLUSÃO

Este trabalho apresenta técnicas da mineração de dados tais como a clusterização por meio de algoritmo *k-means*, árvore de decisão, uso de PCA com o objetivo de caracterizar o perfil sensorial do *cream cheese*. É apresentada uma análise de *clusters* com base em amostras de dados que foram obtidas através de análises sensoriais, as quais incluem os sensores olfato, visão, textura, para que assim se possa ter uma boa descrição do produto e das amostras de cada produto.

A análise de clusters implica na construção do PCA, uso do k-means e construção de árvore de decisão C5.0 para a visualização dos resultados. Para atingir os objetivos de formação dos *clusters*, foi usado a linguagem R, visando assim obter uma visão descritiva e exploratória de todo o conjunto de dados, sendo que o resultado final proporcionou um particionamento adequado.

Os resultados da observação indicam que os atributos das amostras desempenham um papel mais preponderante no processo de classificação do que os atributos do produto. Esta observação sugere uma considerável variação dentro de um mesmo produto, destacando características como nível de leite velho, firmeza, quantidade de manteiga, cremosidade, entre outros como elementos cruciais no resultado. Esse processo resultou em uma acurácia de aproximadamente 94,48% no treinamento final de predito, e 82,11% nos demais testes restantes, para a montagem do PCA final foi usado os algoritmos de clusterização da linguagem R, que oferece bibliotecas prontas para essa tarefa, para a criação da árvore de decisão do tipo c5.0 foi usada a biblioteca c50 do R.

4.3 Trabalhos Futuros

Diante os resultados obtidos, para trabalhos futuros pretende-se: Implementar a funcionalidade de construção da árvore de decisão do tipo cart.

- Implementar uma nova clusterização com as amostras de *cream cheese* para 5 clusters para a avaliação de novos resultados.
- Investigar novas amostras de *cream cheese* para serem adicionadas na planilha para uma nova clusterização de dados apresentados.
- Realizar experimentos de *clusters* com novas amostras com outros tipos de componentes da área alimentícia.

REFERÊNCIAS BIBLIOGRÁFICAS

ALLEM, L. E.; HOPPEN, C.; MARZO, M. M.; SIBEMBERG, L. S. A spectral clustering approach for the evolution of the covid-19 pandemic in the state of rio grande do sul, brazil. *Trends in Computational and Applied Mathematics*, Sociedade Brasileira de Matemática Aplicada e Computacional - SBMAC, v. 23, p. 705–729, 11 2022. ISSN 2676-0029. Disponível em: <https://tema.sbmac.org.br/tema/article/view/1496>.

ANTONIO, S.; FRANCESCA, F.; GIULIA, D. B. T. C.; GATTONE. Cluster analysis as a decision-making tool: A methodological review. In: SHU-HENG; EDGARDO, C. J. M. B.; CHEN (Ed.). [S.l.]: Springer International Publishing, 2018. p. 48–55. ISBN 978-3-319-60882-2.

BRAZ, A. M.; OLIVEIRA, I. J. de; CAVALCANTI, L. C. de S.; ALMEIDA, A. C. de; CHAVEZ, E. S. Cluster analysis for landscape typology. *Mercator*, Universidade Federal do Ceará, v. 19, p. 1–16, 5 2020. ISSN 19842201. Disponível em: <http://www.mercator.ufc.br/mercator/article/view/e19011>.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R.; STONE, C. *1.10. Decision Trees — scikit-learn 1.3.1 documentation*. 2013. <https://scikit-learn.org/stable/modules/tree.html>. (Accessed on 10/11/2023).

BRIGHENTI, M.; GOVINDASAMY-LUCEY, S.; JAEGGI, J. J.; JOHNSON, M. E.; LUCEY, J. A. Behavior of stabilizers in acidified solutions and their effect on the textural, rheological, and sensory properties of cream cheese. *Journal of Dairy Science*, Elsevier Inc., v. 103, p. 2065–2076, 3 2020. ISSN 15253198.

BRIGHENTI, M.; GOVINDASAMY-LUCEY, S.; LIM, K.; NELSON, K.; LUCEY, J. A. Characterization of the rheological, textural, and sensory properties of samples of commercial us cream cheese with different fat contents. *Journal of Dairy Science*, Elsevier, v. 91, p. 4501–4517, 12 2008. ISSN 0022-0302.

CAROLINA, M.; DIAS, D. A. Deteção de estruturas retinianas no diagnóstico da retinopatia diabética. 2012.

CARUSO, G.; GATTONE, S. A.; FORTUNA, F.; BATTISTA, T. D. *Cluster Analysis as a Decision-Making Tool: A Methodological Review*. [S.l.]: Springer, 2017. (International Symposium on Distributed Computing and Artificial Intelligence).

CARVALHO, J. M.; BLENINGER, T. State-transition matrices as an analysis and forecasting tool applied to water quality in reservoirs. *RBRH*, Associação Brasileira de Recursos Hídricos, v. 26, p. e30, 10 2021. ISSN 2318-0331. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2318-03312021000100230&tlng=en>.

