Preditiva e Não Supervisionada da Qualidade Sensorial do Café Arábica: Uma Abordagem Baseada em Dados Sintéticos e Interpretabilidade de Modelos

Luiz Carlos Brandão Junior[1], Carla Simone Araújo Gomes Sarmento[2], Ricardo Rodrigues Magalhães[3]

[1,2,3] Departamento de Engenharia Agrícola/DAG – Universidade Federal de Lavras (UFLA)
Caixa Postal 3037 CEP 37203-202 – Lavras, MG – Brasil

contato@uflaniano.com.br[1], carla.sarmento@estudante.ufla.br[2], ricardorm@ufla.br[3]

Abstract. *This study presents a methodology that uses machine learning and synthetic data, grounded in empirical data (Sarmento 2023), to decode the quality factors of Arabica coffee. From 10,000 samples based on a real-world case in Barra do Choça (BA), the analysis revealed three distinct production profiles (terroirs) and identified altitude as the most determining factor for quality, surpassing all other variables. The main finding was that the differentiation of terroirs is vertical (altimetric) and not horizontal (geographic), refuting the idea of quality micro-regions. To substantiate and validate this finding, the IGV (Interactive Geographic Visualization) tool was developed. It plots the samples on an interactive map, allowing for the detailed inspection of each point's attributes (quality, altitude, terroir profile), confirming the altimetric terroir hypothesis at a micro-level. The IGV functions as a prototype for a decision support system for precision agriculture, capable of assisting producers in identifying areas of excellence on their properties.*

Keywords: *Clustering, Coffee Quality, Machine Learning, Terroir, Synthetic Data, Explainable AI (XAI), K-Means, Precision Agriculture, Altitude, Unsupervised Analysis, SHAP.*

Resumo. *Este estudo apresenta uma metodologia que utiliza aprendizado de máquina e dados sintéticos fundamentados em dados empiticos (Sarmento 2023) para decodificar os fatores de qualidade do café arábica. A partir de 10 mil amostras baseadas em um caso real em Barra do Choça (BA), a análise revelou três perfis de produção distintos (terroirs) e identificou a altitude como o fator mais determinante para a qualidade, superando todas as outras variáveis. A principal descoberta foi que a diferenciação dos terroirs é vertical (altimétrica) e não horizontal (geográfica), refutando a ideia de micro-regiões de qualidade. Para materializar e validar esta descoberta, foi desenvolvida a ferramenta VGI (Visualização Geográfica Interativa). Ela plota as amostras em um mapa interativo, permitindo a inspeção detalhada dos atributos de cada ponto (qualidade, altitude, perfil de terroir), comprovando a hipótese do terroir*

altimétrico em nível micro. A VGI funciona como um protótipo de sistema de apoio à decisão para a agricultura de precisão, capaz de auxiliar produtores a identificar áreas de excelência em suas propriedades.

Palavras-chave: *Clustering, Qualidade do Café, Machine Learning, Terroir, Dados Sintéticos, IA Explicável (XAI), K-Means, Agricultura de Precisão, Altitude, Análise Não Supervisionada, SHAP.*

1. Informações Gerais

A agricultura moderna tem sido progressivamente transformada pela aplicação de técnicas de ciência de dados e inteligência artificial, permitindo uma otimização sem precedentes de processos e uma compreensão mais profunda das interações entre genótipo, ambiente e manejo (Kamilaris, 2018). No setor da cafeicultura, particularmente na produção de cafés especiais, a diferenciação do produto final está intrinsecamente ligada ao conceito de *terroir*, um termo que encapsula a sinergia entre fatores como clima, solo, topografia e práticas culturais (Sarmento, 2023).

Apesar da sua importância comercial, a quantificação da influência relativa de cada componente do *terroir* na qualidade sensorial da bebida permanece um desafio. A complexidade dessas interações muitas vezes impede a formulação de modelos preditivos robustos que possam guiar produtores na busca por lotes de qualidade excepcional.

Este estudo aborda diretamente este problema, buscando responder à seguinte questão de pesquisa: **É possível, através de modelos de *Machine Learning*, desconstruir o conceito de *terroir* em uma hierarquia de variáveis quantificáveis e preditivas para a qualidade do café arábica?**

Para contornar a escassez de grandes bases de dados públicas que associam informações geoespaciais, espectrais e sensoriais, este trabalho adota a abordagem de gerar uma base de dados sintética de alta fidelidade. A estrutura e as distribuições estatísticas desta base foram informadas pelo conhecimento empírico consolidado na dissertação de Sarmento (2023), que estudou a produção de café no município de Barra do Choça, BA. A utilização de dados sintéticos permite um ambiente controlado para o teste e a validação rigorosa de hipóteses e modelos, superando barreiras de privacidade e disponibilidade de dados.

