PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA DE CIÊNCIAS EXATAS E DA COMPUTAÇÃO
GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

CLASSIFICAÇÃO DE ORIGENS E PRINCÍPIOS ATIVOS EM *CINNAMOMUM VERUM* UTILIZANDO IMAGENS HIPERESPECTRAIS

MATHEUS SANCLÉ BUENO BARBOSA

GOIÂNIA
2020

MATHEUS SANCLÉ BUENO BARBOSA

CLASSIFICAÇÃO DE ORIGENS E PRINCÍPIOS ATIVOS EM *CINNAMOMUM VERUM* UTILIZANDO IMAGENS HIPERESPECTRAIS

Trabalho de Conclusão de Curso apresentado à Escola de Ciências Exatas e da Computação, da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para obtenção do título de Bacharel em Engenharia de Computação.

Orientador(a):

Prof. Dr. Arlindo Rodrigues Galvão Filho

Banca examinadora:

Prof. Dr. Clarimar José Coelho

Msc. Isaac Yves Lopes de Macêdo

Prof. Dr. Rafael Viana de Carvalho

GOIÂNIA
2020

MATHEUS SANCLÉ BUENO BARBOSA

CLASSIFICAÇÃO DE ORIGENS E PRINCÍPIOS ATIVOS EM *CINNAMOMUM VERUM* UTILIZANDO IMAGENS HIPERESPECTRAIS

Este Trabalho de Conclusão de Curso julgado adequado para obtenção o título de Bacharel em Engenharia de Computação, e aprovado em sua forma final pela Escola de Ciências Exatas e da Computação, da Pontifícia Universidade Católica de Goiás, em 03/12/2020.

_____

Prof.ª Ma. Ludmilla Reis Pinheiro dos Santos
Coordenador(a) de Trabalho de Conclusão de Curso

Banca examinadora:

_____

Orientador(a): Prof. Dr. Arlindo Rodrigues Galvão Filho

_____

Prof. Dr. Clarimar José Coelho

_____

Msc. Isaac Yves Lopes de Macêdo

_____

Prof. Dr. Rafael Viana de Carvalho

GOIÂNIA
2020

Dedico este trabalho à minha família, aos meus amigos pela estrutura, apoio e os momentos de descontração. Ao meu orientador e coorientador por toda ajuda e conhecimentos fornecidos para que eu chegasse até aqui.

AGRADECIMENTOS

Primeiramente a Deus, por me permitir saúde e determinação para a conclusão do trabalho.

Aos meus familiares e amigos, por todo apoio e por sempre me motivar neste caminho.

Ao Professor Dr. Arlindo Galvão, meu orientador, pelos ensinamentos, paciência e apoio na construção do trabalho.

Ao Msc. Isaac Yves, por juntamente com o LAFAM, disponibilizar as amostras e ser peça importante na realização do estudo.

Aos professores Dr. Clarimar José Coelho e Dr. Rafael Viana de Carvalho por participarem da banca.

A todos os integrantes do LCC e aos colegas de estudo por todo e convivência ímpar durante parte do ciclo acadêmico.

RESUMO

A análise sobre determinadas especiarias que potencialmente são benéficas a saúde é um ramo de pesquisa de grande interesse geral. No entanto, os métodos usuais de extração e análise para identificação e determinação dos componentes responsáveis pelo efeito benéfico, envolvem a utilização de solventes, princípios ativos e outras substâncias na amostra, o que gera um processo oneroso e destrutível. A caneleira-verdadeira (*Cinnamomum verum)* surge como um destas especiarias mais difundidas no mundo, sendo uma das variantes mais consumidas da canela e cultivada em diversas regiões, recebendo vários estudos que relacionam o seu consumo a efeitos anti-inflamatórios, antimicrobianos, antioxidantes, entre outros. Portanto, é proposto neste trabalho o uso de ferramentas e métodos computacionais, mais especificamente a utilização de imagens hiperespectrais, aliados a estratégia de aprendizado de máquina, para determinação da concentração de princípios ativos e sua possível relação com a região de cultivo da amostra. A Máquina de Vetores de Suporte (*SVM*) foi utilizada como estratégia de classificação, para os princípios ativos em amostras de canela, bem como na tentativa de identificação das amostras de canela em relação a sua origem. Os resultados se mostraram promissores para a determinação da concentração dos princípios ativos, tendo sido gerado um modelo com acurácia de 99%, embora sua relação com a região de cultivo não tenha ficado clara, tendo sido gerado um modelo com acurácia de 78%.

Palavras-Chave: *Canela. Cinnamomum verum. Imagem Hiperespectral. Aprendizado de Máquina. Máquina de Vetores de Suporte.*

ABSTRACT

The analysis of certain spices that are potentially beneficial to health is a branch of research of great general interest. However, the usual methods of extraction and analysis to identify and determine the components responsible for the beneficial effect, involve the use of solvents, active ingredients and other substances in the sample, which generates an expensive and destructible process. True cinnamon (Cinnamomum verum) appears as one of these spices most widespread in the world, being one of the most consumed variants of cinnamon and cultivated in several regions, receiving several studies that relate its consumption to anti-inflammatory, antimicrobial, antioxidant effects, among others. Therefore, it is proposed in this work the use of computational tools and methods, more specifically the use of hyperspectral images, combined with the machine learning strategy, to determine the concentration of active principles and its possible relationship with the region where the sample is grown. The Support Vector Machine (*SVM*) was used as a classification strategy for the active ingredients in cinnamon samples, as well as in the attempt to identify Cinnamon in relation to its origin. The results proved to be promising for determining the concentration of active ingredients, with a model with an accuracy of 99% being generated, although its relationship with the growing region has not been clear, having been generated a model with an accuracy of 78%.

Keywords: *Cinnamon. Cinnamomum verum. Hyperspectral image. Machine Learning. Support Vectors Machine.*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Especiarias são cultivadas e difundidas nas mais diversas regiões do mundo, e algumas delas se encaixam na definição de “alimento funcional”, que é definido por TAPSELL et al., (2006) como sendo os alimentos que fornecem benefícios à saúde, além do papel de nutrição básica. Tais especiarias são consumidas para realçar o sabor dos alimentos, atuando como agentes aromáticos, sendo geralmente utilizados em pequenas quantidades.

Várias doenças metabólicas (diabetes, obesidade) e distúrbios degenerativos (hipertensão, alguns tipos de câncer) do desenvolvimento humano ou relacionados à idade, são conhecidos por estarem associados a processos oxidativos no corpo. E conforme descrito por D’SOUZA et al., (2017), especiarias como a canela, o açafrão, a noz-moscada, entre outras, podem combater os danos oxidativos e prevenir a ocorrência de uma série de doenças, desenvolvendo imunidade inata, se consumidas de forma adequada. Portanto, compreender e demonstrar a aplicação destas especiarias por meios científicos continua sendo um desafio, principalmente quando comparados aos padrões aplicados para avaliação de outros princípios ativos benéficos à saúde.

A forma atual como estes componentes são estudados, inclui técnicas de extração dos princípios ativos (compostos responsáveis por efeitos biológicos), de maneira seletiva e sensível. Em que são utilizados métodos como extração em fase líquida (com a utilização de solventes), em fase sólida e extração com fluído supercrítico (com CO2 neste estado), comprovando que diferentes solventes ou a mistura destes, se aplicados em uma mesma amostra, levam a diferentes níveis de eficiência na extração (YASHIN et al., 2017). A aplicação destes métodos além de altamente onerosa, é destrutível à amostra.

Uma alternativa a estas análises, é uso de ferramentas computacionais no processo, como é citado em KIANI et al., (2019), onde grupos de pesquisa desenvolveram vários métodos ágeis e não destrutíveis baseados em visão computacional e espectroscopia. A utilização destes dois métodos pode ser implementada com o sistema de imagens hiperespectrais (FENG; SUN, 2012).

Tais métodos são simples, rápidos, de baixo custo e não destrutivos, permitindo a análise de informações sobre as moléculas em diferentes comprimentos de onda, refletindo a informação sobre as ligações químicas e os constituintes químicos das amostras (ZHANG et al., 2017). Essas técnicas exigem conhecimento profissional e habilidades operacionais voltadas à área computacional, o que facilita aqueles que desconhecem todo o processo que envolve os métodos químicos e cromatógrafos (KIANI et al., 2019).

No entanto, a utilização das imagens hiperespectrais auxilia em parte do processo, pois é necessário ainda que se dê sentido as informações obtidas. Métodos baseados em aprendizado de máquina e reconhecimento de padrões, têm desempenhado um papel importante em tarefas que envolvem a análise de imagens hiperespectrais, pois são capazes de aprender automaticamente a relação entre o espectro de refletância e as informações desejadas, ao mesmo tempo em que são robustos contra ruídos e incertezas presentes na amostra e no processo de aquisição (GEWALI; MONTEIRO; SABER, 2018) .

A canela-verdadeira (*Cinnamomum verum - CV*) é umas destas especiarias mundialmente difundidas, sendo consumida e cultivada em diversas regiões, sendo originária do Sri Lanka. Estudos *in vitro* e *in vivo*, em animais e humanos, de diferentes partes do mundo demonstraram vários efeitos benéficos da *CV* à saúde, como propriedades anti-inflamatórias, atividade antimicrobiana, redução de doenças cardiovasculares, aumento da função cognitiva e redução do risco de câncer de cólon (EMERITUS et al., 2016) .

A canela é uma eficaz fonte de antioxidantes, além de aumentar a eficácia de outros antioxidantes importantes, agindo para reduzir o estresse oxidativo. Tal potencial da canela é atribuído à variedade de compostos polifenólicos que ela possui (SHAHID et al., 2018). Os polifenóis por sua parte, são uma classe de antioxidantes amplamente disponíveis em alguns alimentos e especiarias, como a *CV*. Existem diversos tipos de polifenóis como catequinas, antocianinas, resveratrol, entre outros (ABDALI; SAMSON; GROVER, 2015).

Diante deste cenário, este trabalho propõe o estudo sobre a composição de amostras de *CV*, visando obter uma relação com sua região de origem, bem como determinar a concentração de determinados princípios ativos (resveratrol, catequina e ácido gálico) que auxiliem neste processo. Sendo utilizado o sistema de imagens hiperespectrais para obtenção das informações das amostras, aliado a estratégias de aprendizado de máquina para análise e classificação.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo informa sobre a utilização do sistema de imagens hiperespectrais para captura das imagens e seu conceito de funcionamento. Ainda, são exploradas as ferramentas e técnicas utilizadas no desenvolvimento deste trabalho.

2.1 SISTEMA DE IMAGEM HIPERESPECTRAL

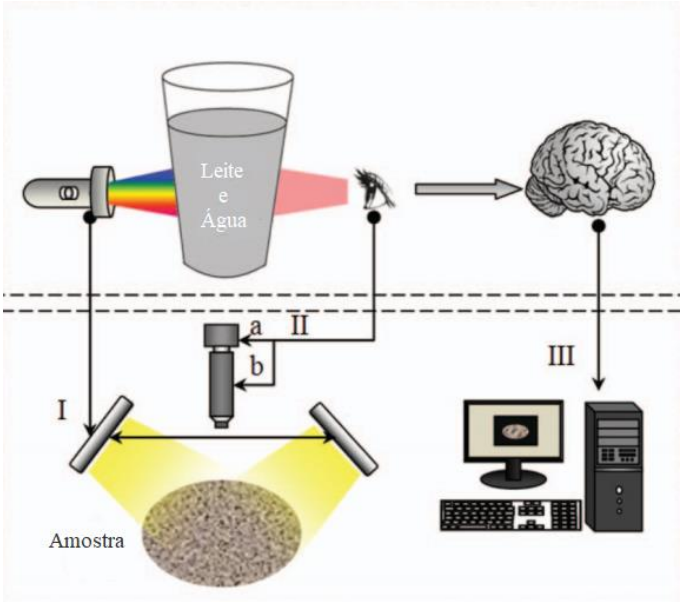
O princípio da imagem hiperespectral pode ser definido pela integração das teorias por trás da espectroscopia e visão computacional, podendo representar uma amostra por meio da combinação de seus espectros e imagens (FENG; SUN, 2012).