CHARRAD, M.; GHAZZALI, N.; BOITEAU, V.; NIKNAFS, A. Nbclust: An r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, v. 61, n. 6, p. 05–18, 2014.

CORNELISSEN, J. *NbClust function - RDocumentation*. 2021. <https://www.rdocumentation.org/packages/NbClust/versions/3.0.1/topics/NbClust>. (Accessed on 10/18/2023).

FRØST, M. *The influence of fat content on sensory properties and consumer perception of dairy products*. Tese (Doutorado), 2002. Copenhagen : Center for Skov, Landskab og Planlægning/Københavns Universitet.

GARG, A.; CHADHA, S. *Step-By-Step Guide to Principal Component Analysis With Example*. 2020. <https://www.turing.com/kb/guide-to-principal-component-analysis>. (Accessed on 10/10/2023).

JOHNSON, M. E. A 100-year review: Cheese production and quality. *Journal of Dairy Science*, v. 100, 2017. ISSN 15253198.

LATOUCHE, G.; RAMASWAMI, V. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial and Applied Mathematics, 1999. ISBN 978-0-89871-425-8. Disponível em: <http://epubs.siam.org/doi/book/10.1137/1.9780898719734>.

MAGEE, J. F. *Árvores de Decisão para Tomada de Decisão*. 1964. <https://hbr.org/1964/07/decision-trees-for-decision-making>. (Accessed on 10/12/2023).

MAIONE, B. F.; KAMINSKI, P. C.; BARALDI, E. C. The automotive recall data search and its analysis applying machine learning. *Production*, Associação Brasileira de Engenharia de Produção, v. 33, p. e20220117, 6 2023. ISSN 1980-5411. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-65132023000100209&tlng=en>.

MAIONE, C.; SOUZA, V. C. de O.; TOGNI, L. R.; COSTA, J. L. da; CAMPIGLIA, A. D.; BARBOSA, F.; BARBOSA, R. M. Using cluster analysis and icp-ms to identify groups of ecstasy tablets in sao paulo state, brazil. *Journal of Forensic Sciences*, Blackwell Publishing Inc., v. 62, p. 1479–1486, 11 2017. ISSN 15564029.

MIGUEL, T. *K-Means Clustering (Agrupamento k-means) - Aprender Ciência de Dados*. 2023. <https://aprenderdatascience.com/k-means-clustering-agrupamento-k-means/>. (Accessed on 10/20/2023).

PASKLAN, A. N. P.; QUEIROZ, R. C. de S.; ROCHA, T. A. H.; SILVA, N. C. da; TONELLO, A. S.; VISSOCI, J. R. N.; TOMASI, E.; THUMé, E.; STATON, C.; THOMAZ, E. B. A. F. Análise espacial da qualidade dos serviços de atenção primária à saúde na redução da mortalidade infantil. *Ciência Saúde Coletiva*, ABRASCO - Associação Brasileira de Saúde Coletiva, v. 26, p. 6247–6258, 12 2021. ISSN 1678-4561. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232021001206247&tlng=pt>.

QANNARI, E. M. Sensometrics approaches in sensory and consumer research. *Current Opinion in Food Science*, v. 15, p. 8–13, 2017. ISSN 2214-7993. Sensory Sciences and Consumer Perception • Food Physics and Material Science. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2214799317300280>.

QURESHI, R. J.; KOVACS, L.; HARANGI, B.; NAGY, B.; PETO, T.; HAJDU, A. Combining algorithms for automatic detection of optic disc and macula in fundus images. *Computer Vision and Image Understanding*, Elsevier Inc., v. 116, n. 1, p. 138–145, 2012. ISSN 10773142. doi:<http://dx.doi.org/10.1016/j.cviu.2011.09.001>.

SAINANI, M. R.; VYAS, H. K.; TONG, P. S. Characterization of particles in cream cheese. *Journal of Dairy Science*, v. 87, 2004. ISSN 00220302.

SHERER, T.; DUNCAN, O. *Seleção de recursos (mineração de dados)*. 2022. Disponível em: <https://learn.microsoft.com/pt-br/analysis-services/data-mining/feature-selection-data-mining?view=asallproducts-allversions>.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. *Introdução à Mineração de Dados - Com Aplicações em R*. 1. ed. [S.l.]: Leandro Augusto da Silva and Sarajane Marques Peres and Clodis Boscarioli, 2016.

SMITH, D.; LEE, D. *C5.0 Node*. 2023. Disponível em: <https://www.ibm.com/docs/en/spss-modeler/18.4.0?topic=trees-c50-node>.

TOURNIER, C.; MARTIN, C.; GUICHARD, E.; ISSANCHOU, S.; SULMONT-ROSSé, C. Contribution to the understanding of consumers' creaminess concept: A sensory and a verbal approach. *International Dairy Journal*, v. 17, p. 555–564, 2007. ISSN 0958-6946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S095869460600197X>.