UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uma abordagem de aprendizado de máquina aplicado em avaliações de qualidade em produção de café

Luana Priscila Dos Santos

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Luana Priscila Dos Santos

Uma abordagem de aprendizado de máquina aplicado em avaliações de qualidade em produção de café

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Profa. Dra. Roseli Aparecida Romero

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

D722a

Dos Santos, Luana Priscila
Uma abordagem de aprendizado de máquina aplicado em avaliações de qualidade em produção de café / Luana Priscila Dos Santos; orientadora Roseli Aparecida Romero . -- São Carlos, 2025.
42 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Aprendizado de máquina. 2. Café. 3. K-means. I. Romero , Roseli Aparecida, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Luana Priscila Dos Santos

A study about machine learning apply in quality review's in coffee production

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Profa. Dra. Roseli Aparecida Romero

Original version

São Carlos

2025

AGRADECIMENTOS

A todos os professores deste curso pelo compartilhamento do conhecimento.

E ao meu esposo por todo apoio nessa jornada.

RESUMO

Dos Santos, L. P. **Uma abordagem sobre aprendizado de máquina aplicado em avaliações de qualidade em produção de café**. 2025. 42p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

RESUMO

Este trabalho apresenta uma análise detalhada da qualidade do café utilizando método de aprendizado de máquina, com o objetivo de aprimorar a classificação dos cafés com base em seus atributos sensoriais e físicos. Os dados utilizado e analisados já possuem a classificação mensurada por pontuação total por xícara, um método tradicional para determinar a qualidade. No entanto, uma abordagem mais avançada foi adotada por meio do algoritmo de clustering KMeans, segmentando os dados em clusters distintos para identificar padrões relevantes em diferentes características do café.

A análise exploratória revelou que atributos como uniformidade, doçura, aroma, sabor, acidez e corpo desempenham papéis cruciais na avaliação da qualidade. O uso do aprendizado de máquina proporcionou uma nova perspectiva, permitindo identificar a relevância de cada atributo e sugerir melhorias específicas para cada cluster.

Além disso, foram aplicadas as métricas de Silhouette, Davies-Bouldin Index e Calinski-Harabasz Index para avaliar a qualidade dos clusters gerados, evidenciando a eficácia da segmentação em termos de coesão e separação dos dados. Essas métricas demonstraram a robustez da segmentação, fornecendo uma avaliação quantitativa da qualidade dos agrupamentos.

Este trabalho propõe, portanto, uma abordagem mais precisa para a classificação do café, útil para orientar melhorias nos processos de produção e facilitar a identificação de amostras com potencial para se destacar nas avaliações qualitativas.

A combinação de técnicas tradicionais de análise de qualidade com métodos de aprendizado de máquina pode fornecer uma visão mais profunda e detalhada sobre a classificação e aprimoramento da qualidade do café, permitindo sugerir ações específicas para a otimização do produto final.

Palavras-chave: Café. K-means. Aprendizado de Máquina.

ABSTRACT

Dos Santos, L. P. **A study about machine learning apply in quality review's in coffee production**. 2025. 42p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

ABSTRACT

This paper presents a detailed analysis of coffee quality using machine learning methods, with the goal of improving the classification of coffees based on their sensory and physical attributes. Initially, the data were analyzed based on the total score per cup, a traditional method for determining quality. However, a more advanced approach was adopted through the KMeans clustering algorithm, segmenting the data into distinct clusters to identify relevant patterns in different coffee characteristics.

Exploratory analysis revealed that attributes such as uniformity, sweetness, aroma, flavor, acidity, and body play crucial roles in quality evaluation. The use of machine learning provided a new perspective, enabling the identification of the relevance of each attribute and suggesting specific improvements for each cluster.

Furthermore, the Silhouette, Davies-Bouldin Index, and Calinski-Harabasz Index metrics were applied to evaluate the quality of the generated clusters, highlighting the effectiveness of the segmentation in terms of cohesion and separation of the data. These metrics demonstrated the robustness of the segmentation, providing a quantitative assessment of the cluster quality.

Therefore, this paper proposes a more precise approach to coffee classification, useful for guiding improvements in production processes and facilitating the identification of samples with the potential to stand out in qualitative evaluations.

The combination of traditional quality analysis techniques with machine learning methods can provide a deeper and more detailed view of the classification and improvement of coffee quality, allowing us to suggest specific actions for the optimization of the final product.

Keywords: Coffee. Kmeans. Machine Learning. Dissertation. Conclusion course paper.

LISTA DE FIGURAS

Figura 1 – Conjunto de Dados 27
Figura 2 – Amostra do conjunto de dados pré-processado 33
Figura 3 – Validação do número de cluster 35
Figura 4 – Gráfico de Coordenadas Paralelas dos Clusters. 35
Figura 5 – Centróides dos Clusters. 36
Figura 6 – Variâncias para Confirmação da Relevância. 37

SUMÁRIO

1 INTRODUÇÃO . . . 17
1.1 Justificativa E Motivação . . . 17
1.2 Questão de Pesquisa e Objetivos . . . 18
1.2.1 Seção 2: Revisão Teórica . . . 18
1.2.2 Seção 3: Metodologia . . . 18
1.2.3 Seção 4: Conclusão . . . 18

2 REVISÃO DE CONCEITOS E TRABALHOS RELACIONADOS . . 19
2.1 Aprendizado De Máquina . . . 19
2.1.1 Clusterização . . . 20
2.1.2 Redução De Dimensionalidade (Generalização) . . . 20
2.1.3 Aprendizagem De Regras De Associação . . . 20
2.1.4 K-means . . . 21
2.2 Café . . . 21
2.3 Trabalhos Relacionados . . . 23
2.4 Cenário, Limitações e Proposta de Solução . . . 24

3 METODOLOGIA . . . 27
3.1 Base De Dados . . . 27
3.1.1 Atributos Sensoriais . . . 28
3.1.2 Atributos Físicos . . . 28
3.2 Ferramenta Utilizada . . . 29
3.3 Pré-Processamento Dos Dados . . . 30
3.3.1 Remoção De Atributos . . . 30
3.3.2 Valores Ausentes E Duplicados . . . 30
3.3.3 Análise Das Variáveis Categóricas . . . 30
3.3.4 Variáveis Númericas - Valores Zerados . . . 31
3.3.5 Análise Sobre Outliers . . . 31
3.3.6 Normalização Dos Dados Numéricos . . . 31
3.3.7 Transformação Dos Dados Categóricos Em Numéricos . . . 32
3.3.8 Descrição Do Conjunto De Dados Após Pré-processamento . . . 32
3.4 Processamento Dos Dados . . . 33
3.4.1 Abordagem Da Clusterização Dados Sensorias E Físico . . . 33
3.4.2 Validação Da Qualidade Dos Clusters Gerados . . . 34
3.4.3 Análise dos Atributos Mais Significativos em Cada Cluster com Base nos Centróides . . . 35

3.4.4 Reflexão da Definição Técnica dos Clusters (agrupamentos) 37
3.4.5 Análise e Interpretação dos Clusters 37
3.4.6 Resumo Da Abordagem 38

4 CONCLUSÃO 39

REFERÊNCIAS 41

1 INTRODUÇÃO

Atualmente, a inteligência artificial (IA) tem assumido um papel cada vez mais relevante no cotidiano das pessoas, influenciando desde aplicativos de uso pessoal até processos industriais complexos e tecnologias avançadas, como carros autônomos. Esse domínio crescente da IA é impulsionado, em grande parte, pelos avanços na Ciência de Dados e pelas técnicas de Aprendizado de Máquina (AM), que permitem a análise de grandes volumes de dados e a extração de insights valiosos para a tomada de decisões. Nesse contexto, o AM surge como uma ferramenta poderosa para interpretar dados complexos e transformá-los em informações significativas, aplicáveis em diversos setores.

No âmbito da produção e consumo de café, a qualidade é um fator essencial, e as práticas atuais de avaliação dependem fortemente de análises físicas e sensoriais. Tradicionalmente, essas avaliações são realizadas por especialistas, que atribuem pontuações com base em critérios específicos. No entanto, a crescente demanda por maior precisão e eficiência no processo de avaliação de qualidade abre espaço para a aplicação de técnicas de AM. Nesse cenário, ao utilizar algoritmos de aprendizado de máquina, como técnicas de agrupamento (clustering), é possível analisar dados históricos de avaliações, identificar padrões e agrupar cafés com características semelhantes, oferecendo uma abordagem inovadora e agregativa de informações para a classificação e melhoria da qualidade do café.