O objetivo principal deste artigo é, portanto, avaliar a aplicabilidade de um *framework* de análise preditiva e não supervisionada para: (1) identificar os principais fatores determinantes da pontuação sensorial do café; (2) quantificar o impacto e a direção da influência de cada fator utilizando técnicas de XAI; e (3) descobrir perfis ou *terroirs* naturais de produção através de algoritmos de *clustering*. Os resultados são consistentemente discutidos em comparação com os achados do estudo de referência, validando a sinergia entre o conhecimento de domínio e as metodologias de ciência de dados.

1.1. Fatores Determinantes da Qualidade do Café e o Conceito de Terroir

A qualidade final da bebida do café é o resultado de uma cadeia complexa de fatores. O conceito de *terroir* postula que as características intrínsecas da bebida são moldadas

pela interação entre o genótipo e o ambiente (Sarmento, 2023). Dentre os fatores ambientais, a altitude é frequentemente citada como um dos mais influentes. Regiões de maior altitude, com temperaturas mais amenas, tendem a promover uma maturação mais lenta e uniforme dos frutos, resultando em um maior acúmulo de açúcares, ácidos e outros compostos precursores de aromas e sabores complexos (Vaast 2006).

O processamento pós-colheita, como os métodos natural (secagem com a casca) e despolpado (remoção da polpa antes da secagem), também modifica significativamente o perfil químico e sensorial do grão (Chalfoun 2013).

1.2. Índices Espectrais de Vegetação como Proxies Ambientais

O sensoriamento remoto por satélites, como o Sentinel-2, tem se consolidado como uma ferramenta essencial para o monitoramento em larga escala da saúde e do estado fisiológico da vegetação (Drusch 2012, Li 2017). Nesse contexto, os índices de vegetação (IVs) são amplamente empregados como proxies ambientais, uma vez que consistem em formulações matemáticas que combinam bandas espectrais específicas para realçar atributos biofísicos e fisiológicos das plantas (Huete 2002, Gamon 1992).

Neste estudo, foram utilizadas como variáveis preditoras algumas métricas consagradas na literatura:

- **NDVI (Normalized Difference Vegetation Index):** indicador robusto do vigor vegetativo e da densidade da biomassa, amplamente utilizado em estudos de monitoramento ambiental (Rouse 1974).
- **EVI (Enhanced Vegetation Index):** semelhante ao NDVI, mas com correções adicionais para minimizar influências atmosféricas e do solo, sendo mais responsivo em áreas de alta biomassa (Huete 2002).
- **PSRI (Plant Senescence Reflectance Index):** sensível à proporção entre carotenoides e clorofila, funcionando como um marcador do estresse fisiológico e da senescência foliar (Merzlyak 1999).
- **PRI (Photochemical Reflectance Index):** relacionado à eficiência do uso da luz no processo fotossintético, sendo um proxy da atividade fotoquímica da vegetação (Gamon 1992).

Estudos anteriores, como a dissertação de (Sarmento 2023), já demonstraram a relevância desses índices no contexto da avaliação da qualidade, observando-se uma correlação inversa entre NDVI e qualidade, enquanto o PSRI apresentou associação positiva, sugerindo que níveis moderados de estresse fisiológico podem estar associados a melhorias qualitativas.

1.3. Machine Learning Aplicado à Agricultura

Modelos de regressão são algoritmos supervisionados cujo objetivo é prever um valor contínuo, como a *Pontuacao_Sensorial_Final*. O **Random Forest Regressor** é um modelo de conjunto (ensemble) que constrói múltiplas árvores de decisão e combina suas previsões, sendo conhecido por sua alta precisão e robustez a outliers (Breiman 2001).

A interpretabilidade de modelos complexos como o Random Forest é alcançada por meio de técnicas de XAI. O **SHAP** é um método baseado na teoria dos jogos que calcula a contribuição de cada feature para cada predição individual, fornecendo insights detalhados e confiáveis sobre o comportamento do modelo (Lundberg 2017).

A análise de **clustering**, por outro lado, é uma técnica não supervisionada. O algoritmo **K-Means** busca particionar o conjunto de dados em k grupos distintos, onde cada amostra pertence ao cluster com a média (centróide) mais próxima. Seu objetivo é descobrir a estrutura intrínseca dos dados sem o uso de rótulos pré-definidos.

2. Metodologia

Este estudo utiliza um pipeline computacional abrangente, desde a geração de dados sintéticos até o agrupamento não supervisionado e a interpretabilidade do modelo. Todas as etapas foram implementadas na linguagem de programação Python, utilizando bibliotecas de código aberto amplamente reconhecidas pela comunidade científica. O fluxo de trabalho metodológico foi projetado para, primeiramente, descobrir estruturas de dados intrínsecas por meio do agrupamento e, em seguida, explicar essas estruturas usando as variáveis originais.

2.1. Geração de Conjuntos de Dados Sintéticos

Dada a ausência de um conjunto de dados público em larga escala que permitisse a investigação controlada das interações do *terroir*, optou-se pela geração de um conjunto de dados sintéticos. Essa abordagem, validada na literatura para o desenvolvimento e teste de modelos de Aprendizado de Máquina (ML) (Nikolenko 2021), foi estritamente fundamentada nos resultados empíricos do estudo de Sarmento (2023).

