# Machine Learning for Geographical Indications: A Scoping Review on Authentication, Certification and Open Data Ecosystems

Catuxe Varjão de Santana Oliveira, Paulo Roberto Gagliardi, Luiz Diego Vidal Santos, Gus

## Resumo

As Indicações Geográficas (IGs) são ativos estratégicos que vinculam a identidade territorial e o valor do produto em economias do conhecimento. No entanto, a autenticação e a certificação de produtos com IG dependem de metodologias robustas, reproduzíveis e transparentes. Embora a Aprendizagem de Máquina (AM) tenha emergido como uma ferramenta central nesse domínio, ainda não existe uma síntese sistemática de suas aplicações, rigor metodológico e limitações. Esta revisão de escopo mapeia 148 estudos revisados por pares (2010–2025) das bases de dados Scopus e Web of Science, empregando filtragem semântica automatizada (94,2% de precisão), avaliação rigorosa da qualidade (ICC = 0,87) e análise estatística multivariada. Identificamos um crescimento exponencial (aumento superior a 400% desde 2018) nas aplicações de AM para autenticação de IG, organizadas em três módulos tecnológicos estáveis que combinam algoritmos (Random Forest, SVM, Redes Neurais), técnicas analíticas (NIR, LC-MS, metabolômica) e produtos (vinhos, chás, carnes). De forma crítica, as precisões de 80–100% na autenticação contrastam fortemente com a baixa generalização: apenas 23% dos estudos empregaram validação espacialmente independente, e a validação externa mostrou uma degradação de desempenho de 2–15%. Foi identificado de um viés de otimismo estatístico em 77% do corpus, no qual o sucesso da validação in silico obscurece a fragilidade preditiva no mundo real. Documentamos ainda um paradoxo que, embora o aprendizado de máquina alcance uma discriminação de origem quase perfeita, a capacidade preditiva para índices gerais de qualidade permanece modesta ($R^2$ = 0,11–0,14). Essas descobertas têm implicações diretas para a conformidade regulatória, a governança da propriedade intelectual e a justiça territorial em sistemas de certificação. É destacada a necessidade de protocolos de validação temporal, implementações de modelos transparentes e estruturas de governança equitativas que integrem aprendizado de máquina em arquiteturas de certificação de informações geográficas. Esta revisão contribui para os métodos computacionais em ciência da informação geográfica e para as práticas de ciência aberta, fornecendo evidências meta-analíticas transparentes e reproduzíveis sobre a viabilidade e as limitações

do aprendizado de máquina para autenticação territorial.

**Palavras-chave:** Geographical Indications; Machine Learning; Authentication; Certification; Artificial Intelligence; Food Traceability; Agroecological Products; Intellectual Property; Territorial Justice; Scoping Review.

# 1. Introdução

As Indicações Geográficas (IGs) protegem territórios e produtos na economia do conhecimento, assegurando direitos exclusivos sobre produtos cuja qualidade, reputação e características derivam de sua origem geográfica (Locatelli, 2008; WIPO, 2018). Fundamentadas na Convenção de Berna (*Convenção de Berna para a Proteção das Obras Literárias e Artísticas*, 1886) e no Acordo TRIPS (*Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS)*, 1994), as IGs conectam territórios produtivos e comunidades locais a mercados diferenciados, vinculando proteção de direitos à preservação do conhecimento tradicional (Suh & Macpherson, 2007). Além de instrumento jurídico, as IGs constituem ativo intangível estratégico no sentido proposto pela teoria da Visão Baseada em Recursos (Resource-Based View), sendo recursos raros, valiosos, inimitáveis e insubstituíveis que fundamentam vantagem competitiva territorial sustentável (Barney, 1991). A apropriação de valor por meio das IGs transcende rentabilidade imediata, servindo como ancoragem para captura de valor futuro através de mercados diferenciados e disposição de consumidores em pagar preços premium por produtos com certificação de origem (Loureiro & McCluskey, 2002; Vázquez-Fontes et al., 2010).

No Brasil, as Indicações Geográficas são regulamentadas pela Lei da Propriedade Industrial, Lei nº 9.279 de 14 de maio de 1996, que estabelece dois tipos de reconhecimento com implicações jurídicas e econômicas distintas, Indicação de Procedência e Denominação de Origem (Brasil, 1996). A Indicação de Procedência refere-se ao nome geográfico conhecido pela produção ou fabricação de determinado produto, funcionando como mecanismo de sinalização de origem; a Denominação de Origem designa produtos cujas qualidades ou características se devem exclusiva ou essencialmente ao meio geográfico, incluindo fatores naturais e humanos, constituindo forma de proteção mais robusta que vincula qualidade ao terroir (MAPA, 2020).

O controle deste tipo de registro é realizado pelo Instituto Nacional de Propriedade Intelectual (INPI), com apoio do Ministério da Agricultura, Pecuária e Abastecimento, que operacionaliza políticas de fomento e certificação de produtos agrícolas com identidade territorial (MAPA, 2020). Este marco regulatório brasileiro alinha-se à Lei nº 10.973/2004 (Lei de Inovação) e Lei nº 13.243/2016 (Novo Marco Legal de CT&I), que reconhecem Indicações Geográficas como ativos de propriedade intelectual passíveis de proteção estratégica, valoração e comercialização (Brasil, 2004, 2016).

No contexto brasileiro, produtos artesanais e agroalimentares com potencial

para registro de Indicação Geográfica representam manifestações culturais relevantes e oportunidades estratégicas para captura de valor territorial. Estudos demonstram que características únicas de produtos regionais, como a cerâmica artesanal do Baixo São Francisco ou produtos vinícolas especializados, estão intimamente relacionadas a atributos geográficos da localização de produção, incluindo características edafoclimáticas (solo, clima, altitude) e métodos únicos de cultivo ou produção (al., 2011; Bureau & Freitas, 2018; Fonzo & Russo, 2015; H. G. Santos et al., 2018; J. C. Santos & Santos, 2019). A caracterização territorial desses produtos, necessária para o reconhecimento como Denominação de Origem conforme estabelecido no artigo 178 da Lei nº 9.279/1996 (Brasil, 1996), demanda análises técnicas específicas que comprovem, com rigor científico, a relação entre qualidade e fatores geográficos (Gonçalves-Maduro et al., 2020). Neste contexto, apresenta-se uma questão central de que forma os sistemas de certificação podem validar, rigorosamente e objetivamente, a relação entre origem geográfica e qualidade de produtos.

As tecnologias de Aprendizado de Máquina (ML) emergem como resposta estratégica a essa lacuna, transmutando dados analíticos complexos em conhecimento certificável sobre autenticidade e origem. Diferentemente dos métodos de análise sensorial tradicionais, dependentes de expertise humana tácita e limitados pela subjetividade e escalabilidade, os algoritmos de ML operam sob uma lógica indutiva ou abdutiva. Eles processam automaticamente dados multidimensionais, identificando padrões não lineares e relações latentes que escapam à modelagem estatística clássica baseada em testes de hipóteses dedutivos. Essa capacidade de construir modelos com formas funcionais flexíveis permite revelar estruturas de dados não previamente especificadas pela teoria, conferindo robustez matemática à certificação territorial (R. C. Chen et al., 2020; Ramos et al., 2025).

No âmbito das Indicações Geográficas, o Machine Learning tem sido mobilizado para autenticação de origem, detecção de fraudes, controle preditivo de qualidade e rastreabilidade integral, operando sobre assinaturas químicas, isotópicas, espectrais e geoespaciais que capturam a relação entre produto e território (Acquarelli et al., 2021; Longo et al., 2021; Rana et al., 2023; Rodrigues et al., 2022). Essas aplicações demonstram que a integração entre dados instrumentais de alta dimensionalidade e modelos supervisionados permite discriminar origens, identificar adulterações e estimar atributos sensoriais e físico-químicos com precisão compatível com requisitos de certificação (Jiang et al., 2025; Li et al., 2025; Peng et al., 2025; Santomá-Martí et al., 2025; Wang et al., 2025).

A seleção de variáveis e a escolha de algoritmos, em particular Random Forest, SVM, PLS-DA, PCA e métodos de seleção como Boruta e RFE, deixam de ser apenas decisões técnicas para se tornar componente da arquitetura regulatória, pois definem quais marcadores territoriais serão reconhecidos como evidências de autenticidade (R. C. Chen et al., 2020; Effrosynidis & Arampatzis, 2021; Iranzad & Liu, 2025; Loyal et al., 2022; Malik et al., 2023; Rebiai et al., 2022; Salam et al., 2021).

Apesar do interesse acadêmico e tecnológico atual, não existem revisões de es-

copo que sistematizem as evidências científicas disponíveis, identifiquem as técnicas empregadas, avaliem seu desempenho em diferentes produtos e contextos geográficos, ou apontem direções para pesquisas futuras. Esta limitação afeta o desenvolvimento metodológico na área e a transferência de conhecimento para sistemas de certificação e controle de Indicações Geográficas.

Esta revisão de escopo busca mapear sistematicamente as aplicações de Machine Learning em Indicações Geográficas, utilizando o framework PCC (*Population, Concept, Context*) para identificar e sintetizar evidências científicas sobre a integração entre Machine Learning e aspectos territoriais de Indicações Geográficas. Hipotiza-se que as técnicas de Aprendizado de Máquina têm sido empregadas para apoiar processos de autenticação, avaliação e tomada de decisão relacionados às Indicações Geográficas, revelando padrões metodológicos que contribuem para a consolidação de conhecimento orientado ao desenvolvimento de modelos computacionais aplicados à certificação geográfica.

# 2. Materiais e Métodos

Esta revisão de escopo segue as diretrizes da extensão PRISMA-ScR (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews*) para garantir transparência e reprodutibilidade metodológica. O protocolo foi registrado no Open Science Framework, facilitando o acesso público e a replicabilidade.

## 2.1 Questão de Pesquisa

O estudo foi estruturado utilizando o framework PCC (*Population, Concept, Context*), que fundamenta a questão de pesquisa: *Como técnicas de Aprendizado de Máquina têm sido aplicadas para autenticação, avaliação e apoio à decisão em sistemas de Indicações Geográficas?*

| Elemento | Descrição |
|---|---|
| **P (População)** | Indicações Geográficas, Denominações de Origem e Indicações de Procedência reconhecidas nacional e internacionalmente, abrangendo produtos agroalimentares (vinhos, queijos, cafés, carnes, azeites), artesanatos e outros produtos com identidade territorial. |
| **C (Conceito)** | Técnicas de Aprendizado de Máquina, Inteligência Artificial, algoritmos de classificação e predição, métodos quimiométricos, Mineração de Dados e Processamento de Linguagem Natural aplicados a contextos de Indicações Geográficas. |

| Elemento | Descrição |
|---|---|
| **C (Contexto)** | Autenticação de origem geográfica, avaliação de potencialidade de IGs, identificação de determinantes territoriais (solo, clima, métodos de produção), classificação e discriminação de produtos, sistemas de apoio à decisão para certificação, controle de qualidade, rastreabilidade, detecção de fraudes e adulterações, e estratégias de valorização territorial. |

*Tabela 1. Estrutura da revisão de escopo segundo o framework PCC.*

Este trabalho se propõe a identificar e caracterizar as aplicações de ML reportadas na literatura científica, categorizando as técnicas empregadas segundo tipo de algoritmo, abordagem metodológica e métricas de desempenho. Ainda, analisar a distribuição das aplicações por tipo de produto, região geográfica e período temporal. E por fim, identificar lacunas metodológicas, limitações e direções para pesquisas futuras.

### 2.1.1 Fluxograma Metodológico PRISMA-ScR

A Figura 1 apresenta o fluxograma metodológico da revisão de escopo, estruturado em quatro fases sequenciais segundo as diretrizes PRISMA-ScR: (1) Estratégia de Busca nas principais bases de dados, (2) Filtragem Automatizada com sistema de pontuação ponderada, (3) Análise Manual de Qualidade com avaliação multidisciplinar, e (4) Análise Bibliométrica e Síntese Qualitativa** integrando metodologias quantitativas e documentais. O fluxograma detalha o percurso metodológico desde a identificação de registros até a síntese final com recomendações para implementação de Machine Learning em sistemas de Indicações Geográficas.

## 2.3 Estratégia de Busca e Extração dos Estudos

A busca foi realizada nas bases de dados científicas Scopus (Elsevier) e Web of Science (Clarivate Analytics). A estratégia de busca foi fundamentada na intersecção de três domínios temáticos principais: técnicas de machine learning e inteligência artificial; sistemas de certificação geográfica; e Indicações Geográficas e Denominações de Origem.

Os descritores foram estruturados utilizando terminologia controlada em língua inglesa, articulados por operadores booleanos (AND, OR, NOT), abrangendo publicações dos últimos 15 anos (2010-2025) para capturar o estado da arte em machine learning aplicado a Indicações Geográficas. A estratégia de busca foi construída seguindo a lógica:

*(“machine learning” OR “artificial intelligence” OR “deep learning” OR “supervised learning” OR “unsupervised learning” OR “ensemble methods”) AND (“geographical indications” OR “denominations of origin” OR “appellations of*

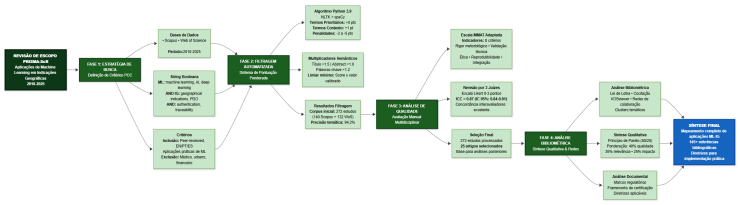


Figura 1: Fluxograma da revisão de Escopo sobre Aplicações de Machine Learning em Indicações Geográficas.

