



Serviço Público Federal
Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul



PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DOS MATERIAIS

**“Diferenciação de espécies de Eucalipto usando análise
multivariada e espectroscopia no infravermelho”**

Miller de Oliveira Lacerda

Orientador: Prof. Dr. Cicero Rafael Cena da Silva

Coorientador: Daniel Araujo Gonçalves

Campo Grande – MS

13/09/2022

INSTITUTO DE FÍSICA- INFI

Cidade Universitária | Unidade 5 |

Fone 67 3345 7485

79070-900 | Campo Grande | MS



Serviço Público Federal
Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul



PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DOS MATERIAIS

“Diferenciação de espécies de Eucalipto usando análise multivariada e espectroscopia no infravermelho”

Miller de Oliveira Lacerda

Orientador: Prof. Dr. Cicero Rafael Cena da Silva

Coorientador: Daniel Araujo Gonçalves

Dissertação apresentada à Universidade Federal de Mato Grosso do Sul – UFMS – Instituto de Física – INFI, para obtenção do título de Mestre em Ciência dos Materiais.

Campo Grande – MS
13/09/2022

INSTITUTO DE FÍSICA- INFI
Cidade Universitária | Unidade 5 |
Fone 67 3345 7485
79070-900 | Campo Grande | MS

AGRADECIMENTO

Gostaria de agradecer a todos que me deram seu apoio para que eu chegasse até aqui, sem eles dificilmente teria ido tão longe.

Devo agradecimentos aos membros do GOF/UFMS (Grupo de Ótica e Fotônica), ao meu orientador o prof.ºDrº Cicero Cena por ter disponibilizado seu tempo e paciência para que este trabalho fosse realizado, levando em consideração as condições adversas enfrentadas ao longo dos anos de 2020 e 2021.

À UFMS e ao Instituto de Física pelo apoio à realização do presente trabalho. Este trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), do CNPq.

À Todos vocês muito obrigado.

RESUMO

A madeira de eucalipto é empregada para a produção de celulose, móveis, construção civil, entre outras aplicações, possuindo espécies mais indicadas para cada uso. A separação e classificação de espécies de madeira baseia-se em características como a cor, o cheiro, o sabor, a textura ou o brilho. Alternativamente, análises laboratoriais mais elaboradas podem ser empregadas, demandando maior tempo e custo para análises. Portanto, é necessário desenvolver um novo método preciso, rápido e de baixo custo para a classificação de espécies de madeira. Neste trabalho, investigamos o uso da espectroscopia no infravermelho por transformada de Fourier (FTIR) associado ao aprendizado de máquina como alternativa para a classificação de 6 diferentes espécies de árvores da família dos eucaliptos: *Eucalyptus camaldulensis*, *Corymbia citriodora*, GG100, *Eucalyptus grandis*, *Eucalyptus saligna* e *Eucalyptus urophylla*. Utilizando algoritmos de aprendizado de máquina nos dados obtidos da análise de componentes principais (PCA) do espectro de FTIR de amostras de alburno em pó, foi possível obter modelos preditores para a classificação das espécies. O espectro de infravermelho foi dividido em três regiões, 4000 – 700, 3000 – 2800 e 2000 – 700 cm^{-1} , para análise da influência da celulose, hemicelulose, lignina e extrativos vegetais nos diferentes classificadores. O algoritmo *Support Vector Machine* (SVM), atingiu uma acurácia superior a 90% - em todos os intervalos estudados – nos testes de validação do tipo *Leave One Out Cross Validation*.

Palavras-chave: Eucalipto; FTIR; Alburno; Análise Multivariada; Aprendizado de Máquina.

ABSTRACT

Eucalyptus wood is used for the production of cellulose, furniture, civil construction, among other applications, with the most suitable species for each use. The separation and classification of wood species is based on characteristics such as color, smell, taste, texture or brightness. Alternatively, more elaborate laboratory analyzes can be used, demanding more time and cost. Therefore, it is necessary to develop a new accurate, fast and low-cost method for classifying wood species. In this work, we investigated the use of Fourier transform infrared spectroscopy (FTIR) associated with machine learning as an alternative for the classification of 6 different tree species of the Eucalyptus family: *Eucalyptus camaldulensis*, *Corymbia citriodora*, GG100, *Eucalyptus grandis*, *Eucalyptus saligna* and *Eucalyptus urophylla*. Using machine learning algorithms on data obtained from principal component analysis (PCA) of the FTIR spectrum of powdered sapwood samples, it was possible to obtain predictive models for species classification. The infrared spectrum was divided into three regions, 4000 – 700, 3000 – 2800 and 2000 – 700 cm^{-1} , to analyze the influence of cellulose, hemicellulose, lignin and plant extractives on the different classifiers. The Support Vector Machine (SVM) algorithm achieved an accuracy greater than 90% - in all studied intervals - in the Leave One Out Cross Validation tests.

Keywords: Eucalyptus; FTIR; Machine Learning; Multivariate Analysis: Sap-wood.

Lista de Figuras

Figura 1: Esquema ilustrativo da composição do tronco de uma árvore. (FONTE: Próprio autor)...	14
Figura 2: Representação de uma célula vegetal, com destaque para a presença da celulose e hemicelulose (FONTE: Próprio autor).....	14
Figura 3: representação da estrutura química de uma molécula de celulose. (Fonte: Próprio autor).	15
Figura 4: Representação da estrutura química da hemicelulose (FONTE: Próprio autor).....	16
Figura 5: Representação da composição química da lignina (FONTE: Próprio Autor).....	16
Figura 6: Esquema dos movimentos de translação (Esquerda) e Rotação (Direita) das moléculas. (FONTE: Próprio Autor).....	17
Figura 7: Esquema simplificado para a obtenção das PC's (<i>Principal Components</i>) e montagem do gráfico de dispersão com as duas Primeiras PC's. O Gráfico de <i>Scores</i> (esquerda) e o Gráfico de <i>Loadings</i> (direita). (FONTE: Próprio Autor).....	19
Figura 8: Esquema do funcionamento da técnica de LDA, onde LD1 se mostra o melhor eixo para se obter uma melhor separação dos dados em relação ao eixo LD2. (FONTE: Próprio autor).....	20
Figura 9: Esquema representando o método de SVM para duas dimensões. Os quadrados verdes e os círculos vermelhos representam as amostras a serem estudadas. O gráfico a esquerda (a) representa o SVM Linear e gráfico a direita (b) o SVM Não Linear. (FONTE: Próprio autor).....	21
Figura 10: Esquema demonstrando o funcionamento da técnica de KNN. As setas pretas definem a distância euclidiana entre os pontos pertencentes as duas classes presentes e um novo ponto a ser classificado. (FONTE: Próprio autor).....	22
Figura 11: Média dos espectros de infravermelho com transformada de Fourier (FTIR) das amostras de pó de Eucalipto das espécies Camaldulencis (CA) (<i>Eucalyptus camaldulensis</i>), Citriodora (CIT) (<i>Corymbia citriodora</i>), GG100 (GG), Grandis (GRA) (<i>Eucalyptus grandis</i>), Saligna (SA) (<i>Eucalyptus saligna</i>) e Urophylla (UR) (<i>Eucalyptus urophylla</i>) a faixa mais clara representa o desvio padrão e a linha escura representa a média dos espectros. (Fonte: Próprio Autor).....	28
Figura 12: Gráficos de <i>Scores</i> a esquerda e o gráfico de <i>Loadings</i> na direita. (a) Gráfico de <i>Scores</i> para o intervalo de 4000 a 700 cm^{-1} . (b) Gráfico de <i>Loadings</i> para a PC1, PC2 e PC3 no intervalo de 4000 a 700 cm^{-1} . (c) Gráfico de <i>Scores</i> para o intervalo de 3000 a 2800 cm^{-1} . (d) Gráfico de <i>Loadings</i> para o intervalo de 3000 a 2800 cm^{-1} . (e) Gráfico de <i>Scores</i> para o intervalo de 2000 a 700 cm^{-1} . (f) Gráfico de <i>Loadings</i> para o intervalo de 2000 a 700 cm^{-1} . (FONTE: Próprio Autor).....	30
Figura 13: Gráfico com a acurácia e o número de PC's (números em cima das barras) que os classificadores Análise de Discriminante, KNN e SVM atingem a maior precisão. Em linhas transversais temos o intervalo de 4000 a 700 cm^{-1} e o desempenho dos respectivos classificadores. Em linhas horizontais temos o intervalo de 3000 a 2800 cm^{-1} e o respectivo desempenho dos	

classificadores. Em linhas cruzadas temos o intervalo de 2000 a 700 cm ⁻¹ com o desempenho dos classificadores. (FONTE: Próprio Autor).....	31
Figura 14: matriz de confusão do SVM (SVM Médio) com 95,42% de acurácia para o intervalo de 2000 a 700 cm ⁻¹ .(FONTE: Próprio Autor).....	32
Figura 15: Espectros de FTIR de diferentes espécies de eucaliptos: <i>Corymbia Citriodora</i> (CIT), <i>Eucalyptus grandis</i> (GRA) e <i>E. Grandis</i> x <i>E. Urophylla</i> (GG). (FONTE: Próprio Autor).....	33
Figura 16: Gráficos de <i>Scores</i> a esquerda e o gráfico de <i>Loadings</i> na direita para amostras da casca. (a) Gráfico de <i>Scores</i> para o intervalo de 4000 a 700 cm ⁻¹ . (b) Gráfico de <i>Loadings</i> para a PC1, PC2 e PC3 no intervalo de 4000 a 700 cm ⁻¹ . (c) Gráfico de <i>Scores</i> para o intervalo de 3800 a 2800 cm ⁻¹ . (d) Gráfico de <i>Loadings</i> para o intervalo de 3800 a 2800 cm ⁻¹ . (e) Gráfico de <i>Scores</i> para o intervalo de 1800 a 800 cm ⁻¹ . (f) Gráfico de <i>Loadings</i> para o intervalo de 1800 a 800 cm ⁻¹ . (FONTE: Próprio Autor).....	34
Figura 17: Gráfico com a acurácia e o número de PC's (números em cima das barras) que os classificadores Análise de Discriminante, KNN e SVM atingem a maior precisão nos intervalos estudados.(FONTE: Próprio Autor).....	35
Figura 18: Matriz de confusão para o LDA (<i>Linear Discriminant Analysis</i>) no intervalo de 3800 a 2800 cm ⁻¹ . (Fonte: Próprio Autor).....	36
Figura 19: Comparação entre o espectro de infravermelho do alburno (esquerda) e casca (direita). (Fonte: Próprio Autor).....	36
Figura 20: Comparação entre os <i>Scores</i> e os <i>Loadings</i> do Alburno (esquerda) e da Casca (direita) de 3 diferentes espécies de eucalipto. (a, b) Gráfico de <i>Scores</i> para o intervalo de 4000 a 700 cm ⁻¹ . (c, d) Gráfico de <i>Loadings</i> para a PC1, PC2 e PC3 no intervalo de 4000 a 700 cm ⁻¹ . (e, f) Gráfico de <i>Scores</i> para o intervalo de 3800 a 2800 cm ⁻¹ . (g, h) Gráfico de <i>Loadings</i> para o intervalo de 3800 a 2800 cm ⁻¹ . (i, j) Gráfico de <i>Scores</i> para o intervalo de 1800 a 800 cm ⁻¹ . (k, l) Gráfico de <i>Loadings</i> para o intervalo de 1800 a 800 cm ⁻¹ . (FONTE: Próprio Autor).....	38
Figura 21: Comparação da acurácia entre o Alburno, listras transversais, e a Casca, linhas horizontais, com o número de PC's acima de cada barra para os três intervalos analisados. (FONTE: Próprio Autor).....	39
Figura 22: Comparação da matriz de confusão do Alburno (esquerda) e da Casca (direita) para as amostras CIT, GG, e GRA para os classificadores KNNF (esquerda) e QDA (direita) para o intervalo de 4000 – 700 cm ⁻¹ . (FONTE: Próprio Autor).....	40

LISTA DE ABREVIACÕES

COPANT – Comissão Panamericana de Normas Técnicas

FTIR – *Fourier-transform Infrared Spectroscopy*

ML – *Machine Learning*

PCA – *Principal Components Analysis*

PC – *Principal Components*

SVM – *Support Vector Machine*

LDA – *Linear Discriminant Analysis*

KNN – *K-Nearest Neighbors*

ROC – *Receiver Operating Characteristic Curve*

AUC – *Area Under the Curve*

LOOCV – *Leave One Out Cross Validation*

CA – *Eucalyptus camaldulensis*

CIT – *Corymbia citriodora*

GG – GG100

GRA – *Eucalyptus grandis*

SA – *Eucalyptus saligna*

UR – *Eucalyptus urophylla*

ATR – *Attenuated Total Reflectance*

SNV – *Standad Normal Variation*

QDA – *Quadratic Discriminant Analysis*

KNNF – KNN Fino

KNNM – KNN Médio

KNNG – KNN Grosso

KNNC – KNN Cosseno

KNN2 – KNN Cúbico

KNNP – KNN Ponderado

KNNS – KNN Subespaço

SVMF – SVM Fino

SVMM – SVM Médio

SVMG – SVM Grosso

SVML – SVM Linear

SVM2 – SVM Quadrático

SVM3 – SVM Cúbico

SUMÁRIO

1	INTRODUÇÃO.....	11
2	REVISÃO BIBLIOGRÁFICA.....	13
2.1	O EUCALIPTO: DESCRIÇÃO E RELEVÂNCIA.....	13
2.2	ESPECTROSCOPIA NO INFRAVERMELHO.....	17
2.3	ANÁLISE MULTIVARIADA.....	18
2.3.1	<i>PRINCIPAL COMPONENT ANALYSIS – PCA</i>	19
2.3.2	<i>LINEAR DISCRIMINANT ANALYSIS – LDA</i>	20
2.3.3	<i>SUPPORT VECTOR MACHINE – SVM</i>	20
2.3.4	<i>K-NEAREST NEIGHBOR – KNN</i>	21
2.4	MÉTODO DE VALIDAÇÃO DOS DADOS.....	22
3	OBJETIVOS.....	23
4	MATERIAIS E MÉTODOS.....	24
4.1	PREPARAÇÃO DAS AMOSTRAS.....	24
4.2	CARACTERIZAÇÃO DAS AMOSTRAS.....	24
4.3	ANÁLISE EXPLORATÓRIA DOS DADOS.....	25
4.4	APRENDIZADO DE MÁQUINA.....	25
5	RESULTADO E DISCUSSÃO.....	27
5.1	CASCA.....	32
5.2	COMPARAÇÃO.....	36
6	CONCLUSÃO.....	41
7	REFERENCIAS.....	42