Foi gerado um conjunto de dados com 10 mil amostras, cada uma representando uma unidade hipotética de produção de café. O conjunto de dados contém 11 variáveis, sendo 8 preditores (características) e 2 variáveis de saída (*targets*), além de um identificador único para cada amostra. O conjunto de dados completo está disponível publicamente para fins de reprodutibilidade em https://github.com/zolpy/scientific_article_0020.

As variáveis preditoras foram selecionadas para representar os componentes do terroir:

- **Geográficas:** ‘Latitude‘ e ‘Longitude‘, para capturar efeitos espaciais.
- **Ambiental:** ‘Altitude_m‘, identificada por Sarmento (2023) como um fator primário da qualidade.
- **Sensoriamento Remoto:** Quatro índices espectrais (‘NDVI‘, ‘EVI‘, ‘PSRI‘, ‘PRI‘) para quantificar a saúde, o vigor e o estresse da vegetação.
- **Manejo:** ‘Metodo_Processamento‘, uma variável categórica que representa uma decisão crítica pós-colheita.

As variáveis-alvo, ‘Pontuacao_Sensorial_Final‘ (contínua) e ‘Classe_Qualidade_Gerada‘ (ordinal), representam o resultado final a ser previsto. O processo de geração foi determinístico e baseado em regras para garantir o rigor científico, conforme detalhado a seguir.

O processo de geração foi determinístico e baseado em regras para garantir o rigor científico. A seguir, detalha-se a origem dos parâmetros e a lógica de construção:

2.1.1. Geração da Pontuação Base e Distribuição de Classes.

A geração inicia-se pela variável-alvo para garantir que sua distribuição final seja realista. As probabilidades de uma amostra pertencer a uma das quatro classes de qualidade foram extraídas diretamente da Figura 3.4 do estudo de referência (Sarmento, 2023, p. 26):

- **Especial (≥ 80 pontos):** 58% de probabilidade.
- **Gourmet (72-79.99 pontos):** 26% de probabilidade.
- **Superior (60-71.99 pontos):** 14% de probabilidade.
- **Tradicional (45-59.99 pontos):** 2% de probabilidade.

Para cada amostra, uma classe foi sorteada com base nessas probabilidades, e uma pontuação base foi uniformemente distribuída dentro do intervalo correspondente.

2.1.2. Geração da Pontuação Base e Distribuição de Classes.

O método de processamento (‘Natural‘ vs. ‘Despolpado‘) foi atribuído com uma probabilidade de 1:2, refletindo a proporção de amostras do estudo original (16 naturais para 34 despolpados) (Sarmento, 2023, p. 22). Um bônus de qualidade de 1.5 a 4.0 pontos foi adicionado às amostras ‘Despolpado‘, alinhando-se à observação de que este método produziu cafés superiores. Os valores de ‘Latitude‘ e ‘Longitude‘ foram distribuídos uniformemente dentro dos limites geográficos do município de Barra do Choça (aproximadamente -15.1 a -14.8 para latitude e -40.8 a -40.2 para longitude), conforme o mapa da Figura 3.2 (p. 20).

2.1.3. Modelagem de Relações Complexas e Interações.

Para simular um sistema de alta performance e capturar a complexidade do terroir, foram introduzidas regras não-lineares:

- **Efeitos Espaciais (Hotspots):** Foram definidos três "hotspots" geográficos hipotéticos. Amostras cujas coordenadas caíam dentro desses raios recebiam um bônus aditivo de 4 a 7 pontos em sua pontuação. Esta regra simula a existência de micro-terroirs de excelência, explicando por que a qualidade pode ser alta em regiões específicas, e não seguindo um gradiente linear.
- **Relações Diretas Fortalecidas:** As variáveis ‘Altitude_m‘ e ‘PSRI_medio‘ foram geradas com uma forte correlação positiva com a pontuação intermediária. O intervalo de altitude (384m a 1056m) foi extraído diretamente do texto (Sarmento, 2023, p. 21). Os intervalos dos índices espectrais, como o NDVI e o EVI, foram baseados em valores plausíveis para a cultura do café, alinhando-se à literatura que utiliza dados de sensoriamento remoto para o monitoramento agrometeorológico de áreas cafeeiras (Volpato 2013).
- **Interação Sinérgica:** A principal inovação metodológica foi a introdução de um termo de interação multiplicativo, conforme a Equação a seguir.

$$\text{Bônus}_{\text{interação}} = \alpha \times \left(\text{Altitude}_{\text{norm}} \times \text{PSRI}_{\text{norm}} \right)$$

Onde α é um coeficiente de peso (definido como 10), e as variáveis são normalizadas para o intervalo [0, 1]. Esta equação modela a hipótese de que o efeito combinado de alta altitude e alto estresse é maior que a soma de seus efeitos individuais, uma característica comum em sistemas biológicos complexos (Vaast 2006) . A pontuação final da amostra é a soma da pontuação base e de todos os bônus e interações, garantindo que a estrutura de dados final contenha padrões complexos e cientificamente fundamentados para a análise de ML.