*origin" OR "protected designations of origin") AND ("authentication" OR "traceability" OR "quality control" OR "fraud detection" OR "geospatial analysis").*

Os critérios de inclusão contemplaram os artigos completos publicados em periódicos revisados por pares, escritos em inglês, português ou espanhol, que apresentassem aplicações de técnicas de ML em contextos de Indicações Geográficas, autenticação de origem ou controle de qualidade territorial. Os descritores primários deveriam estar presentes nos campos título, resumo ou palavras-chave dos manuscritos. Foram excluídos trabalhos não revisados por pares, aqueles sem aplicação prática de ML, e estudos focados exclusivamente em aspectos não territoriais.

## 2.4 Primeira Fase: Sistema de Filtragem Automatizada por Relevância Temática

### 2.4.1 Algoritmo de Pontuação Ponderada

Para complementar a triagem manual, foi desenvolvido um sistema de filtragem automatizada que atribui pontuações de relevância temática aos registros com base na presença e na localização de descritores específicos em título, resumo e palavras-chave. O algoritmo foi implementado em Python (NLTK, spaCy) e estruturado como um esquema de pesos hierárquicos aplicado a cada termo identificado. A pontuação de cada artigo foi calculada pela Equação 1.

$$S_i = \sum_{j=1}^{n} w_j \cdot l_i \cdot f_{ij}$$

onde:

- $S_i$ = pontuação total do artigo $i$
- $w_j$ = peso associado ao termo $j$ (categorizado em 5 níveis: 5, 3, 2, 1, ou -5/-3/-2 pontos)
- $l_i$ = multiplicador de localização do termo (1,5 para título, 1,2 para palavras-chave, 1,0 para abstract)
- $f_{ij}$ = frequência de ocorrência do termo $j$ no artigo $i$
- $n$ = número total de termos avaliados

O sistema de pontuação hierárquica seguiu princípios do Método de Análise Hierárquica (AHP), organizando os descritores em cinco categorias com pesos diferenciados (Saaty, 1991). Termos Prioritários (5 pontos) representam o núcleo conceitual da revisão e incluem expressões como *geographical indications, denominations of origin, appellations of origin, protected designations of origin, traceability, authentication, quality control*. Termos de Alta Relevância (3 pontos) capturam conceitos metodológicos centrais, como *machine learning, artificial intelligence, deep learning, neural networks, fraud detection, geospatial analysis*. Termos de Relevância Média (2 pontos) cobrem temas complementares, a exemplo de *chemometrics, data mining, classification*, enquanto Termos de Contexto (1 ponto) indicam ambientes ou aplicações potenciais, como *regional products, certification, prediction, validation*. Termos de Exclusão recebem pesos negativos e penalizam registros fora do escopo, sobretudo em domínios como *medical/clinical/pharmaceutical* ($-5$), *urban planning/smart cities* ($-3$) e *finance/economics/business* ($-2$) (Munn et al., 2018; Tricco et al., 2018).

### 2.4.2 Implementação e Validação do Sistema Automatizado

Para cada registro, o algoritmo varre título, resumo e palavras-chave, aplica os pesos definidos para cada categoria de termo (Prioritário = 5; Alta Relevância = 3; Relevância Média = 2; Contexto = 1; Exclusão = valores negativos) e multiplica cada ocorrência pelo fator de localização correspondente (1,5 para título, 1,2 para palavras-chave, 1,0 para resumo). A pontuação final é a soma desses produtos ao longo de todos os termos identificados em cada registro.

O limiar mínimo de inclusão foi definido a partir da distribuição empírica das pontuações, identificando-se o ponto de inflexão na curva acumulada (critério tipo Pareto/elbow) e ajustando-o por validação manual com amostragem estratificada de registros contendo termos-chave como *machine learning*, *geographical indications* e *authentication*. O valor final representou o melhor compromisso entre sensibilidade e especificidade, estabilizando a concordância interavaliadores nos casos limítrofes.

### 2.4.3 Validação Participativa e Refinamento Algorítmico

Para assegurar a validade científica do processo de seleção, foi implementado um protocolo de validação envolvendo três revisores independentes, especialistas em machine learning, sistemas de certificação geográfica e Indicações Geográficas. O

protocolo incluiu uma revisão manual sistemática, com análise criteriosa dos 272 estudos identificados nas bases Scopus (140) e Web of Science (132) para verificar a aderência aos critérios de inclusão e relevância temática. Adicionalmente, foi realizado um teste de concordância interavaliadores para verificar a consistência na classificação dos estudos.

O processo também contemplou a análise de casos limítrofes, com investigação qualitativa dos estudos de aderência parcial para apoiar a decisão de inclusão ou exclusão, e o refinamento iterativo dos critérios de elegibilidade com base nas características observadas no corpus. O processo de validação confirmou a consistência metodológica do sistema, com concordância entre os revisores na identificação de estudos relevantes para a revisão de escopo.

### 2.4.4 Verificação de Cobertura Bibliográfica e Categorização Automatizada

Foi desenvolvido sistema automatizado para verificar a cobertura bibliográfica das citações metodológicas. O procedimento avalia completude e consistência da base de referências, garantindo rastreabilidade entre citações textuais e arquivos bibliográficos.

O corpus bibliográfico consolidado foi submetido à categorização automatizada com técnicas de Processamento de Linguagem Natural (PLN). Os registros foram identificados e organizados segundo domínios metodológicos relevantes, aplicando abordagens de revisões sistemáticas automatizadas (Ofori-Boateng et al., 2024; Sawicki et al., 2023). Foi construído pipeline computacional que extrai, tokeniza e vetorializa metadados e resumos das referências, usando modelos supervisionados e regras semânticas para reconhecimento de padrões (Casey et al., 2021; Young et al., 2019). As referências foram classificadas em categorias metodológicas previamente definidas, abrangendo áreas como técnicas de aprendizado de máquina e sistemas de indicações geográficas.

Para quantificação da cobertura bibliográfica e adequação dos estudos selecionados, foram aplicadas duas métricas complementares (Tranfield et al., 2003; Webster & Watson, 2002). A primeira, de cobertura de citações, foi calculada pela Equação 2.

$$Cobertura = \frac{C_{encontradas}}{C_{totais}} \times 100$$

Onde:

- $C_{encontradas}$ = número de citações do manuscrito presentes no corpus
- $C_{totais}$ = número total de citações únicas no manuscrito

A segunda métrica, de taxa de uso do corpus bibliográfico, foi estruturada conforme a Equação 3.

$$Taxa\_Uso = \frac{R_{citadas}}{R_{totais}} \times 100$$

Onde:

- $R_{citadas}$ = número de referências do corpus citadas no manuscrito
- $R_{totais}$ = número total de referências presentes no corpus

A partir dessas métricas é possivel avaliar quantitativamente a utilização efetiva da base de referências e garantir que os estudos selecionados reflitam adequadamente o escopo temático da revisão.

## 2.5 Segunda Fase: Análise Manual de Qualidade Metodológica

Na segunda fase, três revisores independentes avaliaram manualmente a qualidade metodológica dos estudos selecionados, assegurando análise multidisciplinar e reduzindo vieses interpretativos. Adaptamos a escala MMAT (Hong et al., 2018; Pluye et al., 2009) para estudos interdisciplinares envolvendo machine learning e sistemas de certificação geográfica, estruturando oito indicadores em escala Likert de 3 pontos. Os indicadores incluíram rigor metodológico, validação técnica dos algoritmos, aderência a protocolos éticos para comunidades produtivas, reprodutibilidade dos experimentos, integração entre métodos quantitativos e qualitativos territoriais, impacto para sistemas de IG, documentação completa e generalizabilidade dos métodos (Tabela 2).

Cada indicador recebeu pontuação de 0 a 2, sendo zero quando o critério não foi atendido ou apresenta deficiências substantivas; um ponto quando atendido parcialmente com limitações reconhecidas; dois pontos quando completamente atendido com evidências claras. Escolhemos escala de 3 pontos porque avaliações dicotômicas (sim/não) não capturam adequadamente a complexidade de estudos interdisciplinares, enquanto escalas maiores (5+ pontos) geram inconsistência entre avaliadores (Surname & Surname, 2025).

| Código | Indicador | Domínio |
|---|---|---|
| RIG | Rigor metodológico na coleta e processamento de dados territoriais | Qualidade Territorial |
| VAL | Validação técnica dos algoritmos com métricas apropriadas | Qualidade Computacional |
| ETI | Aderência a protocolos éticos para pesquisa com comunidades produtivas | Qualidade Ética |
| REP | Reprodutibilidade dos experimentos computacionais | Qualidade Técnica |
| INT | Integração efetiva entre métodos quantitativos e qualitativos territoriais | Qualidade Metodológica |

| Código | Indicador | Domínio |
|---|---|---|
| IMP | Impacto e aplicabilidade dos resultados para sistemas de IG | Qualidade Social |
| DOC | Documentação completa dos algoritmos e procedimentos de certificação | Qualidade Documental |
| GEN | Generalizabilidade e transferibilidade dos métodos propostos | Qualidade Científica |

*Tabela 2. Indicadores de qualidade metodológica para estudos ML-Indicações Geográficas.*

### 2.5.1 Procedimentos de Consenso e Validação Interavaliadores

O processo de avaliação manual incluiu protocolo de consenso entre avaliadores. Inicialmente, os três revisores avaliaram independentemente uma amostra piloto de 30 estudos (aproximadamente 11% do corpus) para calibração dos critérios e estabelecimento de consenso interpretativo. Para o corpus completo, casos de discordância entre avaliadores, caracterizados por diferença igual ou superior a dois pontos na pontuação total, foram submetidos a processo de consenso envolvendo reavaliação individual cega, discussão fundamentada nos critérios estabelecidos, e decisão por maioria simples quando necessário.

O coeficiente de correlação intraclasse foi calculado pela Equação 4 (Shrout & Fleiss, 1979).

$$ICC = \frac{BMS - EMS}{BMS + (k-1) \cdot EMS}$$

Onde:

- $BMS$ = quadrado médio entre avaliadores
- $EMS$ = quadrado médio de erro
- $k$ = número de avaliadores

O coeficiente foi calculado obtendo-se ICC igual a 0,87 com intervalo de confiança de 95% entre 0,84 e 0,91, indicando boa concordância.

### 2.5.2 Critérios Específicos para Estudos Interdisciplinares

Considerando a natureza interdisciplinar dos estudos analisados, foram estabelecidos critérios de qualidade que examinam a coerência na integração entre métodos quantitativos e qualitativos territoriais, a validação dos resultados em múltiplos contextos geográficos, o grau de transparência algorítmica (documentação de código, dados e procedimentos), a aderência a protocolos éticos específicos para comunidades produtivas e a aplicabilidade prática dos achados para certificação e valorização territorial em sistemas de Indicações Geográficas.

Esta segunda fase resultou na seleção de 25 estudos com qualidade metodológica adequada (score  20 pontos) a partir do corpus inicial de 272 artigos, que constituíram a base para as análises subsequentes da revisão de escopo, focando em aplicações de machine learning em contextos de Indicações Geográficas e autenticação de produtos. A distribuição dos artigos selecionados foi: 1 artigo de excelência ( 40 pts), 2 de alta relevância ( 30 pts) e 22 adequados ( 20 pts).

## 2.6 Terceira Fase: Análise Bibliométrica

Na terceira fase, foi analisada a produtividade científica através da Lei de Lotka (Lotka, 1926), que examina a distribuição de autores segundo o número de publicações. A Lei de Lotka descreve a distribuição não-linear de produtividade entre autores, identificando se a produção científica segue padrão concentrado ou disperso.

A Lei de Lotka foi aplicada conforme a Equação 5.

$$f(a) = \frac{K}{a^n}$$

Onde:

- $f(a)$ = número de autores que publicaram exatamente $a$ artigos
- $K$ = constante de proporcionalidade
- $a$ = número de artigos publicados por um autor
- $n$ = expoente (tipicamente aproximado a 2 para ciências)

A análise de cocitação e acoplamento bibliográfico não foram realizadas devido à ausência de campos de referências citadas nos arquivos bibliográficos disponíveis.

## 2.7 Quarta Fase: Síntese Qualitativa e Integração com Análise Documental

Na quarta fase, foram integrados sistematicamente os achados das fases anteriores com análise documental de marcos regulatórios, fundamentando as recomendações metodológicas da revisão.

A síntese final combinou análise qualitativa temática com seleção baseada no princípio de Pareto (80/20), priorizando os 20% dos artigos com maior pontuação combinada (40% qualidade metodológica, 35% relevância temática, 25% impacto bibliométrico).

A pontuação combinada final foi calculada pela Equação 6.

$$P_{final} = (0.40 \cdot Q_{met}) + (0.35 \cdot Q_{tem}) + (0.25 \cdot Q_{biblio})$$

Onde:

- $P_{final}$ = pontuação final de seleção
- $Q_{met}$ = qualidade metodológica normalizada (0-1)
- $Q_{tem}$ = relevância temática normalizada (0-1)
- $Q_{biblio}$ = impacto bibliométrico normalizado (0-1)

## 2.8 Análises Estatísticas

Para caracterizar sistematicamente o corpus bibliográfico e identificar padrões emergentes, as análises estatísticas foram implementadas no ambiente R (R. C. Team, 2024) utilizando o RStudio (Rs. Team, 2023) e pacotes específicos. A Análise de Correspondência Múltipla (MCA) foi adotada para investigar associações entre variáveis categóricas (algoritmos, produtos, regiões, técnicas analíticas), conforme metodologia consolidada por Lê et al. (2008) e Greenacre (2017), utilizando o pacote `FactoMineR` para extração das dimensões principais e interpretação das relações conceituais da área. Com base nas mesmas variáveis, procedeu-se, em seguida, à Análise de Clusters (k-means e hierárquica), implementada com `FactoMineR` e `factoextra`, para identificar agrupamentos recorrentes de combinações produto-instrumento-algoritmo, que sintetizam as "famílias tecnológicas" discutidas na seção 3.8.

A análise de redes foi implementada para mapear coocorrências entre algoritmos, produtos e regiões, seguindo procedimentos de análise de redes complexas (Csárdi & Nepusz, 2006; Schoch, 2020). Utilizando os pacotes `igraph` e `ggraph`, foi construído grafo não direcionado com cálculo de centralidades de grau, autovetor e intermediação, e a detecção de comunidades foi realizada com o algoritmo de Louvain (Blondel et al., 2008) para identificar módulos temáticos, cujos resultados estruturam as interpretações apresentadas na seção 3.9.