A espectroscopia no infravermelho próximo (*NIR*) é amplamente descrita como uma técnica rápida, econômica e não destrutiva, capaz de retornar estimativas confiáveis de várias propriedades físico-químicas em diferentes amostras (BAGCHI; SHARMA; CHATTOPADHYAY, 2016). A abordagem recente da espectroscopia *NIR* e outras técnicas espectroscópicas aplicadas à ciência alimentar se baseia na exploração dos espectros como uma impressão digital a ser analisada por quimiometria (VARRÀ et al., 2020).

Na visão computacional são extraídas informações quantitativas de cores de imagens digitais usando processamento e análise de imagens, resultando na obtenção de medições de cores rápidas e sem contato (WU; SUN, 2013).

Portanto, a técnica de imagem hiperespectral estende a capacidade da espectroscopia adicionando a dimensão espacial, utilizando imagens digitais convencionais no mesmo sistema, para fornecer ambas as informações espaciais e espectrais simultaneamente (KAMRUZZAMAN; SUN, 2016). O sistema de imagem hiperespectral consiste em três partes principais, uma fonte de luz, um dispositivo de dispersão de luz e unidades para captura e processamento das imagens, agindo como o olho e o cérebro humano (FENG; SUN, 2012). A Figura 1 exemplifica o processo:

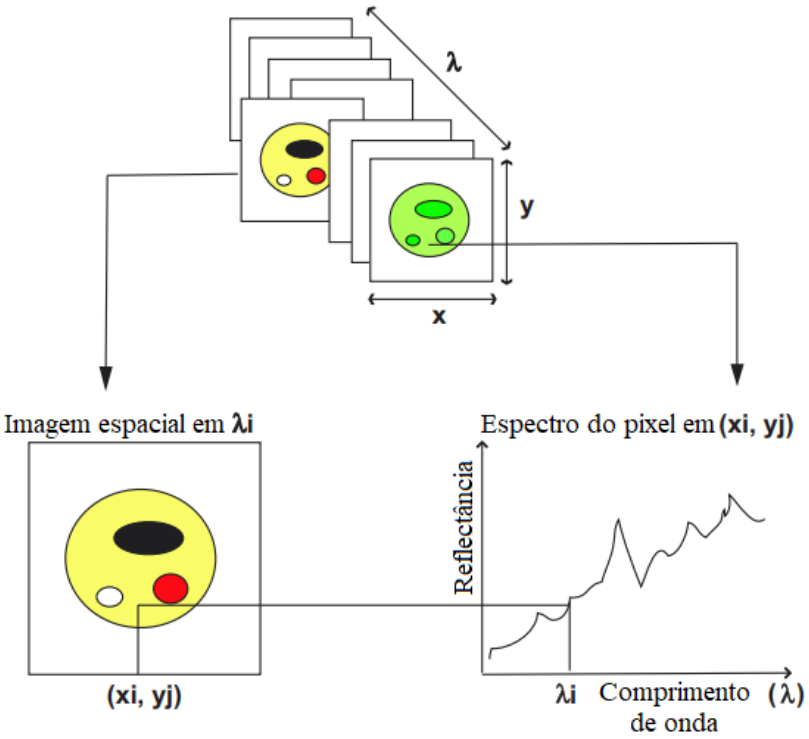
Figura 1: Configuração de um sistema de imagem hiperespectral. I: Fonte de Luz. II: Espectrógrafo (a) unidade de imagem; b) aparato para dispersão dos comprimentos de onda). III: Informação processada.



Fonte: Adaptado de (FENG; SUN, 2012).

A estrutura de dados resultante da aquisição das imagens hiperespectrais é chamada de "hipercubo", pois pode ser ilustrada como contendo três dimensões, com duas para coordenadas espaciais e a outra para valores espectrais. Consequentemente, I (x, y, λ) é denotado, onde x e y são referentes as posições dos pixels na imagem, com λ_i para o comprimento de onda, sendo i o comprimento específico normalmente medido em nanômetros (FENG; SUN, 2012). A Figura 2 demonstra a estrutura de dados resultante:

Figura 2: Representação esquemática do hipercubo, demonstrando a relação entre as dimensões espaciais e espectrais.

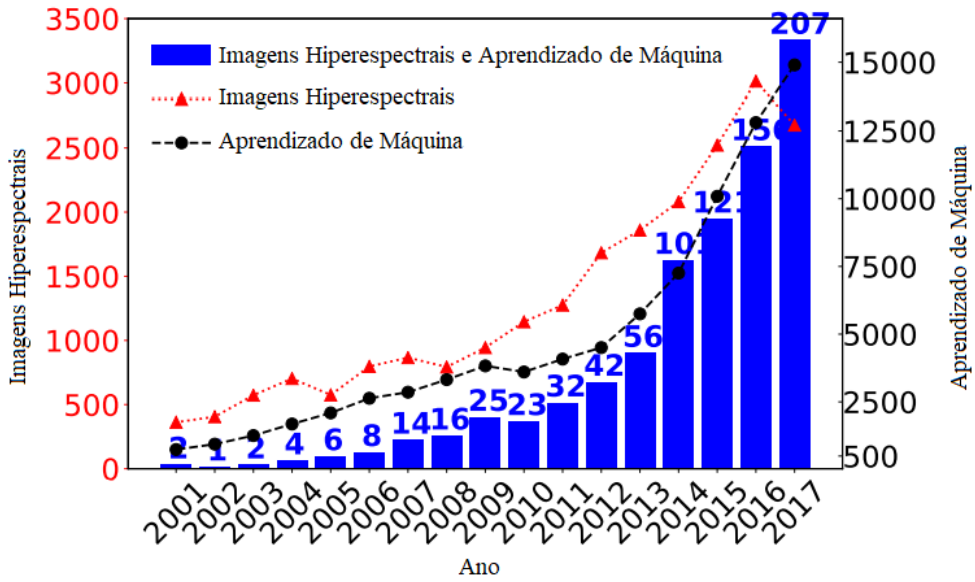


Fonte: Adaptado de (GOWEN et al., 2007).

Dependendo do comprimento de onda λ_i em questão, pré-processamentos são necessários pois, os espectros dos objetos de interesse podem conter ruído ou serem confundidos com o fundo da imagem, resultando em dificuldades para a identificação. Portanto, métodos de análise de imagem eficientes devem ser cuidadosamente desenvolvidos e empregados para filtrar as bandas úteis para a análise (FENG; SUN, 2012).

Se mostra importante citar como a comunidade de imageamento hiperespectral têm demonstrado interesse em estratégias de aprendizado de máquina, para análise e classificação destes espectros (GEWALI; MONTEIRO; SABER, 2018). A Figura 3 demonstra a produção de um grande número de estudos sobre novos métodos para as duas áreas nos últimos anos:

Figura 3:Número de publicações ao longo dos anos, com estudos relacionados a imagens hiperespectrais e aprendizado de máquina. Dados obtidos a partir de base de dados *Web of Science* da *Clarivate Analytics*.



Fonte: Adaptado de (GEWALI; MONTEIRO; SABER, 2018).

2.2 APRENDIZADO DE MÁQUINA

O aprendizado de máquina pode ser colocado como um subcampo da ciência da computação e é considerado como um método de inteligência artificial. Os métodos que implementam o aprendizado de máquina podem ser utilizados nos mais diversos domínios, pois encontram relações entre entradas e saídas de um conjunto de dados que represente determinado domínio, mesmo se a representação não for possível, essa característica permite o uso em reconhecimento de padrões, problemas de classificação, filtragem de *spam* e também em mineração de dados e problemas de previsão (VOYANT et al., 2017).

Com base no tipo de aprendizado, GEWALI; MONTEIRO; SABER, (2018) define que métodos de aprendizado de máquina podem ser categorizados em cinco grupos: aprendizado supervisionado, aprendizado não-supervisionado, aprendizado semi-supervisionado, aprendizado ativo e aprendizado por transferência, sendo os dois primeiros os mais popularmente utilizados.

Na aprendizagem supervisionada, a relação entre as variáveis de entrada e saída é estabelecida usando um conjunto de exemplos rotulados, ou seja, os exemplos para os quais os valores das variáveis de saída correspondem são conhecidos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). O problema é denominado regressão se a variável de saída for real, e é denominada classificação se a variável de saída for discreta.

Na aprendizagem não-supervisionada, a estrutura ou as características dos dados de entrada são descobertos usando exemplos não rotulados (exemplos para os quais os valores de saída correspondentes não estão disponíveis). Por exemplo, *K-means* (*clustering*) é um algoritmo de aprendizagem não supervisionado que agrupa os dados de entrada em grupos homogêneos. A análise de componente principal (*PCA*) é outro algoritmo de aprendizado não supervisionado que pode ser usado para encontrar uma representação linear de baixa dimensão, não correlacionada dos dados de entrada (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.2.1 Supervisionado

Na aprendizagem supervisionada como citado anteriormente, são apresentados ao algoritmo, exemplos de entradas e seus valores de saída desejados, sendo estes conhecidos a priori. O objetivo é então aprender uma regra geral que mapeia entradas em saídas, sendo cada padrão um par que inclui o objeto de entrada e seu valor de saída esperado (VOYANT et al., 2017).

2.2.1.1 Support Vector Machine (SVM)

O *SVM* (*Support Vector Machine)* é um exemplo de abordagem de aprendizado supervisionado, tendo sido amplamente utilizado para a classificação de dados hiperespectrais, devido à sua capacidade de lidar com dados de alta dimensão com um número limitado de amostras de treinamento (GHAMISI, P. et al., 2017).