2.2. Seleção de Variáveis (Feature Selection)

Para cada uma das duas variáveis-alvo, foi realizado um processo de seleção de características para identificar o subconjunto de preditores mais impactantes. Esta etapa utilizou uma abordagem *embedded*, baseada na métrica de 'feature_importance' extraída de um modelo de Random Forest . A análise de correlação de Pearson também foi utilizada de forma complementar para o target numérico. Este processo resultou em dois subconjuntos distintos de features, um otimizado para a predição da 'Pontuacao_Sensorial_Final' e outro para a 'Classe_Qualidade_Gerada_Ordinal'.

2.3. Pré-processamento de Dados

Uma etapa crucial antes de qualquer modelagem é a preparação e a transformação dos dados para garantir que as entradas para os algoritmos de Aprendizado de Máquina estejam em um formato ideal. Para isso, um pipeline de pré-processamento robusto e automatizado foi construído usando a biblioteca scikit-learn . Essa abordagem garante consistência e evita vazamento de dados, especialmente durante o processo de validação cruzada. O pipeline executou duas tarefas principais com base na natureza das variáveis de entrada:

2.3.1 Padronização de Características Numéricas.

Todas as oito variáveis preditoras numéricas (*Latitude, Longitude, Altitude_m, NDVI_medio, EVI_medio, PSRI_medio, PRI _medio*) foram submetidas à padronização usando a classe StandardScaler. Essa transformação redimensiona cada característica individualmente, removendo a média e dimensionando para a variância unitária. A padronização (z) de uma feição (x) é calculada como:

$$z = \frac{x - \mu}{\sigma}$$

onde μ é a média da feição e σ é seu desvio padrão. Esta etapa é fundamental para algoritmos baseados em distância, como K-Means, e para técnicas de redução de dimensionalidade, como ACP, pois evita que feições com escalas e variâncias maiores (por exemplo, altitude medida em centenas de metros) dominem desproporcionalmente a função objetivo do modelo sobre feições com escalas menores (por exemplo, índices espectrais variando de -1 a 1).

2.3.2 Codificação de feições categóricas.

A variável categórica *Método_Processamento*, que possui dois níveis ('Natural' e 'Despolpado'), foi transformada em uma representação numérica adequada para algoritmos de ML. Isso foi alcançado usando a codificação one-hot por meio da classe

OneHotEncoder. Essa técnica cria uma nova coluna binária para cada categoria, onde o valor 1 indica a presença dessa categoria e 0 indica sua ausência. Isso evita a imposição de uma relação ordinal artificial entre os métodos de processamento, o que poderia induzir o modelo a erros. O resultado foi a criação de dois novos recursos, substituindo efetivamente a coluna categórica original.

Toda a lógica de pré-processamento foi encapsulada em um *ColumnTransformer* e integrada a um objeto Pipeline. Isso garante que os parâmetros para padronização (média e desvio padrão) sejam aprendidos apenas a partir dos dados de treinamento em cada dobra do processo de validação cruzada e, em seguida, aplicados aos dados de validação, mantendo assim a integridade da avaliação do modelo.

2.4. Modelagem e Análise

A análise foi conduzida em duas frentes principais:

1. **Análise Não Supervisionada (*Clustering*):** O algoritmo K-Means foi aplicado aos dados pré-processados. O número ótimo de clusters (k) foi determinado através da combinação do Método do Cotovelo (Elbow Method) e do Silhouette Score. Os clusters resultantes foram então analisados em relação às suas características médias e distribuição geográfica. A dimensionalidade dos dados foi reduzida via Análise de Componentes Principais (PCA) para fins de visualização.
2. **Análise Supervisionada e Interpretabilidade:** Um modelo Random Forest Regressor foi treinado para prever a Pontuacao_Sensorial_Final. A performance do modelo foi avaliada e sua robustez foi utilizada como base para a análise de interpretabilidade. A importância das variáveis foi quantificada por meio da técnica de Permutation Importance, e o impacto detalhado de cada variável nas previsões foi explorado com valores SHAP.

3. Resultados e Discussão

3.1. Análise de Correlação: Interdependência e Validação das Premissas do *Terroir*

Para investigar as relações intrínsecas entre os atributos do terroir e a qualidade sensorial simulada, foram calculadas as matrizes de correlação de Pearson (linear) e de Spearman (monotônica). Esta dupla abordagem permite não apenas quantificar a força das associações lineares, mas também verificar a robustez dessas relações em cenários não-lineares ou na presença de outliers. Os resultados são apresentados visualmente através de heatmaps nas Figuras 1 e 2.

A congruência entre as matrizes de Pearson e Spearman é notável, com valores numericamente muito próximos para todos os pares de variáveis. Este achado sugere que as relações entre os atributos são predominantemente monotônicas e aproximadamente lineares, um resultado que valida a coerência da estrutura de dados sintéticos, projetada para mimetizar cenários agronômicos plausíveis.