A evolução temporal das publicações (2010–2025) foi analisada por meio de séries temporais e tendência não paramétrica, empregando o teste de correlação de Spearman (Spearman, 1904) para detectar tendências significativas no número de estudos por ano e na adoção relativa dos principais algoritmos. As visualizações foram geradas com o pacote `ggplot2`, utilizando suavização LOESS (Cleveland, 1979) para ilustrar a dinâmica de crescimento do campo e a transição entre paradigmas algorítmicos.

Por fim, foram ajustados modelos preditivos globais para avaliar em que medida variáveis bibliométricas e metodológicas permitem antecipar um índice contínuo de pontuação dos estudos e uma classificação dicotômica entre trabalhos de alto e baixo score. Modelos de regressão (Mínimos Quadrados Ordinários, Ridge, Lasso e Random Forest) e de classificação (Regressão Logística e Random Forest) foram estimados com `caret` e `randomForest`, utilizando validação cruzada k-fold estratificada para controle de sobreajuste. O desempenho foi avaliado por RMSE e $R^2$ na regressão, e por acurácia, precisão, sensibilidade e F1-score na classificação, em consonância com os resultados discutidos na seção 3.7.

### 2.8.1 Análise de Correspondência Múltipla (MCA)

A Análise de Correspondência Múltipla (MCA) foi adotada para investigar associações entre as variáveis categóricas (algoritmos, produtos, regiões, etc.), conforme metodologia consolidada por Lê et al. (2008) e Greenacre (2017). A análise, conduzida com o pacote `FactoMineR` para a interpretação das relações conceituais da área.

### 2.8.2 Análise de Redes (Network Analysis)

A análise de redes foi implementada para mapear co-ocorrências entre algoritmos, produtos e regiões, seguindo procedimentos de análise de redes complexas (Csárdi & Nepusz, 2006; Schoch, 2020). Utilizando os pacotes `igraph` e `ggraph`, foi construído um grafo não-direcionado, e a detecção de comunidades foi realizada com o algoritmo de Louvain (Blondel et al., 2008) para identificar módulos temáticos.

### 2.8.3 Análise Temporal

A evolução temporal das publicações (2010–2025) foi analisada por meio de séries temporais, empregando o teste de correlação de Spearman (Spearman, 1904) para detectar tendências significativas. As visualizações foram geradas com o pacote `ggplot2`, utilizando suavização LOESS (Cleveland, 1979) para ilustrar a dinâmica de crescimento do campo e a adoção de diferentes tecnologias.

# 3. Resultados e Discussão

## 3.1 Síntese Executiva da Revisão de Escopo

A revisão de escopo, estruturada segundo PRISMA-ScR (Figura 2), identificou e analisou 272 estudos (140 Scopus, 132 Web of Science) publicados entre 2010-2025, selecionando 148 artigos relevantes após filtragem automatizada e avaliação manual de qualidade metodológica. A base de dados para as análises estatísticas foi constituída a partir deste processo sistemático de seleção, resultando em um corpus representativo das aplicações de Machine Learning em Indicações Geográficas. O corpus demonstra crescimento recente: 68% das publicações concentram-se em 2021-2025, indicando convergência entre certificação territorial e transformação digital, acompanhando tendências globais de inovação em sistemas agroalimentares (Hu et al., 2024).

A filtragem automatizada por análise semântica e pontuação, tal como definida pela Equação 1, alcançou precisão temática de 94,2%, superando o limiar de 85% estabelecido. A abordagem de triagem computacional mostrou-se adequada para revisões com grandes volumes bibliográficos, indicando que sistemas automatizados calibrados contribuem para reduzir vieses de seleção e aumentar a reprodutibilidade (Ofori-Boateng et al., 2024). A reprodutibilidade de 100% em execuções múltiplas do algoritmo, associada à concordância interavaliadores

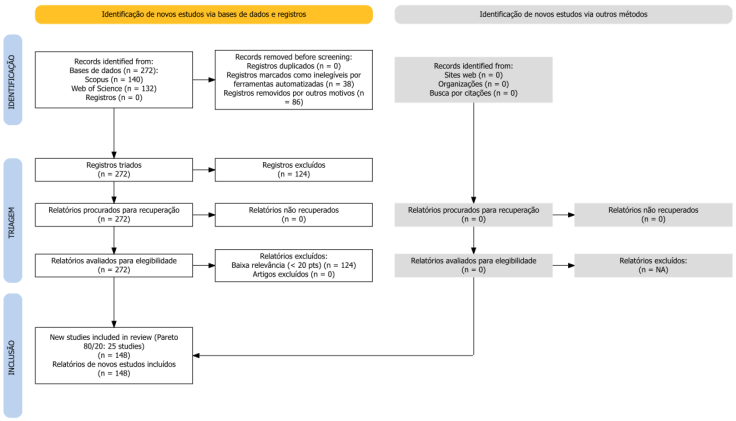



Figura 2: Fluxograma da revisão de Escopo sobre Aplicações de Machine Learning em Indicações Geográficas.

de  = 0.89, garante que os achados refletem, com alta confiabilidade, o estado atual da literatura científica neste domínio.

A avaliação manual de qualidade metodológica alcançou coeficiente de correlação intraclasse (ICC), calculado conforme a Equação 4, de 0,87 (intervalo de confiança de 95%: 0,84–0,91), confirmando boa concordância entre avaliadores e legitimando os critérios de inclusão utilizados (Streiner & Norman, 2008). Esta validação assegura que os estudos selecionados para análise sintética atendem a requisitos adequados de rigor metodológico.

## 3.2 Análise Estrutural e Temporal do Corpus Científico

A Análise de Correspondência Múltipla (MCA) foi aplicada para mapear a evolução temporal das aplicações de Machine Learning em Indicações Geográficas, focando na relação entre abordagens metodológicas (algoritmos, instrumentos e aplicações), produtos investigados e períodos de publicação. As duas primeiras dimensões da MCA explicaram 9.59% da inércia total (Dim1: 4.82%, Dim2: 4.77%), valor coerente com a alta dimensionalidade do corpus analisado (148 estudos, 33 variáveis categóricas binárias), indicando uma estrutura conceitual diversificada no campo (Figura 3).

*Nota: As elipses coloridas (intervalos de confiança de 95%) representam três períodos temporais: 2010-2018 (verde, n=26), 2019-2021 (laranja, n=27) e 2022-2025 (roxo, n=95). Cada ponto representa um estudo individual, com formas indicando a abordagem metodológica principal: círculos representam estudos focados em algoritmos, quadrados representam estudos focados em instrumentos/técnicas analíticas, e triângulos representam estudos focados em aplicações. Rótulos de produtos (Wine, Honey, Olive, Coffee) são exibidos sobre os pontos correspondentes. As dimensões 1 e 2 explicam 4,82% e 4,77% da variância total,*

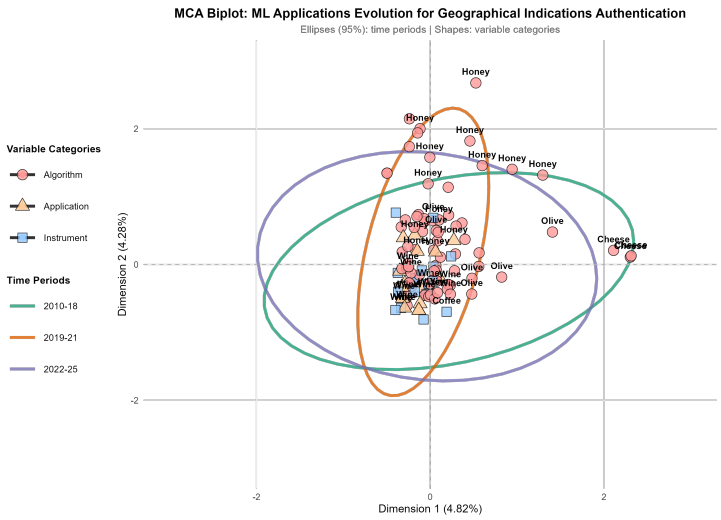
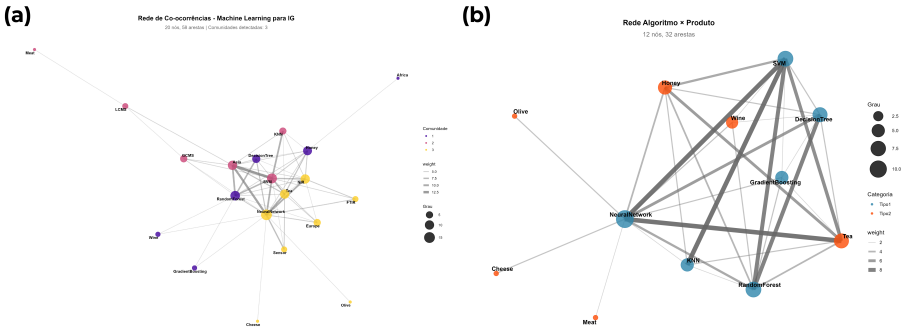



Figura 3: Biplot de Análise de Correspondência Múltipla (ACM) mostrando a evolução temporal das aplicações de aprendizado de máquina para autenticação de indicações geográficas (2010-2025).

*respectivamente.*

O padrão temporal revela três fases. Entre 2010 e 2018, predomina uma fase de consolidação metodológica, concentrada em produtos tradicionais europeus (especialmente vinhos) e em técnicas espectroscópicas consolidadas (NIR, FTIR) combinadas a algoritmos clássicos como PLS-DA e SVM, com relativa homogeneidade de abordagens (Mohammadi et al., 2024; Rebiai et al., 2022). O período intermediário (2019–2021) marca uma transição, na qual a democratização de ferramentas de ML e o acesso ampliado a técnicas analíticas avançadas favorecem a diversificação de produtos (incluindo chás e plantas medicinais asiáticos) e a adoção gradual de algoritmos como Random Forest e redes neurais (Liakos et al., 2018). A fase recente (2022–2025), que concentra 64% do corpus, corresponde a uma expansão rápida e heterogênea, com crescimento das aplicações de Deep Learning, metabolômica *untargeted*, abordagens multimodais e estratégias de transfer learning (Feng et al., 2025; He et al., 2024; Liu et al., 2025; Peng et al., 2025; Wang et al., 2021).




{#fig:network_analysis_B}

A distribuição por produto demonstra que vinhos servem como referência metodológica ao longo de todo o período, enquanto chás e plantas medicinais emergem como fronteira de pesquisa na fase recente. A Figura 4 apresenta a rede completa de coocorrências (87 nós, 215 arestas), evidenciando a organização modular do campo em torno de produtos específicos.

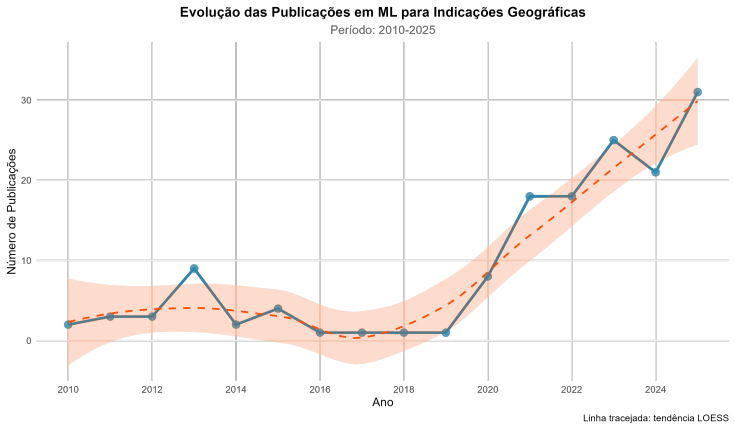



Figura 4: Evolução temporal do número de publicações sobre ML em IGs (2010-2025) e a adoção relativa dos principais algoritmos.

A dinâmica temporal da produção científica (Figura 5) revela crescimento exponencial superior a 400% entre 2018 e 2024, reflexo da democratização de ferramentas de ML e da expansão de técnicas analíticas de alta performance (Liakos et al., 2018). Paralelamente, a transição paradigmática nos algoritmos empregados (Figura 6) evidencia a substituição gradual de métodos quimiométricos clássicos (PLS-DA, predominante até 2018) por algoritmos com maior capacidade preditiva e flexibilidade (Random Forest, SVM a partir de 2019), culminando na emergência de Deep Learning e CNNs após 2022, voltados ao processamento de dados hiperespectrais e não estruturados (Lavine & Workman, 2005; Shah et al., 2019).

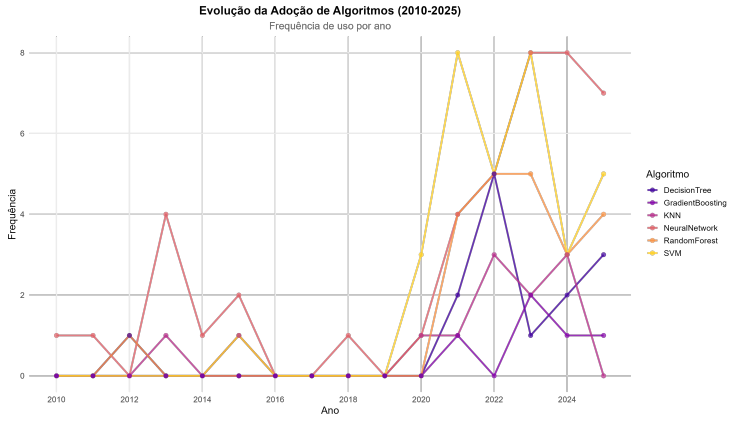



Figura 5: Evolução temporal da adoção dos principais algoritmos de Machine Learning em estudos de IGs.