Alinhado às tendências tecnológicas atuais e como proposta de pesquisa sobre os conteúdos abordados no curso de Inteligência Artificial e Big Data, este trabalho tem como objetivo realizar uma abordagem sobre o uso do Aprendizado de Máquina para a descoberta de novas informações a partir de dados referentes a avaliações qualitativas de café. Com isso, busca-se obter insights agregativos que possam contribuir para uma definição mais precisa e eficiente da qualidade do café, alinhando teoria e prática no contexto da IA e da Ciência de Dados.

1.1 Justificativa E Motivação

O café é amplamente reconhecido como a segunda bebida mais consumida no mundo, ficando atrás apenas da água (Governo do Brasil, 2025). A qualidade do café é de suma importância para toda a cadeia produtiva, desde os produtores até os consumidores finais.

Para os produtores, manter um alto padrão de qualidade assegura competitividade no mercado e atende às exigências dos consumidores que buscam produtos de excelência (Splabor, 2025). Por outro lado, para os consumidores, a qualidade do café está diretamente ligada à confiança e à experiência sensorial proporcionada pela bebida (CoffeesBR, 2025).

Na análise dos dados sobre avaliações, entender melhor se há combinações de características que podem influenciar na qualidade degustativa do café, torna-se um conhecimento enriquecedor para o produtor e para uma tomada de decisão do consumidor.

1.2 Questão de Pesquisa e Objetivos

Com a aplicação do Aprendizado de Máquina na avaliação qualitativa do café, o que é possível obter ao descobrir novos agrupamentos de dados? Este trabalho tem como objetivo analisar quais agrupamentos de características sensoriais e físicas do grão de café exercem influência significativa na definição da qualidade do café. A partir disso, busca-se compreender como esses agrupamentos podem contribuir para uma avaliação mais precisa e eficiente da qualidade, orientando melhor os produtores e auxiliando na tomada de decisão dos consumidores.

Para o desenvolvimento e aplicação prática desse objetivo, a estrutura do trabalho foi organizada da seguinte forma:

1.2.1 Seção 2: Revisão Teórica

Realizou-se uma revisão dos conceitos fundamentais para contextualização das informações de apoio. Foram abordados temas como Inteligência Artificial, com foco nas definições dos algoritmos e as características do produto café. Além disso, foram mencionados trabalhos relacionados ao tema, proporcionando um embasamento teórico sólido para a pesquisa.

1.2.2 Seção 3: Metodologia

Descreve-se a metodologia aplicada. Inicialmente, é detalhada a base de dados utilizada e as ferramentas empregadas, bem como todo o pré-processamento realizado para garantir a qualidade e a pertinência dos resultados. Em seguida, é apresentado o processamento realizado com o algoritmo K-means, incluindo uma análise da eficácia do agrupamento e a definição dos perfis dos grupos obtidos.

1.2.3 Seção 4: Conclusão

Apresenta-se a conclusão do trabalho, com a síntese dos resultados obtidos, as contribuições do estudo e sugestões para pesquisas futuras. Essa seção consolida os insights gerados e destaca o potencial da aplicação do Aprendizado de Máquina na avaliação da qualidade do café.

2 REVISÃO DE CONCEITOS E TRABALHOS RELACIONADOS

Neste capítulo serão apresentados os conceitos principais para o estudo, abordando aprendizado de máquina e conceitos sobre especificações de cafés para melhor entendimento sobre os dados analisados.

2.1 Aprendizado De Máquina

Na literatura há diversas interpretações sobre a definição de aprendizado de máquina.

Segundo GÉRON (GÉRON, 2021), aprendizado de máquina é a ciência (e a arte) da programação de computadores de modo que eles possam aprender com os dados.

Já na definição de CARVALHO (CARVALHO, 2022) diz que em aprendizado de máquina, dispositivos computacionais são programados para aprender a partir de experiências passadas. Para tal, frequentemente empregam um princípio de inferência denominado indução, que permite extrair conclusões genéricas a partir de um conjunto particular de exemplos.

Comparando os algoritmos tradicionais e algoritmos de aprendizado de máquina, entende-se que o primeiro é realizado com a programação sequencial ou de regras de decisão e o segundo as regras são aprendidas a partir da análise de dados exemplares.

O aprendizado de máquina se viabiliza por meio de duas abordagens principais: a análise de dados históricos e o método de tentativa e erro. Essas técnicas formam a base das três categorias fundamentais do aprendizado de máquina: Aprendizado Supervisionado, Aprendizado Não Supervisionado e Aprendizado por Reforço. Conforme destacado por GOODFELLOW; BENGIO; COURVILLE em 2016, o aprendizado supervisionado depende de dados históricos rotulados para treinar modelos, enquanto o aprendizado por reforço utiliza um sistema de recompensas e penalidades, baseado em tentativa e erro, para guiar o processo de aprendizado. Já o aprendizado não supervisionado busca identificar padrões e estruturas em dados não rotulados, sem a necessidade de intervenção direta. Essas categorias representam os pilares das técnicas modernas de aprendizado de máquina e deep learning.

No contexto do aprendizado supervisionado, LUDERMIR ressalta que, para cada exemplo apresentado ao algoritmo, é necessário indicar a resposta desejada. Cada exemplo é descrito por um vetor de valores (atributos) e pelo rótulo da classe associada. O objetivo é construir um classificador capaz de determinar corretamente a classe de novos exemplos ainda não rotulados. Quando os rótulos são discretos, o problema é chamado de classificação; para valores contínuos, denomina-se regressão.

Por outro lado, na abordagem de aprendizado por reforço, o algoritmo não recebe a resposta correta diretamente, mas sim um sinal de reforço, como recompensa ou punição. Com base nesse sinal, o algoritmo formula hipóteses a partir dos exemplos e avalia sua eficácia. Esse processo iterativo permite que o modelo aprenda a tomar decisões ótimas ao longo do tempo (LUDERMIR, 2021).

Já o aprendizado não supervisionado, como o nome sugere, não conta com um supervisor para definir a saída do modelo, utilizando apenas os dados de entrada. Segundo SOUSA em 2023, o objetivo principal é identificar padrões, regularidades ou estruturas presentes no conjunto de dados. Complementando essa definição, ESCOVEDO (ESCOVEDO, 2020) ressalta que, nessa abordagem, o algoritmo não recebe resultados esperados durante o treinamento. Em vez disso, o foco está na exploração dos dados para descobrir possíveis relacionamentos. Como não há dados rotulados, não é necessário realizar a divisão em conjuntos de treino e teste, simplificando o processo de análise.

2.1.1 Clusterização

Entre os diversos algoritmos não supervisionados, a clusterização (também conhecida como agrupamento) é um problema fundamental em áreas como aprendizado de máquina, reconhecimento de padrões e análise de dados. Seu objetivo principal é particionar um conjunto de itens em grupos (clusters), de modo que os itens dentro de um mesmo grupo sejam mais semelhantes entre si do que os itens de grupos distintos. Conforme destacado por MOHRI; ROSTAMIZADEH; TALWALKAR (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018), a clusterização consiste em um problema de particionamento de itens em regiões homogêneas sem supervisão.

2.1.2 Redução De Dimensionalidade (Generalização)

A redução de dimensionalidade é uma tarefa relacionada ao aprendizado não supervisionado, cujo objetivo é simplificar os dados sem perder informações significativas. De acordo com GÉRON, essa técnica permite combinar características correlacionadas em uma única representação, reduzindo a complexidade dos dados e facilitando a análise e a visualização.

2.1.3 Aprendizagem De Regras De Associação

O aprendizado de regras de associação é uma técnica que visa explorar grandes volumes de dados para identificar relações significativas entre atributos. Conforme apontado por GÉRON, essa abordagem é amplamente utilizada em aplicações como análise de cestas de compras e mineração de dados, onde o objetivo é descobrir padrões ou associações relevantes.

2.1.4 K-means

O K-means é um dos algoritmos mais populares no aprendizado não supervisionado, especialmente para tarefas de clusterização. Segundo BORHMKE (BORHMKE, 2019), o K-means particiona um conjunto de dados em (k) grupos (ou clusters), onde (k) é um número que pode ser pré-definido pelo analista ou determinado automaticamente por meio de métodos como o Método do Cotovelo ou o Coeficiente de Silhueta. O algoritmo busca agrupar objetos de modo que os itens dentro de um mesmo cluster sejam altamente semelhantes (alta similaridade intraclasse), enquanto itens de clusters diferentes sejam o mais distintos possível (baixa similaridade interclasse). Cada cluster é representado por um centróide, que corresponde à média dos pontos atribuídos ao grupo.