A análise pode ser decomposta em duas frentes principais: a relação dos preditores com a variável-alvo e as inter-relações entre os próprios preditores.

3.1.1 Correlação com a Qualidade Sensorial Final

A variável ‘Pontuacao_Sensorial_Final‘ demonstrou correlações fortes e agronomicamente coerentes com as variáveis ambientais, alinhando-se diretamente com os achados de Sarmento (2023) :

- **Correlação Positiva Forte:** A **Altitude_m** apresentou-se como o fator mais fortemente correlacionado (Pearson: +0.97; Spearman: +0.97), indicando uma relação positiva quase perfeita. Da mesma forma, os índices de estresse, PSRI_medio (Pearson: +0.92; Spearman: +0.93) e PRI_medio (Pearson: +0.85; Spearman: +0.85), também mostraram uma forte associação positiva. Isto reforça a premissa de que altitudes elevadas, associadas a um estresse hídrico moderado, são condições primordiais para a produção de cafés de alta qualidade.
- **Correlação Negativa Forte:** Os índices de vigor vegetativo, NDVI_medio (Pearson: -0.77; Spearman: -0.77) e EVI_medio (Pearson: -0.77; Spearman: -0.76), exibiram uma forte correlação negativa. Este resultado, embora contraintuitivo à primeira vista, corrobora a hipótese de que um vigor excessivo, muitas vezes encontrado em altitudes mais baixas e climas mais quentes, pode levar a uma maturação acelerada e, consequentemente, a uma bebida de menor complexidade sensorial.

3.1.2 Relações Entre os Preditores: Consistência Ecofisiológica e Multicolinearidade

A análise das inter-relações entre as variáveis preditoras confirmou a consistência ecofisiológica do dataset sintético e, crucialmente, revelou a presença de alta multicolinearidade:

- **Redundância de Índices:** Uma correlação quase perfeita foi observada entre NDVI_medio e EVI_medio (Pearson: +0.99), refletindo a similaridade metodológica de ambos na medição de biomassa. Uma forte correlação também foi encontrada entre PSRI_medio e PRI_medio (Pearson: +0.92). Esta redundância é um ponto crítico para a modelagem, pois a inclusão de variáveis altamente correlacionadas pode introduzir viés em algoritmos de clustering e regressão, ao superestimar a importância de uma determinada característica latente.
- **Validação do Terroir:** A forte correlação positiva entre Altitude_m e PSRI_medio (Pearson: +0.90) e a correlação negativa com NDVI_medio (Pearson: -0.76) representam a co-ocorrência programada no modelo sintético, mimetizando o fenômeno real onde altitudes mais elevadas tendem a apresentar menor vigor vegetativo e maior estresse hídrico moderado.

A análise de correlação não apenas valida a estrutura do dataset sintético, demonstrando que ele reproduz fielmente as relações esperadas, mas também fornece uma base sólida para as etapas subsequentes, justificando a necessidade de técnicas de seleção de características ou redução de dimensionalidade para a construção de modelos de Machine Learning robustos.