1. Introdução

O eucalipto é uma espécie de planta de grande interesse econômico, comumente empregada para a produção de celulose, móveis, construção civil, entre outras aplicações, possuindo espécies mais indicadas para cada uso, só no Brasil há uma produtividade média estimada de 39 m³/ha/ano [1]. Esse gênero de planta apresenta rápido crescimento e grande adaptabilidade a climas quentes e secos [2]. Dentre as espécies mais comuns da família Myrtaceae temos: a *Eucalyptus Camaldulensis* (CA), *Corymbia Citriodora* (CIT), *Eucalyptus Grandis* (GRA), *Eucalyptus Saligna* (SA), *Eucalyptus Urophylla* (UR), e a espécie híbrida entre *Eucalyptus Grandis* x *Eucalyptus Urophylla* conhecida como GG100 (GG). Usualmente, a separação e classificação destas espécies baseia-se em métodos subjetivos como a cor, o cheiro, o sabor, a textura ou o brilho da madeira [3-6]. Alternativamente, análises laboratoriais mais elaboradas podem ser empregadas, com a desvantagem de demandar maior tempo e custo para análises [7-9]. Portanto, o desenvolvimento de um novo método laboratorial preciso, rápido e de baixo custo para a classificação de espécies de madeira se faz necessário. Nesta vertente, o emprego de técnicas fotônicas combinada com técnicas de aprendizado de máquina (*Machine Learning* – ML) tem demonstrado alto potencial para a classificação de materiais de diferentes natureza, tais como sementes [10,11], materiais biológicos [12,13], materiais forenses [14,15].

Neste contexto, como a madeira é composta por celulose ((C₆H₁₀O₅)_n), hemicelulose ((C₅H₈O₄)_n e (C₆H₁₀O₅)_n), lignina (C₉H₁₀O₂, C₁₀H₁₂O₃ ou C₁₁H₁₄O₄) [28,29,30] e extrativos vegetais em diferentes concentrações, com modos ativos no infravermelho médio o emprego de técnicas de espectroscopia molecular e ML apresenta potencial para o desenvolvimento de um método prático para a classificação de espécies de eucalipto.

Estudos dedicados a classificação de madeira, possibilitam a identificação de 22 espécies de madeira (das famílias: Aceraceae, Anacardiaceae, Araucariaceae, Betulaceae, Cupressaceae, Fabaceae, Fagaceae, Juglandaceae, Oleaceae, Pinaceae, Rosaceae e Salicaceae) pela análise da mistura do pó do cerne e alburno utilizando espectroscopia no infravermelho com transformada de Fourier (FTIR) e *Linear Discriminant Analysis* (LDA), com uma acurácia de 87,5% com base na densidade e nos componentes químicos da madeira [16]. Outro trabalho de identificação de espécies pela análise do cerne de duas espécies de pinheiro usando FTIR e *Discriminant Analysis*, obteve uma precisão de 99% que foi atribuída a quantidade de lignina e o grau de cristalização de polissacarídeos [17]. A análise de amostras de duas espécies de nogueira para classificação entre tratada e não tratada termicamente utilizando FTIR e *Partial Least Square Discriminant Analysis*

(PSL-DA), obteve uma acurácia de 87,5% através de diferenças na lignina e celulose [18]. A classificação de oito espécies de louros utilizando *Near-infrared Spectroscopy* (NIR) e PLS – DA, obteve uma precisão superior a 97% com base em diferenças na anatomia da madeira [19]. Por fim, a classificação de *Guibourtia* spp., *Guibouritia*, *Pterocarpus macrocarpus* Kurz, *Pterocarpus erinaceus* Poir, utilizando *Laser-induced breakdown Spectroscopy* (LIBS) e *Artificial Neural Networks* (ANN), com acurácia de 100% utilizando a intensidade dos elementos O e Na como base [20].

Assim, como as espécies de madeira analisadas nos trabalhos citados não há a presença de espécies de eucalipto e por não ser uma espécie nativa é usada para produção de celulose e carvão, onde a correta classificação é necessária para evitar prejuízos, neste trabalho, investigamos o uso de FTIR associado ao aprendizado de máquina como alternativa para a classificação de 6 diferentes espécies de árvores da família dos eucaliptos (CA, CIT, SA, UR e GG). O espectro de FTIR de amostras de alburno em pó, após o pré-tratamento com *Standard Normal Variation* (SNV), foi analisado usando análise de componentes principais (PCA) e algoritmos de aprendizado de máquina, em três regiões distintas do espectro, 4000 – 700, 3000 – 2800 e 2000 – 700 cm⁻¹, para obter modelos preditores para a classificação das espécies.

2. Revisão Bibliográfica

2.1. O Eucalipto: Descrição e Relevância

No Brasil as espécies mais comuns de eucalipto são *Eucalyptus Camaldulensis* (CA), *Corymbia Citriodora* (CIT), *Eucalyptus Grandis* (GRA), *Eucalyptus Saligna* (SA), *Eucalyptus Urophylla* (UR), e a espécie híbrida entre *Eucalyptus Grandis* x *Eucalyptus Urophylla* conhecida como GG100 (GG), onde são utilizadas principalmente na produção de carvão e celulose, e na construção civil. Destacando-se por sua grande adaptabilidade a climas quentes e secos, além de um rápido crescimento [2] e apresentando uma produtividade média estimada de 39 m³/ha/ano [1].

A tabela 1 apresenta algumas características das seis espécies supracitadas, onde é possível notar as diferentes aplicações de cada uma das espécies, sua altura média, e o tempo de cultivo. Embora não exista uma restrição entre espécie e o tipo de aplicação, o tempo de cultivo e o tipo de aplicação apresentam maior correlação, uma vez que o envelhecimento do eucalipto altera suas propriedades físicas e químicas [21,22], como o aumento da resistência mecânica, algo necessário na produção de moveis ou tábuas, diferentemente da produção de celulose e carvão.

Tabela 1: Principais espécies de madeira produzidas no Brasil e respectivas características de cultivo e aplicação. *Eucalyptus Camaldulensis* (CA), *Corymbia Citriodora* (CIT), *Eucalyptus Grandis* (GRA), *Eucalyptus Saligna* (SA), *Eucalyptus Urophylla* (UR), e a espécie híbrida (GRA + UR) GG100 (GG).

Espécie	Altura média (m)	Tempo de cultivo (anos)	Aplicação	Ref.
CA	20	12 a 13	Fabricação de móveis	23,24
CIT	22	6 a 8	Produção de óleos essenciais e carvão	25
GG	23		Produção de celulose	26
GRA	25	7 a 13	Produção de celulose e construção civil	27
SA			Produção de celulose e carvão, construção civil	28
UR	35		Fabricação de móveis, produção de celulose e carvão	29

A figura 1 ilustra uma representação do tronco, onde no centro está localizado o cerne de cor mais escura, e em torno do cerne o alburno, onde é possível visualizar os anéis de crescimento e os raios medulares e, por fim, a casca constituída pela cortiça e o floema vivo.

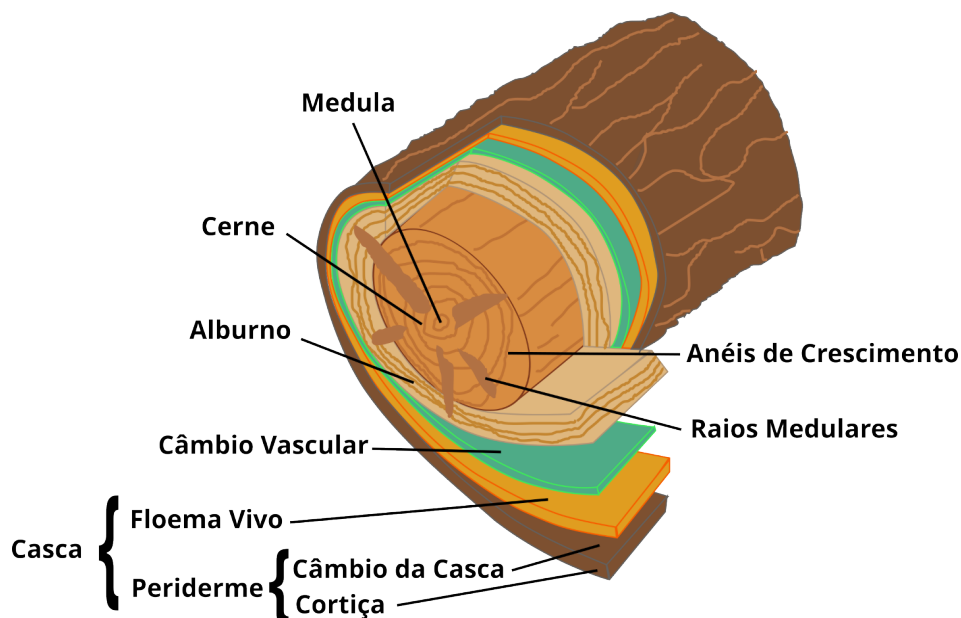


Figura 1: Esquema ilustrativo da composição do tronco de uma árvore. (FONTE: Próprio autor).

A casca tem como função fornecer proteção contra evaporação excessiva, ataque de insetos, clima e também serve como isolante térmico [30] e assim como o cerne é formado por células mortas [22,31]. O alburno tem como função conduzir água e nutrientes das raízes para as folhas [21,22]. Conforme as células do alburno morrem elas dão origem ao cerne [32,33], responsável pelo suporte mecânico da planta [33].

Independentemente da parte da planta, suas células possuirão uma estrutura similar a apresentada na figura 2, que apresenta as principais componentes da parede celular. Como principais componentes destas estruturas podemos citar a celulose que se apresenta na forma de microfibrila (representada por tubos azuis) e a hemicelulose (representada por linhas verdes), que juntas fazem parte da parede celular primária.

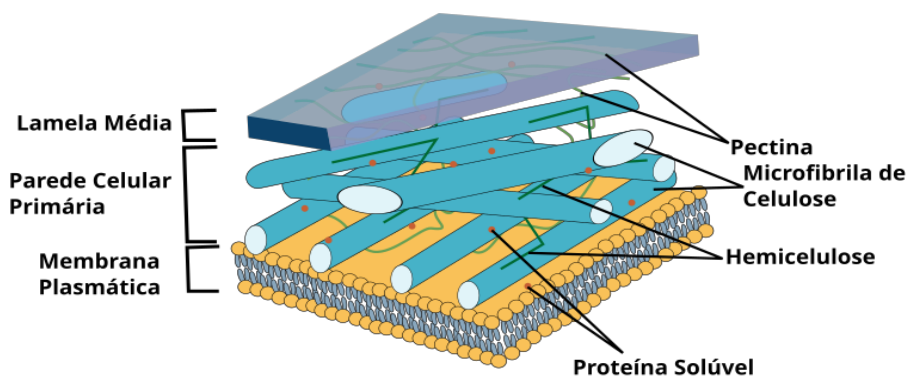


Figura 2: Representação de uma célula vegetal, com destaque para a presença da celulose e hemicelulose (FONTE: Próprio autor).

A concentração de celulose, hemicelulose, lignina e extrativos vegetais não é homogênea, em média variam entre 30 a 50% para a celulose, de 9 a 28% de hemicelulose, de 22 a 40% para a lignina e de 0,2 a 20% de extrativos vegetais. Observa-se variações composicionais com a idade da planta e posição: base ou topo; casca; alburno; e cerne [34].

A celulose $((C_6H_{10}O_5)_n)$ é um polissacarídeo, figura 3, e principal componente da parede celular de vários tipos de plantas, tem como principal função dar resistência as paredes das células ao formar microfibrilas [35]. Na produção de papel é o principal componente, junto com a hemicelulose [36]. Na figura 3, se destaca a presença de longas cadeias carbônicas, as hidroxilas, ligações C – C e C – O, e sua estrutura básica $(C_6H_{10}O_5)$ se une uma com a outra através de ligações do tipo $\beta(1-4)$ Glicosídicas.

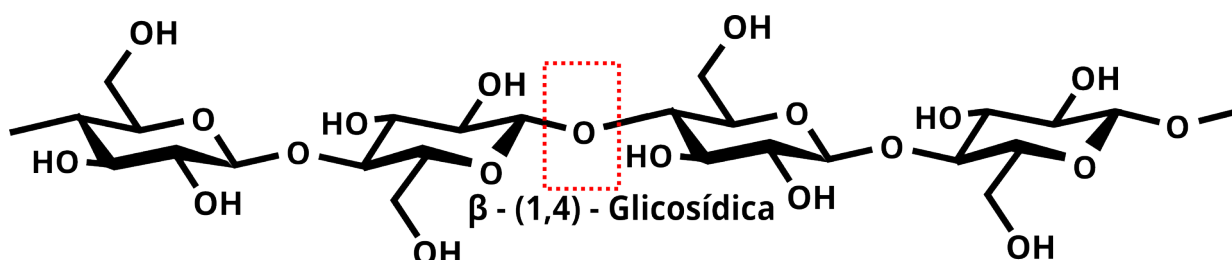


Figura 3: representação da estrutura química de uma molécula de celulose. (Fonte: Próprio autor).