(KASSAMBARA, 2017) lista o algoritmo K-means em cinco etapas:

1. Especificar a quantidade de clusters (k) a ser informada (pelo analista);
2. Selecionar aleatoriamente k objetos do conjunto de dados como os centros de cluster ou médias iniciais deles;
3. Atribuir cada observação ao seu centróide mais próximo, baseado na distância euclidiana entre o objeto e o centróide;
4. Para cada um dos k clusters, atualizar o centróide do cluster calculando o novo valor da média de todos os pontos de dados no cluster. O centróide do k-ésimo cluster é um vetor de tamanho p contendo as médias de todas as variáveis para as observações no k-ésimo cluster, onde p é a quantidade de variáveis;
5. Iterativamente reduzir a soma de quadrados total. Isso é, repetir os passos 3 e 4 até que as atribuições do cluster parem de mudar ou a quantidade máxima de iterações seja alcançada.

2.2 Café

De acordo com HALAL (HALAL, 2017), o café pertence à família Rubiácea e ao gênero *Coffea*, existindo diversas espécies que variam de acordo com o local de cultivo e a região de origem.

Referente a essas espécies, é possível destacar três principais (BAIRROS, 2022):

- **Coffea arabica** – Uma das primeiras espécies produzidas, de origem etíope e propagada pelos árabes. Possui frutos ovais que amadurecem entre 7 e 9 meses e muitas variedades. As mais comuns são Bourbon, Mundo Novo (Brasil), San Ramon (América Central), Blue Mountain (Jamaica) e Sumatra (Indonésia), além de híbridos como a Catuaí, dentre outros.

- **Coffea canephora** – Uma espécie muito popular, porém menos valorizada que a Arábica, com frutos arredondados e menores, que levam até 11 meses para amadurecer. A sua variedade mais comum é a Robusta, cultivada na África ocidental e central, no sudoeste da Ásia e em algumas regiões do Brasil, onde é conhecida como Conilon.
- **Coffea liberica** – Espécie nativa da África, cuja variedade possui menos demanda, sendo chamada de Dewevrei, porém conhecida como Excelsa.

Sobre a ótica econômica, as duas primeiras espécies são mais consumidas, sendo que a Coffea arabica (café Arábica) representa cerca de 60% da produção mundial, seguida da Coffea canephora (café Robusta) e HALAL menciona que o café Arábica tem 50% menos cafeína que o café Robusta.

Há diversas regiões de plantio de café no mundo e segundo ALBUQUERQUE (ALBUQUERQUE, 2022) segue os 10 países com maior produção:

- 1. **Brasil** com 2.295.000 toneladas métricas.
- 2. **Vietnã** com 1.650.000 toneladas métricas.
- 3. **Colômbia** com 810.000 toneladas métricas.
- 4. **Indonésia** com 660.000 toneladas métricas.
- 5. **Etiópia** com 384.000 toneladas métricas.
- 6. **Honduras** com 348.000 toneladas métricas.
- 7. **Índia** com 348.000 toneladas métricas.
- 8. **Uganda** com 288.000 toneladas métricas.
- 9. **México** com 234.000 toneladas métricas.
- 10. **Guatemala** com 204.000 toneladas métricas.

Referente às classificações de qualidade total do café, as mais utilizadas baseiam-se nas características físicas do grão e nas sensoriais da bebida (HALAL, 2017).

Na avaliação da qualidade da bebida, a técnica utilizada é conhecida como a prova da xícara, realizada por degustadores previamente treinados para identificar padrões na bebida. Segundo ROSSETI(ROSSETI, 2007), o processo de avaliação começa com a torrefação e moagem do café, seguido da preparação de uma infusão do pó com 100ml de água em ponto de primeira fervura. Após isso, a mistura é mexida com uma concha, iniciando-se a prova sensorial onde o degustador sente o aroma da infusão e, em seguida,

retira a espuma. O próximo passo envolve a sucção de uma quantidade de líquido com a concha, mantendo-o na boca por um tempo suficiente para a identificação do sabor, sendo o líquido depois expelido em um recipiente chamado "cuspideira".

Sobre os *atributos sensoriais*, correspondem às características do café percebidas pelos sentidos, especialmente pelo olfato e paladar. Estes atributos são importantes para a avaliação de qualidade do café, geralmente realizados em uma prova da xícara (cupping) (BRAVIN, 2021).

Já os *atributos físicos* são as características objetivas e visíveis do grão de café antes da torra. Estes atributos são fundamentais para classificar a qualidade do café com base em sua aparência e outros fatores mensuráveis (ASSOCIATION, 2015).

E além das características sensoriais e físicas, a qualidade do café também é fortemente influenciada pela região de cultivo, incluindo fatores como o clima, a altitude e as práticas de cultivo. A região em que o café é produzido pode afetar as características sensoriais da bebida, como a acidez, o corpo e o sabor, conferindo um perfil único a cada localidade (CoffeesBR, 2025).

Como resultado da avaliação sensorial e física do café, é gerada uma classificação da bebida em diferentes categorias, como *Gourmet*, *Superior*, *Tradicional* e *Não Recomendável para Comercialização*. Essas classificações auxiliam tanto os produtores quanto os consumidores a compreenderem a qualidade do café e facilitam o processo de comercialização (CAFé, 2023).

2.3 Trabalhos Relacionados

O principal trabalho encontrado sobre aplicação de aprendizado de máquina em avaliações de café foi realizado por OLIVEIRA (OLIVEIRA, 2016), onde a pesquisa foi mais especializada na classificação sensorial de cafés especiais com a utilização dos algoritmos Boosting e Bagging. E a conclusão obtida foi que os métodos automáticos de classificação de dados sensoriais permitiram verificar a capacidade de capturar pequenas diferenças de amostras de cafés com a utilização dos algoritmos, mostrando a habilidade simplificada e eficaz de classificação sobre dados de multiclasse.

Há também o trabalho do MESSIAS (MESSIAS, 2009) onde foi investigado o uso de redes neurais artificiais para correlacionar análises sensoriais e químicas de café Arábica, visando classificar a qualidade da bebida. O estudo utilizou amostras das cultivares Acaiá do Cerrado, Topázio, Catuaí Vermelho 99, Acaiá 474-19 e Bourbon, provenientes da região sul de Minas Gerais, com processamento pós-colheita controlado. As análises químicas incluíram açúcar não redutor, açúcar redutor, acidez titulável total, condutividade elétrica, lixiviação de potássio e polifenóis, enquanto a qualidade sensorial foi avaliada através de análise sensorial. Ao comparar a regressão linear múltipla com a rede neural, a regressão

obteve um coeficiente de determinação de 0,6681, enquanto a rede neural atingiu uma taxa de acerto de 85% na classificação sensorial. Testado com amostras de cafés de propriedades da mesma região, mas com processamento pós-colheita sem controle, a rede neural alcançou 75% de precisão. Os resultados destacam a rede neural artificial como uma ferramenta eficaz para correlacionar dados sensoriais e químicos, oferecendo uma abordagem promissora para a classificação da qualidade do café.

No publicação de BAQUETA *et al.* (BAQUETA *et al.*, 2023) analisou a qualidade do café Coffea canephora (café Robusta) usando espectroscopia no infravermelho próximo (NIR) e técnicas quimiométricas. O objetivo foi aplicar essas ferramentas para analisar as características químicas e físicas do café, como umidade, acidez e compostos voláteis, e assim determinar a qualidade da bebida. A espectroscopia NIR é uma técnica rápida e não destrutiva que permite obter informações detalhadas sobre a composição do café. Junto com as técnicas quimiométricas, que ajudam a interpretar os dados, o estudo conseguiu classificar as amostras de café com base nas suas propriedades químicas, identificando padrões que indicam a qualidade do produto. Os resultados mostraram que essa abordagem foi eficaz para diferenciar amostras de café de diferentes qualidades, oferecendo uma maneira mais rápida e objetiva de avaliar o café, sem a necessidade de métodos complexos.

2.4 Cenário, Limitações e Proposta de Solução

O cenário de procura por informação agregativa no contexto da avaliação de café tem se tornado cada vez mais complexo e desafiador. Com a crescente demanda por produtos de alta qualidade e a necessidade de atender às expectativas de consumidores exigentes, a identificação e classificação precisa das características sensoriais e químicas do café se tornaram cruciais.

Tecnologias inovadoras, como inteligência artificial e aprendizado de máquina, têm desempenhado um papel fundamental nesse cenário, oferecendo ferramentas poderosas para análise de dados e extração de informações relevantes. No entanto, a aplicação dessas tecnologias em avaliações de café ainda enfrenta limitações importantes.