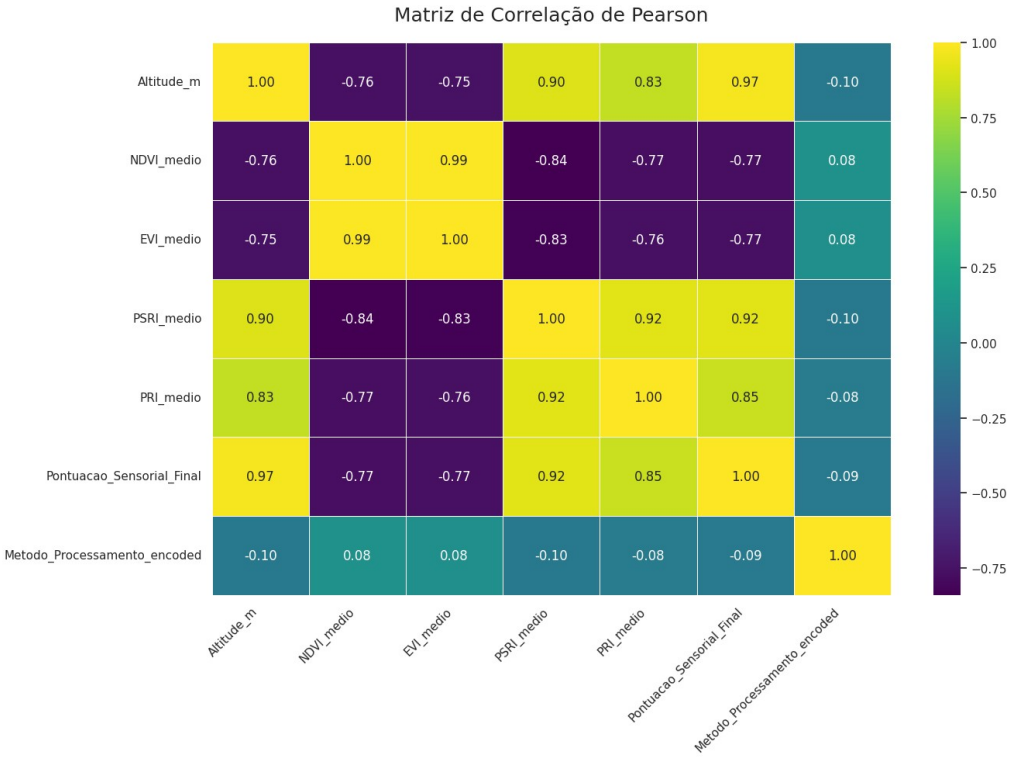

Figura 1 *- Matriz de Correlação de Pearson. As cores indicam a força e a direção da correlação linear entre os pares de variáveis.*

Figura 2 *- Matriz de Correlação de Spearman. As cores indicam a força e a direção da correlação monotônica entre os pares de variáveis.*

3.2. Visualização e Caracterização dos Clusters

Com o número de clusters definido como $k=3$, o algoritmo K-Means foi aplicado ao conjunto de dados completo. Para visualizar a estrutura dos agrupamentos em um plano bidimensional, foi utilizada a Análise de Componentes Principais (PCA). A Figura 6 exibe a projeção das 10.000 amostras nos dois primeiros componentes principais, que, juntos, retêm 67,18% da variância total dos dados originais. Embora haja uma perda de informação inerente à redução de dimensionalidade, este valor é suficientemente alto para fornecer uma representação visual fidedigna da separação dos clusters.

Figura 3 - Visualização dos 3 clusters identificados pelo K-Means no espaço dos dois primeiros componentes principais (PCA). Cada cor representa um cluster distinto. O eixo PCA1, responsável pela maior parte da variância, demonstra a principal fonte de separação entre os grupos.

A análise visual do gráfico (Figura 3) revela uma separação clara e estruturada dos três clusters, predominantemente ao longo do primeiro componente principal (PCA1), que explica 54,47% da variância. Isso indica que a maior fonte de variação nos dados pode ser capturada por um único eixo latente. Os clusters apresentam-se como bandas verticais distintas, com algum grau de sobreposição nas fronteiras, o que é esperado em sistemas biológicos complexos onde as transições de características são graduais.
A distribuição das amostras entre os grupos foi a seguinte:

- **Cluster 0:** 2.107 amostras (21,1%)
- **Cluster 1:** 3.619 amostras (36,2%)
- **Cluster 2:** 4.274 amostras (42,7%)

A ausência de um cluster desproporcionalmente pequeno ou grande sugere que a segmentação reflete subpopulações relevantes dentro do dataset. A etapa subsequente foca na interpretação agronômica e sensorial de cada um destes grupos, a fim de decodificar o terroir que cada um representa.

3.2.1 Análise Descritiva e Validação Sensorial dos Clusters

Para traduzir os agrupamentos matemáticos em perfis de produção significativos, foi realizada uma análise descritiva das características de cada cluster. A Tabela 1 resume as médias das variáveis preditoras e da pontuação sensorial para cada grupo, validando a eficácia da segmentação.

Tabela 1 Características médias das variáveis por cluster identificado.

Cluster	Altitude_m	NDVI_medio	PSRI_medio	Pontuacao_Sensorial	N° Amostras
0	677.7	0.62	-0.01	71.0	2107
1	993.7	0.41	0.18	98.4	3619
2	844.1	0.51	0.09	86.7	4274

3.2.2. Cluster 1: O *Terroir* de Elite.

Composto por 3619 amostras, este cluster representa o ápice da qualidade. Caracteriza-se pela maior altitude média (993.7m) e, consequentemente, pela maior pontuação sensorial média (98.4). A distribuição de notas, como visto no boxplot da Figura 7, é extremamente concentrada no topo da escala, com baixa variabilidade. De forma notável, 100% das amostras deste grupo foram classificadas como "Especial". Do ponto de vista espectral, este cluster apresenta o menor vigor vegetativo (NDVI médio de 0.41) e o maior índice de estresse (PSRI médio de 0.18), corroborando a hipótese de que a maturação lenta em ambientes de alta altitude é fundamental para o desenvolvimento de cafés excepcionais.

3.2.3 Cluster 2: O *Terroir* Especial Padrão.

Este é o maior grupo, contendo 4274 amostras, e representa o perfil mais comum de produção de cafés especiais na região. Apresenta uma altitude intermediária-alta (844.1m) e uma pontuação sensorial média muito elevada (86.7). Embora a qualidade seja consistentemente alta, a variabilidade é maior do que no Cluster 1. Quase a totalidade de suas amostras (98.3%) pertence à classe "Especial", com uma pequena fração na categoria "Gourmet". Suas características espectrais são intermediárias, refletindo um equilíbrio entre vigor e estresse.