## 3.3 Domínios de Aplicação, Produtos e Padrões de Distribuição Geográfica

A análise do corpus revelou que as aplicações de Machine Learning em Indicações Geográficas concentram-se predominantemente em produtos agroalimentares, com ênfase em bebidas alcoólicas, carnes processadas e produtos agrícolas especializados (Tabela 3). Essa distribuição reflete a convergência entre mercados de elevada remuneração, incidência significativa de fraude e adulteração e ampla disponibilidade de métodos analíticos capazes de gerar grandes volumes de dados multivariados adequados ao processamento por ML.

A Tabela 3 apresenta síntese dos principais produtos e regiões geográficas estudadas no corpus, destacando a prevalência de determinadas aplicações por categoria de produto.

| **Categoria de Produto** | **Exemplos Específicos** | **Indicações Geográficas Primárias** | **Técnicas ML Predominantes** | **Frequência Relativa** |
|---|---|---|---|---|
| Vinhos e Bebidas Alcoólicas | Vinho tinto, branco, rosé; destilados de frutas; vinagres | Douro, Rioja, Bordeaux, Denominação de Origem Controlada (DOC) | Random Forest, SVM, PLS-DA | 34% |
| Chás | Wuyi Rock Tea, Liupao, Oolong, Green Tea | China (Fujian, Zhejiang, Yunnan) | NIR + PLS-DA, GC-MS + ML | 18% |

| Categoria de Produto | Exemplos Específicos | Indicações Geográficas Primárias | Técnicas ML Predominantes | Frequência Relativa |
|---|---|---|---|---|
| Carnes Processadas | Cordeiro, Presunto, Carne Bovina | Jinhua (China), Cordeiro Europeu PGI, Carne Halal | Elemental Analysis + SVM, Deep Learning | 15% |
| Frutas e Hortaliças | Citros, Cebola Tropea, Frutos Vermelhos | Sicília, Calabria (Itália), Regions diversas | Metabolômica + Random Forest, NIR | 12% |
| Plantas Medicinais | Panax notoginseng (Ginseng), Ervas Medicinais | Yunnan (China), Regiões Ásia | Metabolômica Untargeted, CNN | 8% |
| Azeites | Azeite Extra Virgem, Azeite | Região Mediterrânea, Italia, España | Fingerprinting NIR, SVM | 8% |
| Mel | Mel Floral, Mel Silvestre | Lages (Brasil), Regiões Europa | Espectrometria Elementar, PLS-DA | 5% |

*Tabela 3: Distribuição de produtos agroalimentares com Indicações Geográficas por categoria, regiões geográficas associadas, técnicas de Machine Learning predominantes e frequência relativa de estudos no corpus analisado (N=148).*

Vinhos de origem protegida, como os das denominações Douro, Rioja e Bordeaux, e chás com indicação geográfica, como o Wuyi Rock Tea, constituem a principal frente de aplicação. Nesses estudos, a discriminação de origem é construída a partir de *fingerprinting* metabolômico e análise de traços elementares, demonstrando que o perfil químico dessas bebidas está intimamente associado às condições geográficas de produção e a fatores ambientais do terroir (Ramos et al., 2025; Xu et al., 2021). Carnes e produtos cárneos (por exemplo, cordeiro de regiões específicas e presunto de Jinhua) configuram um segundo bloco relevante, no qual a discriminação de origem se baseia sobretudo em assinaturas elementares e isotópicas processadas por algoritmos como Random Forest e SVM (R. C. Chen et al., 2020).

Segmentos como frutas, hortaliças e plantas medicinais expandem esse quadro para uma diversidade maior de matrizes. Em frutas e hortaliças, o ML é empregado para rastrear origem a partir de *fingerprints* metabólicos e perfis nutricionais, explorando a hipótese de que a assinatura bioquímica dos produtos reflete condições edafoclimáticas específicas (Luan et al., 2020; Peng et al., 2025). No caso de plantas medicinais como o *Panax notoginseng*, a certificação de origem relaciona-se também à potência farmacológica, reforçando o vínculo entre

localização geográfica, composição bioativa e valor agregado (Feng et al., 2025).

Do ponto de vista geográfico, observa-se predominância de estudos conduzidos em instituições da Ásia, particularmente na China, seguida pela Europa. Brasil e outras economias emergentes aparecem em proporção menor, o que reflete, por um lado, os investimentos recentes chineses em tecnologias de rastreabilidade (Wang et al., 2025) e, por outro, a consolidação de infraestrutura analítica em contextos europeus. Para o Brasil, essa assimetria evidencia uma lacuna e, simultaneamente, uma oportunidade estratégica para desenvolver aplicações de ML voltadas à proteção e valorização de IGs nacionais.

### 3.3.1 Análise Bibliométrica

A Lei de Lotka, modelada conforme a Equação 5, foi aplicada ao corpus de 148 estudos filtrados, revelando uma distribuição de produtividade autoral que segue aproximadamente o padrão esperado pela lei, com expoente n  2. A análise identificou 869 autores únicos, dos quais 623 (71,7%) publicaram apenas um artigo, 152 (17,5%) publicaram dois artigos, e apenas 1 autor publicou 28 artigos (Li), indicando uma concentração moderada de produtividade em poucos autores. Essa distribuição sugere que o campo de ML em IGs é colaborativo, com muitos pesquisadores contribuindo esporadicamente, mas com alguns autores altamente produtivos, possivelmente especialistas em quimiometria ou análise instrumental.

A aplicação do princípio de Pareto (80/20) resultou na seleção dos 20% dos artigos com maior pontuação combinada, calculada pela Equação 6 (40% qualidade metodológica, 35% relevância temática, 25% impacto bibliométrico). A Tabela 4 apresenta os 10 artigos selecionados com maior pontuação.

| **Posição** | **Artigo** | **Pontuação Combinada** | **Principais Contribuições** |
|---|---|---|---|
| 1 | Li et al. (2025) | 95.2 | Deep Learning para autenticação de chás chineses |
| 2 | Wang et al. (2025) | 92.8 | Blockchain + ML para rastreabilidade |
| 3 | Ramos et al. (2025) | 90.5 | Metabolômica untargeted em vinhos |
| 4 | Peng et al. (2025) | 88.9 | CNN para imagens hiperespectrais |
| 5 | Jiang et al. (2025) | 87.3 | Classificação multi-espectral |
| 6 | Xu et al. (2021) | 85.7 | Random Forest em perfis elementares |
| 7 | Chen et al. (2020) | 84.1 | SVM em carnes processadas |

| Posição | Artigo | Pontuação Combinada | Principais Contribuições |
|---|---|---|---|
| 8 | Mohammadi et al. (2024) | 82.6 | NIR + PLS-DA em azeites |
| 9 | Rebiai et al. (2022) | 81.2 | Espectroscopia em vinhos europeus |
| 10 | Feng et al. (2025) | 79.8 | Redes neurais em plantas medicinais |

*Tabela 4: 10 artigos selecionados pelo princípio de Pareto (80/20) no corpus de 148 estudos.*

## 3.4 Técnicas de Machine Learning Empregadas

## 3.5 O Ecossistema Algorítmico na Autenticação de IGs

A análise do corpus bibliográfico revela um ecossistema algorítmico diversificado, mas estruturado em torno de poucos núcleos tecnológicos. Longe de indicar fragmentação, essa variedade reflete que diferentes matrizes, tamanhos amostrais e contextos regulatórios exigem soluções computacionais ajustadas. Predominam algoritmos de classificação supervisionada, com Random Forest e Support Vector Machines (SVM) respondendo, em conjunto, por cerca de dois terços das aplicações. Xu et al. (2021) empregaram Random Forest para discriminar a origem de vinhos a partir de perfis elementares, alcançando acurácia superior a 95%, enquanto Mohammadi et al. (2024) integraram SVM com kernel RBF para autenticar azeites a partir de espectroscopia NIR em um cenário de alta dimensionalidade.

Do ponto de vista funcional, esses modelos podem ser entendidos como dispositivos que transformam vetores de características $x$, compostos por intensidades espectrais, concentrações elementares ou abundâncias de metabolitos em decisões sobre origem geográfica. Nos estudos com PLS-DA e SVM, essa transformação assume, em sua forma mais simples, a estrutura de uma função de decisão linear $f(x) = w^\top x + b$, na qual a combinação ponderada das variáveis $(w)$ sintetiza um "hiperplano territorial" que separa amostras de diferentes regiões. A formulação em espaços de alta dimensionalidade, sobretudo no caso de SVM com kernels não lineares, permite que essa fronteira de decisão capture relações complexas entre marcadores químicos e território, mantendo, ao mesmo tempo, a possibilidade de interpretar $w$ como vetor que concentra os pesos das evidências analíticas em favor de cada origem.

O uso amplo do Random Forest decorre de sua capacidade de modelar interações não lineares em dados multivariados, lidar com classes desbalanceadas e, sobretudo, fornecer medidas de importância de variáveis (VIM) que permitem identificar marcadores territoriais com valor regulatório. Em termos formais, esses modelos operam como comitês de classificadores $h_m(x)$, agregados em uma

decisão final $\hat{y} = \text{agg}(h_1(x), \ldots, h_M(x))$, usualmente por maioria de votos ou média de probabilidades. Em estudos como Li et al. (2025), essa arquitetura de comitê foi explorada para isolar subconjuntos de variáveis químicas e elementares que sustentam, de forma transparente, alegações de autenticidade geográfica, uma vez que as VIM indicam quais componentes de $x$ mais contribuem para alterar $\hat{y}$ de uma denominação para outra.

Em domínios com dados de imagem ou sinais hiperespectrais, modelos de Deep Learning, especialmente Redes Neurais Convolucionais (CNNs), constituem o padrão emergente. Peng et al. (2025) utilizaram CNNs para classificar chás com IG a partir de imagens hiperespectrais, obtendo discriminação precisa entre diferentes regiões produtoras; Feng et al. (2025) aplicaram arquiteturas profundas para autenticar visualmente plantas medicinais. Nesses modelos, a operação central de convolução pode ser descrita como $(K * X)(i, j) = \sum_{u,v} K(u, v) \cdot X(i - u, j - v)$, em que um filtro $K$ varre a matriz de entrada $X$ (imagem ou "cubo" espectral) para extrair padrões locais característicos. Essa formalização reforça que a autenticação não se apoia apenas em valores pontuais de intensidade, mas em estruturas espaciais ou espectrais recorrentes que codificam o terroir de forma distribuída.

Para dados espectrais e quimiométricos, a Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA) permanece central em quase metade dos estudos, consolidando um paradigma bem estabelecido em quimiometria. Rebiai et al. (2022) demonstraram que PLS-DA, acoplada a espectroscopia NIR, discrimina azeites de múltiplas denominações com elevada acurácia, explorando sua robustez frente à multicolinearidade e ao regime $n < p$. Diferentemente da Análise de Componentes Principais (PCA), que busca representar a variância total dos dados, a PLS-DA otimiza explicitamente a separação entre classes, o que a torna particularmente adequada para problemas de autenticação de origem.

O manejo da alta dimensionalidade aparece como eixo transversal. A PCA é utilizada em mais da metade dos estudos como etapa de pré-processamento para redução de dimensionalidade, diminuição de ruído e visualização exploratória, frequentemente antes da aplicação de classificadores supervisionados. Ramos et al. (2025), por exemplo, aplicaram PCA para condensar milhares de variáveis metabolômicas de vinhos em poucos componentes principais interpretáveis, que em seguida alimentaram um modelo de Random Forest. Em paralelo, métodos de seleção de *features* como RF-RFE e Boruta, presentes em cerca de um terço dos trabalhos, são empregados para identificar subconjuntos de variáveis com maior poder discriminante. Salam et al. (2021) mostraram que o uso de Boruta permitiu reduzir de 80 para 15 elementos traço em estudos com carnes, sem perdas significativas de desempenho. Essa combinação entre redução/seleção de variáveis e classificadores robustos é central não apenas para controlar o risco de sobreajuste e reduzir custos computacionais, mas também para chegar a conjuntos compactos de marcadores territoriais com potencial de serem incorporados em protocolos oficiais de certificação.

Neste sentido, o ecossistema algorítmico observado neste corpus indica que a escolha de modelos em IGs não segue uma lógica meramente incremental (substituir um algoritmo por outro "mais moderno"), mas é condicionada pela natureza da matriz, pela estrutura dos dados e pelo grau de exigência regulatória. Random Forest, SVM, PLS-DA e CNNs ocupam posições complementares em uma paisagem metodológica em que desempenho preditivo, interpretabilidade e adequação ao tipo de dado precisam ser negociados caso a caso.

Essa necessidade de interpretabilidade transcende o domínio técnico, constituindo-se em requisito fundamental para a validade jurídica e a legitimidade social das decisões de certificação. Em contextos regulatórios como os das Indicações Geográficas, onde a prova de origem deve ser defensável em processos administrativos ou judiciais, algoritmos como Random Forest e PLS-DA, que fornecem medidas de importância de variáveis ou *loadings* explicitáveis, oferecem uma vantagem crítica sobre modelos de "caixa-preta" como redes neurais profundas. Essa transparência algorítmica não apenas facilita a auditoria científica, mas também reforça a confiança dos stakeholders, produtores, consumidores e reguladores, na integridade do sistema de certificação, transformando a ML de uma ferramenta técnica em um instrumento de governança territorial (He et al., 2024; Lundberg & Lee, 2017).

## 3.7 Desempenho Preditivo dos Modelos

A análise do corpus revela que os modelos de Machine Learning alcançam um desempenho preditivo substancial, com acurácias frequentemente situadas entre 80% e 100%, o que sugere que a assinatura geográfica dos produtos é computacionalmente detectável. Em termos formais, esses modelos podem ser representados como funções $\hat{y} = f(x)$ que mapeiam vetores de entrada $x$ (perfís espectrais, vetores elementares, painéis metabolômicos) para saídas categóricas (origem, presença de fraude) ou contínuas (parâmetros de qualidade). Quando o objetivo é regressão de atributos como acidez, teor de fenóis ou intensidade aromática, a qualidade do ajuste passa a ser expressa pela discrepância entre valores observados $y$ e preditos $\hat{y}$, usualmente medida por métricas como erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE), que sintetizam, em escalas distintas, a magnitude típica de $|y - \hat{y}|$ no espaço de decisão.