Uma das principais limitações é a complexidade e a subjetividade da análise sensorial do café. A avaliação sensorial envolve a percepção humana de atributos como aroma, sabor, acidez e corpo, que são difíceis de quantificar e padronizar. Além disso, a análise sensorial pode ser influenciada por fatores como o estado de espírito do avaliador, o ambiente de degustação e a familiaridade com diferentes tipos de café.

Outra limitação é a falta de dados estruturados e padronizados para treinamento de modelos de aprendizado de máquina. As avaliações sensoriais de café são frequentemente realizadas por especialistas em degustação, que utilizam seus conhecimentos e experiências para classificar a qualidade da bebida. Esses dados podem ser subjetivos e inconsistentes,

o que dificulta o treinamento de modelos precisos e confiáveis.

Diante desse cenário, a proposta de solução para aprimorar a avaliação de café reside na combinação de diferentes abordagens e tecnologias. Uma delas é a aplicação do algoritmo K-means para identificar padrões e agrupamentos em dados sensoriais e químicos de café.

O K-means é um algoritmo de agrupamento não supervisionado que pode ser utilizado para identificar grupos de cafés com características semelhantes, com base em seus atributos sensoriais e químicos. Ao aplicar o K-means em dados de avaliação de café, é possível identificar grupos de cafés com perfis sensoriais e químicos distintos, o que pode auxiliar na classificação e na identificação de tendências e padrões.

Além disso, a utilização do K-means pode contribuir para a padronização e a objetivação da avaliação sensorial do café. Ao identificar grupos de cafés com características semelhantes, é possível criar descritores sensoriais mais precisos e padronizados, que podem ser utilizados por diferentes avaliadores para classificar a qualidade da bebida de forma mais consistente.

A aplicação do K-means em avaliação de café pode ser combinada com outras tecnologias inovadoras, como redes neurais artificiais, para aprimorar a precisão e a confiabilidade da classificação. As redes neurais podem ser utilizadas para modelar a relação entre os dados sensoriais e químicos do café e a qualidade da bebida, permitindo a criação de modelos preditivos mais precisos e robustos.

Em suma, a abordagem proposta para aprimorar a avaliação de café envolve a combinação de diferentes tecnologias, como K-means e redes neurais artificiais, para identificar padrões, padronizar a avaliação sensorial e criar modelos preditivos mais precisos. Essa abordagem pode contribuir para aprimorar a qualidade da avaliação de café, atender às expectativas dos consumidores e impulsionar a inovação no setor.

3 METODOLOGIA

Nesta seção será descrito a estruturação metodológica do trabalho. O processo incluiu a seleção e preparação da base de dados, a definição e análise dos atributos sensoriais e físicos, a escolha da ferramenta para o processamento e análise, e a aplicação de técnicas de pré-processamento para garantir a qualidade dos dados. Além disso, na etapa de processamento, foi implementada uma abordagem de clusterização com o algoritmo K-means, com o objetivo de identificar padrões e agrupamentos naturais nos dados sensoriais e físicos. A seguir, terá o detalhamento em cada uma dessas etapas, iniciando pela base de dados e prosseguindo com as análises e técnicas aplicadas.

3.1 Base De Dados

A base de dados utilizada neste estudo foi extraída da plataforma Kaggle (VOLPATTO, 2019) e contém informações detalhadas sobre a avaliação sensorial, física e regional de amostras de café, fornecidas pelo Instituto de Qualidade do Café (ICQ), uma organização especializada em análises sensoriais. Os dados são referentes ao ano de 2018 e foram utilizados exclusivamente para fins acadêmicos. Esta base foi escolhida devido à riqueza das informações e à sua relevância no contexto da análise da qualidade do café.

O conjunto de dados extraído e utilizado neste estudo contém 1.339 amostras 44 atributos.

Figura 1 – Conjunto de Dados. Fonte: Autor.

Atributos	Tipo	Valores Não Nulos	Valores Nulos
Species	object	1339	0
Owner	object	1332	7
Country.of.Origin	object	1338	1
Farm.Name	object	980	359
Lot.Number	object	276	1063
Mill	object	1021	318
ICO.Number	object	1180	159
Company	object	1130	209
Altitude	object	1113	226
Region	object	1280	59
Producer	object	1107	232
Number.of.Bags	int64	1339	0
Bag.Weight	object	1339	0
In.Country.Partner	object	1339	0
Harvest.Year	object	1292	47
Grading.Date	object	1339	0
Owner.1	object	1332	7
Variety	object	1113	226
Processing.Method	object	1169	170
Aroma	float64	1339	0
Flavor	float64	1339	0
Aftertaste	float64	1339	0
Acidity	float64	1339	0
Body	float64	1339	0
Balance	float64	1339	0
Uniformity	float64	1339	0
Clean.Cup	float64	1339	0
Sweetness	float64	1339	0
Cupper.Points	float64	1339	0
Total.Cup.Points	float64	1339	0
Moisture	float64	1339	0
Category.One.Defects	int64	1339	0
Quakers	float64	1338	1
Color	object	1069	270
Category.Two.Defects	int64	1339	0
Expiration	object	1339	0
Certification.Body	object	1339	0
Certification.Address	object	1339	0
Certification.Contact	object	1339	0
unit_of_measurement	object	1339	0
altitude_low_meters	float64	1109	230
altitude_high_meters	float64	1109	230
altitude_mean_meters	float64	1109	230

Abaixo estão listados os atributos principais do conjunto de dados, organizados em dois grupos com base no objetivo deste trabalho.

3.1.1 Atributos Sensoriais

Os atributos sensoriais refletem a experiência subjetiva na avaliação do café, conforme descrito a seguir:

- **Aroma**: Cheiro do café, uma característica essencial para a análise sensorial (BRAVIN, 2021).
- **Flavor (Sabor)**: Gosto do café, influenciado pela composição química do grão (ALBUQUERQUE, 2022).
- **Aftertaste (Retrogosto)**: Sabor residual percebido após engolir o café (NEWTON, 2020).
- **Acidity (Acidez)**: Sensação refrescante e vibrante associada ao café (BRAVIN, 2021).
- **Body (Corpo)**: Sensação tátil na boca, indicando o peso ou densidade do café (ROSSETI, 2007).
- **Balance (Equilíbrio)**: Harmonia entre os sabores principais, como doce, amargo e ácido (BRAVIN, 2021).
- **Uniformity (Uniformidade)**: Consistência dos atributos sensoriais entre amostras (ALBUQUERQUE, 2022).
- **Sweetness (Doçura)**: Presença de sabores doces naturais no café (ROSSETI, 2007).
- **Clean Cup (Limpeza da xícara)**: Pureza do café, sem defeitos perceptíveis (BRAVIN, 2021).

3.1.2 Atributos Físicos

Os atributos físicos estão relacionados às características intrínsecas do grão de café e seu processamento:

- **Species (Espécie)**: Espécie do café, como Arábica ou Robusta, que influencia o perfil sensorial (BRAVIN, 2021).
- **Color (Cor)**: Indica o grau de maturação e qualidade da colheita (ALBUQUERQUE, 2022).

- **Moisture (Umidade)**: Nível de água nos grãos, importante para conservação e qualidade (BRAVIN, 2021).
- **Category.One.Defects (Defeitos Categoria Um)**: Grãos com defeitos graves que afetam significativamente a qualidade (ROSSETI, 2007).
- **Quakers**: Grãos imaturos que comprometem o perfil sensorial (ALBUQUERQUE, 2022).
- **Category.Two.Defects (Defeitos Categoria Dois)**: Defeitos menores, mas que ainda impactam a qualidade final (NEWTON, 2020).
- **Processing.Method (Método de Processamento)**: Técnica utilizada após a colheita, como via seca ou úmida (BRAVIN, 2021).

É fundamental ressaltar que, embora a base de dados contenha informações sobre a avaliação sensorial do café, incluindo a pontuação total (Total.Cup.Points), este estudo adota uma abordagem de aprendizado não supervisionado com o algoritmo K-means. Dessa forma, o objetivo principal não é prever a qualidade do café com base em seus atributos, mas sim identificar padrões e agrupamentos naturais nos dados, revelando relações entre as características sensoriais e físicas das amostras.

Nesse contexto, a coluna Total.Cup.Points (e outras colunas que possam indicar qualidade ou classificação) não serão utilizadas como rótulos no sentido tradicional de aprendizado supervisionado. O algoritmo K-means será aplicado para agrupar as amostras de café com base na similaridade de seus atributos, sem a necessidade de informações adicionais sobre qualidade ou pontuação.