A hemicelulose $((C_5H_8O_4)_n$ e $(C_6H_{10}O_5)_n$) se refere a uma classe de heteropolímeros que incluem por exemplo: as glucoranoxilanas, arabinoglucoranoxilanas, glucomanamas entre outras [37], sua estrutura química, figura 4, é similar à da celulose com a presença de longas cadeias de carbonos, hidroxilas e ligações C – C, C – O, mas diferente da celulose a hemicelulose não possui uma estrutura linear e é mais suscetível a degradação e possui menor resistência ao calor [38].

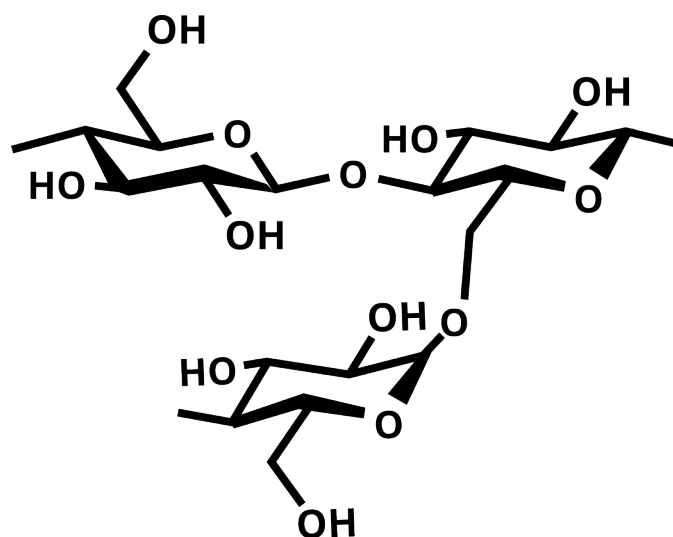


Figura 4: Representação da estrutura química da hemicelulose (FONTE: Próprio autor).

A lignina ($C_9H_{10}O_2$, $C_{10}H_{12}O_3$ ou $C_{11}H_{14}O_4$) é um composto heterogêneo altamente aromático [39] e tem uma importante função estrutural para a parede celular conferindo rigidez e dureza [39], além de preencher os espaços entre a celulose e a hemicelulose [40]. Também tem papel importante no transporte de água pela planta [40]. Na figura 5 podemos ver a estrutura química da lignina que possui grande quantidade de anéis aromáticos, além da presença de hidroxilas e tem uma estrutura mais organizada que a celulose e a hemicelulose.

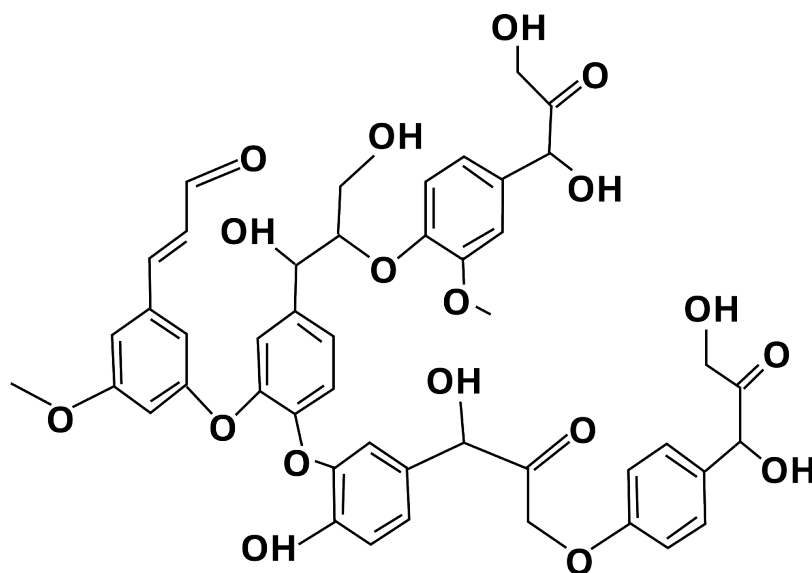


Figura 5: Representação da composição química da lignina (FONTE: Próprio Autor).

2.2. Espectroscopia no infravermelho

A espectroscopia é uma área da ciência que estuda a interação da matéria com o espectro eletromagnético, essa interação pode ocorrer por absorção, emissão ou espalhamento [41], dependendo da faixa do espectro eletromagnético utilizado pode-se obter informações sobre a composição atômica e a organização dos átomos dentro do material [42].

A espectroscopia no infravermelho, que compreende a faixa de 14000 cm^{-1} até 10 cm^{-1} , é muito utilizada por ser a faixa do espectro eletromagnético que interage com os modos vibracionais das moléculas, permitindo a análise dos elementos químicos presentes no material [43]. Em especial a região de 1800 cm^{-1} a 400 cm^{-1} é onde se encontram as assinaturas espectrais, região das frequências espectrais fundamentais [44]. A interação entre o espectro de infravermelho e o material só é possível quando as moléculas que o compõem possuem um momento de dipólo elétrico [41].

As frequências de vibração que interagem com o espectro de infravermelho podem ser de tipo simétrico, quando a vibração das moléculas está em fase, ou assimétricas, quando a vibração das moléculas está fora de fase [43]. Usando um sistema de coordenadas como referência, figura 6, as vibrações simétricas ou assimétricas podem ser explicadas em composições de movimentos relacionados aos eixos do sistema de coordenadas, representados pelas setas roxas. Outra forma de interação da radiação infravermelha é com as rotações das moléculas [42,43], que podem ser explicados através de rotações em torno de eixos coordenados, setas pontilhadas. Alguns tipos de vibração e rotação podem não estar disponíveis para todas as moléculas devido a sua simetria.

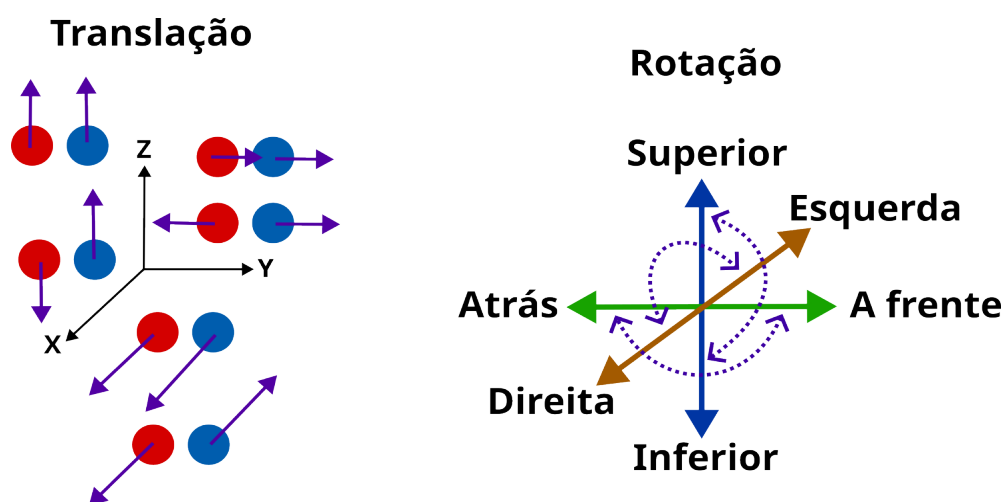


Figura 6: Esquema dos movimentos de translação (Esquerda) e Rotação (Direita) das moléculas.(FONTE: Próprio Autor).

A energia associada as vibrações e rotações das moléculas são quantidades quantizadas, possuindo um estado fundamental e estados excitados que podem ou não ser degenerados [45]. Para baixas energias pode se considerar os níveis de energia como próximos ao de um oscilador harmônico. Com a resolução da equação diferencial para o oscilador harmônico em baixa energia é possível estimar qual o número de onda para a banda absorvida e assim determinar a qual átomo ou molécula ele pertence.

A vibração de uma molécula pode ser dividida entre a frequência estrutural, que envolve a vibração de todos os átomos, e a frequência de grupo, que envolve um pequeno número de átomos [42,43,45]. Cada grupo funcional, como ligações C – C, C = C, OH, possui frequências bem definidas permitindo sua identificação, analisando a posição e intensidade dos picos, no espectro de infravermelho. Uma vez que parte da radiação infravermelha é absorvida, ao comparar o espectro absorvido e as frequências características dos grupos funcionais é possível identificar a molécula.

Os modos vibracionais da celulose incluem vibrações de C – C, C – O – C, C – H e OH, os mesmos componentes são encontrados na hemicelulose, com a adição de C – O, a lignina apresenta anéis aromáticos e C = O, além daquelas encontradas na celulose.

2.3. Análise Multivariada

A análise multivariada é um ramo da estatística que lida com múltiplas variáveis de forma conjunta [46] com o objetivo de reconhecer padrões em grandes bases de dados de natureza complexa, que se tornou mais relevante com a melhora da tecnologia [47].

O *Machine Learning* (ML) utiliza da análise multivariada para os métodos de aprendizagem dos algoritmos, que podem ser divididos entre o Aprendizado Supervisionado, Aprendizado Não Supervisionado e Reforço, sendo a diferença entre o aprendizado supervisionado e não supervisionado é que no aprendizado supervisionado os dados precisam estar rotulados, cada conjunto de dados deve ter seu “nome” indicado para o classificador [48,49].

Os métodos não supervisionados podem ser divididos em análise de *clusters* ou redução de dimensão [48], podemos citar como exemplo o PCA (*Principal Component Analysis*). Já os métodos supervisionados podem ser divididos em métodos de regressão ou classificação [50] entre esses modelos de classificação temos o LDA (*Linear Discriminant Analysis*), KNN (*K-Nearest Neighbor*) e SVM (*Support Vector Machine*).

2.3.1. Principal Component Analysis – PCA

O PCA é um método de análise multivariada não supervisionada que permite a redução na dimensão dos dados a serem estudados e cria um novo conjunto de variáveis não correlacionadas entre si que carregam a variação dos dados e preservam a informação estatística [51], as PC's (*Principal Components*), onde a primeira PC carrega a maior variação dos dados, a segunda PC a segunda maior variação dos dados e assim por diante de forma decrescente, sendo que as PC's não são correlacionadas entre si [51].

Na figura 7 temos um esquema simplificado do processo de obtenção das PC's. Primeiramente se monta uma matriz ($n \times p$), onde p representa as variáveis e n as amostras, a partir dessa matriz é preciso encontrar a matriz de variância – covariância dos dados, ao resolver a equação característica da matriz de variância – covariância se obtêm os autovalores e os autovetores, onde cada PC é igual a um autovalor correspondente. Através das PC's é possível encontrar os *Scores*, que representam os valores das PC's, e o *Factor Loading*, que demonstram a relação de uma variável com um determinado fator, que serão utilizados para montar gráficos de dispersão onde se espera encontrar a formação de grupos com fronteiras bem delineadas e com o suporte do gráfico de *Loadings* encontrar os intervalos que ajudam a explicar essa separação.

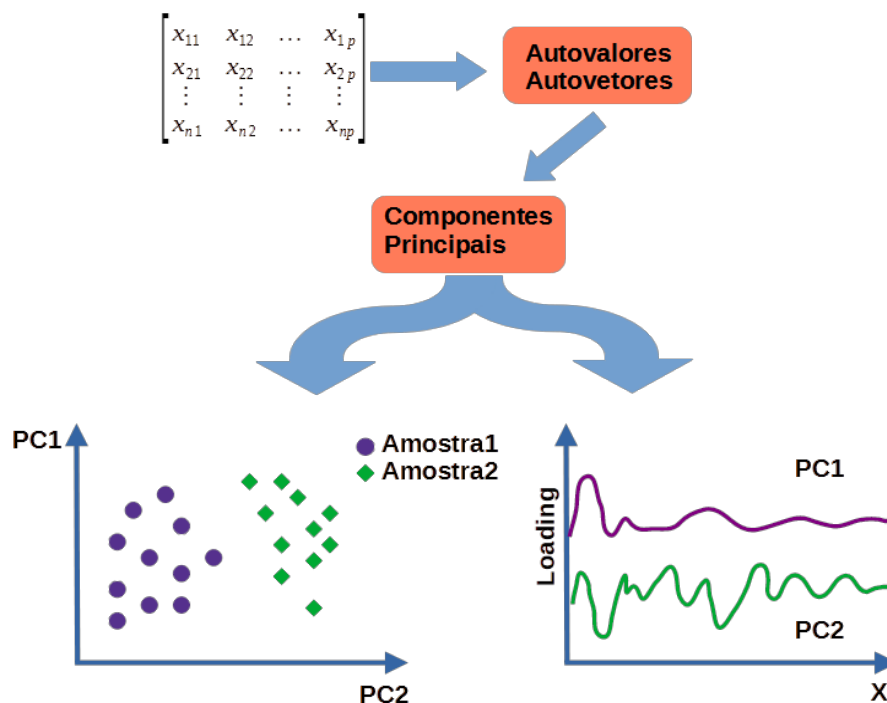


Figura 7: Esquema simplificado para a obtenção das PC's (*Principal Components*) e montagem do gráfico de dispersão com as duas Primeiras PC's. O Gráfico de *Scores* (esquerda) e o Gráfico de *Loadings* (direita). (FONTE: Próprio Autor).

2.3.2. Linear Discriminant Analysis – LDA

No LDA o objetivo é reduzir a dimensionalidade dos dados e manter a separação dos grupos, através da criação de fronteiras que maximizam a diferença entre as classes dos dados analisados [52]. Para fazer isso é preciso saber a média das diferenças entre as classes e também dentro de uma mesma classe, para se montar a matriz de espalhamento. Resolvendo a equação característica se obtêm os autovetores e os autovalores, e no subespaço definido por eles encontra-se o eixo em que as classes apresentem uma maior separação [52].