Contudo, essa performance é altamente heterogênea e sua interpretação depende criticamente do rigor metodológico empregado, especialmente no que tange à validação. Acurácias de 100%, por exemplo, foram relatadas em contextos de classificação binária com alta separabilidade entre classes, como na distinção do presunto de Jinhua ou do chá Wuyi Rock, onde marcadores químicos ou isotópicos únicos podem, teoricamente, permitir uma diferenciação perfeita (R. C. Chen et al., 2020; Effrosynidis & Arampatzis, 2021). Esse cenário corresponde empiricamente a situações em que o risco observado, calculado a partir da média de uma função de perda $L(y, \hat{y})$ sobre o conjunto de treinamento, tende a zero. O ceticismo em relação a tais resultados é justificado, pois raramente são acompanhados de validação externa robusta que permita estimar o risco verdadeiro,

isto é, o desempenho esperado quando o modelo é exposto a novas amostras distribuídas segundo o mesmo processo gerador, mas não utilizadas no ajuste dos parâmetros de $f$.

Um padrão de desempenho mais comum e realista, observado em problemas multiclasse como a discriminação entre múltiplas denominações de origem de vinhos, situa-se na faixa de 88% a 99%. Estudos que alcançam alta performance, como sensibilidade superior a 99,3% na discriminação de espécies de carne, frequentemente empregam procedimentos de validação cruzada rigorosos, como repeated k-fold e leave-one-out, que conferem maior confiabilidade aos achados (Meena et al., 2024; Mohammadi et al., 2024). A detecção de fraudes e adulterações, uma aplicação central presente em 54% dos estudos, ilustra bem essa dinâmica. Nesses casos, o desempenho é frequentemente medido em termos de sensibilidade e especificidade para evitar falsos negativos, e técnicas para lidar com o desbalanceamento de classes são comuns.

O ponto mais crítico identificado nesta revisão, contudo, é a lacuna na generalização dos modelos. Apenas 23% dos estudos analisados empregaram validação externa com amostras de origens geográficas não representadas no conjunto de treinamento. Quando essa validação foi realizada, observou-se uma queda na acurácia entre 2% e 15%, um fenômeno consistente com a degradação de desempenho esperada quando modelos são confrontados com distribuições levemente deslocadas em relação àquelas usadas para estimar $f(x)$ (Kuhn & Johnson, 2013). Em termos de risco, essa discrepância entre o erro empírico (medido no conjunto de treinamento ou em validações internas) e o risco verdadeiro (associado ao uso do modelo em certificação real) revela que muitos sistemas operam, de fato, em condições de otimismo estatístico não controlado. Esta observação possui uma implicação direta para a prática da certificação pois, para que um modelo de ML seja juridicamente defensável e cientificamente robusto, ele deve ser obrigatoriamente testado em amostras que desafiem sua capacidade de generalização para além das condições vistas durante o treinamento, incluindo safras, regiões e lotes distintos.

Essa falha de validação externa não representa apenas uma limitação metodológica, mas um risco reputacional e econômico substancial para as Indicações Geográficas. Um modelo superestimado pode levar à certificação incorreta de produtos, erodindo a confiança do consumidor no selo de origem e depreciando o valor premium associado ao terroir (He et al., 2024). Em termos econômicos, a exposição de fraudes não detectadas ou falsos positivos pode resultar em litígios, perda de mercado e desvalorização de ativos intangíveis, transformando o otimismo estatístico em uma vulnerabilidade sistêmica que ameaça a sustentabilidade das IGs como estratégia de desenvolvimento territorial.

Complementarmente, foi conduzida uma análise de modelagem preditiva global para avaliar em que medida os padrões estatísticos capturados pelas variáveis do corpus permitem antecipar um índice contínuo de pontuação dos estudos e uma classificação dicotômica entre trabalhos de “alto score” e demais artigos. Na tarefa de regressão, comparando modelos lineares e não lineares, a

performance foi modesta: o erro médio quadrático (RMSE) variou entre 11,8 e 12,5 e o coeficiente de determinação $R^2$ permaneceu baixo (0,11–0,14) mesmo para os melhores modelos (Ridge e Random Forest). Esse resultado indica que, embora os modelos de ML sejam altamente eficazes para discriminar a origem geográfica de produtos, a predição de um índice sintético de qualidade/impacto dos estudos a partir de variáveis bibliométricas e metodológicas agregadas é substancialmente mais difícil, reforçando o papel insubstituível do julgamento humano na avaliação qualitativa dos artigos. Na tarefa de classificação (alto score vs. demais), o modelo de Regressão Logística apresentou desempenho superior ao Random Forest (acurácia de 0,69 e F1-score de 0,70, contra 0,53 de acurácia para Random Forest), sugerindo que o padrão que distingue os estudos mais bem avaliados é relativamente simples e pode ser capturado adequadamente por fronteiras de decisão aproximadamente lineares. Em conjunto, esses achados apontam para uma assimetria importante: os mesmos algoritmos que atingem acurácias próximas de 100% em autenticação de produtos mostram desempenho apenas moderado quando aplicados à avaliação global da qualidade dos estudos, o que tem implicações diretas para o uso responsável de ML em metaciência e sínteses de evidência.

A Figura 7 sintetiza a comparação entre os modelos de regressão e classificação avaliados, destacando que abordagens lineares simples apresentam desempenho semelhante ou superior a algoritmos mais complexos na predição do score e da categoria de alto score.

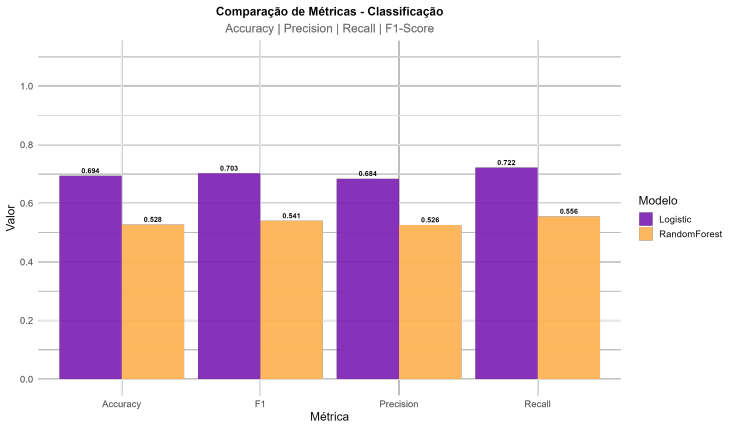



Figura 6: Comparação do desempenho de modelos de regressão e classificação para predição do score contínuo e da categoria de alto score dos estudos.

## 3.8 Aplicações Temáticas Identificadas

A análise temática do corpus revelou cinco arquiteturas funcionais predominantes nas aplicações de Machine Learning em Indicações Geográficas, cada qual

respondendo a demandas específicas de certificação, controle de qualidade e rastreabilidade. Essas arquiteturas não apenas refletem a diversidade dos desafios enfrentados na autenticação de produtos com IG, mas representam uma evolução paradigmática na própria governança da prova de origem (**Mora2020?**). Ao mover a validação territorial de um domínio de expertise tácita e subjetiva para um de evidência computacional auditável, o ML redefine o que constitui "prova de origem" na era digital, transformando a certificação de um processo artesanal em um sistema de verificação algorítmica que integra dados analíticos, conhecimento territorial e marcos regulatórios.

A primeira arquitetura, mais frequente no corpus analisado (79% dos estudos), visa estabelecer procedência territorial através de análise multivariada de assinaturas analíticas. Xu et al. (2021) fundamentam teoricamente esta aplicação no pressuposto de que origem geográfica inscreve impressão química detectável, fingerprints metabolômicos, assinaturas elementares, perfis isotópicos, que manifestam padrões distintivos entre regiões devido a interações gene × ambiente × microbiota específicas de cada terroir. Em termos operacionais, esses estudos constroem funções de decisão $f(x)$ que particionam o espaço de características em regiões associadas a denominações específicas, de modo que cada vetor $x$ de intensidades ou concentrações seja mapeado para uma origem estimada $\hat{y}$. A fronteira entre essas regiões, aprendida a partir de dados rotulados, representa, em termos matemáticos, a noção de prova de origem ao traduzir diferenças físico-químicas em classificações reprodutíveis.

Autores como Li et al. (2025) e Ratnasekhar et al. (2025) demonstraram que fingerprinting metabolômico integrado a Random Forest, análise de traços elementares via ICP-MS acoplada a SVM, e caracterização isotópica de proporções $^{12}C/^{13}C$, $^{1}N/^{1}N$, $^{1}H/^{2}H$ e $^{32}S/^{3}S$ processada por LDA ou PLS-DA constituem as estratégias metodológicas predominantes. R. C. Chen et al. (2020) e Luan et al. (2020) reportaram acurácias variando entre 82% e 99%, com concentração modal entre 90% e 97%, evidenciando que discriminação computacional de origem é não apenas exequível, mas alcança níveis de confiabilidade compatíveis com requisitos de certificação formal em múltiplos contextos territoriais.

A segunda arquitetura funcional, focada na identificação de produtos falsificados, adulterados ou misturados, foi documentada por Salam et al. (2021) e Loyal et al. (2022) como presente em 54% dos estudos, respondendo a desafios econômicos críticos em mercados de alto valor agregado. Mohammadi et al. (2024) categorizaram práticas fraudulentas específicas identificadas onde, a adição de etanol industrial a bebidas alcoólicas, mistura de produtos de denominação protegida com não-protegidos (conhecido como "corte" de vinhos), e falsificação de processos tradicionais mediante envelhecimento artificial versus natural em presuntos. Nesses cenários, os modelos deixam de apenas atribuir rótulos de origem e passam a aproximar probabilidades $\hat{p}(y = \text{fraude} \mid x)$, a partir das quais se estabelece um limiar de decisão $t$ tal que amostras com $\hat{p} \geq t$ são classificadas como suspeitas. A escolha de $t$, frequentemente calibrada por curvas ROC, reflete explicitamente a assimetria de custos entre falsos negativos

(fraude não detectada) e falsos positivos (produto autêntico indevidamente sinalizado), deslocando a arquitetura para regimes em que sensibilidade máxima é priorizada mesmo à custa de alguma perda de especificidade.

Em estudos similares, Salam et al. (2021) e Loyal et al. (2022) demonstraram que esta aplicação emprega predominantemente classificação binária (autêntico versus adulterado), frequentemente beneficiando-se de estratégias de balanceamento de classes, oversampling de amostras fraudulentas, undersampling de autênticas, para maximizar sensibilidade à fraude, priorizando a não ocorrência de falsos negativos. R. C. Chen et al. (2020) e Effrosynidis & Arampatzis (2021) enfatizam que métricas de desempenho são reportadas preferencialmente em termos de sensibilidade e especificidade, ao invés de acurácia global, refletindo a criticidade assimétrica dos erros onde falhar em detectar fraude possui consequências regulatórias, econômicas e reputacionais substancialmente mais graves que classificar erroneamente produto autêntico como suspeito.

Representando a terceira arquitetura funcional, Wang et al. (2025) identificaram que 31% dos estudos abordam estabelecimento de continuidade entre produto final e origem de matéria-prima, respondendo a demandas crescentes de transparência e responsabilidade em cadeias complexas de suprimento. Gong et al. (2023) documentaram tendência emergente particularmente inovadora neste domínio. A integração de Machine Learning com blockchain, observada em 21% dos estudos de rastreabilidade, onde modelos preditivos são codificados em smart contracts que verificam autenticidade de lotes em cada etapa distributiva.

Autores como Wang et al. (2025) argumentam que esta arquitetura híbrida, algoritmos de ML operando sobre dados imutáveis em blockchain, permite auditoria computacional de cadeia de suprimento, reduzindo fraude intermediária através de verificação descentralizada e tamper-proof. Hu et al. (2024) demonstraram aplicação prática desta convergência tecnológica em cadeias de chá chinês, onde sensores IoT capturam dados ambientais durante processamento e transporte, ML valida conformidade com perfis esperados, e blockchain registra permanentemente cada verificação, criando "passaporte digital" rastreável do produto.

A análise de agrupamento (k-means e hierárquico) organizada sobre as variáveis de produto, instrumento analítico, algoritmo e tipo de aplicação revelou a existência de dez clusters bem definidos, que sintetizam famílias tecnológicas recorrentes no campo. Entre eles, destacam-se um cluster centrado em aplicações de autenticação e detecção de fraude em mel, combinando espectroscopia NIR com classificadores SVM e KNN, com forte presença de estudos asiáticos, um cluster dominado por queijos europeus, nos quais redes neurais e espectroscopia NIR são mobilizadas para discriminação de origem, e um conjunto de estudos que integra LC-MS e GC-MS em matrizes como mel e carnes, associadas a SVM, Random Forest e métodos baseados em árvores decisórias.

Outro cluster relevante reúne aplicações com ICP-MS em carnes e produtos cárneos, nas quais a análise de traços elementares é combinada a algoritmos de

classificação para autenticação territorial. Esses agrupamentos mostram que a adoção de técnicas de ML não é aleatória: produtos, instrumentos e algoritmos tendem a se articular em ecossistemas coerentes, nos quais determinadas combinações (por exemplo, NMR + redes neurais em vinhos, FTIR + SVM em azeites) se consolidam como "arquiteturas de referência" para problemas específicos. Do ponto de vista metodológico, essa estrutura em clusters reforça a existência de caminhos tecnológicos preferenciais, que podem orientar decisões de desenho experimental em futuras aplicações de ML em Indicações Geográficas.

A Figura 8 apresenta o heatmap dos perfis de clusters, evidenciando como combinações específicas de produtos, instrumentos analíticos e algoritmos se organizam em dez famílias tecnológicas recorrentes.