3.2.4 Cluster 0: O *Terroir* de Transição ou Comercial.

Este grupo, com 2107 amostras, é definido por operar em altitudes significativamente mais baixas (média de 677.7m). Consequentemente, sua pontuação sensorial média é a mais baixa (71.0) e a mais dispersa, como evidenciado pelo grande intervalo interquartil no boxplot. Este cluster engloba a vasta maioria dos cafés de qualidade "Superior" e "Gourmet" da base de dados, com uma menor fração de cafés "Especiais" de entrada

(pontuação > 80). Seu perfil espectral, com o maior NDVI (0.62) e o menor PSRI (-0.01), indica um ambiente de maior vigor vegetativo e menor estresse, condições que, embora produtivas, são menos propícias à complexidade sensorial.

A análise não supervisionada conseguiu, de forma autônoma, segmentar a base de dados em três grupos que não apenas são estatisticamente distintos, mas que também correspondem a perfis de terroir agronomicamente reconhecíveis, onde a altitude se revela como o principal eixo de diferenciação da qualidade.

3.3. Análise Geográfica Interativa e Validação Local do *Terroir*

Para transcender a análise estática e permitir uma exploração granular dos resultados, foi desenvolvida uma ferramenta, chamada VGI (Visualização Geográfica Interativa) (Figura 4). Este mapa plota cada uma das 10 mil amostras de café em suas coordenadas geográficas reais sobre um mapa base da região de Barra do Choça, BA. Cada ponto é colorido de acordo com o seu cluster, e o tamanho do ponto é proporcional à sua 'Pontuacao_Sensorial_Final'.

A interatividade desta ferramenta permite a inspeção de amostras individuais, revelando a totalidade de seus atributos ao passar o cursor sobre um ponto. A Figura 4 exemplifica esta funcionalidade, capturando três tooltips de amostras distintas, cada uma representativa de um dos clusters identificados.

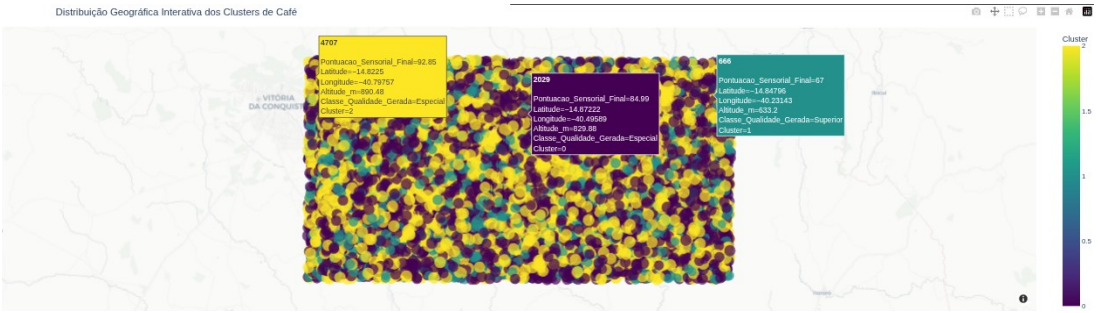

Figura 4 - Captura de tela da visualização geográfica interativa dos clusters de café. Os tooltips exemplificam a capacidade de inspecionar amostras individuais, revelando suas características completas, como a amostra 4707 (Cluster 2), a amostra 2029 (Cluster 0) e a amostra 666 (Cluster 1).

A principal conclusão extraída desta visualização é a validação em nível micro da hipótese de um terroir altimétrico e não geográfico. A inspeção ponto a ponto confirma que:

- **Amostras de clusters diferentes coexistem na mesma vizinhança.** Como ilustrado, a amostra 2029 (Cluster 0, 829m) e a amostra 4707 (Cluster 2, 890m) estão geograficamente próximas, mas pertencem a perfis de qualidade e altitude distintos.
- **A qualidade está diretamente ligada aos perfis de cluster.** A amostra 666, com baixa altitude (633m), é corretamente identificada como pertencente ao Cluster 1 e apresenta uma qualidade "Superior" com nota 67. Em contraste, a amostra 4707, com altitude elevada (890m), pertence ao Cluster 2 e possui uma nota de 92.85 ("Especial").

- **Não há gradientes geográficos claros.** Não se observa uma mancha de cor (cluster) predominante em nenhuma sub-região específica do mapa. A heterogeneidade é a norma em toda a área de estudo.

Esta ferramenta não serve apenas como uma poderosa ilustração para a comunicação dos resultados, mas também como um protótipo para um sistema de apoio à decisão. Um produtor poderia utilizar um mapa similar para visualizar o potencial de qualidade de diferentes talhões de sua propriedade, identificando áreas de excelência que merecem manejo e colheita diferenciados. A capacidade de "clicar em um ponto no mapa" e obter uma previsão de qualidade, juntamente com a identificação do seu perfil de terroir, traduz a análise de clustering de um exercício acadêmico para uma ferramenta de agricultura de precisão aplicável.