Na figura 8 temos duas amostras, cruzes laranjas e triângulos roxos, distribuídos entre dois eixos LD1 e LD2, e as linhas laranjas e roxas presentes em cada eixo representam a projeção dos dados nos respectivos eixos, e por isso é possível observar que o eixo LD1 é capaz de separar melhor os dados comparado ao eixo LD2.

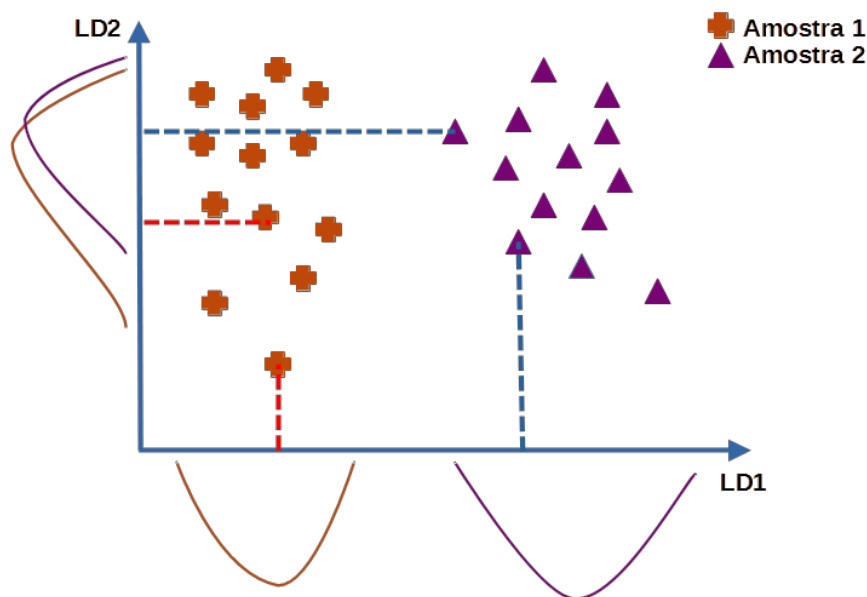


Figura 8: Esquema do funcionamento da técnica de LDA, onde LD1 se mostra o melhor eixo para se obter uma melhor separação dos dados em relação ao eixo LD2. (FONTE: Próprio autor).

2.3.3. Support Vector Machine – SVM

A princípio o SVM foi criado para lidar com problemas de separação binária, mas depois foi adaptado para lidar com problemas multiclasse, através de uma escolha adequada da função de transformação ou função *kernel* [53]. O *Kernel* pode ser uma função linear, cúbica, seno, entre

outras, cada uma dessas funções produz diferentes fronteiras de decisão para os dados analisados [54].

Para que o SVM encontre a fronteira de decisão que melhor separa os dados, é preciso primeiramente selecionar os vetores de suporte que percorreram os dados até que se atinja a distância máxima de separação entre os dados [53,54]. Essa fronteira de decisão é denominada de hiperplano. É possível ver o resultado do SVM, figura 9, onde temos duas amostras, verdes e vermelhas, separadas pelo hiperplano, obtido por um *kernel* linear na figura (a) e por um *kernel* não-linear na figura (b), que atinge a margem máxima de separação entre as amostras.

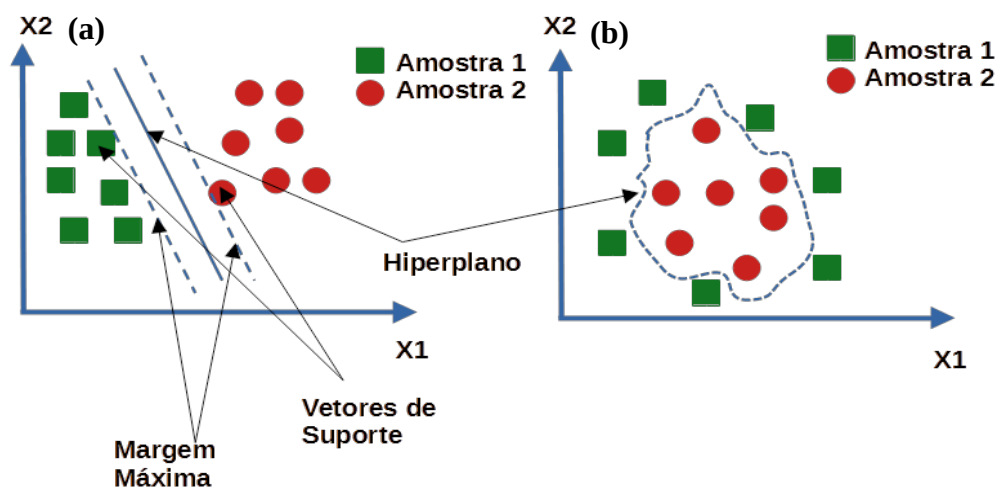


Figura 9: Esquema representando o método de SVM para duas dimensões. Os quadrados verdes e os círculos vermelhos representam as amostras a serem estudadas. O gráfico a esquerda (a) representa o SVM Linear e gráfico a direita (b) o SVM Não Linear. (FONTE: Próprio autor).

2.3.4. *K-Nearest Neighbor – KNN*

O KNN analisa a distância entre um par de dados com um terceiro e dependendo da distância em que este se encontra, agrupa esse dado ao grupo já formado ou o considera pertencente a outro grupo de dados [55,56]. Para fazer isso primeiro é necessário selecionar a quantidade de vizinhos a ser usada para calcular a distância entre eles, seja através da distância euclidiana ou mahalanobis, com a distância determinada se separa os dados em grupos através da proximidade, dados mais próximos pertencem ao mesmo grupo e dados mais distantes pertencem a outro grupo, como pode ser visto na figura 10.

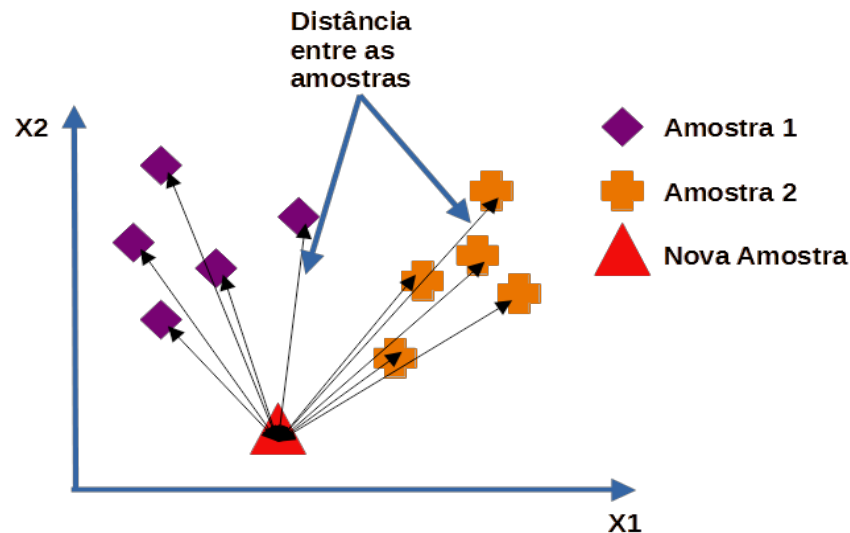


Figura 10: Esquema demonstrando o funcionamento da técnica de KNN. As setas pretas definem a distância euclidiana entre os pontos pertencentes as duas classes presentes e um novo ponto a ser classificado. (FONTE: Próprio autor).

Um motivo para o uso de diferentes tipos de ML é o fato de cada deles possuir uma precisão diferente e um custo computacional, tempo de execução do algoritmo, por causa das diferentes abordagens utilizada por cada um desses métodos [57].

2.4. Métodos de Validação dos Dados

Um método de validação muito usado é o LOOCV (*Leave One Out Cross Validation*) para validação interna, onde não é feita essa reserva dos dados [58]. Os dados são divididos de forma aleatória em tamanhos iguais e separados em partições (*Folds*) onde uma dessas partições é usada para a validação e as outras para o treinamento dos modelos, a partição usada para validação muda até que todas as partições sejam usadas para a validação, assim é possível ter uma visão de qual é o desempenho de cada um dos modelos para cada conjunto de dados [59].

3. Objetivos

Gerais:

Classificar/identificar diferentes espécies de eucalipto utilizando espectroscopia no infravermelho com métodos de classificação multivariada.

Específicos:

Classificar diferentes espécies de eucalipto utilizando FTIR associado a análise multivariada.

Investigar o espectro obtido com algoritmos de ML.

Identificar os grupos funcionais que contribuem para a classificação.

Determinar qual parte da madeira que permite a melhor comparação.

4. Materiais e Métodos

4.1. Preparação das amostras

As amostras de eucalipto utilizadas pertencem a 4 lotes de seis espécies diferentes: *Corymbia Citriodora* (CIT), *Eucalyptus Grandis* (GRA), *Eucalyptus Saligna* (SA), *Eucalyptus Urophylla* (UR), e a espécie híbrida entre *Eucalyptus Grandis* x *Eucalyptus Urophylla* conhecida como GG100 (GG).

De cada espécie analisada foi coletado o alburno, parte interna do tronco da árvore, e a casca, ilustrado na figura 1 temos as principais estruturas do tronco. As amostras de alburno foram trituradas e sua granulometria controlada utilizando uma peneira analítica de 710 μm . O material retido foi macerado com auxílio de gral e pistilo, e novamente peneirado – esse processo é repedido até completa passagem do material pela peneira – assegurando a uniformidade das amostras.

A casca das espécies *Corymbia Citriodora* (CIT), *Eucalyptus Grandis* (GRA), e a espécie híbrida GG100 (GG), foram primeiramente cortadas com uma tesoura em pequenos pedaços e depois foram levados ao liquidificador para serem trituradas, a granulometria foi controlada com uma peneira analítica de 710 μm . O material que ficou retido na peneira foi macerado no almofariz e novamente foi colocada no liquidificador e peneirado, para assegurar que as amostras sejam uniformes esse processo é repetido até a completa passagem do material pela peneira.

4.2. Caracterização das Amostras

Os espectros no infravermelho por transformada de Fourier das amostras em pó foram obtidos em um espectrômetro Perkin-Elmer modelo Spectrum 100N FT-NIR com o acessório de ATR (*Attenuated Total Reflectance*) com cristal de germânio. Os parâmetros experimentais para cada uma das medições foram 10 varreduras no intervalo de 4000 a 700 cm^{-1} , com resolução de 4 cm^{-1} e intervalo de 0,5 cm^{-1} , e 10 *scan*, o *background* foi realizado antes de cada amostra e as análises foram realizadas em duplicata no modo de transmitância, gerando um total de 40 medidas para cada amostra.

4.3. Análise exploratória dos dados

Os espectros obtidos foram submetidos a um pré-tratamento para remoção de *offset* usando a técnica de SNV (*Standard Normal Variation*), que consiste em subtrair do espectro sua média e o resultado ser dividido pelo seu desvio padrão, isso faz com que cada espectro obtido tenha média zero e desvio padrão igual a um. Além disso a normalização por SNV permite que os espectros sejam proporcionais ao remover o desvio vertical (*offset*) existente entre as amostras mantendo a forma geral do espectro e ajudando a remover problemas como espalhamento e mudanças no caminho óptico [60], evitando que esses parâmetros contribuam para separação dos grupos observados em nossas análises.

O espectro após esse pré-tratamento foi estudado utilizando o método de análise de componentes principais (PCA), que é um método matemático usado para a redução de dimensionalidade dos dados em componentes que carregam a maior variação do conjunto de dados originais, tornando assim mais fácil a compreensão dos dados através das Componentes Principais (PC) [51].

4.4. Aprendizado de Máquina

Foi utilizado nesse trabalho o programa Matlab versão R2018a (Mathworks, USA) para a realização tanto da análise de PCA quanto das técnicas de aprendizado de máquina supervisionado (*Machine Learning* – ML). O número de PC's utilizado nos algoritmos de ML foi escolhido de forma a evitar *Overfitting/Underfitting*.

O aprendizado de máquina supervisionado compreende vários tipos de análises, neste trabalho usamos o *Support Vector Machine* (SVM), *Linear Discriminant Analysis* (LDA), e *K-Nearest Neighbors* (KNN), descritos em mais detalhes no capítulo 2.

A principal diferença entre as técnicas de ML é a forma como os dados são usados por esses classificadores e o gasto computacional envolvido em cada um deles. Outra diferença é com relação a distribuição dos dados, para o LDA se supõe que os dados estão em uma distribuição normal [52], já para o SVM não é necessário fazer nenhuma suposição [54], e para o KNN é necessário o número correto de vizinhos próximos que foi escolhido para que a análise funcione [56].

Além de cada um desses classificadores possuir um desempenho diferente é necessário ter um método para se calcular a precisão desses algoritmos. O cálculo da precisão dos algoritmos utilizados neste trabalho foi realizado através do *Leave One Out Cross Validation* (LOOCV), esse teste nos permite avaliar a confiabilidade de cada um dos modelos de predição [58].

5. Resultado e Discussão

Analisando todo o intervalo do espectro de infravermelho, de 4000 a 700 cm^{-1} , na figura 11 vemos uma média dos espectros para cada uma das espécies estudadas, a linha escura representa a média dos dados e a parte sombreada representa a variação dos dados. Como pode ser observado na figura os espectros médios das amostras são similares.

Há um largo pico em torno de 3382 cm^{-1} que está associada aos diferentes modos vibracionais do O – H, usualmente associado a presença de água [61,62].