A metodologia de classificação do *SVM* busca separar amostras pertencentes a classes diferentes definindo um hiperplano, com margem máxima no espaço onde as amostras são mapeadas (CAMPS-VALLS; BRUZZONE, 2005). Dado um conjunto de dados com pares classe-instância $((x_i, y_i), i) = 1, \dots, 1$ onde $x_i \in R^n$ e $y \in \{1, -1\}^l$, o *SVM* busca a solução para o seguinte problema de otimização (WANG; LU, 2006):

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2} \mathrm{w}^T w + C \sum_{i=1}^{l} \xi_i \tag{1}$$

Sujeito a

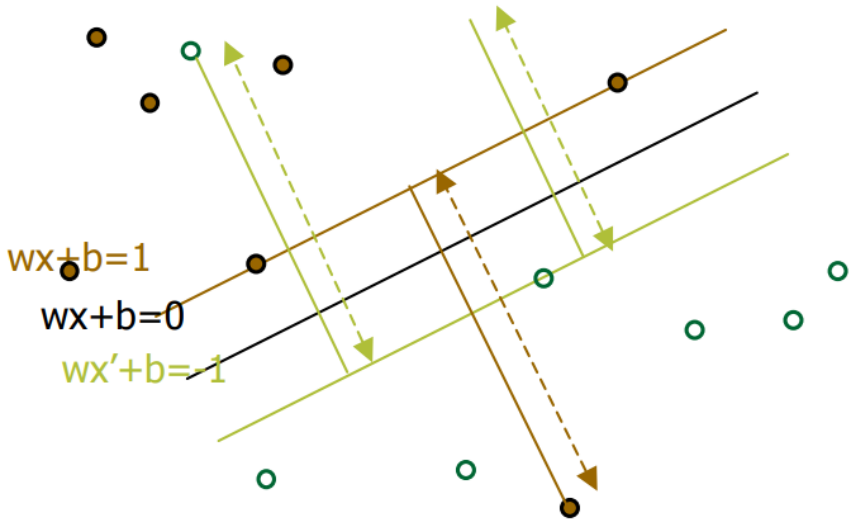
$$y_i(w^T \phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \;\; \xi_i \geq 0 \tag{2}$$

, onde w é a norma para a decisão ótima do hiperplano e b representa a distância mais próxima da origem do sistema de coordenadas, sendo $\boldsymbol{x}_i$ um dado recurso no espaço original e ϕ uma função de mapeamento não linear, que garante que as amostras transformadas são mais

prováveis de serem linearmente separáveis. Estes parâmetros definem um classificador linear no espaço de recursos de um determinado kernel (CAMPS-VALLS; BRUZZONE, 2005).

Há um mapeamento no espaço dimensional dos vetores de treinamento x_i pela função ϕ, e então pela natureza da técnica, o *SVM* traça um hiperplano linear separador, com uma margem máxima no seu maior espaço dimensional encontrado. A estratégia utiliza-se de alguns parâmetros, como o C, onde C > 0 se refere a penalidade do erro na classificação de uma determinada instância. A Figura 4 exemplifica a definição de um hiperplano baseando-se nos vetores de suporte:

Figura 4: Representação de um Hiperplano.



Fonte: Adaptado de (JAKKULA, 2011).

Outro parâmetro muito importante da técnica são os *kernels*, que dependendo o domínio do problema, retornam diferentes resultados. As equações 3, 4, 5, e 6 demonstram matematicamente alguns exemplos de *kernels:*

- Linear:

$$K(x_i, x_j) = x_i \cdot x_j \tag{3}$$

- Polinomial (*Poly*):

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d \tag{4}$$

- Função de base radial gaussiana (*RBF*):

$$K(x_i, x_j) = \exp\left(-\frac{\left|\left|x_i - x_j\right|\right|^2}{2\sigma^2}\right) \quad (5)$$

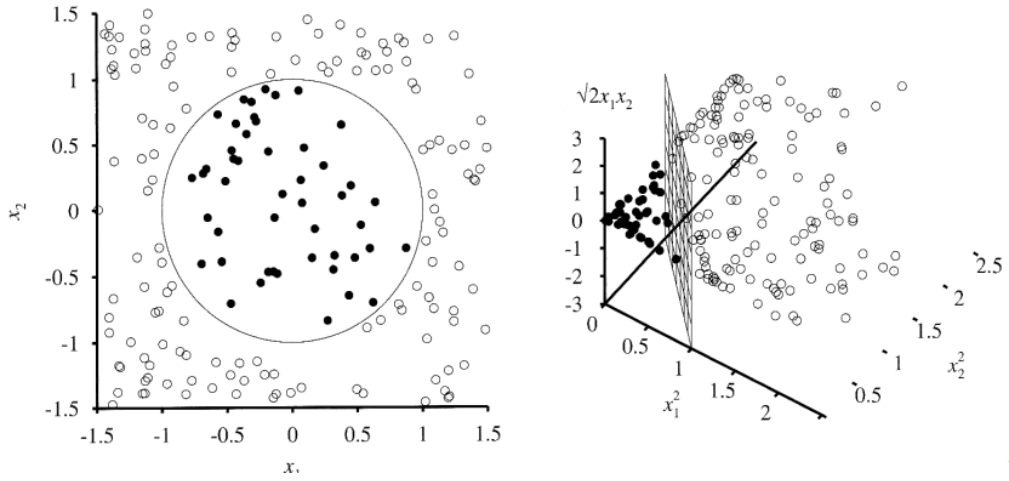
- Sigmoide:

$$K(x_i, x_j) = \tanh\left(\alpha x^T y + c\right) \quad (6)$$

Onde i e j são coordenadas para um determinado recurso (vetor de características) x dentro do espaço de entrada do conjunto de dados. Para o kernel polinomial *d* é especificado como parâmetro de grau do polinômio, no kernel *RBF* o *gamma* pode ser descrito como $\frac{1}{2\sigma^2}$, sendo um parâmetro livre para definição do quanto de influência um recurso terá. Para o kernel sigmoide tem-se uma equivalência a uma estratégia de Rede Neural *2-layer perceptron*.

Quanto a este parâmetro, o *kernel* de função de base radial Gaussiana (*RBF)* é tido como um dos mais amplamente utilizados, podendo lidar com distribuições de classes não lineares mais complexas em comparação com um *kernel* linear simples (GHAMISI, P. et al., 2017). Sendo assim, o *kernel RBF* é utilizado pela maioria dos algoritmos *SVM* para dados hiperespectrais, tendo inclusive alguns *kernels* projetados especificamente para modelar este tipo de dado (GEWALI; MONTEIRO; SABER, 2018). A Figura 5 demonstra um espaço de recursos separável:

Figura 5: Representação do espaço de características, onde recursos em preto e branco se diferem como duas classes separáveis pelo hiperplano (círculo) traçado.



Fonte: Adaptado de (JAKKULA, 2011).

3 MATERIAIS E MÉTODOS

Neste capítulo serão apresentados os materiais e métodos utilizados no desenvolvimento do trabalho.

3.1 ESTUDO DE CASO

Devido aos seus potenciais benéficos a saúde, dietas que incluem a canela-verdadeira (*Cinnamomum verum)* sofrem quanto a falta de clareza sobre as espécies utilizadas em estudos, juntamente com a identidade dos materiais comercializados. Na maioria destes produtos encontrados em mercados, o ingrediente é apenas referido pelo nome comum de "canela", mas a realidade é que esses ingredientes possuem graus de benefícios, tipos e variedades, dependendo das espécies de *Cinnamomum* presentes no produto (OKETCH-RABAH; MARLES; BRINCKMANN, 2018).

Como estudo de caso para este trabalho, foram utilizadas imagens hiperespectrais de amostras de *CV* e princípios ativos disponibilizadas pelo Laboratório de Análise Farmacêutica e Ambiental da Universidade Federal de Goiás (LAFAM – UFG), onde foram utilizadas ferramentas de aprendizado de máquina como o *SVM*, para determinação de uma relação entre a composição das amostras e sua região de cultivo.

3.2 MATERIAIS

Ao todo foram utilizadas amostras de *CV* de 17 diferentes marcas e regiões de origem, juntamente com amostras dos padrões (alta concentração) de 3 princípios ativos, Ácido Gálico, Catequina e Resveratrol. Para cada amostra de canela, não se sabe nenhuma informação acerca da concentração destes princípios em sua composição, apenas em qual região foi cultivada e em alguns casos, a marca que a comercializa. A Tabela 1 expõe essas informações (Amostra – Marca – Região) e as Figuras 6 e 7 apresentam as amostras de *CV* e dos padrões dos princípios ativos fornecidas pelo LAFAM.

Tabela 1: Relação das amostras de *CV*, marcas comerciais e suas regiões de cultivo.

Amostra	Marca	Região	Amostra	Marca	Região
1	Denner	Indonésia	9	MF Ervas Medicinais e Especiarias	Brasil/Trindade-Goiás
2	Fairway	Estados Unidos	10	Qualitá	Brasil/São Paulo-SP
3	-	Brasil/Goiânia-GO	11	Kayia	Grécia
4	Kitano	Brasil/Paraná e Minas Gerais	12	Di Cheff	Brasil/Aparecida de Goiânia-GO
5	Velly	Brasil/ Aparecida de Goiânia–GO	13	Escazu	Costa Rica
6	Mais Sabor	Brasil/Goiânia–GO	14	Pastéis de Belém	Portugal
7	Tempero Baiano	Brasil/Paraná	15	Premium	Brasil/Inhumas-GO
8	Paladar	Brasil/Goiânia-GO	16	Junco	Brasil/Uberlândia-MG
			17	MG Canyella Molta	Espanha

Fonte: Cedido pelo LAFAM-UFG (2020).

Figura 6: Amostras de *CV* cedidas pelo LAFAM-UFG.

Fonte: Autoria própria (2020).

Figura 7: Amostras dos padrões dos princípios ativos (Ácido Gálico, Catequina e Resveratrol)

Fonte: Autoria própria (2020).

Para o processamento dos dados e geração de resultados, foram utilizados "*notebooks*" que são ambientes de desenvolvimento em linguagem Python, com processamento computacional em nuvem na plataforma Kaggle, que por sua vez é uma comunidade online de cientistas de dados e profissionais da área de aprendizado de máquina. As especificações dos recursos disponibilizados foram de 16 GB de RAM, 20 GB de disco rígido, CPU AMD EPYC 7B12 de 2250MHz e 9 horas de cota para consumo contínuo.

3.3 AMOSTRAGEM

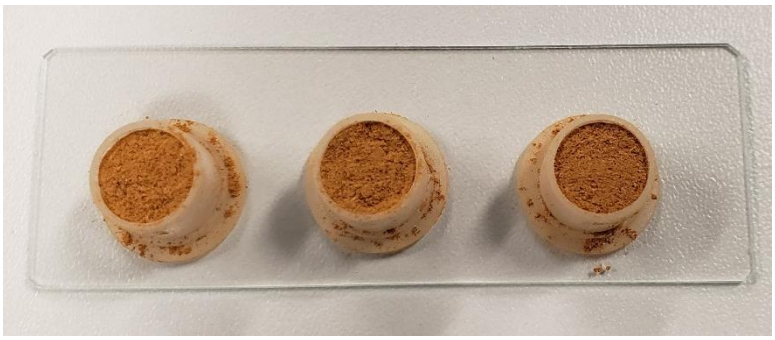
O processo de preparação das amostras para captura das imagens hiperespectrais foi realizado com a transferência do conteúdo da amostra para um batoque (espécie de tampa utilizada em certos recipientes em experimentos químicos) de tamanho médio, preenchendo-o sem sua totalidade e mantendo sua superfície o mais retilínea possível. Para manuseio e reaproveitamento da amostra, os batoques foram posicionados sobre placas de Petri (recipientes cilíndricos em plástico ou vidro, amplamente utilizado para cultura de micro-organismos). A Figura 8 demonstra parte do processo de amostragem:

Figura 8: Processo de transferência de uma amostra para os batoques.

Fonte: Autoria própria (2020).