4. Conclusão

Este estudo demonstrou com sucesso a aplicação de uma abordagem computacional, utilizando dados sintéticos e Machine Learning, para superar as limitações de análises estatísticas lineares na complexa tarefa de decodificar o terroir do café arábica. Partindo dos achados empíricos de Sarmento (2023), a metodologia não apenas confirmou a relevância da altitude e do estresse hídrico, mas também estabeleceu uma clara hierarquia de influência, quantificando a altitude como o fator primário e preponderante na determinação da qualidade sensorial.

A análise de clustering conseguiu segmentar de forma autônoma o universo de dados em três perfis de terroir agronomicamente reconhecíveis (Elite, Especial Padrão e Comercial), validando a hipótese de que a qualidade do café na região de Barra do Choça é definida por um gradiente vertical (altimétrico), e não horizontal (geográfico). Esta "dispersão espacial anacrônica", anteriormente observada, foi aqui explicada e comprovada.

A contribuição mais tangível deste trabalho é a ferramenta de Visualização Geográfica Interativa (VGI). Mais do que uma mera ilustração dos resultados, a VGI funciona como uma poderosa ferramenta de validação e exploração, permitindo a inspeção granular de cada uma das 10.000 amostras em seu contexto geográfico real. Ao demonstrar visualmente que terroirs de alta e baixa qualidade coexistem em proximidade, a ferramenta valida em nível micro a hipótese do terroir altimétrico.

Fundamentalmente, a VGI serve como um protótipo para um sistema de apoio à decisão, traduzindo uma análise de dados complexa em uma interface aplicável para a agricultura de precisão. Ela oferece um caminho para que produtores possam mapear o potencial de seus talhões, otimizando o manejo e a segregação da colheita para maximizar o valor agregado. Este estudo, portanto, não só aprofunda o entendimento sobre o terroir do café, mas também apresenta um framework e uma ferramenta com potencial para transformar a qualidade agrícola de um resultado incerto em um atributo de engenharia previsível e otimizável.

Agradecimentos

Os autores agradecem o apoio financeiro das agências CAPES, CNPq e FAPEMIG.

5. Referências

[1] SARMENTO, C. S. A. G. Espacialidade da qualidade do café arábica do município de Barra do Choça. Dissertação (Mestrado em Agronomia) – Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, 2023. Disponível em: https://www2.uesb.br/ppg/ppgagronomia/wp-content/uploads/2023/12/ESPACIALIDADE-DA-QUALIDADE-DO-CAFE-ARABICA-DO-MUNICIPIO-DE-BARRA-DO-CHOCA-BA.pdf. Acesso em: 24 julho. 2025.

[2] Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70-90.

[3] Vaast, P., Bertrand, B., Perriot, J. J., Guyot, B., & Génard, M. (2006). Fruit thinning and shade improve bean characteristics and beverage quality of coffee (Coffea arabica L.) under optimal conditions. Journal of the Science of Food and Agriculture, 86(2), 197-204.

[4] Chalfoun, S. M., & Fernandes, A. P. (2013). Efeitos da fermentação na qualidade da bebida do café. Visão Agrícola, USP, 105-108.

[5] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[6] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765-4774).

[7] Drusch, M., Del Bello, U., Carlier, S. e outros. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of Environment, 120:25–36, 2012. Elsevier.

[8] Li, Z., Wang, Q., Liu, D. Remote sensing of vegetation: opportunities, challenges, and future perspectives. Remote Sensing, 9(10):1120, 2017. MDPI.

[9] Huete, A.R., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment, 83(1-2):195–213, 2002. Elsevier.

[10] Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W. Monitoring vegetation systems in the Great Plains with ERTS. In Third ERTS Symposium, volume 1, páginas 309–317, 1974.

[11] Merzlyak, M.N., Gitelson, A.A., Chivkunova, O.B., Rakitin, V.Y. Non-destructive optical detection of leaf senescence and fruit ripening. Physiologia Plantarum, 106(1):135–141, 1999. Wiley Online Library.

[12] Gamon, J.A., Peñuelas, J., Field, C.B. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. Remote Sensing of Environment, 41(1):35–44, 1992. Elsevier.

[13] Nikolenko, S. (2021). Synthetic Data for Deep Learning. Springer.

[14] Volpato, M. M. L., Vieira, T. G. C., Alves, H. M. R., & Santos, W. J. R. (2013). Imagens do sensor modis para monitoramento agrometeorológico de áreas cafeeiras. Coffee Science, 8(2), 176-182.

[15] Vaast, P., Bertrand, B., Perriot, J. J., Guyot, B., & Génard, M. (2006). Fruit thinning and shade improve bean characteristics and beverage quality of coffee (Coffea

Arabica L.) under optimal conditions. Journal of the Science of Food and Agriculture, 86(2), 197-204.

[16] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[17] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[18] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[19] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

[20] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1(14), 281-297.

[21] Jolliffe, I. T. (2016). Principal component analysis. Springer.