A escolha por uma abordagem não supervisionada justifica-se pela busca por insights sobre a estrutura inerente dos dados e pela identificação de grupos de cafés com perfis sensoriais e físicos distintos. Essa análise pode contribuir para um melhor entendimento da qualidade do café, revelando relações entre seus atributos e auxiliando na tomada de decisões em diversas etapas da cadeia de produção. A determinação do número ideal de clusters (k) será feita com base em critérios como a inércia, o coeficiente de silhueta ou outros métodos de avaliação de agrupamentos, buscando um equilíbrio entre a granularidade da análise e a interpretabilidade dos resultados.

3.2 Ferramenta Utilizada

A implementação da análise foi realizada utilizando a linguagem Python, devido ao seu amplo suporte para bibliotecas de análise de dados e aprendizado de máquina, como Pandas, NumPy e Scikit-learn. O ambiente de desenvolvimento escolhido foi o Google Colab, que oferece uma interface interativa e recursos computacionais gratuitos.

3.3 Pré-Processamento Dos Dados

O pré-processamento é crucial para garantir a integridade e a qualidade dos dados. Nessa etapa, foram realizadas transformações e exclusões para remover inconsistências, redundâncias e atributos irrelevantes.

3.3.1 Remoção De Atributos

Visando um bom desempenho e o objetivo de analisar agrupamentos com base nos aspectos sensoriais, somados aos principais aspectos físicos, foi realizada a remoção de atributos irrelevantes que forneciam informações administrativas e regionais.

Em relação aos atributos sensoriais e físicos, foi realizada uma análise que resultou nas seguintes conclusões:

Atributo "Espécie": Por ser um atributo categórico binário, sua inclusão poderia gerar uma segmentação excessiva dos clusters. Como o foco da análise está nas características sensoriais, que têm maior relevância para a avaliação da qualidade do café, optou-se por não incorporar esse atributo ao conjunto de dados.

Atributos "Quaker"e "Defeito de Primeira Categoria (Graves)": A análise de outliers revelou que a maioria das amostras apresentava valor zero para esses atributos, indicando uma ausência significativa desses defeitos na maior parte dos cafés avaliados. Portanto, esses atributos foram considerados pouco informativos para a análise.

A remoção desses atributos foi necessária para otimizar o desempenho do algoritmo de agrupamento e garantir que o foco da análise permanecesse nos atributos mais relevantes para a avaliação da qualidade do café.

3.3.2 Valores Ausentes E Duplicados

Durante a verificação dos valores ausentes, observou-se que apenas o atributo "Color"apresentava valores ausentes em 270 instâncias. Após análise, decidiu-se pela remoção dessas instâncias, em vez de utilizar técnicas de imputação, com o objetivo de evitar qualquer alteração ou viés nos resultados.

Quanto à verificação de valores duplicados, foi identificada apenas uma instância duplicada, a qual foi removida para preservar a integridade dos dados restantes.

3.3.3 Análise Das Variáveis Categóricas

Em relação à variável categórica "Color", foram identificados três valores distintos: 'Blue-Green', 'Bluish-Green' e 'Green'. Observou-se que 'Blue-Green' e 'Bluish-Green' têm o mesmo significado, e, portanto, foi realizada a substituição de 'Blue-Green' por 'Bluish-Green'. Dessa forma, foi atualizada os valores da coluna, passando a apresentar apenas dois valores categóricos distintos.

Além disso, foi realizada a uniformização dos valores da coluna "Color", com a conversão para minúsculas e a remoção de espaços em branco extras.

3.3.4 Variáveis Númericas - Valores Zerados

Na análise estatística descritiva, foram identificadas instâncias com todos os atributos zerados. Para tratar esse problema, os atributos foram separados em três grupos: Sensoriais, Físicos e Total Cup Points. Após a segmentação, foi realizada uma consulta para identificar instâncias com valores zerados em cada grupo. O resultado indicou que apenas uma instância apresentava valores zerados nos atributos sensoriais e em Total Cup Points. Diante disso, essa instância foi removida.

3.3.5 Análise Sobre Outliers

A identificação e remoção de outliers é uma etapa crucial no pré-processamento dos dados, pois esses valores extremos podem distorcer os resultados e afetar negativamente a análise subsequente.

Para a identificação dos outliers, utilizou-se a metodologia do Intervalo Interquartílico (IQR). Primeiramente, foram calculados o Q1 (primeiro quartil) e o Q3 (terceiro quartil) para cada um dos atributos sensoriais analisados. O IQR foi então obtido pela diferença entre Q3 e Q1. Com esse valor, foi possível determinar os limites inferiores e superiores para a detecção de outliers, utilizando o seguinte cálculo:

Limite inferior = Q1 - 1.5 $\times$ IQR Limite superior = Q3 + 1.5 $\times$ IQR

A priorização dos atributos para verificação de outliers foi realizada com base na observação de altos desvios padrão nos atributos Defeitos Categoria 2 (Leves) e Umidade, identificados por meio da análise estatística descritiva, e considerando informações da literatura (ASSOCIATION, 2022), que indicam que a umidade ideal do café deve variar entre 10% e 12%.

Com base nesses dados, foi realizada a verificação de outliers, o que revelou a necessidade de remoção das instâncias com valores discrepantes, a fim de assegurar a integridade dos dados para análises subsequentes.

3.3.6 Normalização Dos Dados Numéricos

Para a normalização dos dados numéricos, iniciou-se com a verificação da ausência de valores NaN. Em seguida, foi aplicado o método Pipeline, uma ferramenta para a criação de fluxos de trabalho de aprendizado de máquina, combinado com o Min-Max Scaler para normalizar os valores, escalando-os para o intervalo entre 0 e 1. Após a transformação, foi realizada uma nova verificação de valores NaN gerados pela normalização. Essa estratégia de verificação tanto antes quanto depois da transformação foi adotada para garantir que

nenhuma instância fosse impactada de forma indesejada durante o processo, assegurando a integridade dos dados.

3.3.7 Transformação Dos Dados Categóricos Em Numéricos

Para o único atributo categórico Color, foi adotada a função map devido à simplicidade da variável, que possui apenas dois valores possíveis, tornando-a binária. Ou seja, os valores 'green' foram mapeados para 1 e 'bluish-green' foram mapeados para 0. A escolha do mapeamento binário se baseia no fato de que a presença de apenas dois valores distintos não justifica a necessidade de técnicas mais complexas, como One-Hot Encoding, que seriam desnecessárias e poderiam aumentar a dimensionalidade do conjunto de dados sem trazer benefícios reais para a análise. Dessa forma, a transformação direta utilizando o map garante que os dados estejam preparados de forma adequada para o modelo, mantendo a eficiência do processo sem introduzir complexidade adicional.

3.3.8 Descrição Do Conjunto De Dados Após Pré-processamento

Após o pré-processamento, o conjunto de dados contém 823 entradas e 12 atributos. Destas, 11 são numéricas e 1 é categórica. O conjunto de dados foi reduzido de 1339 para 823 instâncias após a remoção de valores ausentes, outliers e a exclusão de atributos irrelevantes para a análise. A listagem a seguir apresenta as colunas restantes e seus tipos de dados:

- Aroma: 823 entradas, tipo numérico (float64)
- Flavor: 823 entradas, tipo numérico (float64)
- Aftertaste: 823 entradas, tipo numérico (float64)
- Acidity: 823 entradas, tipo numérico (float64)
- Body: 823 entradas, tipo numérico (float64)
- Balance: 823 entradas, tipo numérico (float64)
- Uniformity: 823 entradas, tipo numérico (float64)
- Clean.Cup: 823 entradas, tipo numérico (float64)
- Sweetness: 823 entradas, tipo numérico (float64)
- Moisture: 823 entradas, tipo numérico (float64)
- Color: 823 entradas, tipo categórico (int64), com valores mapeados de 'green' (1) e 'bluish-green' (0)

- Category.Two.Defects: 823 entradas, tipo numérico (float64)

Após o tratamento de dados ausentes e a remoção de outliers, as variáveis restantes passaram a estar em uma escala uniforme e sem dados faltantes, o que garante a integridade e a qualidade das análises subsequentes. O uso de técnicas como normalização e mapeamento de valores categóricos possibilitou a preparação adequada do conjunto de dados para os algoritmos de aprendizado de máquina, com destaque para o K-means, que será aplicado para identificar relacionamentos orgânicos entre os atributos sensoriais e físicos das amostras de café, buscando padrões que influenciam na avaliação da qualidade. O tamanho final do conjunto de dados, com 823 instâncias, está adequado para a realização das análises e modelagem propostas neste estudo.

A Figura 2 abaixo mostra as 5 primeiras entradas do conjunto de dados após o pré-processamento, permitindo visualizar a disposição dos dados.