O processo para cada amostra foi realizado em triplicata, onde quantidade total de uma amostra era suficiente para o enchimento de 3 batoques. Em seguida a amostra era retornada para o recipiente original, os materiais eram limpos e higienizados, e repetia-se o processo para uma nova amostra. A aquisição das imagens em triplicata visa evitar problemas de perda das informações e quaisquer outros empecilhos que levassem a uma nova coleta. A Figura 9 exemplifica o processo final de amostragem de uma das amostras de *CV* em triplicata:

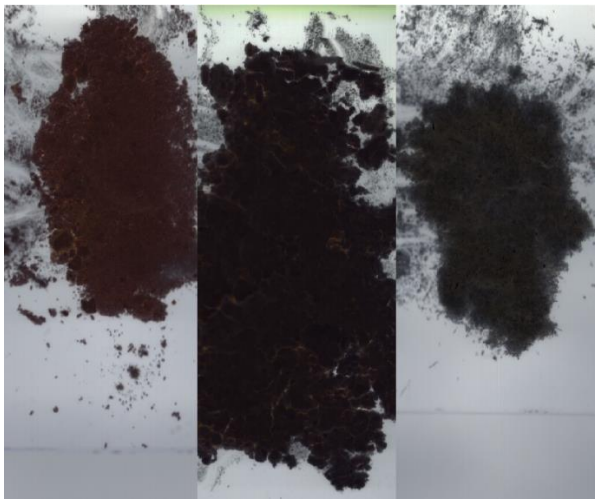
Figura 9: Batoques (triplicata) após a transferência da amostra.



Fonte: Autoria própria (2020).

Para as amostras dos padrões dos princípios ativos, foi coletada uma porção do recipiente original, sendo feita a dispersão da amostra sobre uma lâmina de vidro. A Figura 10 representa os padrões dos princípios ativos sobre lâminas de vidro após a aquisição das imagens hiperespectrais:

Figura 10: Exemplo da dispersão feita sobre as amostras dos padrões (imagem gerada em RGB após a aquisição das imagens hiperespectrais)



Fonte: Autoria própria (2020).

3.4 AQUISIÇÃO DE IMAGENS HIPERESPECTRAIS

Após a preparação das amostras, foi realizada a aquisição das imagens por meio da câmera/analisador de imagens hiperespectrais SisuCHEMA (Figura 11), pertencente ao Laboratório de Computação Científica (LCC) da Pontifícia Universidade Católica Goiás.

Figura 11: Câmera hiperespectral SisuCHEMA com o batoque contendo uma amostra para aquisição das imagens.

Fonte: Autoria própria (2020).

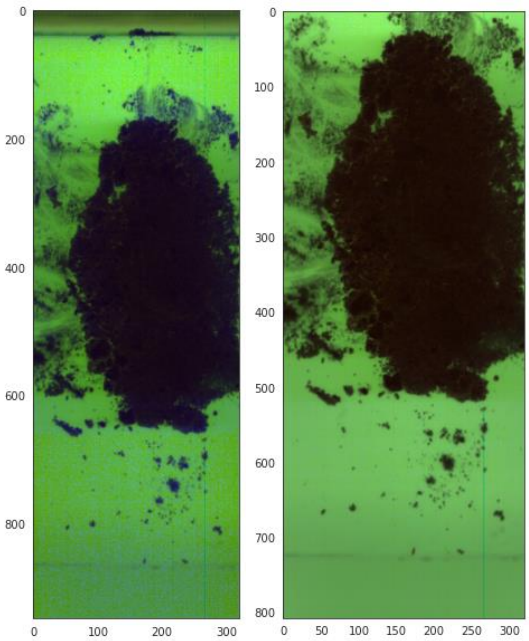
3.5 PRÉ-PROCESSAMENTO DOS DADOS

Este tópico trata dos métodos utilizados no pré-processamento dos dados gerados a partir da aquisição das imagens hiperespectrais.

3.5.1 Remoção de região espectral ruidosa

De maneira natural, durante o processo de aquisição das imagens, pode haver a presença de ruídos, com a adição de informações indesejadas ao conjunto de dados. Portanto foi realizada a remoção das regiões espectrais ruidosas de duas maneiras distintas. Inicialmente consistiu na remoção dos 16 últimos comprimentos de onda do eixo λ do conjunto de dados, uma vez que por análise visual nestes comprimentos estavam presentes o maior número de ruídos, sendo inclusive recomendado pelo fabricante da câmera hiperespectral a remoção também por este motivo. Em seguida foram retirados (*cropped)* os 140 pixels iniciais que compunham a dimensão espacial do hipercubo, nos eixos (x, y), uma vez que era visível sobre a dimensão espacial a presença de ruídos de captura. A Figura 12 representa em uma imagem RGB (espaço de cores *Red*, *Blue* e *Green)* a amostra de Ácido Gálico antes e depois da remoção espacial de ruídos gerados durante a aquisição dos dados da amostra.

Figura 12: a) Amostra de Ácido Gálico sem remoção de região superior ruidosa. b) Amostra de Ácido Gálico com remoção de região superior ruidosa.



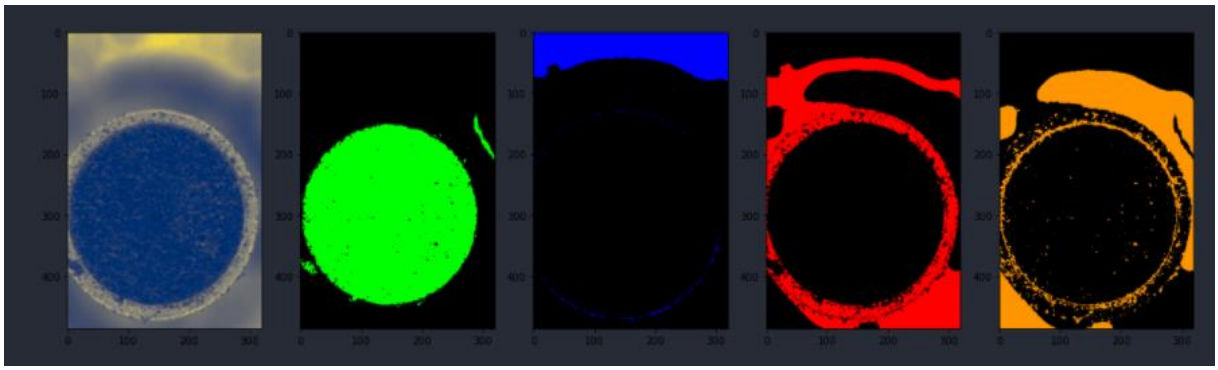
Fonte: Autoria própria (2020).

3.5.2 Remoção de fundo com *K-means*

Após a remoção inicial de regiões e espectros ruidosos, foi realizada a remoção de fundo das imagens, para que o conjunto de dados a ser manipulado passasse a ser somente os pixels correspondentes as amostras. O *K-means* (*clustering*) é uma estratégia de aprendizado de máquina não-supervisionado, amplamente utilizado para o agrupamento de regiões baseado em suas características e valores. Assume-se que pixels semelhantes formam *clusters* (grupos) no espaço de recursos, quando aplicados a imagens hiperespectrais, esses métodos podem fornecer resultados satisfatórios. (HAUT et al., 2017).

O *K-means* foi aplicado em todas as amostras, para indexação dos pixels correspondentes as regiões de interesse (amostra e fundo) na amostra. Devido a alguns ruídos inerentes presentes nas imagens das amostras de *CV*, foi utilizado o valor de 4 *clusters* na aplicação do algoritmo, enquanto para as amostras dos padrões apenas 2 *clusters* foram suficientes para a separação. A Figura 13 demonstra o resultados da execução do algoritmo *K-means* na amostra 1 de *CV*, com as diferentes cores representando os diferentes *clusters* encontrados:

Figura 13: Resultado da aplicação do algoritmo de *K-means* na amostra 1 de *CV*, para remoção de fundo.



Fonte: Autoria própria (2020).

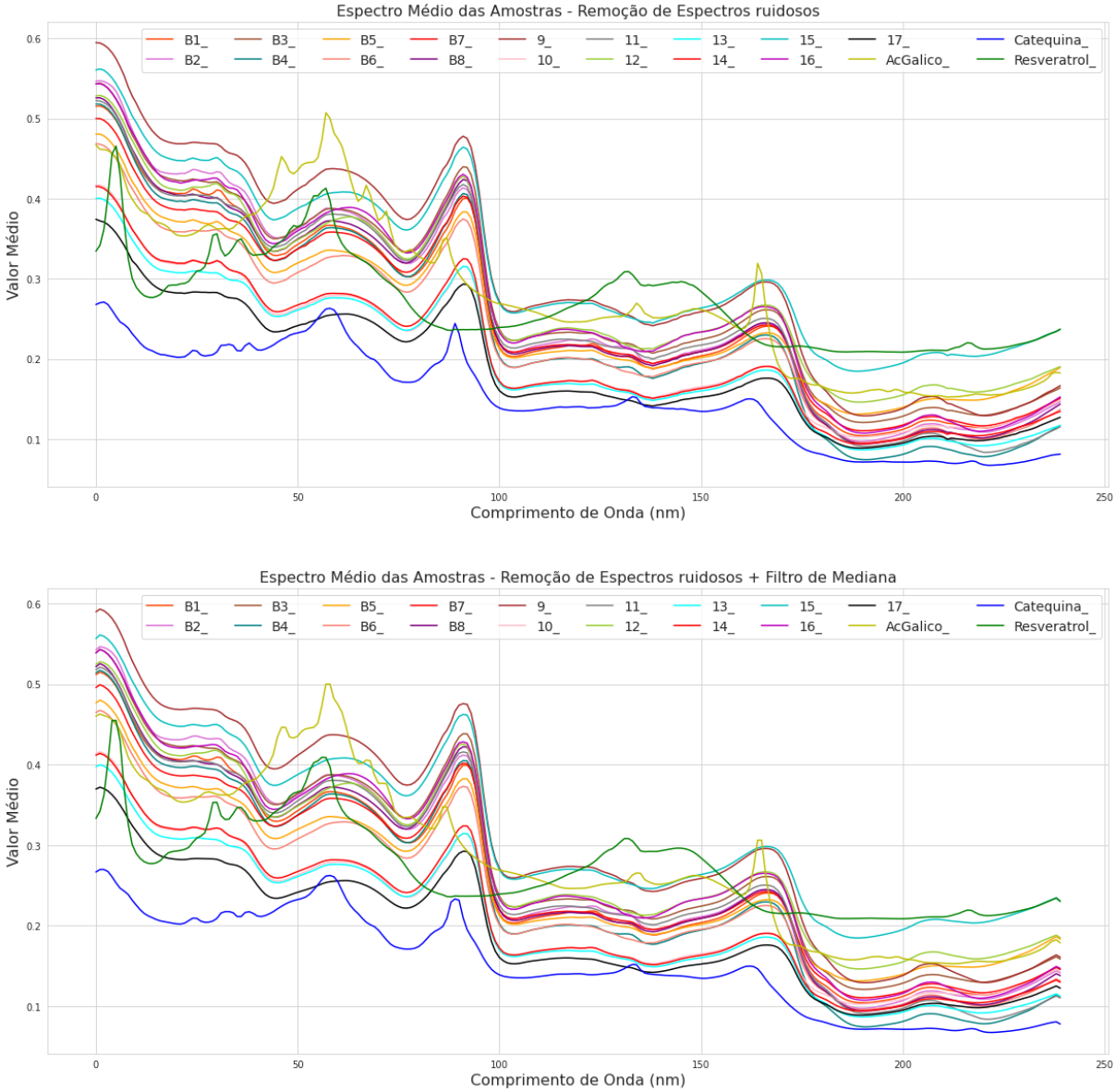
3.5.3 Remoção de *outliers (spikes)* com filtro de mediana

Com a remoção inicial de ruídos e remoção de fundo, foi realizada a plotagem dos espectros médios de cada uma das amostras. A partir de uma análise visual ficou evidente a presença de *spikes* nos espectros das amostras. As *spikes* podem ser definidas como um aumento repentino e acentuado seguido por um declínio acentuado no espectro, elas frequentemente mascaram detalhes da imagem, levando à identificação incorreta de um sinal de interesse. Tais ruídos podem aparecer devido a um comportamento anormal do detector e/ou

por imperfeições de circuitos eletrônicos e até mesmo condições ambientais (VIDAL; AMIGO, 2012).