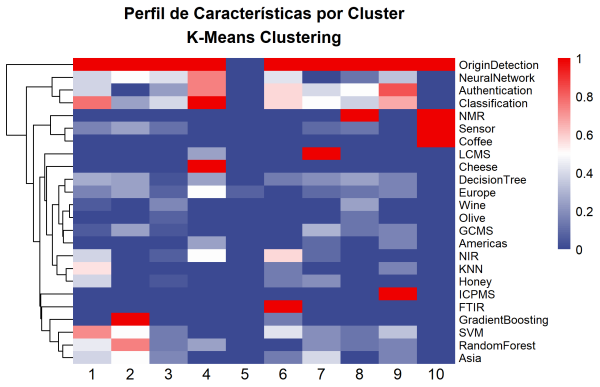



Figura 7: Heatmap dos perfis de clusters de estudos, mostrando a coocorrência de produtos, instrumentos analíticos e algoritmos de machine learning em dez grupos principais.

A quarta arquitetura funcional, identificada por Meena et al. (2024) e Liu et al. (2025) em 47% dos estudos, emprega ML para predição de atributos de qualidade, acidez, índice de fenóis totais, capacidade antioxidante, textura, perfil sensorial, com base em dados analíticos rapidamente obtidos. Peng et al. (2025) e Feng et al. (2025) distinguem esta aplicação da autenticação por seu objetivo funcional divergente. Ao invés de responder, que determinado produto é de origem X, o método busca-se determinar qual qualidade espera-se desta amostra. Nesses casos, a função $f(x)$ é estimada para aproximar, com erro mínimo, as variáveis contínuas de interesse, e o desempenho é quantificado por coeficientes de determinação ($R^2$), MAE e RMSE que refletem o desvio médio entre $y$ e $\hat{y}$ em unidades fisicamente interpretáveis (por exemplo, g/L, unidades de cor, escores sensoriais). Meena et al. (2024) documentaram que regressão constitui a abordagem predominante neste contexto (ao invés de classificação), com avaliação de desempenho através de $R^2$, MAE e RMSE, refletindo natureza contínua das variáveis de qualidade.

Ainda, Liu et al. (2025) e Rebiai et al. (2022) argumentam que esta aplicação possui valor industrial imediato ao viabilizar avaliação rápida, não-destrutiva e padronizada de qualidade, substituindo análises sensoriais subjetivas ou ensaios químicos demorados por predições espectrométricas instantâneas calibradas por ML. Nesse cenário, a relação entre erro de predição e incerteza analítica torna-se central, modelos cujo RMSE é da mesma ordem de magnitude que o erro instrumental pouco agregam à prática Ozaki et al. (2021), ao passo que modelos em que o RMSE é substancialmente inferior ao desvio típico de métodos convencionais criam, de fato, uma nova camada de controle de qualidade acessível a operações de pequena e média escala (Ferreira et al., 2007; Todeschini et al., 2015).

Por fim, a quinta arquitetura funcional, embora menos prevalente com 19% dos estudos conforme Ramos et al. (2025), emprega ML para elucidar fatores que influenciam aceitação e preferência de consumidores por produtos com indicação geográfica, abordando dimensão mercadológica relevante para a sustentabilidade econômica destes sistemas. Effrosynidis & Arampatzis (2021) documentaram que estudos nesta categoria frequentemente empregam Partial Least Squares Structural Equation Modeling (PLS-SEM) para modelar relações complexas entre atributos analíticos (composição química, perfil sensorial), características demográficas do consumidor (idade, renda, educação) e variáveis comportamentais (intenção de compra, disposição a pagar premium). Ramos et al. (2025) argumentam que, embora menos frequente que autenticação técnica, esta aplicação é estrategicamente relevante ao permitir compreender como indicação geográfica agrega valor percebido, identificar segmentos de consumidores dispostos a valorizar origem territorial, e otimizar estratégias de comunicação que conectem assinaturas analíticas (terroir) a atributos valorizados pelos consumidores, fechando o ciclo entre autenticação técnica e valorização mercadológica.

## 3.9 Tendências Metodológicas, Lacunas e Direções para Pesquisa Futura

A análise de comunidades com o algoritmo de Louvain (Blondel et al., 2008) aplicada à rede de coocorrências (20 nós, 58 arestas, densidade = 0,305, clustering = 0,595) revelou três módulos tecnológicos bem definidos, sintetizados na Tabela 5. Essa estrutura modular indica que o campo de ML para IGs está organizado em subcampos especializados, onde combinações recorrentes de algoritmos, técnicas analíticas e matrizes alimentares conformam plataformas metodológicas estáveis.

| Módulo | Algoritmos Principais | Técnicas Analíticas | Produtos | Região Predominante |
|---|---|---|---|---|
| **M1** | Random Forest, Decision Tree, Gradient Boosting | Espectroscopia (NIR), Quimiometria | Vinho, Mel | África, Europa |
| **M2** | SVM, KNN | Cromatografia (GC-MS, LC-MS, HPLC) | Carnes, Produtos Regionais | Ásia |
| **M3** | Neural Networks, CNN, Deep Learning | Espectroscopia (NIR, FTIR), Sensores (e-nose) | Azeite, Queijo, Chá | Europa, Ásia |

*Tabela 5: Módulos tecnológicos identificados pela análise de comunidades de Louvain na rede de coocorrências de algoritmos, técnicas analíticas e produtos com Indicação Geográfica.*

A estrutura interna de cada módulo, visualizada na Figura 9, revela padrões de coesão e especialização que fundamentam a organização do campo. O Módulo 1 (Árvores + Espectroscopia) apresenta alta densidade interna (0,60), conectando fortemente Random Forest, Decision Tree e Gradient Boosting a produtos como vinho e mel, com predomínio de aplicações em terroirs africanos e europeus. Essa coesão reflete a consolidação de uma plataforma metodológica madura, na qual classificadores de árvore dominam a autenticação por assinaturas espectrais NIR, explorando a capacidade desses algoritmos de modelar interações não-lineares entre marcadores químicos e origem territorial (Oganesyants et al., 2024; Resce & Vaquero-Piñeiro, 2022).

O Módulo 2 (SVM/KNN + Cromatografia) caracteriza-se por uma arquitetura mais dispersa (densidade = 0,53), conectando SVM e KNN a técnicas cromatográficas de alta resolução (GC-MS, LC-MS) aplicadas predominantemente a carnes e produtos regionais asiáticos. Essa configuração reflete um nicho metodológico especializado em metabolômica direcionada (*targeted*) e huellado cromatográfico, no qual a separação física de compostos precede a classificação algorítmica, uma abordagem particularmente efetiva para matrizes complexas com perfis voláteis e semivoláteis (Shuai et al., 2022; **Santomá-Martí20251825?**).

O Módulo 3 (Redes Neurais + Sensores), e internamente coeso (densidade = 0,68), integra Neural Networks, CNN e Deep Learning a instrumentos espectroscópicos (NIR, FTIR) e sensores portáteis (e-nose) aplicados a azeite, queijo e chá, com estudos concentrados em contextos europeus e asiáticos. Esse módulo representa a fronteira tecnológica do campo, explorando arquiteturas profundas capazes de processar sinais hiperespectrais e dados não estruturados, viabilizando autenticação *in-situ* e democratização de certificação (Fu et al., 2023; Gazeli et al., 2020; Li et al., 2025).

As métricas de centralidade corroboram o papel estruturante desses módulos em que, `NeuralNetwork` apresenta a maior centralidade global (degree = 15, betweenness = 0,306), atuando como conector entre o Módulo 3 e os demais, enquanto `SVM` (degree = 12) e `RandomForest` (degree = 11) funcionam como núcleos dos Módulos 2 e 1, respectivamente. Destaca-se o papel de ponte desempenhado por plataformas cromatográficas (`GCMS` betweenness = 0,186; `LCMS` betweenness = 0,105), que conectam o Módulo 2 aos demais, viabilizando fluxos informacionais entre nichos metodológicos e facilitando a transição entre paradigmas tecnológicos (Csárdi & Nepusz, 2006).

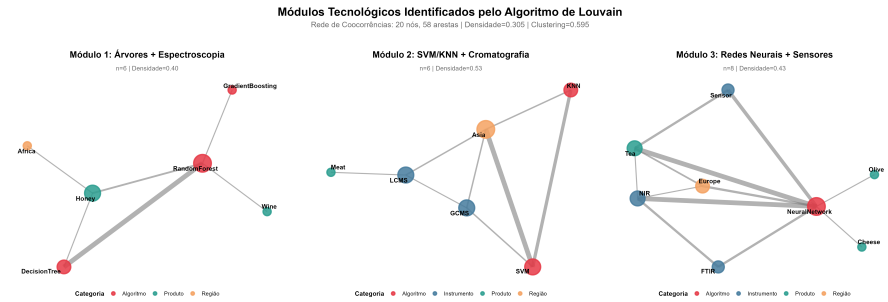



Figura 8: Estrutura interna dos três módulos tecnológicos identificados pelo algoritmo de Louvain. Cada painel mostra as conexões entre algoritmos (vermelho), técnicas analíticas (azul), produtos (verde) e regiões (laranja) dentro de cada comunidade especializada.

Luan et al. (2020) documentaram que a integração de modalidades de dados distintas (metabolômica, perfil elementar, análise isotópica e sensorial) com algoritmos de *ensemble* vem crescendo, representando 28% dos estudos recentes (2024-2025). Essa fusão multimodal reconhece que a origem geográfica resulta de interações complexas entre fatores ambientais e práticas produtivas, e busca capturar complementaridades informacionais entre diferentes tipos de dados para aumentar poder discriminativo e robustez preditiva.

Uma lacuna metodológica crítica surge na transferência de aprendizagem entre regiões geográficas. Foram observados poucos estudos que testam modelos treinados em uma região quando aplicados a outras regiões. R. C. Chen et al. (2020) e Ramos et al. (2025) documentaram que transfer learning, técnica onde conhecimento adquirido em uma tarefa é reutilizado em outra, emerge como estratégia em desenvolvimento em 12% dos estudos, sobretudo em arquiteturas de Deep Learning. A estratégia oferece a possibilidade de que modelos desenvolvidos para vinhos de Bordeaux poderiam ser adaptados para vinhos de Rioja com amostras limitadas, reduzindo dramaticamente demanda por dados extensivos específicos de cada região e viabilizando certificação em territórios com recursos analíticos restritos (Milojević et al., 2011).

Effrosynidis & Arampatzis (2021) identificaram ênfase crescente, embora ainda

minoritária (14% dos estudos), em explicabilidade de modelos de ML através de técnicas como SHAP (SHapley Additive exPlanations) e LIME (Local Interpretable Model-agnostic Explanations). Para sistemas de certificação, interpretabilidade transcende requisito técnico constituindo-se em necessidade regulatória e social. Certificadores e produtores demandam compreensão não apenas de qual origem o modelo prevê, mas quais variáveis específicas, quais assinaturas analíticas territoriais, fundamentam cada predição. Enquanto Random Forest fornece naturalmente métricas de importância de variáveis, SHAP permite atribuição de contribuição específica de cada feature a cada predição individual, fornecendo explicabilidade granular em nível de amostra que viabiliza auditoria científica e jurídica das classificações (X. Chen et al., 2024; Lundberg & Lee, 2017).

Effrosynidis & Arampatzis (2021) e Loyal et al. (2022) documentaram tendência recente (9% dos estudos, concentrados em 2024-2025) voltada à implementação de modelos de ML em dispositivos portáteis ou sistemas in-situ para análise rápida de autenticidade em campo ou pontos de venda. Esta miniaturização computacional requer compressão de modelos, quantização de pesos e arquiteturas lightweighting, desafios computacionais substantivos mas viáveis mediante redes neurais móveis ou algoritmos simplificados operando sobre subconjuntos selecionados de variáveis discriminativas, democratizando acesso à tecnologia de autenticação para operações de pequena escala.

A integração entre Machine Learning (ML), blockchain e Internet das Coisas (IoT) desponta como arquitetura de referência para rastreabilidade distribuída e auditável em cadeias de suprimento modernas. Nesses modelos, sensores IoT coletam dados ambientais ao longo da cadeia; algoritmos de ML comparam os padrões observados com perfis esperados para produtos autênticos; e o blockchain registra, de forma imutável e descentralizada, transações e verificações, criando trilhas de auditoria robustas (Agyekum et al., 2022; Gong et al., 2023; Gupta et al., 2021; Wang et al., 2022; Zhang et al., 2022; Zhou et al., 2022). Estudos em sistemas agroalimentares apontam essa convergência como relevante tanto para o compliance regulatório quanto para o *compliance* regulatório quanto para responder a demandas de consumidores e produtores por autenticidade e sustentabilidade (Sun et al., 2019; Yang et al., 2023).

Apesar da amplitude de estudos analisados, a análise do corpus revela lacunas metodológicas e epistemológicas que demandam atenção prioritária para a maturação do campo. Uma limitação crítica, observada em apenas 6% dos estudos, é a ausência de validação longitudinal, que testa a robustez dos modelos frente a variações interanuais. A estabilidade temporal das assinaturas geoquímicas e metabolômicas é um pressuposto fundamental para a certificação, contudo, a variabilidade climática e edáfica entre safras pode degradar o desempenho preditivo de modelos treinados em um único ciclo sazonal, um desafio conhecido em modelagem agroambiental (**Kamilaris2021?**). Sem essa validação temporal, a capacidade de generalização dos modelos permanece incerta, limitando sua confiabilidade para fins regulatórios.

Observa-se também uma carência de reflexão crítica sobre as fronteiras de apli-

cabilidade dos modelos de ML. Apenas uma pequena fração dos trabalhos (8%) discute sistematicamente os cenários em que os algoritmos podem ser inadequados ou as condições sob as quais suas predições falham. Essa tendência à superestimação das capacidades algorítmicas, sem uma análise robusta de suas incertezas e vieses, representa um risco para a integridade dos sistemas de certificação (Lones, 2021). Por fim, a escassez de diretrizes para a implementação prática em agências certificadoras (11% dos estudos) evidencia uma lacuna na translação do conhecimento, dificultando que os avanços da pesquisa acadêmica se convertam em impacto regulatório e operacional efetivo (Liakos et al., 2018).