Figura 2 – Amostra do conjunto de dados pré-processado

	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Clean.Cup	Sweetness	Moisture	Color	Category.Two.Defects
0	0.978202	1.000000	1.000	1.000	1.000000	0.936	1.0	1.0	1.0	0.666667	1	0.0
1	1.000000	0.941818	0.932	0.932	0.963134	0.936	1.0	1.0	1.0	0.666667	1	0.1
2	0.841962	0.909091	0.900	0.868	1.000000	0.868	1.0	1.0	1.0	0.500000	1	0.2
3	0.863760	0.880000	0.832	0.900	0.963134	0.900	1.0	1.0	1.0	0.666667	1	0.2
4	0.953678	0.850909	0.900	0.900	0.884793	0.900	1.0	1.0	1.0	0.500000	0	0.1

Fonte: Autor.

3.4 Processamento Dos Dados

3.4.1 Abordagem Da Clusterização Dados Sensorias E Físico

A segmentação dos dados foi realizada utilizando o algoritmo K-Means, com o objetivo de identificar padrões e agrupar amostras de café com características sensoriais semelhantes. Inicialmente, optou-se por 5 clusters, baseando-se na experiência do analista e no conhecimento prévio do comportamento dos dados, visando uma segmentação que refletisse adequadamente a diversidade das características sensoriais.

O modelo foi configurado com o parâmetro de semente estado randomico definido como 1224, garantindo a reprodutibilidade dos resultados. Além disso, a opção verbose foi ativada para monitorar o progresso do treinamento do modelo durante a execução.

Após o treinamento do modelo, o K-Means atribuiu a cada amostra de café um rótulo correspondente ao cluster ao qual ela pertence. Esse rótulo foi adicionado ao conjunto de dados original, em uma nova coluna chamada Cluster. Com isso, cada instância foi categorizada em um dos cinco clusters gerados, o que facilita a análise de padrões sensoriais comuns dentro de cada grupo.

A escolha de 5 clusters permitiu uma segmentação eficiente, criando grupos com características sensoriais semelhantes e possibilitando uma avaliação detalhada da qualidade

do café. No entanto, para validar a escolha do número de clusters, foram aplicadas métricas de avaliação, conforme descrito na seção seguinte.

3.4.2 Validação Da Qualidade Dos Clusters Gerados

A decisão sobre o número de clusters a ser utilizado no K-Means é uma parte decisiva da análise, pois influencia diretamente a qualidade e a utilidade dos agrupamentos gerados. Durante o processo de avaliação dos diferentes valores de K, foi possível observar que, apesar de ser possível testar uma faixa ampla de valores de K, os resultados das métricas de avaliação (Silhouette Score, Davies-Bouldin e Calinski-Harabasz) mostraram que o número de 5 clusters, escolhido inicialmente, já apresentava uma segmentação robusta e consistente.

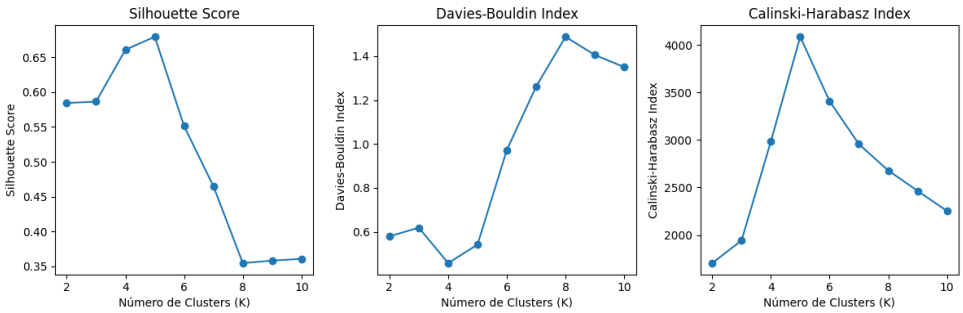
1. Resultados Consistentes com K = 5: Ao testar o número de clusters variando de K = 2 até K = 10, as métricas de avaliação (Silhouette Score, Davies-Bouldin e Calinski-Harabasz) mostraram um desempenho satisfatório para K = 5, com pontuações consistentes e bem equilibradas em todas as métricas. O Silhouette Score, que mede a coesão e a separação dos clusters, apresentou um valor próximo ao ideal (entre -1 e 1), indicando que os pontos estão bem agrupados e distantes de clusters vizinhos. O Índice Davies-Bouldin, que avalia a relação entre a dispersão intra-cluster e a separação inter-cluster, mostrou valores baixos, sugerindo que os clusters são compactos e bem distanciados entre si. Por fim, o Índice Calinski-Harabasz, que quantifica a razão entre a dispersão inter-cluster e intra-cluster, apresentou valores elevados, reforçando a robustez da segmentação.

2. Saturação Das Métricas: Com o aumento do número de clusters além de K = 5, observou-se uma estabilização ou até mesmo uma pequena deterioração das métricas. Por exemplo, o Silhouette Score pode ter mostrado um leve aumento, mas os ganhos não foram significativos, sugerindo que o número ótimo de clusters já havia sido atingido. Da mesma forma, as métricas de Davies-Bouldin e Calinski-Harabasz não indicaram uma melhora consistente com valores de K maiores que 5.

3. Evitar Overfitting: Testar um número maior de clusters poderia resultar em uma segmentação excessivamente detalhada, onde os grupos se tornariam muito pequenos e menos representativos, o que pode gerar um modelo mais complexo sem necessariamente refletir a verdadeira estrutura dos dados. Esse fenômeno é conhecido como overfitting onde o modelo tenta "forçar"uma segmentação que não é realmente significativa.

4. Simplicidade E Interpretabilidade: A simplicidade e a interpretabilidade dos modelos são aspectos importantes na análise de dados. Utilizar 5 clusters oferece um equilíbrio entre a segmentação precisa dos dados e a facilidade de interpretação dos resultados. Segmentações com um número maior de clusters podem ser mais difíceis de interpretar, especialmente se os clusters adicionais não representarem padrões reais e importantes dentro dos dados.

Figura 3 – Validação do número de cluster



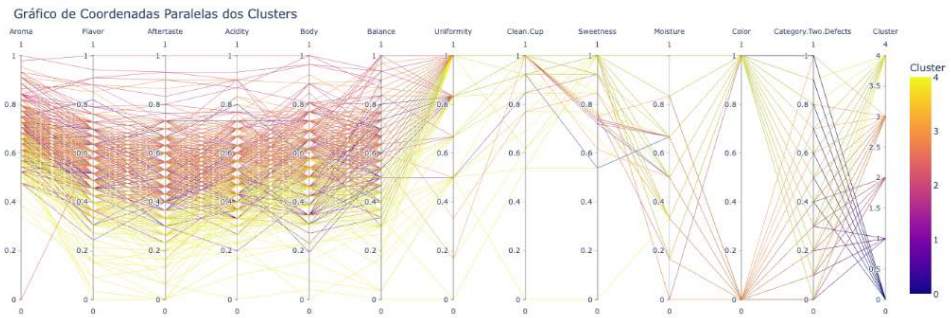
Fonte: Autor.

Com base no resultado apresentado pela Figura 3, é possível concluir que a escolha de 5 clusters foi validada pelas métricas de avaliação e pela análise qualitativa do modelo. Não houve a necessidade de testar valores de K maiores, uma vez que o número de clusters já proporcionava uma segmentação eficaz e balanceada, além de garantir um modelo mais simples e interpretável. A partir disso, conclui-se que a segmentação com 5 clusters é adequada para os objetivos do estudo e para uma análise mais robusta dos dados.

Com o número de clusters definido, prossegue-se para a análise detalhada dos resultados da segmentação. Nesta próxima etapa, investigaremos as características de cada cluster, com foco nos atributos mais significativos, identificados com base nos centróides. Essa análise permitirá extrair padrões e insights relevantes sobre os dados.

3.4.3 Análise dos Atributos Mais Significativos em Cada Cluster com Base nos Centróides

Figura 4 – Gráfico de Coordenadas Paralelas dos Clusters.



Fonte: Autor.

A Figura 4 apresenta o gráfico de coordenadas paralelas dos clusters gerados após a validação do número de clusters. Este gráfico permite visualizar a relação entre os atributos e os clusters, mostrando como os dados foram agrupados. Cada linha representa uma amostra de café, com a cor indicando o cluster correspondente. Os eixos paralelos representam os atributos, e a posição da linha em cada eixo reflete o valor do atributo para aquela amostra.

A análise visual do gráfico revela padrões e características distintas entre os clusters, destacando como os atributos contribuem para a formação de cada grupo. Para uma interpretação mais aprofundada, é essencial identificar os atributos mais relevantes em cada cluster. Para isso, analisamos os centróides gerados pelo K-means, que representam as médias das características de cada grupo. A Figura 5 exibe os centróides, ou seja, a média de cada atributo por cluster.