Diferentes técnicas e algoritmos têm sido propostos a fim de remover ou interpolar as *spikes* com base em seus nos valores de seus vizinhos mais próximos, em uma metodologia de comparação de pixels. Umas destas técnicas é a aplicação do filtro de mediana, na qual cada pixel de saída é definido como a mediana dos valores de pixel de seus vizinhos (dependendo do tamanho da janela definida) (BEHREND; TARNOWSKI; MORRIS, 2002). Foi então aplicado o filtro de mediana em todos os espectros para remoção dos *spikes* com janela de tamanho 3, tamanho este que é o padrão da função em Python que implementa o processo do filtro de mediana, sendo mantido por ter sido eficiente no processo de remoção das *spikes*. A Figura 14 demonstra os espectros médios das amostras de *CV* e dos princípios ativos antes e depois da aplicação do filtro de mediana:

Figura 14: Espectro médio das amostras de *CV* e dos padrões, comparativo do antes (*plot 1*) e depois (*plot 2*) da aplicação do filtro de mediana.



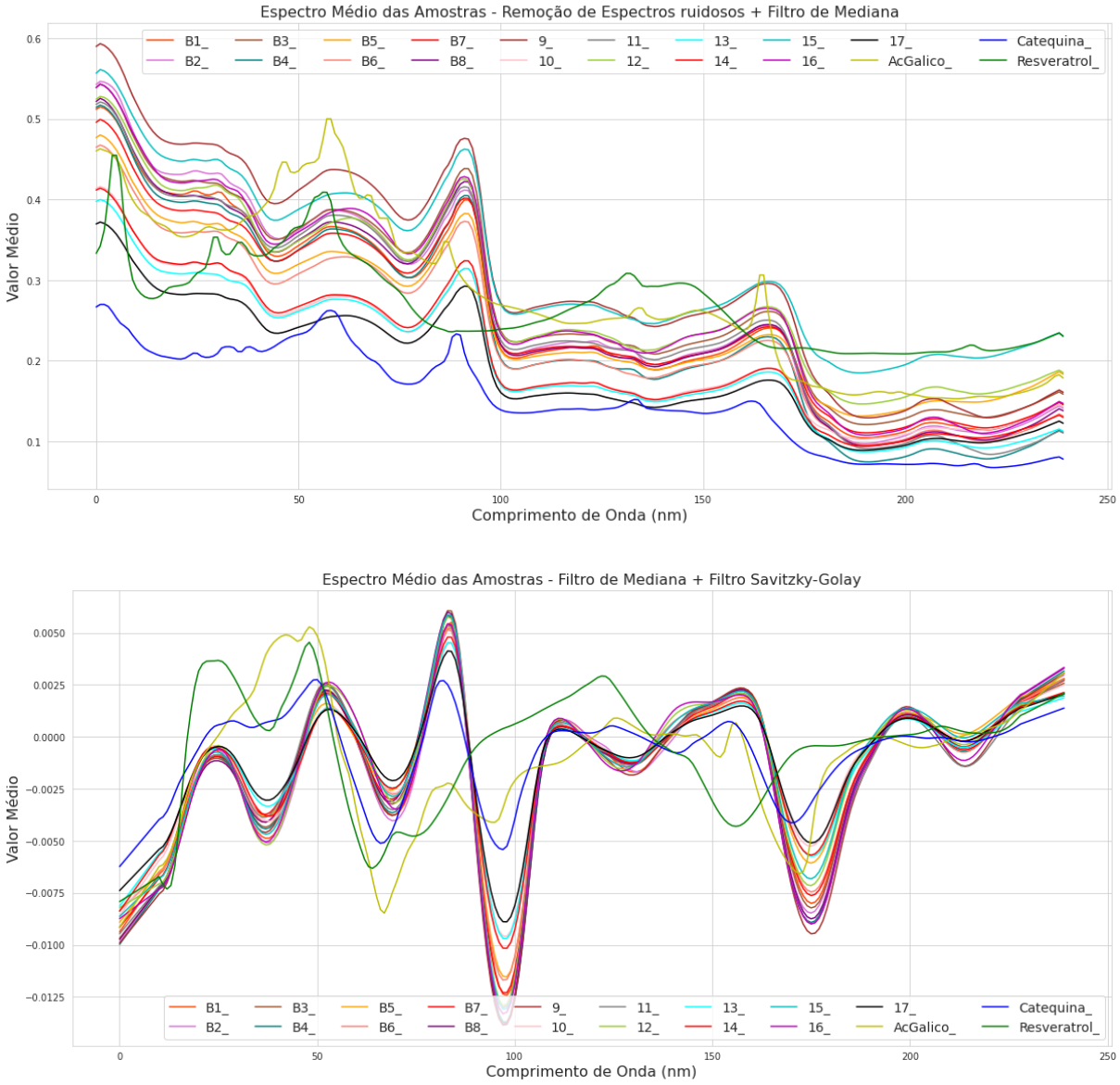
Fonte: Autoria própria (2020).

3.5.4 Filtro de suavização Savitzky-Golay

O filtro Savitzky-Golay é uma técnica de suavização comum utilizada em dados hiperespectrais, sendo baseado na aproximação de mínimos quadrados, que determina coeficientes de suavização a partir da aplicação de uma equação polinomial, com grau e tamanho de janela a serem definidos como parâmetros da técnica. O filtro é ideal para dados espectroscópicos, pois minimiza o ruído do sinal enquanto preserva a originalidade e a forma dos espectros de entrada (LOGGENBERG et al., 2018). Foi aplicado o filtro derivativo, com polinômio de segundo grau e janela de tamanho 21. A escolha desta parametrização para o filtro

foi baseada em exemplos encontrados na literatura. A Figura 15 demonstra os espectros médios das amostras de *CV* e dos princípios ativos antes e depois da aplicação do filtro de suavização Savitzky-Golay.

Figura 15: Espectro médio das amostras de *CV* e dos padrões, comparativo do antes (*plot* 1) e depois (*plot* 2) da aplicação do filtro Savitzky-Golay.



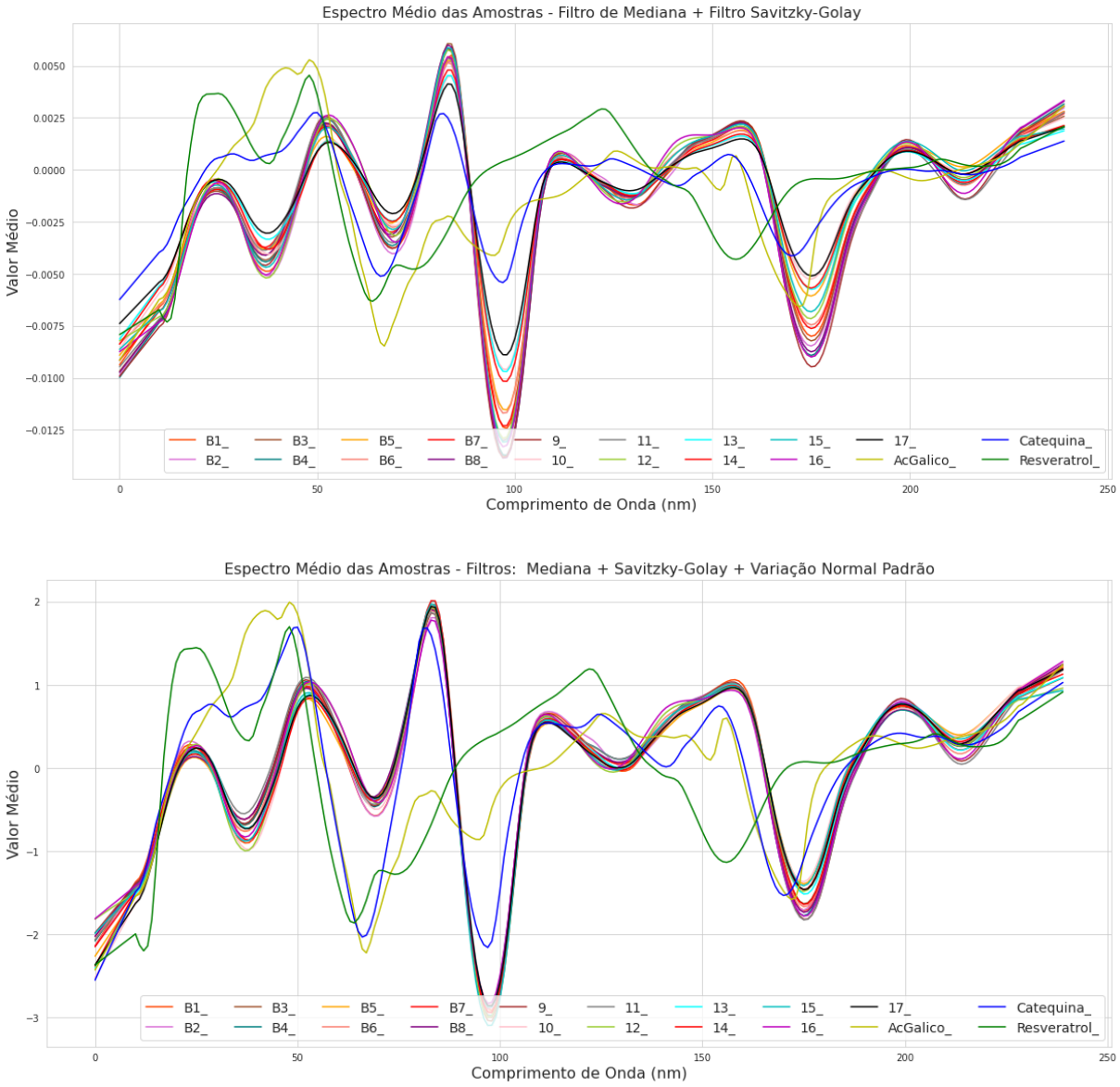
Fonte: Autoria própria (2020).

3.5.5 Filtro de Variação Normal Padrão (*SNV*)

A maioria dos dados hiperespectrais carregam consigo uma certa dificuldade de interpretação, isso se deve ao ruído e colinearidade entre as bandas espectrais. Métodos de pré-processamento, como a transformação para Variação Normal Padrão (*SNV*), permitem minimizar eficientemente a interferência no sinal (ou seja, outros tipos de ruído) e simplificar

os processos de interpretação das informações (ZENG et al., 2016). Desta forma, foi aplicado o filtro *SNV* nos espectros de todas as amostras. A Figura 16 demonstra os espectros médios das amostras de *CV* e dos princípios ativos antes e depois da aplicação do filtro de Variação Normal Padrão (*SNV).*

Figura 16: Espectro médio das amostras de *CV* e dos padrões, comparativo do antes (*plot* 1) e depois (*plot* 2) da aplicação do filtro de transformação de Variação Normal Padrão (*SNV*).