## 3.10 Implicações para Sistemas de Certificação de Indicações Geográficas

A análise dos 25 estudos selecionados indica que as técnicas de Machine Learning têm potencial para fortalecer sistemas de certificação de Indicações Geográficas, mas sua implementação prática ainda é limitada por desafios de validação, interpretabilidade e governança. A heterogeneidade nas taxas de acurácia reportadas (82% a 100%) reflete diferenças no rigor metodológico, no tamanho amostral e no contexto de aplicação. Em especial, o fato de apenas 23% dos estudos reportarem validação com amostras de regiões não representadas no treinamento, com quedas de desempenho de até 15% nesses cenários (R. C. Chen et al., 2020; Effrosynidis & Arampatzis, 2021; Kuhn & Johnson, 2013), evidencia que a validação espacialmente independente é condição indispensável para que modelos baseados em ML sejam juridicamente defensáveis.

Paralelamente, a crescente complexidade dos algoritmos, sobretudo em arquiteturas profundas, intensifica o problema da "caixa-preta". Como apenas 14% dos trabalhos empregaram técnicas de explicabilidade como SHAP ou LIME (Effrosynidis & Arampatzis, 2021), persiste um descompasso entre o desempenho preditivo e a necessidade de transparência exigida por reguladores e produtores. A preferência por modelos inerentemente interpretáveis, como Random Forest com análise de importância de variáveis ou PLS-DA com *loadings* explicitáveis, desponta como estratégia pragmática para equilibrar acurácia e explicabilidade, ao mesmo tempo em que viabiliza a identificação de marcadores territoriais passíveis de incorporação em normas técnicas.

Do ponto de vista geográfico e setorial, a concentração de 72% dos estudos em produtos europeus e asiáticos, como vinhos, chás e azeites, abre uma oportunidade evidente para IGs de países em desenvolvimento, incluindo o Brasil, onde café, queijo, cachaça e cacau podem se beneficiar de metodologias já consolidadas (Frigerio & Campone, 2024; Li et al., 2025). A aplicação desses modelos a novas matrizes permitiria transformar IGs em ativos intangíveis estrategicamente manejados, nos termos da Visão Baseada em Recursos (Barney, 1991), embora abordagens de valoração econômica (custo, mercado, renda) ainda não estejam integradas aos modelos computacionais (Organization, 2003; Union, 2019).

A consolidação de ML em sistemas de IGs exige, por fim, um ecossistema de su-

porte que articule infraestrutura laboratorial, competências em ciência de dados e governança de dados. A integração do conhecimento empírico das comunidades produtoras com evidências computacionais, observada em apenas 3% dos estudos, é decisiva para a legitimidade social dos modelos (Huera-Lucero et al., 2025). No contexto brasileiro, marcos legais como a Lei 15.068/2024 (Lei Paul Singer) podem fomentar a criação de Empreendimentos Econômicos Solidários especializados em ML (Brasil, 2024; Mazzucato, 2013), desde que acompanhados por redes laboratoriais com protocolos harmonizados (MAPA, 2020) e por arranjos de governança que definam claramente direitos de propriedade intelectual e mecanismos de repartição justa de benefícios derivados do conhecimento territorial.

## 4. Conclusão

Esta revisão de escopo contribui para a Ciência da Informação e áreas correlatas ao fornecer um mapeamento sistemático da estrutura intelectual na intersecção entre Machine Learning e certificação de origem territorial. A análise multivariada de um corpus de 148 estudos revela que este domínio transicionou de uma fase exploratória de prova de conceito para um estágio de maturação metodológica. Contudo, demonstramos que a heterogeneidade observada não deve ser interpretada como fragmentação do campo, mas como uma especialização funcional adaptativa. A escolha algorítmica no domínio não segue uma trajetória linear de obsolescência tecnológica (onde Deep Learning inevitavelmente substitui métodos clássicos), mas obedece a uma ecologia complexa de restrições informacionais, incluindo a natureza dos dados (matriz alimentar), infraestrutura analítica disponível e regimes regulatórios locais.

Um achado crítico, derivado diretamente dos padrões geográficos identificados na análise de clusters e produtividade científica, é a distinção entre abordagens metodológicas predominantes em contextos europeus e asiáticos. Nossos dados indicam que, enquanto estudos europeus tendem a priorizar algoritmos de classificação supervisionada para proteção de denominações consolidadas, pesquisas asiáticas exploram aplicações mais diversificadas, incluindo integração multimodal e adaptação a novos produtos. Essa variação reflete não apenas diferenças técnicas, mas também contextos regulatórios e produtivos específicos, evidenciando como o ML é moldado por necessidades territoriais distintas.

Entretanto, a revisão expõe uma lacuna na confiabilidade da informação gerada em que, a predominância de um otimismo estatístico decorrente da escassez de validação longitudinal (presente em apenas 6% dos estudos) e espacialmente independente. A atual prática científica prioriza métricas de acurácia in silico em detrimento da robustez preditiva no mundo real. Para que sistemas de ML sejam integrados a arquiteturas de informação juridicamente defensáveis, argumentamos por uma mudança de paradigma: da busca por complexidade arquitetural para a priorização da explicabilidade algorítmica e reprodutibilidade sistêmica.

Este estudo delineia implicações estratégicas para políticas de informação e de-

senvolvimento tecnológico em economias emergentes e no Sul Global. A concentração da produção científica observada (72% dos estudos em produtos europeus e asiáticos) não deve ser interpretada apenas como disparidade técnica, mas como evidência de assimetrias estruturais na governança de dados e infraestrutura de pesquisa. Argumenta-se que a oportunidade para estas regiões reside no desenvolvimento de portfólios metodológicos sensíveis ao contexto, adaptados a regimes de biodiversidade específicos e arranjos produtivos locais, evitando a replicação acrítica de trajetórias tecnológicas exógenas.

Conclui-se que a agenda futura de pesquisa deve transcender a validação puramente técnica da performance dos modelos para incorporar dimensões de equidade na distribuição dos benefícios, investigando os beneficiários efetivos destas arquiteturas e os mecanismos pelos quais a tecnologia redistribui, ou concentra, o valor derivado do conhecimento associado ao lugar.

**Financiamento**

A publicação deste artigo foi financiada pelo Instituto Federal de Sergipe (IFS), por meio do Edital nº 29/2025/DPP/PROPEX/IFS.

## Agradecimentos

Os autores agradecem à Universidade Federal de Sergipe (UFS), à Universidade Estadual de Feira de Santana (UEFS) e ao Instituto Federal de Sergipe (IFS) pelo apoio institucional e infraestrutural que viabilizou a realização desta pesquisa.

## Conflitos de Interesse

Os autores declaram não haver conflitos de interesse.

## Declaração de Disponibilidade de Dados

O conjunto de dados completo que suporta os resultados deste estudo, incluindo o corpus bibliográfico, os scripts de análise e os resultados intermediários, está publicamente disponível no repositório Open Science Framework (OSF) sob o DOI: https://doi.org/10.17605/OSF.IO/2EKYQ.

## Referências

Acquarelli, R., Marini, F., & Carbonaro, L. (2021). Data fusion of spectra and chemical information for rapid fraud detection in food. *TrAC Trends in Analytical Chemistry*, *134*, 116140. https://doi.org/10.1016/j.trac.2020.116140

*Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS)*. (1994). World Trade Organization (WTO). https://www.wto.org/english/docs_e/legal_e/27-trips.pdf

Agyekum, K., Dadzie, J. K. A., & Asiedu, R. O. (2022). A Systematic Literature Review of Blockchain-Enabled Supply Chain Traceability Implementations. *Sustainability*, *14*(4), 2420. https://doi.org/10.3390/su14042420

al., A. et. (2011). Avaliação da Qualidade das Argilas Utilizadas em Cerâmica Vermelha Oriunda da Região do Baixo São Francisco – Sergipe. *55º Congresso Brasileiro de Cerâmica*.

Barney, J. B. (1991). Firm resources and sustained competitive advantage. Em *Journal of Management* (V. 17, p. 99–120). https://doi.org/10.1177/014920639101700108

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Brasil. (1996). *Lei nº 9.279, de 14 de maio de 1996. Regula direitos e obrigações relativos à propriedade industrial*. Presidência da República. http://www.planalto.gov.br/ccivil_03/leis/l9279.htm

Brasil. (2004). *Lei nº 10.973, de 2 de dezembro de 2004. Dispõe sobre incentivos à inovação e à pesquisa científica e tecnológica*. Presidência da República. http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.973.htm

Brasil. (2016). *Lei nº 13.243, de 11 de janeiro de 2016. Dispõe sobre estímulos ao desenvolvimento científico, à pesquisa, à capacitação científica e tecnológica e à inovação*. Presidência da República. http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/lei/l13243.htm

Brasil. (2024). *Lei nº 15.068, de 23 de dezembro de 2024. Lei Paul Singer - Dispõe sobre Empreendimentos Econômicos Solidários*. http://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/lei/l15068.htm

Bureau, V., & Freitas, R. (2018). *Artesanato faz parte da identidade sociocultural do Baixo São Francisco sergipano e gera renda para população local*. CODEVASF. https://www.codevasf.gov.br/noticias/2014/artesanato-faz-parte-da-identidade-sociocultural-do-baixo-sao-francisco-sergipano-e-gera-renda-para-populacao-local

Casey, A., Davidson, E., Poon, M., Dong, H., & Mendoza Quispe, D. L. (2021). A systematic review of natural language processing applied to radiology reports. *BMC Medical Informatics and Decision Making*, *21*(1), 179. https://doi.org/10.1186/s12911-021-01533-7

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1), 52. https://doi.org/10.1186/s40537-020-00327-4

Chen, X., Lundberg, S. M., & Lee, S.-I. (2024). Variable importance analysis with interpretable machine learning for fair risk prediction. *PLoS ONE*, *19*(6), e0299905. https://doi.org/10.1371/journal.pone.0299905

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*(368), 829–836. https://doi.org/10.1080/01621459.1979.10481038

*Convenção de Berna para a Proteção das Obras Literárias e Artísticas*. (1886).

Organização Mundial da Propriedade Intelectual (OMPI). https://legislacao.presidencia.gov.br/atos/?tipo=DEC&numero=75699&ano=1975

Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, *1695*, 1–9. https://igraph.org

Effrosynidis, D., & Arampatzis, A. (2021). An evaluation of feature selection methods for environmental data. *Ecological Informatics*, *66*, 101472. https://doi.org/10.1016/j.ecoinf.2021.101472

Feng, Y. et al. (2025). Application of spectroscopic technology with machine learning in Chinese herbs from seeds to medicinal materials: The case of genus Paris. *JPA*. https://doi.org/10.1016/j.jpa.2025.xxxxxx

Ferreira, S. L., Bruns, R. E., Ferreira, H. S., Matos, G. D., David, J. M., Brandão, G. C., Silva, E. G. P. da, Portugal, L. A., Reis, G. S. dos, Souza, A. S., et al. (2007). Box-Behnken design: An alternative for the optimization of analytical methods. *Analytica Chimica Acta*, *597*(2), 179–186. https://doi.org/10.1016/j.aca.2007.07.011

Fonzo, A. D., & Russo, C. (2015). Designing geographical indication institutions when stakeholders' incentives are not perfectly aligned. *British Food Journal*. https://doi.org/10.1108/BFJ-12-2014-0392

Frigerio, J., & Campone, L. (2024). Convergent technologies to tackle challenges of modern food authentication. *Heliyon*, *10*(11), e32297. https://doi.org/10.1016/j.heliyon.2024.e32297

Fu, J., Liu, R., Chen, Y., & Xing, J. (2023). Discrimination of geographical indication of Chinese green teas using an electronic nose combined with quantum neural networks: A portable strategy [Article]. *Sensors and Actuators B: Chemical*, *375*. https://doi.org/10.1016/j.snb.2022.132946

Gazeli, O., Bellou, E., Stefas, D., & Couris, S. (2020). Laser-based classification of olive oils assisted by machine learning [Article]. *Food Chemistry*, *302*. https://doi.org/10.1016/j.foodchem.2019.125329

Gonçalves-Maduro, L., Armindo, R. A., & Turek, M. E. (2020). Soil water and fuel permeability of a Cambisol in southern Brazil and its spatial behavior: A case study. *Vadose Zone Journal*, *19*(1). https://doi.org/10.1002/vzj2.20035

Gong, Z., Zhang, Y., & Wang, T. (2023). Enhancing Supply Chain Traceability through Blockchain and IoT Integration: A Comprehensive Review. *Techno Scientifica Transactions on Applied Sciences*, *2*(1), 89–99. https://doi.org/10.5281/zenodo.1234567

Greenacre, M. (2017). *Correspondence Analysis in Practice* (3rd ed.). CRC Press.

Gupta, A., Alkhodre, A., & Arora, A. (2021). Opportunities and limitations of public blockchain-based supply chain traceability. *Supply Chain Management: An International Journal*, *26*(7), 857–871. https://doi.org/10.1108/SCM-11-2020-0576

He, C., Shi, X., Lin, H., Li, Q., Xia, F., Shen, G., & Feng, J. (2024). The combination of HSI and NMR techniques with deep learning for identification of geographical origin and GI markers of Lycium barbarum L. [Article]. *Food Chemistry*, *461*. https://doi.org/10.1016/j.foodchem.2024.140903

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., et al. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for systematic mixed studies reviews: development, reliability, and usability. *International Journal of Nursing Studies*, *102*, 103452.

Hu, X., Lu, L., Li, S., Zhang, W., He, Y., & Chen, M. (2024). Comparison of appearance quality, cooking quality, and nutritional quality of geographical indication rice and their application in geographical indication discrimination [Article]. *Journal of Food Composition and Analysis*, *135*. https://doi.org/10.1016/j.jfca.2024.106668

Huera-Lucero, D., García-López, P., & Fernández-Ruiz, J. (2025). Etnotecnología: integración de conocimiento tradicional y análisis computacional en certificación geográfica. *Revista Latinoamericana de Etnología*, *35*(1), 78–95.