Figura 5 – Centróides dos Clusters.

Cluster	Aroma	Flavor	Aftertaste	Acidity	Body	Balance
0	0.681143	0.538751	0.499864	0.520163	0.548199	0.575510
1	0.654306	0.484553	0.447826	0.472730	0.485314	0.521200
2	0.732036	0.621050	0.594197	0.602836	0.648839	0.674085
3	0.697198	0.551469	0.527487	0.537154	0.566791	0.605231
4	0.578577	0.326706	0.251065	0.366805	0.385062	0.356883

Cluster	Uniformity	Clean.Cup	Sweetness	Moisture	Color
0	0.984065	0.997905	0.989337	0.557823	1.0
1	0.982565	0.996981	0.995687	0.555072	1.0
2	0.985094	0.993859	0.987854	0.500000	1.0
3	0.970000	0.992592	0.989449	0.520299	0.0
4	0.854968	0.922078	0.955661	0.536797	1.0

Cluster	Category.Two.Defects	Count
0	0.607483	147
1	0.107826	230
2	0.146479	213
3	0.169231	156
4	0.455844	77

Fonte: Autor.

A análise dos centróides permite identificar quais variáveis estão mais associadas a cada segmento, destacando os atributos que mais contribuem para a formação dos clusters. Ao comparar os valores dos centróides, é possível determinar as características mais proeminentes em cada grupo de cafés.

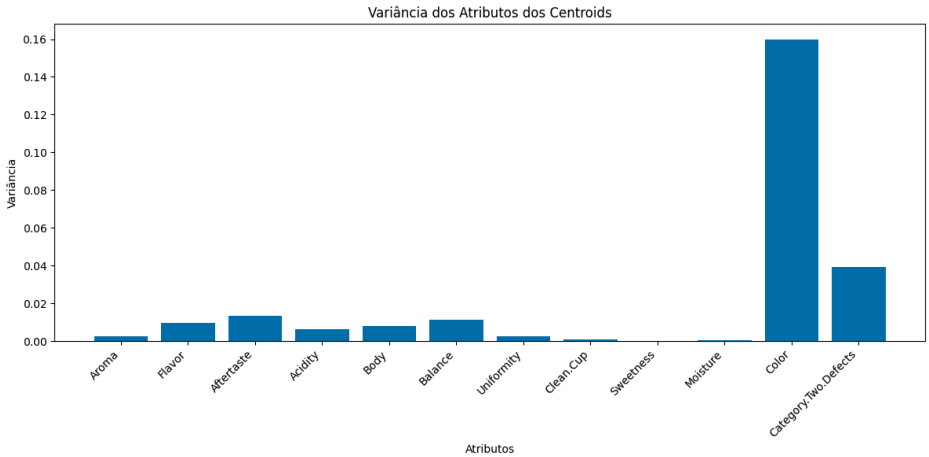
Variâncias Para Confirmação Da Relevância

A análise da variância de cada atributo dentro dos centróides complementa a análise dos centróides, permitindo identificar os atributos com maior dispersão entre os clusters e, consequentemente, maior influência na diferenciação dos grupos. A variância, calculada para cada atributo com base nos valores dos centróides, revela quais variáveis têm maior variação entre os clusters, indicando sua relevância na formação dos grupos.

A ordenação decrescente das variâncias destaca os atributos mais relevantes para a diferenciação entre os clusters, revelando aqueles com maior impacto na segmentação.

Na Figura 6, é possível visualizar os resultados obtidos sobre a variância dos centróides:

Figura 6 – Variâncias para Confirmação da Relevância.



Fonte: Autor.

3.4.4 Reflexão da Definição Técnica dos Clusters (agrupamentos)

Os resultados da análise de variância revelam que a cor do café (atributo *Color*) e a presença de defeitos (atributo *Category.Two.Defects*) são os atributos mais relevantes para a diferenciação dos clusters, apresentando as maiores variâncias. Atributos sensoriais como *Aftertaste* e *Balance* também contribuem significativamente, enquanto atributos como *Moisture* e *Sweetness* apresentam variações mínimas, indicando sua menor influência na segmentação.

A análise detalhada dos clusters, apresentada a seguir, permite identificar os perfis específicos de cada grupo de cafés, complementando a análise de variância e oferecendo uma compreensão mais aprofundada da segmentação.

3.4.5 Análise e Interpretação dos Clusters

E conforme demonstrado na análise da variância com base nos centróides, a categoria da cor do grão é o atributo com maior relevância para a diferenciação dos clusters. Esse fator tornou-se a base motivadora para a definição dos grupos, uma vez que a cor reflete características intrínsecas do café, como maturação, processamento e qualidade. Portanto, abaixo segue a interpretação detalhada de cada cluster rotulando incialmente com base na cor do grão de café seguido das caracterísricas sensorias e físicas:

Cluster 0: Grão Verde Este cluster se caracteriza pelas maiores pontuações em limpeza da xícara, umidade e presença de defeitos leves. Isso sugere que os grãos possuem boa qualidade sensorial e são menos propensos a defeitos visíveis. No entanto, a alta presença de defeitos leves pode comprometer outras qualidades sensoriais do café, impactando negativamente o perfil geral do produto.

Cluster 1: Grão Verde Azulado Os grãos deste cluster se destacam pela alta

pontuação em doçura, equilíbrio, aroma, retrogosto e sabor, refletindo um perfil sensorial equilibrado e agradável. No entanto, as baixas pontuações em uniformidade, defeitos leves e umidade indicam maior variabilidade na qualidade do grão, afetando a consistência e a confiabilidade do produto final.

Cluster 2: Grão Verde Este cluster é o melhor em termos de qualidade sensorial, com altas pontuações em aroma, sabor, retrogosto, acidez, corpo, equilíbrio e uniformidade. A umidade é o único atributo que se mantém baixo, sugerindo que os grãos passaram por um processo de secagem eficiente, preservando a qualidade sem comprometer os atributos sensoriais. Esses grãos representam a excelência em termos de características sensoriais.

Cluster 3: Grão Verde Os grãos deste cluster apresentam pontuações mais baixas na maioria dos atributos sensoriais, exceto em umidade e presença de defeitos leves, que são relativamente mais altos. Isso sugere que os grãos possuem deficiências sensoriais significativas, como baixa qualidade de sabor e aroma. No entanto, a umidade controlada pode ser um ponto positivo, favorecendo a conservação do produto por mais tempo.

Cluster 4: Grão Verde Neste cluster, os grãos têm pontuações baixas na maioria dos atributos sensoriais, com exceção de uniformidade e limpeza da xícara. A doçura é o atributo mais alto, enquanto a presença de defeitos leves é a menor. Isso indica que, embora algumas características sensoriais sejam menos marcantes, os grãos são bem cuidados, com poucos defeitos e boa uniformidade e limpeza, sugerindo um produto de qualidade consistente.

3.4.6 Resumo Da Abordagem

Conclui-se que os clusters analisados apresentam uma variedade de perfis sensoriais e físicos, que vão desde cafés de alta qualidade sensorial e equilíbrio (Cluster 2), até cafés com defeitos e qualidades inferiores (Cluster 3). A análise revela que, para melhorar a qualidade geral do café, é importante focar no controle de defeitos, na uniformidade do processo e na umidade dos grãos. A doçura é um atributo positivo, mas não é suficiente para compensar a falta de outros aspectos essenciais, como sabor e aroma. Melhorias no processamento e armazenamento podem ter um impacto direto na qualidade final do produto.

4 CONCLUSÃO

Este trabalho teve como objetivo analisar e segmentar dados sensoriais de amostras de café por meio do algoritmo de clustering K-Means, com o intuito de identificar padrões relevantes nas características sensoriais como sabor, aroma, acidez, umidade, entre outros, e, assim, compreender melhor a relação entre esses atributos e a qualidade do café.

A primeira etapa envolveu uma análise rigorosa dos dados, na qual valores discrepantes foram identificados e removidos para garantir a qualidade da segmentação. Em seguida, aplicaram-se técnicas de normalização para uniformizar os atributos em uma escala comum, essencial para o bom desempenho do algoritmo de clustering. A transformação dos dados categóricos em variáveis numéricas também foi realizada, permitindo a correta aplicação do modelo K-Means.

Após o pré-processamento, o algoritmo K-Means foi utilizado para segmentar as amostras de café em cinco clusters (K=5), escolha validada pelas métricas Silhouette Score, Davies-Bouldin Index e Calinski-Harabasz Index. Esses índices indicaram que cinco clusters eram suficientes para representar os diferentes grupos de amostras, sem a necessidade de testar maiores valores de K. A segmentação resultante revelou clusters com perfis sensoriais distintos, que refletiram características como doçura, acidez, aroma, presença de defeitos e qualidade da xícara.