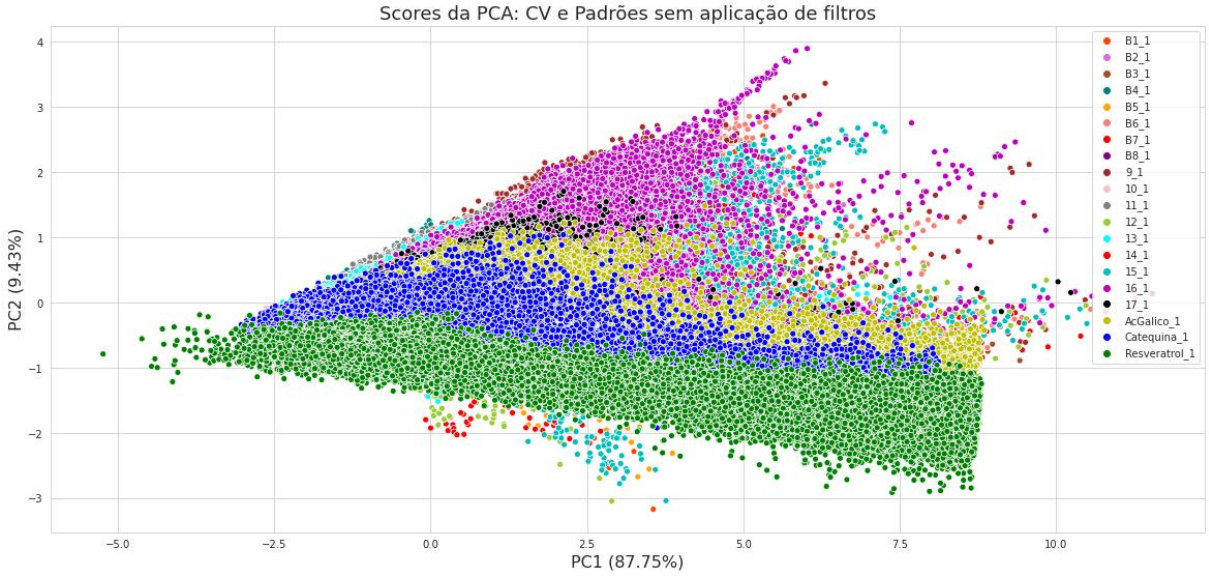
Fonte: Autoria própria (2020).

3.5.6 Análise de Componentes Principais (*PCA*)

Para uma análise exploratória acerca das características do conjunto de dados, a *PCA* foi utilizada no intuito de diminuir a alta dimensionalidade dos dados e permitir uma visualização mais simples de como os dados estão distribuídos. A *PCA* é uma estratégia de

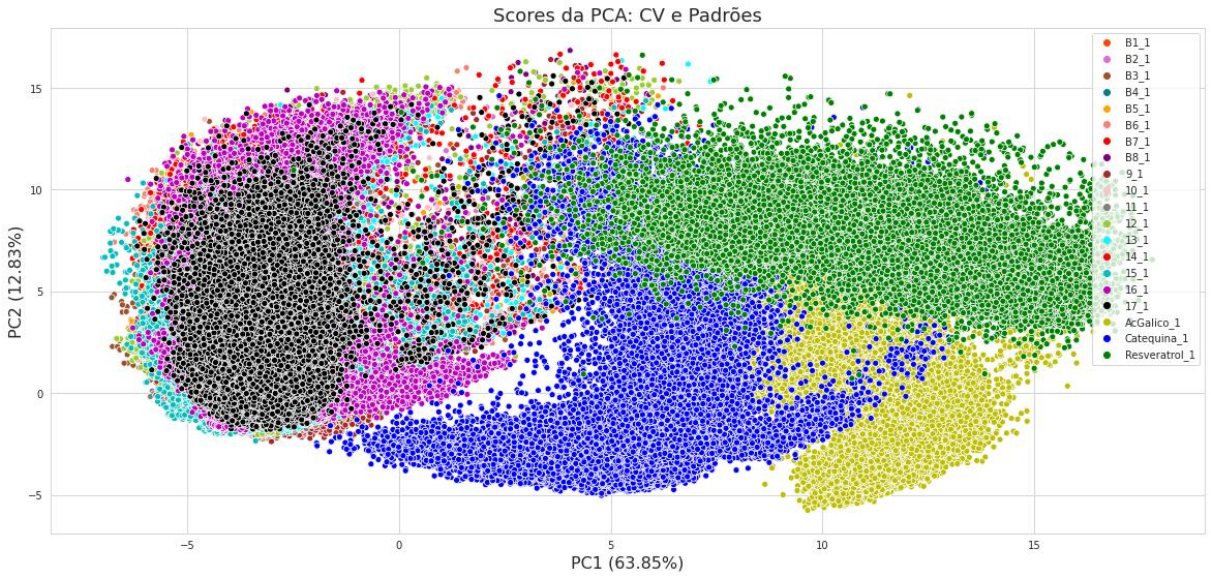
aprendizado de máquina não-supervisionado, sendo eficaz na junção de informações complementares e redução de informações redundantes, possibilitando a diminuição de influência de algum ruído restante nos dados, pois estando estes em pequenas quantidades serão ignorados no processo. Portanto a *PCA* busca extrair as informações mais relevantes e efetivamente maximizar as diferenças espectrais entre os pixels (KANG; DUAN; LI, 2020). Vale ressaltar que a *PCA* possui um uso diversificado, e para este caso, foi aplicada para auxiliar no processamento e análise exploratória dos dados. As Figuras 17 e 18 demonstram o resultado da aplicação do algoritmo *PCA* sobre o conjunto de amostras de *CV* e princípios ativos, antes e depois de todo o pré-processamento. As porcentagens presentes nas Figuras 17 e 18 são referentes aos índices de variância explicada obtidos.

Figura 17: Scores da *PCA* sobre amostras de *CV* e dos padrões dos princípios ativos sem aplicação dos filtros de pré-processamento.



Fonte: Autoria própria (2020).

Figura 18: Scores da *PCA* sobre amostras de *CV* e dos padrões dos princípios ativos após todo o pré-processamento.



Fonte: Autoria própria (2020).

3.5.7 Treinamento e avaliação do modelo utilizado

A análise sobre os conjuntos de dados após a classificação, foi dividida em duas vertentes. A primeira trata da classificação apenas das amostras de *CV*, sendo utilizadas as 17 amostras, o que totalizou pouco mais de 1.000.000 de pixels a serem envolvidos no processo de treinamento e teste, o que tornou inviável a análise geral sobre a *CV*, uma vez que o *SVM* possui um já alto custo computacional. Desta forma, se utilizando da grande habilidade da estratégia de conseguir desempenhar bem mesmo com um número reduzido de amostras, foram realizados testes exploratórios onde a classificação utilizando o *SVM* foi realizada em cima de um conjunto de dados subamostrado, totalizando um conjunto de 100.000 pixels a serem envolvidos no processo de treinamento e teste. A escolha do conjunto subamostrado foi realizado de maneira pseudoaleatória, o que garante um peso equivalente de todas as classes no processo.

A segunda análise envolve o treinamento em 4 classes, rotulados como os 3 princípios ativos e *CV*. Para tal, foram utilizadas as amostras dos padrões juntamente com a amostra 1 de *CV*, para que desta forma o classificador pudesse avaliar e aprender sobre as 4 diferentes classes. Propositalmente, dentro da amostra 1, foram indexados todos os pixels como sendo *CV*, uma vez que não se tem conhecimento destes rótulos a priori. Para este processo, o classificador teve de lidar com um conjunto menor de dados, algo em torno de 400.000 pixels para o processo de treinamento e teste, portanto não foram necessárias estratégias de subamostragem.

Vale ainda ressaltar que a primeira análise consiste na tentativa de determinar o quanto o classificador conseguirá separar as classes de *CV* entre si, encontrando características suficientes que as diferenciem em relação apenas as suas regiões de origem. Já para a segunda análise, será realizada a predição sobre todas as outras amostras de *CV* para que a partir do treinamento e teste iniciais com as 4 classes, o classificador consiga pixel-a-pixel rotular cada classe nas outras amostras, e assim determinar a concentração dos princípios ativos em cada uma delas.

Para a escolha da parametrização do modelo, foram realizados testes para as duas análises, variando-se *kernel*, *C*, *gamma* e tamanho do conjunto de dados geral, para treino e teste. Para que pudesse ser metrificado o quanto cada um dos parâmetros influenciava para a melhor ou piora do modelo. O resultado foi a utilização do *kernel RBF*, com *C = 100 e gamma =10*, a escolha dos valores de parametrização foi baseada no resultado de testes empíricos realizados sobre o conjunto de treinamento e teste, variando-se os valores dos parâmetros e partindo inicialmente dos valores padrão das funções em Python utilizadas. Quanto ao tamanho do conjunto de dados, ficou claro a partir de testes de subamostragem, como com 10% do conjunto original o classificador já se estabilizava em relação a acurácia máxima para a análise das amostras de *CV*, enquanto para o segundo caso a utilização de subamostragem não se aplica.

Para avaliação do modelo nos testes exploratórios e nos resultados finais, foram considerados o erro de raiz quadrático médio (*RMSE*), erro médio absoluto (*MAE*) e coeficiente de determinação (R^2), bem como a matriz de confusão, precisão e acurácia do modelo. As equações 7, 8 e 9 representam matematicamente as métricas avaliadas:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{7}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{9}$$

Onde têm-se N como o tamanho do conjunto de teste, $\hat{y}_i$ o i-ésimo valor predito, y_i o i-ésimo valor esperado e $\bar{y}_i$ como a média dos valores de teste esperados. Quanto aos valores, o *RMSE e MAE* tem seu ótimo em 0, e o R^2em 1.

A acurácia corresponde a quantos valore foram corretamente preditos em relação a todo o conjunto de teste, sendo dado em valor proporcional entre 0 e 1. A matriz de confusão adiciona os valores de *recall*, *f1_score* e precisão.

O *recall* pode ser definido como a proporção dos verdadeiros positivos em relação ao número de falsos negativos, sendo intuitivamente a capacidade do classificador de encontrar as amostras positivas. O *f1_score* pode ser interpretado como uma média ponderada da precisão e o *recall*, atingindo seu ótimo em 1 e o pior em 0.

Por fim, a precisão é a proporção dos verdadeiros positivos em relação aos falsos positivos, sendo a capacidade do classificador de não rotular como positiva uma amostra negativa. A matriz de confusão traz também a relação de todas estas métricas de maneira expressa, contabilizando cada um destes valores por classe. Todas as informações referentes as métricas utilizadas, foram baseadas no material da biblioteca de aprendizado de máquina *sklearn* presente na linguagem de programação Python (SCIKIT-LEARN DEVELOPERS, 2020).

4 RESULTADOS

Este capítulo apresentará os resultados das duas análises propostas, de acordo com cada uma das métricas de avaliação propostas.