A banda entorno do pico em 2926 cm^{-1} indica a presença de vibrações de C – H que pode ser associado principalmente a molécula de celulose $((\text{C}_6\text{H}_{10}\text{O}_5)_n)$ [62,63]. Em 1737 cm^{-1} temos a presença de modos vibracionais de C = O que pode ser associado a molécula de hemicelulose $((\text{C}_5\text{H}_8\text{O}_4)_n)$ e $((\text{C}_6\text{H}_{10}\text{O}_5)_n)$ [64,65]. Já entorno dos picos em 1737, 1245, 1155 e 1046 cm^{-1} temos a vibração de C = O, C – H, C – O – C e C – O que estão presentes em diferentes tipos de carboidratos [65]. Ainda em 1737 cm^{-1} podemos destacar a inversão do pico na amostra de CIT. Em 1327 e 1106 cm^{-1} indicam os modos vibracionais do C – H e C – O que se encontram na lignina $(\text{C}_9\text{H}_{10}\text{O}_2, \text{C}_{10}\text{H}_{12}\text{O}_3)$ ou $\text{C}_{11}\text{H}_{14}\text{O}_4)$ e carboidratos [63].

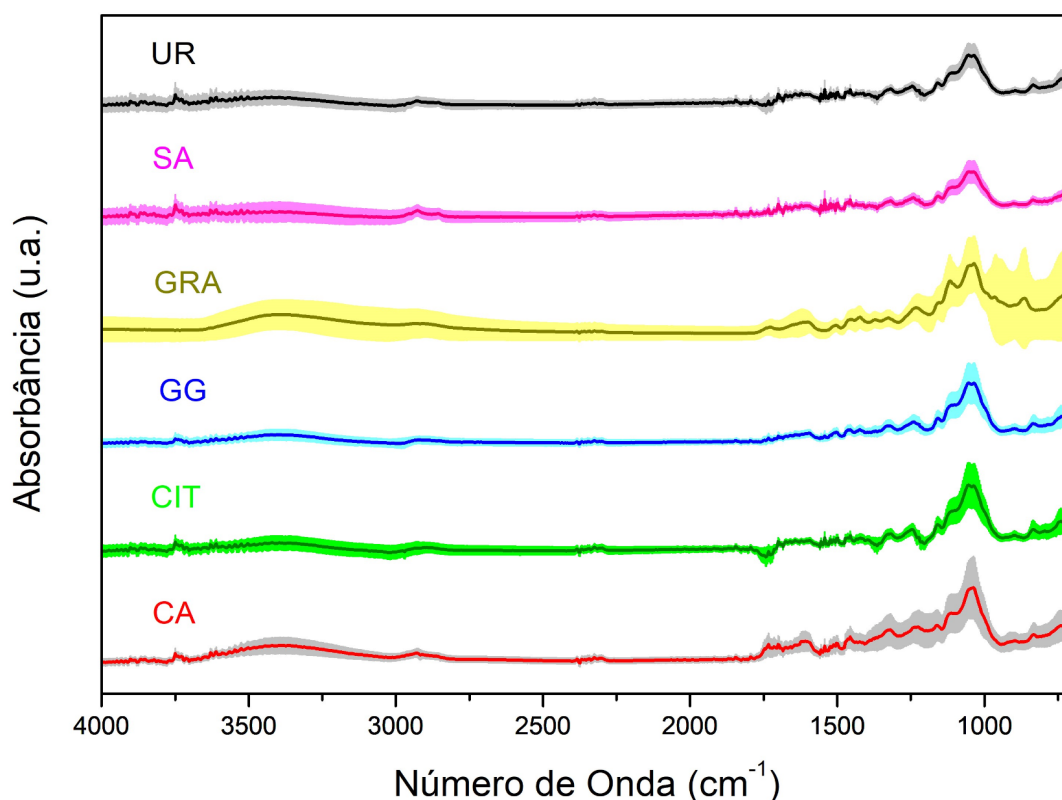


Figura 11: Média dos espectros de infravermelho com transformada de Fourier (FTIR) das amostras de pó de Eucalipto das espécies Camaldulencis (CA) (*Eucalyptus camaldulensis*), Citriodora (CIT) (*Corymbia citriodora*), GG100 (GG), Grandis (GRA) (*Eucalyptus grandis*), Saligna (SA) (*Eucalyptus saligna*) e Urophylla (UR) (*Eucalyptus urophylla*) a faixa mais clara representa o desvio padrão e a linha escura representa a média dos espectros. (Fonte: Próprio Autor).

Com o método de PCA obtivemos as curvas dos *loadings* e o gráfico de *scores* para as primeiras componentes principais (PC), na figura 12. As diferentes partes do espectro selecionadas para uma análise mais profunda foram os intervalos de 4000 a 700, 3000 a 2800 e 2000 a 700 cm^{-1} , esses intervalos possuem os modos vibracionais que podem ser associados a presença de lignina, celulose, hemicelulose e outros extrativos vegetais [61 – 65].

No intervalo de 4000 a 700 cm^{-1} podemos ver no gráfico de *Scores*, figura 12 (a), que a uma tendência para a formação de grupos, mas há uma grande dispersão nos dados, mesmo com 79,81% da variação total dos dados. Olhando para o gráfico de *Loadings*, figura 12 (b), vemos que a PC1 e a PC2 são similares até em torno do pico em 1000 cm^{-1} . Temos também certa similaridade entre a PC1 e a PC2 que correspondem a celulose e lignina respectivamente.

Já no intervalo de 3000 a 2800 cm^{-1} no gráfico de *Scores*, figura 12 (c), ainda não temos a formação de grupos separados e pouco dispersos, entretanto vemos que as amostras mostram uma tendência a se agrupar, mas se encontram muito dispersos, sendo a amostra UR a mais dispersa. No

gráfico de *Loading*, figura 12 (d), temos que as PC's juntas representam 38,76% da variação total dos dados e podemos notar que a PC1 e a PC2 possuem picos nos mesmos intervalos com diferentes intensidades, podendo ser associadas a celulose, e a PC3 possui picos invertidos nas mesmas posições da PC1 e PC2.

No intervalo de 2000 a 700 cm^{-1} temos no gráfico de *Scores*, figura 12 (e), temos a formação de grupos individuais, mas os grupos se encontram muito próximos entre si, com a amostra GRA formando dois grupos distintos. As PC's carregam 86,46% da variação total dos dados o que pode indicar a menor dispersão nos dados quando comparados as amostras anteriores. O gráfico de *Loadings*, figura 12 (f), temos que a PC1 pode ser associada a lignina, e a PC2 a celulose, apresentam maior diferença na região de 1050 cm^{-1} com a inversão dos picos e a PC3 apresenta um pico mais pronunciado na região de 1850 cm^{-1} .

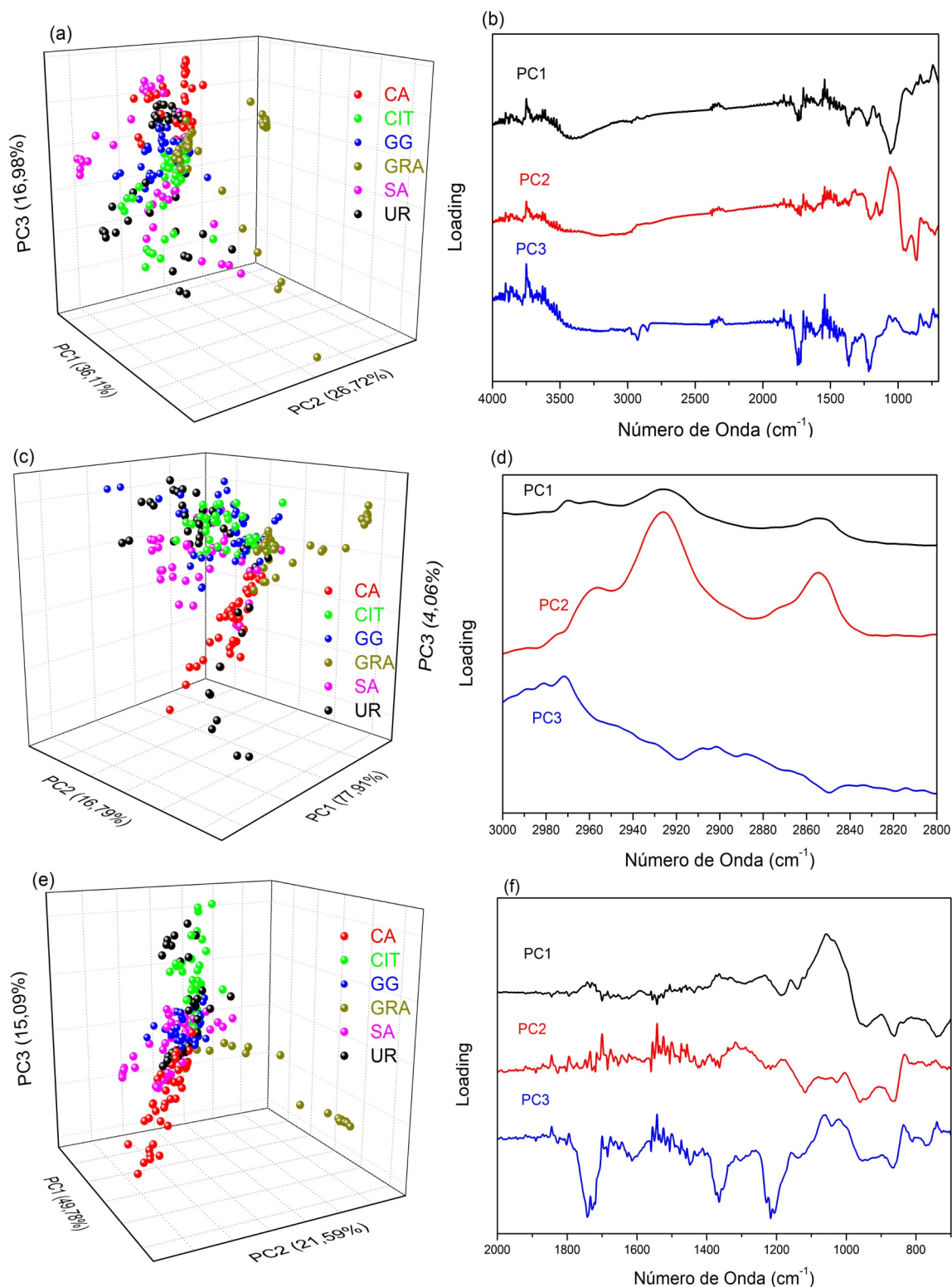


Figura 12: Gráficos de Scores a esquerda e o gráfico de Loadings na direita. (a) Gráfico de Scores para o intervalo de 4000 a 700 cm^{-1} . (b) Gráfico de Loadings para a PC1, PC2 e PC3 no intervalo de 4000 a 700 cm^{-1} . (c) Gráfico de Scores para o intervalo de 3000 a 2800 cm^{-1} . (d) Gráfico de Loadings para o intervalo de 3000 a 2800 cm^{-1} . (e) Gráfico de Scores para o intervalo de 2000 a 700 cm^{-1} . (f) Gráfico de Loadings para o intervalo de 2000 a 700 cm^{-1} . (FONTE: Próprio Autor).

Os dados de PCA foram então usados para os vários algoritmos de aprendizado de máquina. Usamos as técnicas de Análise de Discriminante Linear (LDA) e Quadrática (QDA), KNN Fino (KNNF), Médio (KNNM), Grosso (KNNG), Cosseno (KNNC), Cúbico (KNN3), Ponderado (KNNP) e Subespaço (KNNs), SVM Fino (SVMF), Médio (SVMM), Grosso (SVMG), Linear (SVML), Quadrático (SVM2) e Cúbico (SVM3).

Na figura 13 temos o desempenho dos melhores classificadores analisados em termos de sua acurácia para os intervalos de 4000 a 700, listras transversais, 3000 a 2800, listras horizontais, e 2000 a 700 cm^{-1} , listras cruzadas. Acima das barras temos o número de PC's que geram a precisão indicada para cada classificador.

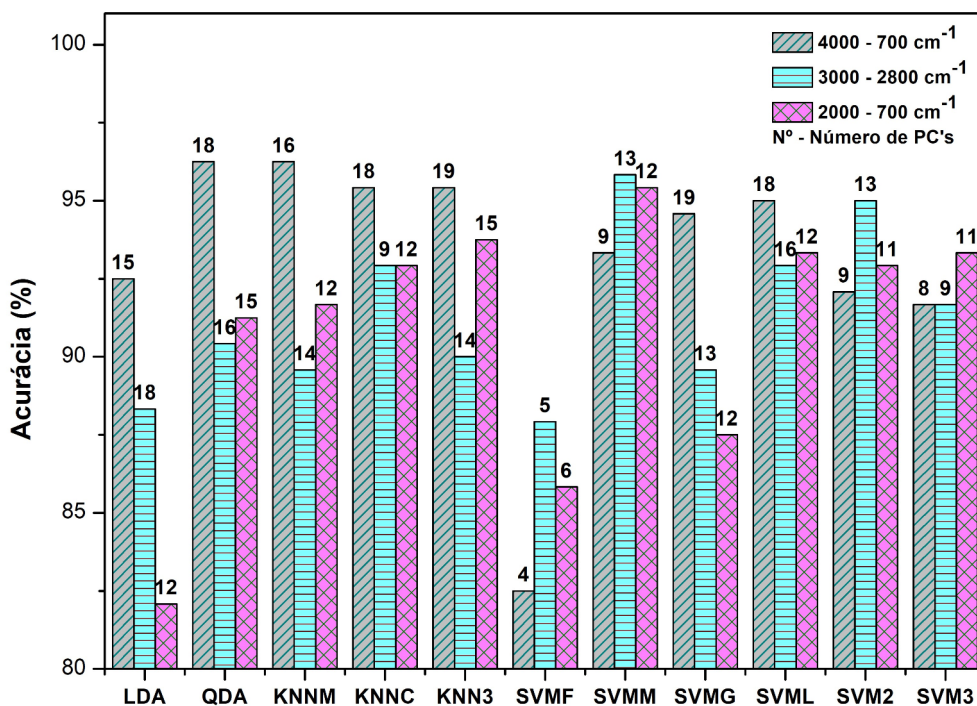


Figura 13: Gráfico com a acurácia e o número de PC's (números em cima das barras) que os classificadores Análise de Discriminante, KNN e SVM atingem a maior precisão. Em linhas transversais temos o intervalo de 4000 a 700 cm^{-1} e o desempenho dos respectivos classificadores. Em linhas horizontais temos o intervalo de 3000 a 2800 cm^{-1} e o respectivo desempenho dos classificadores. Em linhas cruzadas temos o intervalo de 2000 a 700 cm^{-1} com o desempenho dos classificadores. (FONTE: Próprio Autor).