A interpretação dos clusters forneceu insights valiosos sobre os diferentes tipos de grãos de café presentes nas amostras:

Cluster 0 (Grão Verde): Grãos com alta limpeza da xícara e umidade, mas com defeitos leves significativos, o que indica boa qualidade inicial, mas com impacto nos atributos sensoriais devido à presença desses defeitos.

Cluster 1 (Grão Verde Azulado): Grãos com excelente perfil sensorial em doçura, equilíbrio, aroma e sabor, porém com baixa uniformidade e umidade. Isso aponta para uma qualidade sensorial notável, mas com variabilidade nos grãos que pode afetar a consistência do café.

Cluster 2 (Grão Verde): Este cluster representa grãos com alto desempenho sensorial em todos os atributos, exceto a umidade, que é controlada. Indica que os grãos têm excelente qualidade sensorial, com uma secagem eficiente que não compromete o sabor nem o aroma.

Cluster 3 (Grão Verde): Grãos com defeitos evidentes, refletindo baixa pontuação em quase todos os atributos, exceto na umidade. A alta presença de defeitos leves compromete fortemente a qualidade sensorial desses grãos.

Cluster 4 (Grão Verde): Apesar de uma pontuação baixa em atributos sensoriais,

esse grupo se destaca pela uniformidade e limpeza da xícara, com uma doçura notável e presença mínima de defeitos leves. Isso sugere que, embora o perfil sensorial seja modesto, a qualidade do processamento dos grãos é bem mantida.

A validação dos clusters através das métricas indicou que a segmentação foi adequada para representar a variabilidade nas amostras de café e oferecer insights sobre a qualidade sensorial dos grãos. A análise dos resultados sugere que, embora a segmentação tenha sido eficaz, existem aspectos a serem melhorados, como a redução de defeitos e a melhoria da uniformidade.

Propostas para Trabalhos Futuros Para expandir os resultados obtidos e melhorar ainda mais a classificação sensorial do café, uma análise mais detalhada dos atributos correlacionados, como a relação entre acidez, corpo e umidade, pode ser realizada utilizando técnicas como análise de componentes principais (PCA) e modelos de regressão multivariada. Além disso, a integração de técnicas de aprendizado de máquina avançadas, como redes neurais e modelos de ensemble, pode ser explorada para refinar a segmentação e melhorar a precisão nas previsões.

Outras oportunidades incluem a análise comparativa entre os rótulos de pontuação por xícara, utilizados por especialistas da Specialty Coffee Association (SCA), e os resultados obtidos a partir da segmentação dos dados sensoriais. Isso permitirá validar a capacidade do modelo de refletir corretamente os padrões de qualidade estabelecidos pelas metodologias tradicionais de avaliação.

Por fim, seria relevante considerar os aspectos regionais na análise da qualidade do café. Diferentes regiões podem apresentar características únicas em termos de sabor, aroma e outros atributos sensoriais, influenciados por fatores climáticos, geográficos e agronômicos. Incorporar esses dados pode enriquecer ainda mais a análise e fornecer uma visão mais precisa da qualidade do café em contextos específicos.

Essas direções para novos trabalhos possibilitam uma compreensão mais profunda sobre os fatores que determinam a qualidade do café, ajudando a refinar as práticas de cultivo e processamento, além de oferecer um novo caminho para melhorar a classificação de cafés com base em técnicas de aprendizado de máquina.

REFERÊNCIAS

ALBUQUERQUE, D. Os 10 maiores produtores de café do mundo. **Socientífica**, 2022. Available at: https://socientifica.com.br/maiores-produtores-de-cafe-do-mundo/.

ASSOCIATION, S. C. **SCA - Specialty Coffee Association**. 2015. Acesso em: 18 jan. 2025. Available at: https://sca.coffee.

ASSOCIATION, S. C. **Atributos para avaliação de café: Diretrizes e normas**. 2022. Acessado: 2022-12-15. Disponível em: https://sca.coffee/.

BAIRROS, D. Café arábica: O que é, principais características, variações e blends. **Review Café**, 2022. Available at: https://reviewcafe.com.br/dicas-e-receitas/cafe-arabica/.

BAQUETA, M. R. *et al.* Brazilian canephora coffee evaluation using nir spectroscopy and discriminant chemometric techniques. **Journal of Food Composition and Analysis**, Elsevier, 2023. Available at: https://www.elsevier.com/locate/jfca.

BORHMKE, B. **K-means Cluster Analysis.** 2019. Available at: Availableat:https://uc-r.github.io/kmeans_clustering.

BRAVIN, N. P. **Caracterização sensorial dos cafés da região das Matas de Minas selecionados em concurso de qualidade**. 2021. Apud SCAA 2015. Dissertação apresentada à Universidade Federal de Viçosa.

CAFé, E. **Sistema de Classificação de Café**. 2023. Acesso em: 18 jan. 2025. Available at: https://www.embrapa.br/cafe.

CARVALHO, A. C. P. d. L. F. E. A. **Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina**. [*S.l.: s.n.*]: "LTC", 2022.

CoffeesBR. **A conexão entre produtores e amantes do café: a importância do selo de qualidade**. [*S.l.: s.n.*]: https://coffeesbr.com/a-conexao-entre-produtores-e-amantes-do-cafe-a-importancia-do-selo-de-qualidade, 2025. Acesso em: 18 jan. 2025.

ESCOVEDO, T. Machine learning: Conceitos e modelos — parte ii: Aprendizado não-supervisionado. **Medium**, 2020. Available at: https://medium.com/@tatianae_79457/machine-learning-conceitos-e-modelos-f0373bf4f445.

GÉRON, A. **Mãos A Obra: Aprendizado De Máquina Com Scikit-Learn, Keras TensorFlow**. [*S.l.: s.n.*]: "Alta Books", 2021.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [*S.l.: s.n.*]: MIT Press, 2016.

Governo do Brasil. **Brasil é o maior produtor mundial e o segundo maior consumidor de café**. [*S.l.: s.n.*]: https://www.gov.br/agricultura/pt-br/assuntos/noticias/brasil-e-o-maior-produtor-mundial-e-o-segundo-maior-consumidor-de-cafe, 2025. Acesso em: 18 jan. 2025.

HALAL, S. L. M. E. Composição, processamento e qualidade do café. **Revista de Qualidade do Café**, 2017.

KASSAMBARA, A. **Practical guide to cluster analysis in R: Unsupervised machine learning**. [*S.l.: s.n.*]: Sthda, 2017. v. 1.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. **Estudos Avançados**, SciELO Brasil, v. 35, p. 85–94, 2021.

MESSIAS, J. A. T. **Redes neurais artificiais: Nova abordagem para identificação da qualidade da bebida do café Arábica utilizando-se resultados de análises químicas**. 2009. Dissertação (Mestrado) — Universidade Federal de Viçosa, Viçosa, MG, 2009.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. [*S.l.: s.n.*]: MIT press, 2018.

NEWTON, T. Altitude no café, o quanto ela afeta a qualidade? **Perfect Daily Grind**, 2020. Available at: https://perfectdailygrind.com/pt/2020/06/01/cafe-de-altitude-o-quanto-ela-afeta-a-qualidade/.

OLIVEIRA, L. M. D. **CLASSIFICAÇÃO DE DADOS SENSORIAIS DE CAFÉS ESPECIAIS COM RESPOSTA MULTICLASSE VIA ALGORITMO BOOSTING E BAGGING**. [*S.l.: s.n.*]: Dissertação apresentada à Universidade Federal de Lavras, 2016.

ROSSETI, R. P. **Determinação de fenóis totais em frutos do café: Avaliações em diferentes fases de maturação.** [*S.l.: s.n.*]: Dissertação de Mestrado apresentado ao Institudo de Química da USP de São Carlos, 2007.

SOUSA, M. C. C. Uma análise do algoritmo k-means como introdução ao aprendizado de máquinas. **Universidade Federal do Tocantins**, 2023. Trabalho de Conclusão de Curso.

Splabor. **Controle de qualidade do café: importância e desafios**. [*S.l.: s.n.*]: https://www.splabor.com.br/blog/equipamentos-para-laboratorio/controle-de-qualidade-do-cafe-importancia-e-desafios, 2025. Acesso em: 18 jan. 2025.

VOLPATTO, D. **Coffee Quality database from CQI**. 2019. Kaggle. Acesso em: 16 maio 2024. Available at: https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi?select=merged_data_cleaned.csv.