4.1 CLASSIFICAÇÃO DA *CINNAMOMUM VERUM*

As tabelas 2 e 3 apresentam os valores das métricas correspondentes ao processo de treinamento e teste do modelo do classificador *SVM*, para os parâmetros de *kernel='rbf' e C = 100* e *gamma = 10*. Como já foi citado, para o processo foi utilizado um conjunto subamostrado de 125.000 pixels das 17 amostras de canela, com 25.000 sendo para teste e 100.000 para treino. A divisão foi realizada baseando-se em uma das proporções (20% do conjunto de dados para teste e 80% para treino) mais usualmente utilizada na literatura, em relação a classificação por aprendizado de máquina. As Tabelas 2 e 3 contém os valores das métricas de avaliação para o modelo gerado e a Figura 19 apresenta a Matriz de Confusão resultante.

Tabela 2: Métricas gerais de avaliação do modelo. Classificação realizada no conjunto subamostrado das amostras de *CV*.

Métrica	Acurácia	RMSE	MAE	R^2
Valor	0.78	3.04	1.19	0.62

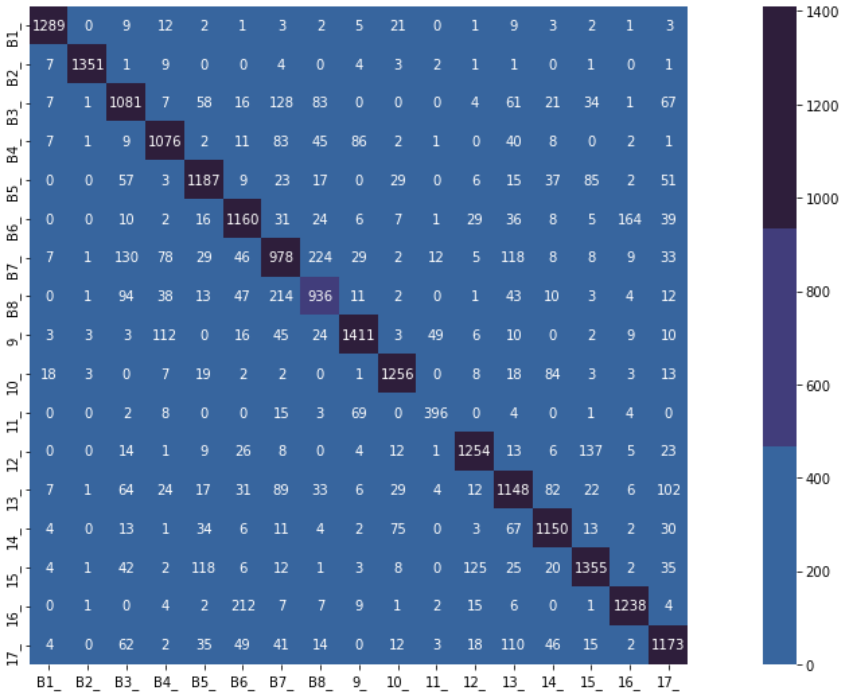
Fonte: Autoria própria (2020).

Tabela 3: Métricas por classe para avaliação do modelo. Classificação realizada no conjunto subamostrado das amostras de *CV*.

Classe	Precisão	*Recall*	*F1-score*	Número de Pixels preditos
1	0.95	0.95	0.95	1363
2	0.99	0.98	0.98	1385
3	0.68	0.69	0.68	1569
4	0.78	0.78	0.78	1374
5	0.77	0.78	0.78	1521
6	0.71	0.75	0.73	1538
7	0.58	0.57	0.57	1717
8	0.66	0.66	0.66	1429
9	0.86	0.83	0.84	1706
10	0.86	0.87	0.87	1437
11	0.84	0.79	0.81	502
12	0.84	0.83	0.84	1513
13	0.67	0.68	0.68	1677
14	0.78	0.81	0.79	1415
15	0.80	0.77	0.79	1759
16	0.85	0.82	0.84	1509
17	0.73	0.74	0.74	1586

Fonte: Autoria própria (2020).

Figura 19: Matriz de Confusão com os valores dos pixels preditos para o modelo gerado para classificação subamostrada das amostras de CV.



Fonte: Autoria própria (2020).

4.2 CLASSIFICAÇÃO DOS PRINCÍPIOS ATIVO EM *CV*

As tabelas 4 e 5 apresentam os valores das métricas correspondentes ao processo de treinamento e teste do modelo do classificador *SVM*, para os parâmetros de *kernel='rbf' e C = 100* e *gamma = 10*. Como citado anteriormente, foram utilizadas as 3 amostras de princípios ativos juntamente com a amostra 1 de *CV*, com a divisão do conjunto em 70% para treino e 30% para teste, uma vez que não foi necessária a subamostragem. A divisão foi mais uma vez escolhida baseando-se em proporções usualmente utilizadas na literatura para este tipo de classificação. As Tabelas 4 e 5 contém os valores das métricas de avaliação para o modelo gerado e a Figura 20 apresenta a Matriz de Confusão resultante.

Tabela 4: Métricas gerais de avaliação do modelo. Classificação realizada no conjunto das 3 amostras de princípio ativo e amostra 1 de *CV*.

Métrica	Acurácia	RMSE	MAE	R^2
Valor	0.99	0.035	0.0007	0.99

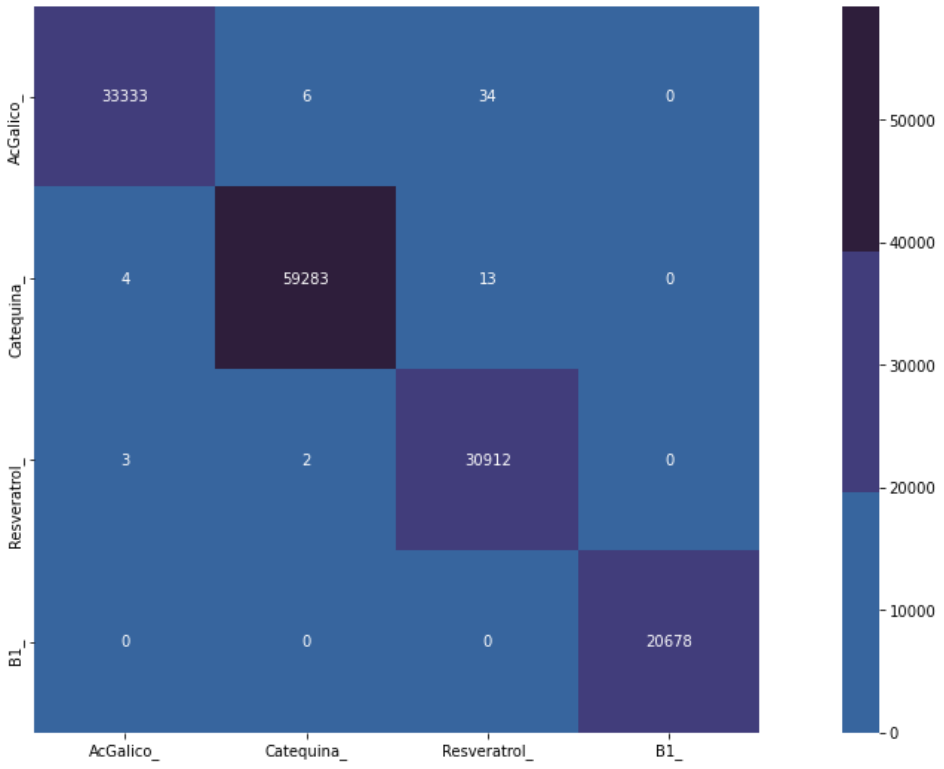
Fonte: Autoria própria (2020).

Tabela 5: Métricas por classe para avaliação do modelo. Classificação realizada no conjunto das 3 amostras de princípio ativo e amostra 1 de *CV*.

Classe	Precisão	*Recall*	*F1-score*	Número de Pixels preditos
Ác. Gálico	1.00	1.00	1.00	33373
Catequina	1.00	1.00	1.00	59300
Resveratrol	1.00	1.00	1.00	30917
Amostra 1 (*CV)*	1.00	1.00	1.00	20678

Fonte: Autoria própria (2020).

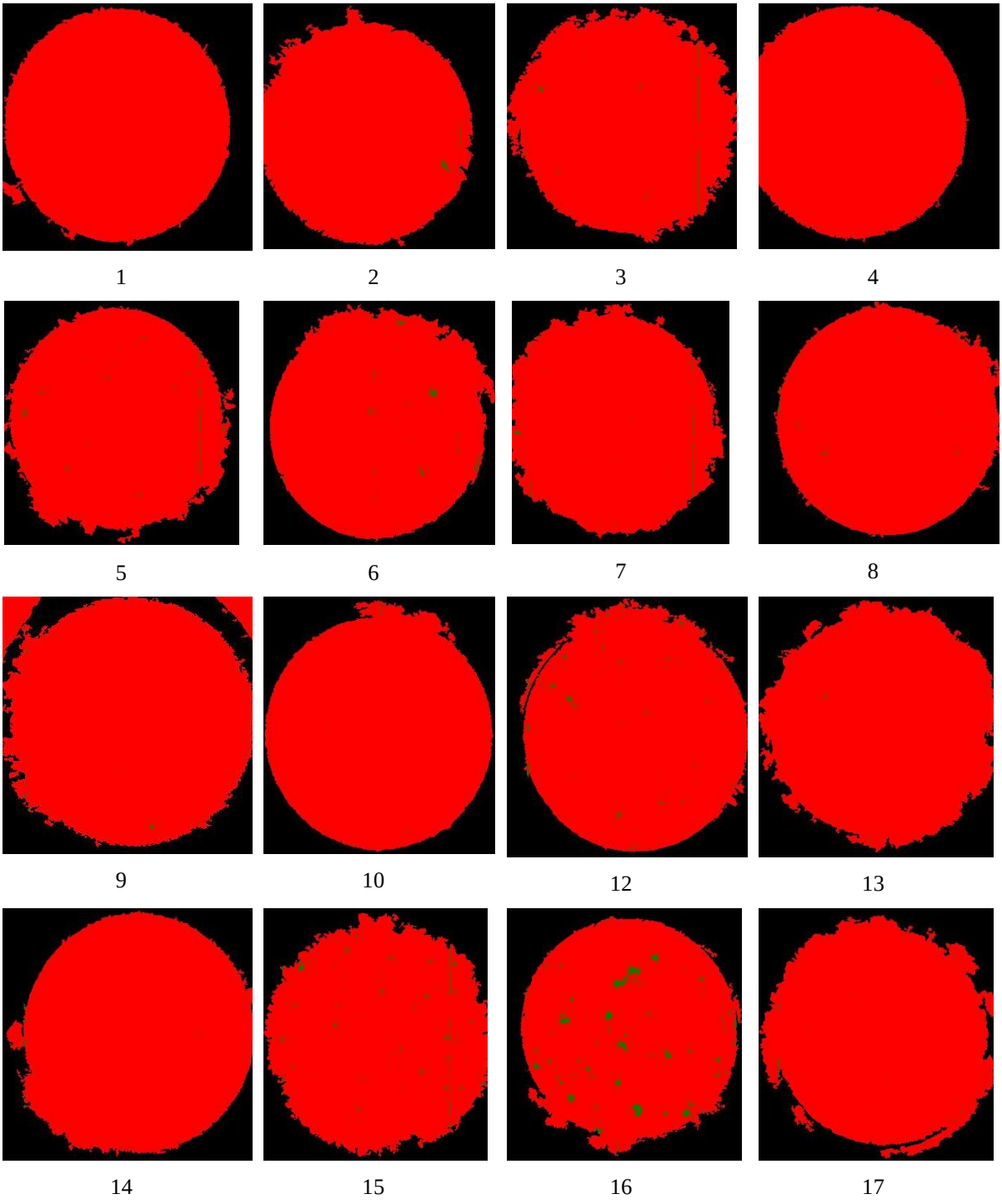
Figura 20: Matriz de Confusão com os valores dos pixels preditos para o modelo gerado para classificação das 3 amostras de princípio ativo e amostra 1 de *CV*.