Após o treinamento foi realizado a validação dos dados com o método de LOOCV para verificar a precisão de cada algoritmo utilizado. Temos que o intervalo de 2000 a 700 cm^{-1} apresenta o melhor resultado com o SVMM obtendo uma precisão de 95,42% com 12 PC's. Na matriz de confusão, figura 14, e como pode ser observado na figura 13 o SVMM atinge mais de 90% de

precisão com o mesmo número de PC's usados pelo LOOCV servindo de indicativo de ausência de viés.

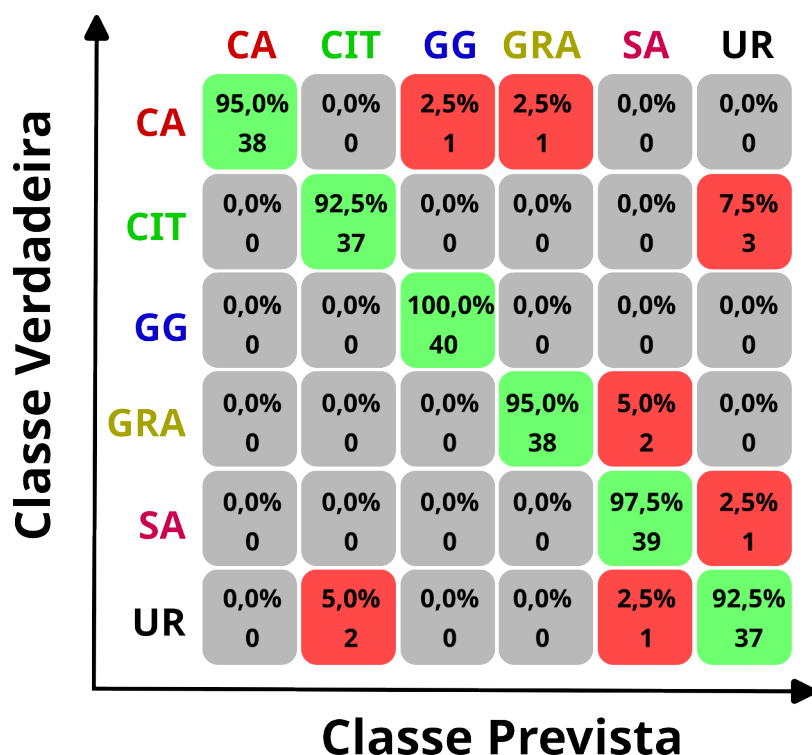


Figura 14: matriz de confusão do SVM (SVM Médio) com 95,42% de acurácia para o intervalo de 2000 a 700 cm^{-1} . (FONTE: Próprio Autor).

Nos intervalos estudados é possível encontrar classificadores com mais de 90% de precisão e até um mesmo algoritmo para vários intervalos, como o SVM que conseguiu validar 3 intervalos diferentes, sendo um deles com o menor número de PC's utilizados. Isso demonstra o potencial da técnica desenvolvida.

5.1. Casca

Na figura 15 temos o espectro de infravermelho médio para as três espécies de eucaliptos estudadas para o intervalo de 4000 a 700 cm^{-1} . No intervalo de 3800 cm^{-1} à 2700 cm^{-1} temos uma banda acentuada em 3400 cm^{-1} que está associada ao estiramento de diferentes modos vibracionais de O – H, o intervalo de 2920 e 2840 cm^{-1} possui duas bandas associadas ao estiramento simétrico e assimétrico de grupos metil e metileno que estão presentes na celulose [66].

Na região entre 1800 à 800 cm^{-1} temos um pico em 1595 cm^{-1} referente a C = C de estruturas aromáticas na lignina [67], próximos aos picos em 1460, 1425 e 1375 cm^{-1} temos a deformação de C – H que são associados a contribuições de vários modos vibracionais de carboidratos e lignina

[64]. Os picos em 1240 e 1040 cm^{-1} estão associadas ao estiramento de C – O, relacionados a Polissacarídeos e estruturas aromáticas [66]. Da mesma forma que para o alburno não é possível distinguir as amostras apenas com o espectro.

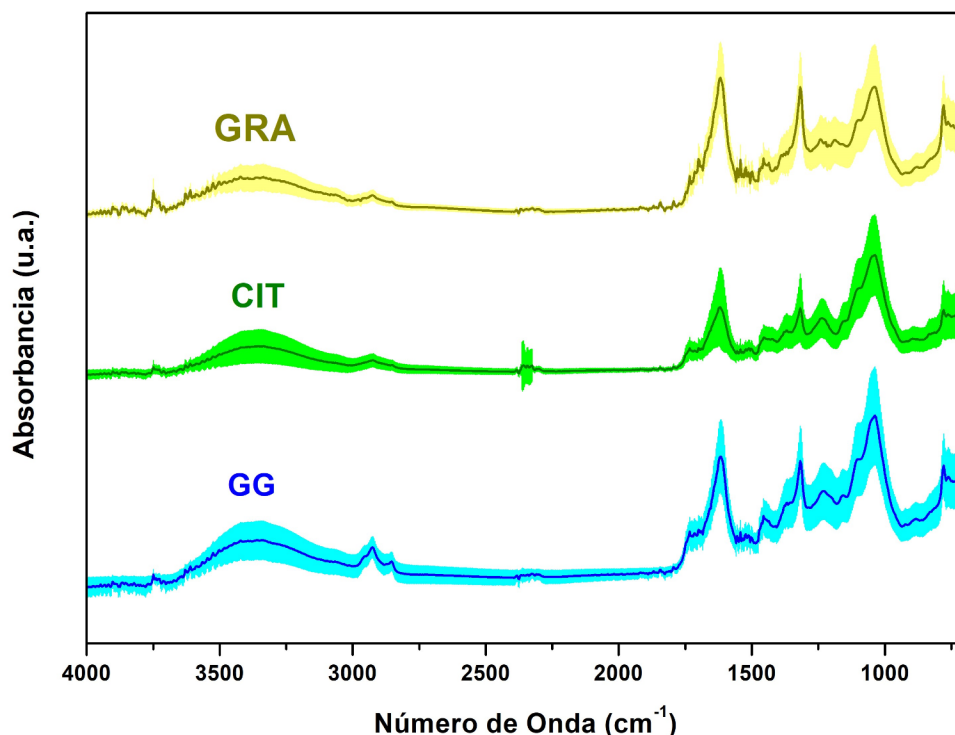


Figura 15: Espectros de FTIR de diferentes espécies de eucaliptos: *Corymbia Citriodora* (CIT), *Eucalyptus grandis* (GRA) e *E. Grandis* x *E. Urophylla* (GG). (FONTE: Próprio Autor).

A figura 16 mostra os resultados da análise PCA dos espectros obtidos através do FTIR nos intervalos de 4000 a 700, 3800 a 2800 e 1800 a 800 cm^{-1} . A CIT é representada pelas bolas verdes, a GG foi representada pelas bolas azuis e GRA pelas amarelas nos gráficos de scores ((a), (c) e (e)). Percebemos que na figura 16 (a), que corresponde a 92,74 %, do espectro inteiro (4000 – 700 cm^{-1}), a tendência dos dados a aglomerar por espécie, porém as amostras GRA e GG se encontram em um único ponto, com a amostra CIT formando 3 grupos.

Na figura 16 (c) referente a faixa de 3800 à 2800 cm^{-1} , com 96,48%, as amostras apresentam uma melhor separação uma vez que não há entrelaçamento entre os grupos. A figura 16 (d) apresenta diferenças em relação a lignina, PC1, e celulose, PC2.

A figura 16 (e) na faixa de 1800 à 800 cm^{-1} , que corresponde a 95,63%, temos uma tendência a separação em três grupos, mas também há uma mistura entre as amostras. Na figura 16 (f) indica que as espécies de eucaliptos podem apresentar diferenças relacionadas a lignina [66,67].

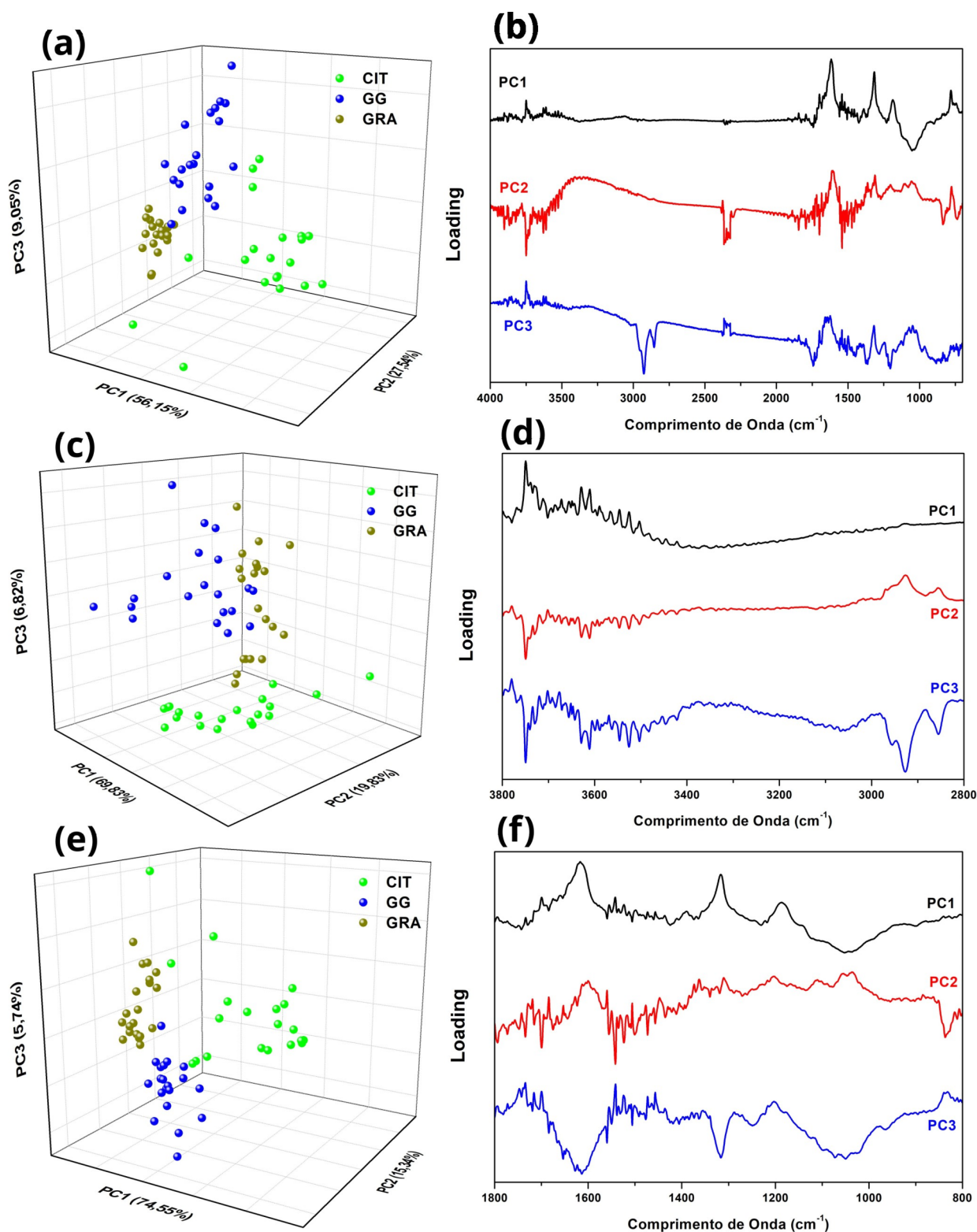


Figura 16: Gráficos de Scores a esquerda e o gráfico de Loadings na direita para amostras da casca. (a) Gráfico de Scores para o intervalo de 4000 a 700 cm^{-1} . (b) Gráfico de Loadings para a PC1, PC2 e PC3 no intervalo de 4000 a 700 cm^{-1} . (c) Gráfico de Scores para o intervalo de 3800 a 2800 cm^{-1} . (d) Gráfico de Loadings para o intervalo de 3800 a 2800 cm^{-1} . (e) Gráfico de Scores para o intervalo de 1800 a 800 cm^{-1} . (f) Gráfico de Loadings para o intervalo de 1800 a 800 cm^{-1} . (FONTE: Próprio Autor).

A acurácia dos métodos de aprendizagem de máquina para os intervalos de 4000 – 700, 3800 – 2800 e 1800 – 800 cm^{-1} , figura 17, apresentou 100% de precisão em vários algoritmos. O LDA, QDA, KNNF, KNNM, KNNC e SVML obtiveram a mesma acurácia de 100%, para diferentes intervalos, demonstrando o potencial de separação usando a casca.