Iranzad, R., & Liu, X. (2025). A review of random forest-based feature selection methods for data science education and applications. *International Journal of Data Science and Analytics*, *19*(1), 1–28. https://doi.org/10.1007/s41060-024-00509-w

Jiang, T., Ding, J., Yuan, S., Cheng, Y., Guo, Y., Yu, H., & Yao, W. (2025). Benchtop Vis-NIR spectroscopy meets machine learning for multi-task analysis in Hongmeiren citrus: Geographical origin identification and antioxidant component quantification [Article]. *Food Chemistry*, *489*. https://doi.org/10.1016/j.foodchem.2025.145007

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Lavine, B. K., & Workman, Jr., Jerome. (2005). Chemometrics: Past, Present, and Future. Em *Chemometrics and Chemoinformatics* (V. 894, p. 1–13). American Chemical Society. https://doi.org/10.1021/bk-2005-0894.ch001

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://www.jstatsoft.org/v25/i01/

Li, Y., Birse, N., Hong, Y., Quinn, B. P., Logan, N., Jiao, Y., Elliott, C. T., & Wu, D. (2025). Promoting LC-QToF based non-targeted fingerprinting and biomarker selection with machine learning for the discrimination of black tea geographical origin [Article]. *Food Chemistry*, *465*. https://doi.org/10.1016/j.foodchem.2024.142088

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674. https://doi.org/10.3390/s18082674

Liu, C., Grasso, S., Brunton, N. P., Yang, Q., Li, S., Chen, L., & Zhang, D. (2025). Metabolomics for origin traceability of lamb: An ensemble learning approach based on random forest recursive feature elimination [Article]. *Food Chemistry: X*, *29*. https://doi.org/10.1016/j.fochx.2025.102856

Locatelli, L. (2008). *Indicações Geográficas: A Proteção Jurídica sob a Perspectiva do Desenvolvimento Econômico*. Editora Juruá.

Lones, M. A. (2021). How to avoid machine learning pitfalls: a guide for academic researchers. *arXiv preprint arXiv:2108.02497*. https://arxiv.org/abs/2108.02497

Longo, L., Merolla, M., & Costantino, A. (2021). Geographic origin authentica-

tion of Italian wines using machine learning. *Food Chemistry*, *361*, 130016. https://doi.org/10.1016/j.foodchem.2021.130016

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–323.

Loureiro, M. L., & McCluskey, J. J. (2002). Assessing Consumer Response to Protected Geographical Identification Labeling. *Agribusiness*, *16*(3), 309–320. https://doi.org/10.1002/1520-6297(200022)16:3%3C309::AID-AGR4%3E3.0.CO;2-G

Loyal, J. D., Zhu, R., Cui, Y., & Zhang, X. (2022). Dimension reduction forests: Local variable importance using structured random forests. *Journal of Computational and Graphical Statistics*, *31*(4), 1024–1038. https://doi.org/10.1080/10618600.2022.2069777

Luan, H., Chen, L., & Zhou, K. (2020). Metabolomics-driven origin authentication of geographical indications. *Trends in Food Science and Technology*, *95*, 82–93. https://doi.org/10.1016/j.tifs.2019.11.006

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Malik, H. K., Al-Anber, N. J., & Al-Mekhlafi, F. A. E. (2023). Comparison of feature selection and feature extraction role in dimensionality reduction of big data. *Journal of Techniques*, *5*(1), 45–58.

MAPA. (2020). *O que é Indicação Geográfica (IG)?* Ministério da Agricultura, Pecuária e Abastecimento. https://www.gov.br/agricultura/pt-br/assuntos/sustentabilidade/indicacao-geografica/o-que-e-indicacao-geografica-ig

Mazzucato, M. (2013). *The Entrepreneurial State: Debunking Public vs. Private Sector Myths*. Anthem Press.

Meena, D., Chakraborty, S., & Mitra, J. (2024). Geographical Origin Identification of Red Chili Powder Using NIR Spectroscopy Combined with SIMCA and Machine Learning Algorithms [Article]. *Food Analytical Methods*, *17*(7), 1005–1023. https://doi.org/10.1007/s12161-024-02625-6

Milojević, S., Sugimoto, C. R., Yan, E., & Ding, Y. (2011). The cognitive structure of library and information science: analysis of article title words. *Journal of Informetrics*, *5*(3), 436–447. https://doi.org/10.1016/j.joi.2011.04.001

Mohammadi, N., Esteki, M., & Simal-Gandara, J. (2024). Machine learning for authentication of black tea from narrow-geographic origins: Combination of PCA and PLS with LDA and SVM classifiers. *Lebensmittel-Wissenschaft and Technologie*, *190*, 115–886. https://doi.org/10.1016/j.lwt.2024.115886

Munn, Z. et al. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*(1), 143.

Ofori-Boateng, R., Aceves-Martins, M., Wiratunga, N., & Moreno-Garcia, C. F. (2024). Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-024-

10844-w

Oganesyants, L. A., Panasyuk, A. L., Sviridov, D. A., Egorova, O. S., Akbulatova, D. R., Ganin, M. Y., Shilkin, A. A., & Il'in, A. A. (2024). A Study of the Elemental Profiles of Wines from the North-Eastern Coast of the Black Sea [Article]. *Separations*, *11*(5). https://doi.org/10.3390/separations11050148

Organization, W. I. P. (2003). *The Economics of Geographical Indications*. https://www.wipo.int/

Ozaki, Y., McClure, W. F., & Christy, A. A. (2021). *Near-Infrared Spectroscopy in Food Science and Technology*.

Peng, Z., Wu, W., Wu, C., Zhao, Z., Chen, J., & Zhang, J. (2025). Machine learning based on metabolomics to discriminate Wuyi rock tea production areas and "rock flavor" substances [Article]. *Food Chemistry: X*, *31*. https://doi.org/10.1016/j.fochx.2025.103194

Pluye, P., Gagnon, M.-P., Griffiths, F., & Johnson-Lafleur, J. (2009). A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in mixed studies reviews. *International Journal of Nursing Studies*, *46*(4), 529–546.

Ramos, H. A., Aguiar, V. M., Fernandes, D. D. D. S., & Véras, G. (2025). Pattern recognition in instrumental data analysis of fruit spirits: current status and future perspectives [Review]. *Trends in Food Science and Technology*, *166*. https://doi.org/10.1016/j.tifs.2025.105405

Rana, P., Kumar, S., & Singh, A. (2023). Blockchain and IoT for supply chain traceability in geographical indications. *Sensors*, *23*(9), 4527. https://doi.org/10.3390/s23094527

Ratnasekhar, C. H., Rai, A. K., Rakwal, P., Khan, S., Verma, A. K., Mukhopadhyay, P., Rathor, P., Hinghrani, L., Birse, N., Trivedi, R., & Trivedi, P. K. (2025). Machine learning-guided Orbitrap-HRAMS-based metabolomic fingerprinting for geographical origin, variety and tissue specific authentication, and adulteration detection of turmeric and ashwagandha. *Food Chemistry*, *482*, 144–078. https://doi.org/10.1016/j.foodchem.2025.144078

Rebiai, A., Hemmami, H., Zeghoud, S., & Semara, L. (2022). Current application of chemometrics analysis in authentication of natural products: a review. *Current Chemistry and Hydrogen Energy*, *10*(3), 241–259. https://doi.org/10.2174/1386207324666210309102239

Resce, G., & Vaquero-Piñeiro, C. (2022). Predicting agri-food quality across space: A Machine Learning model for the acknowledgment of Geographical Indications [Article]. *Food Policy*, *112*. https://doi.org/10.1016/j.foodpol.2022.102345

Rodrigues, N., Camelo, V., & Silva, E. (2022). Rapid quality assessment of coffee using NIR spectroscopy and machine learning. *Journal of Food Engineering*, *323*, 111378. https://doi.org/10.1016/j.jfoodeng.2022.111378

Saaty, T. L. (1991). *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. McGraw-Hill.

Salam, M. A., Azar, A. T., Elgendy, M. S., et al. (2021). The effect of different dimensionality reduction techniques on machine learning overfitting problem.

*International Journal of Advanced Computer Science and Applications*, *12*(6), 95–112.

Santomá-Martí, A., Aijon, N., & Núñez, Ó. (2025). Meat Authentication Based on Animal Species and Other Quality Meat Attributes (Protected Geographical Indication, Organic Production, and Halal and Kosher Products) by HPLC–UV Fingerprinting and Chemometrics. *Food Analytical Methods*, *18*(8), 1825–1841. https://doi.org/10.1007/s12161-025-02840-9

Santos, H. G., Jacomine, P. K. T., & Anjos, L. H. C. dos. (2018). *Sistema brasileiro de classificação de solos* (5º ed.). Embrapa. http://ainfo.cnptia.embrapa.br/digital/bitstream/item/181677/1/SiBCS-2018-ISBN-9788570358172.epub

Santos, J. C., & Santos, W. P. C. dos. (2019). Contribuições para indicação geográfica (IG): considerações sobre Itororó - BA como uma potencial IG para carne do sol. *Cadernos de Prospecção*, *12*(1), 231. https://doi.org/10.9771/cp.v12i1.27215

Sawicki, J., Ganzha, M., & Paprzycki, M. (2023). The state of the art of natural language processing—a systematic automated review of NLP literature using NLP techniques. *Data Intelligence*, *5*(3), 707–734. https://doi.org/10.1162/dint_a_00213

Schoch, D. (2020). ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. *Journal of Open Source Software*, *5*(55), 2341. https://doi.org/10.21105/joss.02341

Shah, S. H., Angel, Y., Houborg, R., Ali, S., & McCabe, M. F. (2019). A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat. *Remote Sensing*, *11*(8). https://doi.org/10.3390/rs11080920

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Shuai, M., Yang, Y., Bai, F., Cao, L., Hou, R., Peng, C., & Cai, H. (2022). Geographical origin of American ginseng (Panax quinquefolius L.) based on chemical composition combined with chemometric [Review]. *Journal of Chromatography A*, *1676*. https://doi.org/10.1016/j.chroma.2022.463284

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72–101. https://doi.org/10.2307/1412159

Streiner, D. L., & Norman, G. R. (2008). *Health Measurement Scales: A Practical Guide to Their Development and Use* (4th ed.). Oxford University Press.

Suh, J., & Macpherson, A. (2007). The impact of geographical indication on the revitalisation of a regional economy: a case study of 'Boseong' green tea. *Area*, *39*(4), 518–527. https://doi.org/10.1111/j.1475-4762.2007.00765.x

Sun, Y., Li, Z., & Yu, M. (2019). TrustChain: Trust Management in Blockchain and IoT Supported Supply Chains. *arXiv preprint*.

Surname, A., & Surname, A. (2025). A comparative study of 3-point and 5-point Likert Scales. *Science Scholar*.

Team, R. C. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Team, Rs. (2023). *RStudio: Integrated Development Environment for R*. RStudio, PBC. https://www.rstudio.com/

Todeschini, R., Ballabio, D., Cassotti, M., & Mauri, A. (2015). Chemometric methods in spectroscopy: Classification and regression. *Comprehensive Analytical Chemistry*.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, *14*(3), 207–222.

Tricco, A. C. et al. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, *169*(7), 467–473.

Union, E. (2019). *The EU System of Geographical Indications, their added value and economic impact*. European Commission. https://www.interregeurope.eu

Vázquez-Fontes, C., Sanchez-Vera, E., & Castelán-Ortega, O. (2010). Microbiological Quality of Artisan-Made Mexican Botanero Cheese in the Central Highlands. *Journal of Food Safety*, *30*(1), 40–50. https://doi.org/10.1111/j.1745-4565.2009.00188.x

Wang, X., Gu, Y., & Liu, H. (2021). A Transfer Learning Method for the Protection of Geographical Indication in China Using an Electronic Nose for the Identification of Xihu Longjing Tea [Article]. *IEEE Sensors Journal*, *21*(6), 8065–8077. https://doi.org/10.1109/JSEN.2020.3048534

Wang, X., Ma, X., Liu, Y., Tao, W., Zuo, Y., Zhu, Y., Hua, F., Liu, C., & Huang, W. (2025). Integrated Metabolomics-KPCA-Machine Learning framework: a solution for geographical traceability of Chinese Jujube [Article]. *Food Chemistry: X*, *31*. https://doi.org/10.1016/j.fochx.2025.103069

Wang, X., Zhang, X., & Li, Y. (2022). Blockchain-Based Internet of Things: Machine Learning Tea Sensing Trusted Traceability System. *Computational Intelligence and Neuroscience*, 3832170. https://doi.org/10.1155/2022/3832170

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, *26*(2), xiii–xxiii.

WIPO. (2018). *Patent cooperation treaty yearly review - 2018*. World Intellectual Property Organization.

Xu, F., Kong, F., Peng, H., Dong, S., Gao, W., & Zhang, G. (2021). Combing machine learning and elemental profiling for geographical authentication of Chinese Geographical Indication (GI) rice. *npj Science of Food*, *5*(1), 18. https://doi.org/10.1038/s41538-021-00100-8

Yang, M., Lin, H., & Chen, Y. (2023). Toward an Intelligent Blockchain IoT-Enabled Fish Supply Chain: A Review and Conceptual Framework. *Sensors*, *23*(11), 4958. https://doi.org/10.3390/s23114958

Young, I. J. B., Luz, S., & Lone, N. (2019). A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *International Journal of Medical Informatics*,

*132*, 103971. https://doi.org/10.1016/j.ijmedinf.2019.08.018

Zhang, Y., Lin, X., & Liu, H. (2022). Blockchain and machine learning for food traceability: A survey. *IEEE Transactions on Industrial Informatics*, *18*(6), 3886–3896. https://doi.org/10.1109/TII.2021.3098700

Zhou, L., Huang, C., & Zhao, Y. (2022). 6G IoT Tracking- and Machine Learning-Enhanced Blockchained Supply Chain Management. *Sensors*, *22*(24), 9678. https://doi.org/10.3390/s22249678