Fonte: Autoria própria (2020).

A Figura 21 (*grid*) e a Tabela 6 apresentam os resultados da predição do modelo gerado sobre todas as outras amostras de *CV*. A amostra 11 de *CV*, foi a única a ter sua aquisição de imagens prejudicada pela presença de ruídos em excesso no momento da aquisição, o que inviabilizou a utilização da segunda e terceira amostra de sua triplicata em relação a retirada do fundo. A primeira é a que será mostrada no grid, em que mesmo tendo sido cortada boa parte das informações no momento da aquisição, ela permitiu a retirada de fundo.

Figura 21: Grid de imagens contendo os pixels preditos por classe para todas as amostras de CV.



Fonte: Autoria própria (2020).

Tabela 6: Relação dos percentuais e quantidades de princípios em amostras CV.

Amostra	Ác. Gálico		Catequina		Resveratrol		*Cin. verum*	
	Percentual	Qtd. Pixels	Percentual	Qtd. Pixels	Percentual	Qtd. Pixels	Percentual	Qtd. Pixels
1	0.0%	0	0.0%	0	0.01%	1	99.99%	68522
2	0.0%	0	0.0%	0	0.16%	105	99.84%	68783
3	0.0%	0	0.0%	0	0.32%	242	99.68%	76802
4	0.0%	0	0.0%	0	0.02%	8	99.98%	66906
5	0.0%	0	0.03%	20	0.32%	233	99.65%	72239
6	0.0%	0	0.00%	0	0.35%	255	99.65%	74082
7	0.0%	0	0.01%	2	0.3%	242	99.69%	80579
8	0.0%	0	0.01%	3	0.15%	108	99.84%	71006
9	0.0%	0	0.01%	4	0.09%	78	99.90%	85757
10	0.0%	0	0.01%	1	0.03%	27	99.96%	70433
11	0.0%	0	0.01%	1	0.06%	15	99.93%	25785
12	0.0%	0	0.00%	0	0.48%	346	99.52%	73169
13	0.0%	0	0.01%	2	0.1%	85	99.89%	80811
14	0.0%	0	0.00%	0	0.12%	86	99.88%	72885
15	0.0%	0	0.01%	2	0.7%	589	99.29%	82503
16	0.0%	0	0.00%	0	2.08%	1503	97.92%	71073
17	0.0%	0	0.01%	7	0.15%	112	99.84%	76359

Fonte: Autoria própria (2020).

5 CONCLUSÃO

O trabalho em questão propôs o estudo acerca de amostras de *CV*, juntamente com amostras de Ácido Gálico, Catequina e Resveratrol, para determinação da relação entre a composição das amostras *CV* e sua região de origem, baseando-se ou não na quantificação dos princípios ativos em sua composição. Considerando os resultados, para a análise da classificação utilizando apenas as amostras de *CV*, é válido considerar uma possível limitação do classificador, seja pela parametrização ou pela própria estratégia que ele implementa. A hipótese de limitação se sustenta pela baixa probabilidade de que de fato não há a presença do princípio ativo Ácido Gálico em nenhuma das amostras de *CV*, uma vez que até mesmo pela métrica de erro do modelo, alguns pixels poderiam erroneamente serem considerados como tal. Portanto se mostrou uma classificação razoável a partir da análise das métricas do modelo, enfatizando como as amostras de *CV* são altamente semelhantes até mesmo âmbito espectral, o que não permite definir uma relação clara quanto as regiões de cultivo.

Quanto a análise da quantificação dos princípios ativos nas amostras de *CV*, o modelo obteve métricas muito boas, o que valida o modelo inicial gerado. No entanto, a utilização de uma amostra de *CV* específica durante o treinamento abre brechas sobre se ela realmente era capaz de representar ao classificador como se comporta a classe *CV*. Vale ressaltar ainda sobre a presença de bandas espectrais ruidosas mesmo após todo o pré-processamento, o que de certa forma influenciou no resultado.

Entretanto, o classificador foi capaz de determinar quantidades relevantes de Resveratrol em todas as amostras, ao passo que a Catequina foi encontrada em algumas amostras, mas em quantidade menor. O Ácido Gálico não foi classificado em nenhuma das amostras, o que sugere uma fraca presença na composição das amostras, ou uma grande semelhança espectral com a própria *CV*, tendo uma classificação errônea como *CV* de pixels que possuam um comportamento espectral característico do Ácido Gálico. Por fim, a partir dos resultados desta parte específica da composição, não é possível de maneira clara, definir a relação com sua quantificação e a região de cultivo, fornecendo apenas uma direção sobre potenciais análises, como o caso da amostra 16 de *CV* que obteve uma quantidade considerável de Resveratrol.

Para trabalhos futuros, aconselha-se a utilização de outras estratégias de classificador, como outras opções de aprendizado de máquina e até mesmo redes neurais. Sendo válida a análise acerca do peso das bandas espectrais na separação das classes, a partir da extração de características. E por fim, uma análise exploratória sobre outras etapas a serem incluídas no processo de pré-processamento das amostras. Uma revisão sobre as bandas espectrais ruidosas

se mostra útil, que aliada a extração de características, possa fornecer uma relação apenas dos comprimentos de ondas mais importantes no processo de classificação.

REFERÊNCIAS

ABDALI, D.; SAMSON, S. E.; GROVER, A. K. How effective are antioxidant supplements in obesity and diabetes? **Medical Principles and Practice**, v. 24, n. 3, p. 201–215, 2015.

BAGCHI, T. B.; SHARMA, S.; CHATTOPADHYAY, K. Development of NIRS models to predict protein and amylose content of brown rice and proximate compositions of rice bran. **Food Chemistry**, v. 191, p. 21–27, 2016.

BEHREND, C. J.; TARNOWSKI, C. P.; MORRIS, M. D. Identification of outliers in hyperspectral Raman image data by nearest neighbor comparison. **Applied Spectroscopy**, v. 56, n. 11, p. 1458–1461, 2002.

CAMPS-VALLS, G.; BRUZZONE, L. Kernel-based methods for hyperspectral image classification. **IEEE Transactions on Geoscience and Remote Sensing**, v. 43, n. 6, p. 1351–1362, 2005.

D'SOUZA, S. P. et al. Pharmaceutical Perspectives of Spices and Condiments as Alternative Antimicrobial Remedy. **Journal of Evidence-Based Complementary and Alternative Medicine**, v. 22, n. 4, p. 1002–1010, 2017.

EMERITUS, E. et al. Health benefits of Ceylon cinnamon. **the Ceylon Medical Journal**, v. 61, n. 1, p. 1–5, 2016.

FENG, Y. Z.; SUN, D. W. Application of Hyperspectral Imaging in Food Safety Inspection and Control: A Review. **Critical Reviews in Food Science and Nutrition**, v. 52, n. 11, p. 1039–1058, 2012.

GEWALI, U. B.; MONTEIRO, S. T.; SABER, E. Machine learning based hyperspectral image analysis: A survey. **arXiv**, 2018.

GOWEN, A. A. et al. Hyperspectral imaging - an emerging process analytical tool for food quality and safety control. **Trends in Food Science and Technology**, v. 18, n. 12, p. 590–598, 2007.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Springer Series in Statistics. In: **The Elements of Statistical Learning**. [s.l: s.n.]. v. 27p. 83–85.

HAUT, J. M. et al. Cloud implementation of the K-means algorithm for hyperspectral image analysis. **Journal of Supercomputing**, v. 73, n. 1, p. 514–529, 2017.

JAKKULA, V. Tutorial on Support Vector Machine (SVM). **School of EECS, Washington State University**, p. 1–13, 2011.

KAMRUZZAMAN, M.; SUN, D. W. **Introduction to Hyperspectral Imaging Technology**. [s.l.] Elsevier Inc., 2016.

KANG, X.; DUAN, P.; LI, S. Hyperspectral image visualization with edge-preserving filtering and principal component analysis. **Information Fusion**, v. 57, p. 130–143, 2020.

KIANI, S. et al. Hyperspectral imaging as a novel system for the authentication of spices: A nutmeg case study. **Lwt**, v. 104, n. January, p. 61–69, 2019.

LOGGENBERG, K. et al. Modelling water stress in a Shiraz vineyard using hyperspectral imaging and machine learning. **Remote Sensing**, v. 10, n. 2, p. 1–14, 2018.

OKETCH-RABAH, H. A.; MARLES, R. J.; BRINCKMANN, J. A. Cinnamon and Cassia Nomenclature Confusion: A Challenge to the Applicability of Clinical Data. **Clinical Pharmacology and Therapeutics**, v. 104, n. 3, p. 435–445, 2018.

P. GHAMISI et al. Advanced Spectral Classifiers for Hyperspectral Images: A review. **IEEE Geoscience and Remote Sensing Magazine**, v. 5, n. 1, p. 8–32, 2017.

SCIKIT-LEARN DEVELOPERS. **Scikit Learn:** metrics. Metrics. 2020. Disponível em: https://scikit-learn.org/stable/modules/classes.html?highlight=metrics#modulesklearn.metrics. Acesso em: 29 nov. 2020.

SHAHID, M. Z. et al. Antioxidant capacity of cinnamon extract for palm oil stability. **Lipids in Health and Disease**, v. 17, n. 1, p. 1–8, 2018.

TAPSELL, L. C. et al. Health benefits of herbs and spices: the past, the present, the future. **The Medical journal of Australia.**, v. 185, n. 4 Suppl, 2006.

VARRÀ, M. O. et al. Use of near infrared spectroscopy coupled with chemometrics for fast detection of irradiated dry fermented sausages. **Food Control**, v. 110, n. September 2019, p.

107009, 2020.

VIDAL, M.; AMIGO, J. M. Pre-processing of hyperspectral images. Essential steps before image analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 117, p. 138–148, 2012.

VOYANT, C. et al. Machine learning methods for solar radiation forecasting: A review. **Renewable Energy**, v. 105, p. 569–582, 2017.

WANG, X. Z.; LU, S. X. Improved Fuzzy Multicategory Support Vector Machines Classifier. **Proceedings of the 2006 International Conference on Machine Learning and Cybernetics**, v. 2006, n. August, p. 3585–3589, 2006.

WU, D.; SUN, D. W. Colour measurements by computer vision for food quality control - A review. **Trends in Food Science and Technology**, v. 29, n. 1, p. 5–20, 2013.

YASHIN, A. et al. Antioxidant activity of spices and their impact on human health: A review. **Antioxidants**, v. 6, n. 3, p. 1–18, 2017.

ZENG, W. Z. et al. Hyperspectral reflectance models for soil salt content by filtering methods and waveband selection. **Ecological Chemistry and Engineering S**, v. 23, n. 1, p. 117–130, 2016.

ZHANG, C. et al. Rapid and non-destructive measurement of spinach pigments content during storage using hyperspectral imaging with chemometrics. **Measurement: Journal of the International Measurement Confederation**, v. 97, p. 149–155, 2017.