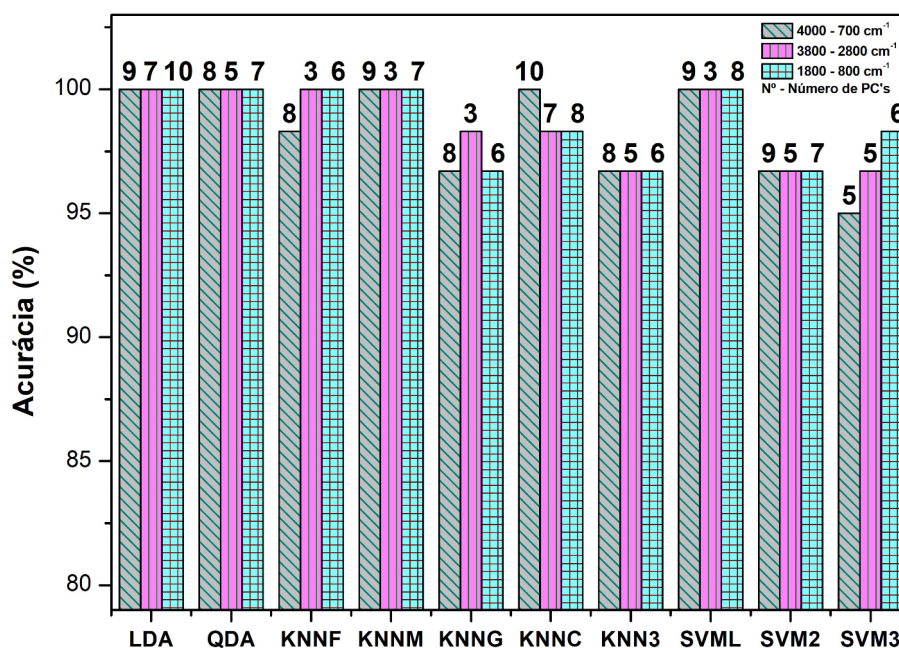


Figura 17: Gráfico com a acurácia e o número de PC's (números em cima das barras) que os classificadores Análise de Discriminante, KNN e SVM atingem a maior precisão nos intervalos estudados.(FONTE: Próprio Autor).

A matriz de confusão, figura 18, do melhor método de aprendizagem de máquina obtido no comprimento de onda de 3800 – 2800 cm^{-1} a partir do LOOCV. O modelo apresentou um total de 100% de acurácia para as três espécies analisadas. A Figura 16 mostra que no comprimento de 3800 – 2800 cm^{-1} as bandas das espécies de eucalipto apresentam intensidades diferentes em regiões associadas a lignina e a celulose, essas diferenças podem ser relacionadas a resistência mecânica da casca das espécies. A lignina é uma molécula responsável pela rigidez e proteção mecânica da planta, com isso a discriminação de espécies de eucalipto através da aprendizagem de máquina pode estar relacionada principalmente a moléculas de lignina.

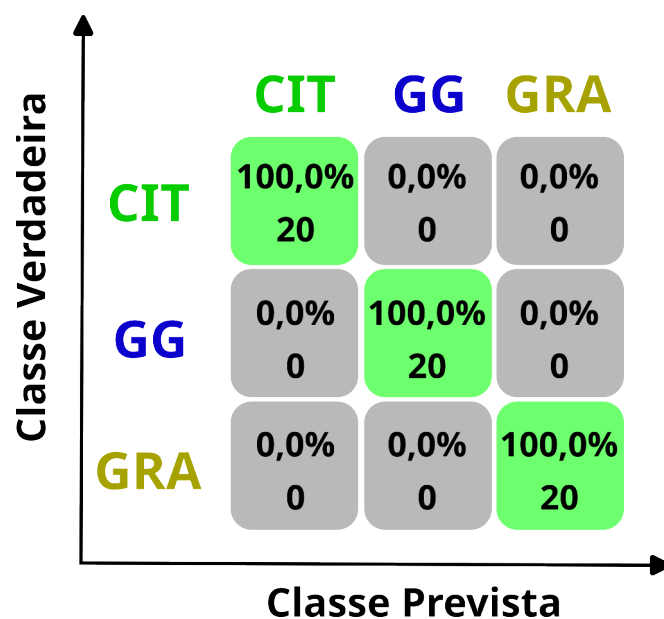


Figura 18: Matriz de confusão para o LDA (*Linear Discriminant Analysis*) no intervalo de 3800 a 2800 cm^{-1} . (Fonte: Próprio Autor).

5.2. Comparação

Quando comparamos o espectro de infravermelho do alburno e da casca é possível ver diferenças, figura 19, em especial no intervalo de 2500 – 700 cm^{-1} , com a amostra GRA apresentando a maior diferença nesse intervalo, é preciso considerar que a amostra GRA possui a maior variação entre as amostras de alburno. As amostras CIT e GG apresentam maiores diferenças no intervalo de 2500 – 1500 cm^{-1} .

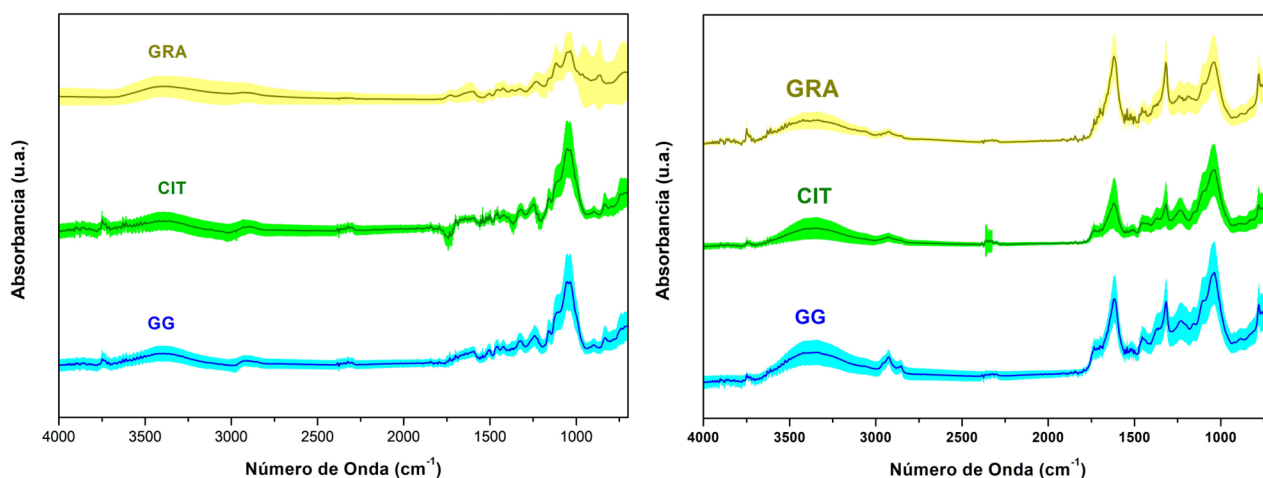


Figura 19: Comparação entre o espectro de infravermelho do alburno (esquerda) e casca (direita). (Fonte: Próprio Autor).

Os *Scores* e *Loadings* apresentam nítidas diferenças, figura 20, com a casca apresentando a melhor separação. Nos gráficos de *Score*, figura 20 (a, b, e, f, i, j), em todos os intervalos analisados a casca apresenta melhor separação, enquanto alburno tem sua melhor separação no intervalo de $2000 - 700 \text{ cm}^{-1}$. Os gráficos de *Loadings*, figura 20 (c, d, g, h, k, l), também apresentam diferenças sendo as figuras 20 g e h os mais similares e as figuras k e l as que apresentam a maior diferença.

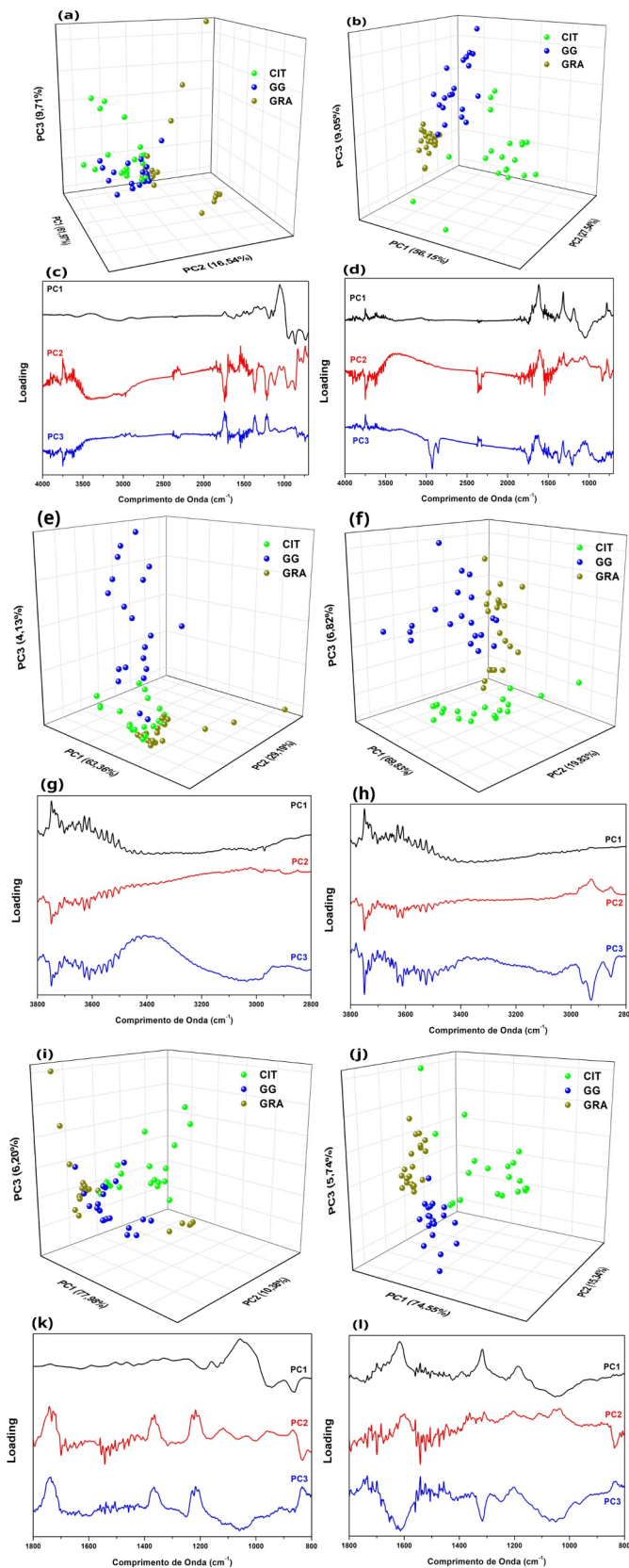


Figura 20: Comparação entre os Scores e os Loadings do Alburno (esquerda) e da Casca (direita) de 3 diferentes espécies de eucalipto. (a, b) Gráfico de Scores para o intervalo de 4000 a 700 cm^{-1} . (c, d) Gráfico de Loadings para a PC1, PC2 e PC3 no intervalo de 4000 a 700 cm^{-1} . (e, f) Gráfico de Scores para o intervalo de 3800 a 2800 cm^{-1} . (g, h) Gráfico de Loadings para o intervalo de 3800 a 2800 cm^{-1} . (i, j) Gráfico de Scores para o intervalo de 1800 a 800 cm^{-1} . (k, l) Gráfico de Loadings para o intervalo de 1800 a 800 cm^{-1} . (FONTE: Próprio Autor).

Comparando a acurácia para o intervalo de 4000 a 700 cm^{-1} , onde o alburno atinge seu melhor desempenho, figura 21, temos que a casca atinge uma precisão de 100% em todos os classificadores, exceto o KNNF, e com uma menor quantidade de PC's quando comparado com o alburno. O melhor desempenho fica com o QDA, para a casca, e KNNF, para o alburno, com 100% e 8 PC's e 96,7% e 9 PC's respectivamente.

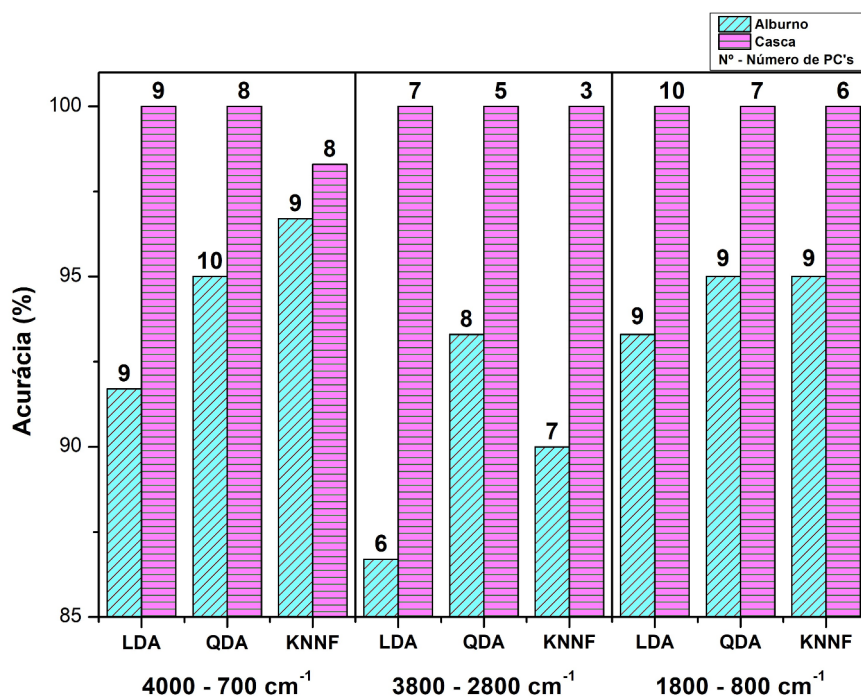


Figura 21: Comparação da acurácia entre o Alburno, listras transversais, e a Casca, linhas horizontais, com o número de PC's acima de cada barra para os três intervalos analisados. (FONTE: Próprio Autor).

Por fim comparando a matriz de confusão para os melhores classificadores para a casca e o alburno, figura 22, vemos que a casca não apresenta erros de classificação, para um total de 20 amostras. Já para o alburno apresenta erros para GG, 2 amostras são classificadas erradas como CIT e GRA respectivamente.

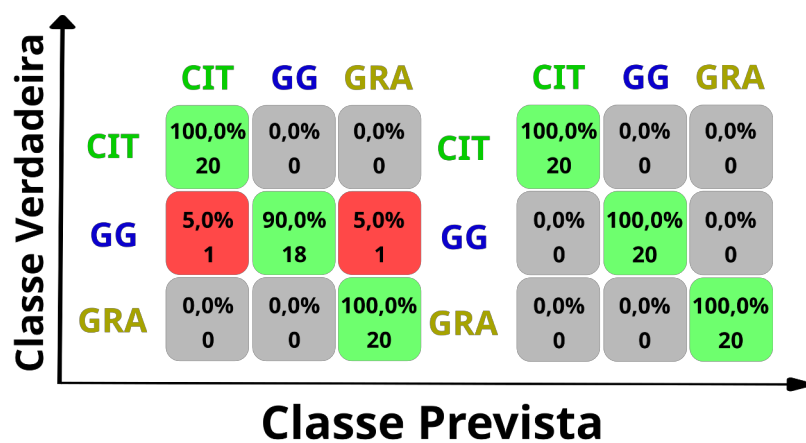


Figura 22: Comparação da matriz de confusão do Alburno (esquerda) e da Casca (direita) para as amostras CIT, GG, e GRA para os classificadores KNNF (esquerda) e QDA (direita) para o intervalo de 4000 – 700 cm^{-1} . (FONTE: Próprio Autor).

6. Conclusão

Podemos concluir que o uso do FTIR combinado com as técnicas de ML podem ser usadas para diferenciar o alburno em pó de 6 espécies de eucalipto com precisão de mais de 90% em três diferentes intervalos, com o SVM tendo uma performance significativa em todos os intervalos estudados, atingindo o melhor resultado no intervalo de $2000 - 700 \text{ cm}^{-1}$ atribuindo a diferença na lignina. A casca o que apresentou os melhores resultados atingindo 100% de acurácia para vários classificadores em todos os intervalos, que igual ao alburno pode ser atribuída a lignina. Na comparação entre a casca e o alburno, a casca obtém melhores resultados e o alburno tem seu melhor resultado no intervalo de $4000 - 700 \text{ cm}^{-1}$, que igual à casca e ao alburno pode ser atribuído a diferenças na lignina.

7. Referencias

- [1] EUCALIPTO. Portal Embrapa, 2019. Disponível em: <https://www.embrapa.br/florestas/transferencia-de-tecnologia/eucalipto>. Acesso em: 19/04/2022.
- [2] Viera, M.; Ruíz Fernández, F.; Rodríguez-Soalleiro, R. “Nutritional Prescriptions for Eucalyptus Plantations: Lessons Learned from Spain”. *Forests* 2016, 7, 84. <https://doi.org/10.3390/f7040084>.
- [3] Color, B. Munsell. “Munsell color charts for plant tissues”. Munsell Color Company, 1963.
- [4] Schreiner, L., Bauer, P. & Buettner, A. “Resolving the smell of wood - identification of odour-active compounds in Scots pine (*Pinus sylvestris* L.)”. *Sci Rep* 8, 8294 (2018). <https://doi.org/10.1038/s41598-018-26626-8>.
- [5] Wei, L. Jinglu, T. “Grain Pattern Characterization and Classification of Walnut by Image Processing”. *Wood and Fiber Science* 3 (2004).
- [6] Tou, J. Y., Tay, Y. H., & Lau, P. Y. “A Comparative Study for Texture Classification Techniques on Wood Species Recognition Problem”. 2009 Fifth International Conference on Natural Computation. 2009. <https://doi.org/10.1109/icnc.2009.594>.
- [7] de Norma Técnicas – COPANT, Comisión Panamericana. “COPANT 30: 1-004: Maderas: método de determinación del peso específico aparente”. Caracas: COPANT (1971).
- [8] Chen, H.; Ferrari, C.; Angiuli, M.; Yao, J.; Raspi, C.; Bramanti, E. “Qualitative and quantitative Analysis of wood samples by Fourier transform infrared spectroscopy and multivariate Analysis”. *Carbohydrate Polymers*. Vol. 82, p. 772 - 778, 2010. <https://doi.org/10.1016/j.carbpol.2010.05.052>.
- [9] VAN SOEST, P. J. “Use of detergents in the Analysis of fibrous feeds. II. A rapid method for the determination composition of fiber and lignin”. *Journal of the Association of Official Analytical Chemists*. Vol. 46, p. 829 - 835, 1990. <https://doi.org/10.1093/jaoac/73.4.491>.
- [10] Larios, Gustavo S., et al. “Laser-induced breakdown spectroscopy as a powerful tool for distinguishing high-and low-vigor soybean seed lots”. *Food Analytical Methods*, 13.9 (2020): 1691 - 1698.
- [11] Oliveira, I. C., Franca, T., Nicolodelli, G., Morais, C. P., Marangoni, B., Bacchetta, G., ... Cena, C. “Fast and Accurate Discrimination of *Brachiaria brizantha* (A.Rich.) Stapf Seeds by Molecular Spectroscopy and Machine Learning”. *ACS Agricultural Science & Technology*. 2021. <https://doi.org/10.1021/acsagscitech.1c00067>.
- [12] González-Solís, J. L., Villafan-Bernal, J. R., Martínez-Zérega, B. E., & Sánchez-Enríquez, S. “Type 2 diabetes detection based on serum sample Raman spectroscopy”. *Lasers in Medical Science* (2018). <https://doi.org/10.1007/s10103-018-2543-4>.
- [13] Larios, Gustavo, et al. “A new strategy for canine visceral leishmaniasis diagnosis based on FTIR spectroscopy and machine learning”. *Journal of Biophotonics* 14.11 (2021): e202100141. <https://doi.org/10.1002/jbio.202100141>.

- [14] Sharma, A., Verma, R., Kumar, R., Chauhan, R., & Sharma, V. "Chemometric analysis of ATR-FTIR spectra of fingernail clippings for classification and prediction of sex in forensic context". *Microchemical Journal*, 105504, 2020. <https://doi.org/10.1016/j.microc.2020.105504>.
- [15] Carneiro, C. R., Silva, C. S., de Carvalho, M. A., Pimentel, M. F., Talhavini, M., & Weber, I. T. "Identification of Luminescent Markers for Gunshot Residues: Fluorescence, Raman Spectroscopy, and Chemometrics". *Analytical Chemistry*, 91(19), 12444–12452, 2019. <https://doi.org/10.1021/acs.analchem.9b03079>.
- [16] Sharma, Vishal, et al. "On the rapid and non-destructive approach for wood identification using ATR-FTIR spectroscopy and chemometric methods". *Vibrational Spectroscopy*, 110 (2020): 103097.
- [17] Traoré, M., Kaal, J., & Martínez Cortizas, A. "Differentiation between pine woods according to species and growing location using FTIR-ATR". *Wood Science and Technology*, 52(2), 487–504, 2017. <https://doi.org/10.1007/s00226-017-0967-9>.
- [18] Hobro, A. J., Kuligowski, J., Döll, M., & Lendl, B. "Differentiation of walnut wood species and steam treatment using ATR-FTIR and partial least squares discriminant analysis (PLS-DA)". *Analytical and Bioanalytical Chemistry*, 398(6), 2010, 2713 – 2722. <https://doi.org/10.1007/s00216-010-4199-1>.
- [19] Dos Santos, J. X., Vieira, H. C., Souza, D. V., de Menezes, M. C., de Muñiz, G. I. B., Soffiatti, P., & Nisgoski, S. "Discrimination of "Louros" wood from the Brazilian Amazon by near-infrared spectroscopy and machine learning techniques". *European Journal of Wood and Wood Products*, 79(4), 989–998, 2021. <https://doi.org/10.1007/s00107-021-01685-3>.
- [20] Cui, X., Wang, Q., Zhao, Y., Qiao, X., & Teng, G. "Laser-induced breakdown spectroscopy (LIBS) for classification of wood species integrated with artificial neural network (ANN)". *Applied Physics B*, 125(4), 2019. <https://doi.org/10.1007/s00340-019-7166-3>.
- [21] Rowell, Roger M. "Handbook of wood chemistry and wood composites". CRC press, 2005.
- [22] Ek, Monica, Göran Gellerstedt, and Gunnar Henriksson, eds. "Wood chemistry and biotechnology". Vol. 1. Walter de Gruyter, 2009.
- [23] Elaissi, Ameer, et al. "Antibacterial activity and chemical composition of 20 Eucalyptus species' essential oils". *Food Chemistry*, 129.4 (2011): 1427 - 1434.
- [24] Asao, S., Hayes, L., Aspinwall, M.J., Rymer, P.D., Blackman, C., Bryant, C.J., Atkin, O.K. "Leaf trait variation is similar among genotypes of Eucalyptus camaldulensis from differing climates and arises in plastic responses to the seasons rather than water availability". *New Phytol.* 2020. <https://doi.org/10.1111/nph.16579>.
- [25] Souza, B.M., Freitas, M.L.M., Sebbenn, A.M., Gezan, S.A., Zanatto, B., Zulian, D.F., de Aguiar, A.V. "Genotype-by-environment interaction in *Corymbia citriodora* (Hook.) KD Hill, & LAS Johnson progeny test in Luiz Antonio, Brazil". *For. Ecol. Manage.* 2020, 460, 117855 <https://doi.org/10.1016/j.foreco.2019.117855>.
- [26] Chaín, J.M., Tubert, E., Graciano, C., Castagno, L.N., Recchi, M., Pieckenstain, F.L., Baroli, I. "Growth promotion and protection from drought in Eucalyptus grandis seedlings inoculated with

- beneficial bacteria embedded in a superabsorbent polymer”. *Sci. Rep.* 10 (1), 2020, 1–17. <https://doi.org/10.1038/s41598-020-75212-4>.
- [27] Biazzon, J.C., Lima Jr., M.P., Munis, R.A., De Araújo, V.A., Morales, E.A.M., Gonçalves, M.T.T., Valarelli, I.D.D. “Shear strength of eucalyptus saligna wood joints bonded with polyvinyl acetate adhesive”. *BioResources* 14 (2), 2019, 4590–4602. <https://doi.org/10.15376/biores.14.2.4590-4602>.
- [28] de Oliveira, B. R., et al. “Eucalyptus growth recognition using machine learning methods and spectral variables”. *Forest Ecology and Management* 497 (2021): 119496.
- [29] Costa, Julia de Almeida. “Qualidade da madeira de Eucalyptus urograndis, plantado no Distrito Federal, para produção de celulose kraft”. (2011).
- [30] Lawes, Michael J, et al. “Bark Functional Ecology and Its Influence on the Distribution of Australian Half-Butt Eucalypts.” *Austral Ecology*, vol. 46, no. 7, 2021, pp. 1097–1111. <https://doi/10.1111/aec.13045>.
- [31] Vane, C. H., Drage, T. C., & Snape, C. E. “Bark decay by the white-rot fungus *Lentinula edodes*: Polysaccharide loss, lignin resistance and the unmasking of suberin”. *International Biodeterioration & Biodegradation*, 57(1), 14–23, 2006. <https://doi/10.1016/j.ibiod.2005.10.004>.
- [32] Cutler, David F., Ted Botha, and Dennis Wm Stevenson. “Anatomia vegetal: uma abordagem aplicada”. Artmed Editora, 2009.
- [33] Costa, Ana Carolina Silva, et al. “Propriedades da madeira de cerne e alburno de Eucalyptus camaldulensis”. *Revista Ciência da Madeira (Brazilian Journal of Wood Science)* 8.1 (2017).
- [34] Fonte, Ana Paula Namikata, et al. “Propriedades físicas e químicas da madeira de cerne e alburno de *Cryptomeria japonica*”. *Revista de Ciências Agroveterinárias*, 16.3 (2017): 277-285.
- [35] Bidhendi, A. J., & Geitmann, A. “Relating the mechanics of the primary plant cell wall to morphogenesis”. *Journal of Experimental Botany*, 67(2), 2015, 449–461. <https://doi/10.1093/jxb/erv535>.
- [36] Klemm, D., Heublein, B., Fink, H.-P., & Bohn, A. “Cellulose: Fascinating Biopolymer and Sustainable Raw Material”. *Angewandte Chemie International Edition*, 44(22), 3358–3393, (2005). <https://doi/10.1002/anie.200460587>.
- [37] DESHAVATH, N. N.; VEERANKI, V. D.; GOUD, V. V. “Lignocellulosic feedstocks for the production of bioethanol availability, structure, and composition”. *Sustainable Bioenergy*. p. 1-19, 2019.
- [38] GÍRIO, F.M.; FONSECA, C.; CARVALHEIRO, F.; DUARTE, L. C.; MARQUES, S.; BOGEL-LUKASIK, R. “Hemicelluloses for fuel ethanol: a review”. *Bioresour Technol.* Vol. 101, p. 4775-4800, 2010.
- [39] WATIKINS, D.; NURUDDIN, MD.; MAHESH, H.; TCHERBI-NARTEH, A.; JEELANI, S. “Extraction and characterization of lignina from diferent biomass resources”. *Journal of Materials Research and Tecnology*. Vol. 4, p.26-32, 2015.

- [40] Khalil, H. P. S. A., Alwani, M. S., & Omar, A. K. M. “Chemical composition, anatomy, lignin distribution, and cell wall structure of Malaysian plant waste fibers”. *BioResources*, 1(2), 220–232, 2006. <https://doi/10.15376/biores.1.2.220-232>.
- [41] MIRABELLA, Francis M. (Ed.). “Modern techniques in applied molecular spectroscopy”. John Wiley & Sons, 1998.
- [42] DEMTRÖDER, Wolfgang. “Laser Spectroscopy: Basic Concepts and Instrumentation”. 3. ed. Springer, 2003.
- [43] BANWELL, Colin N. “Fundamentals of Molecular Spectroscopy”. 3. ed. UK: McGraw-Hill Book Company, 1983.
- [44] CANEVAROLO JR, Sebastião V. et al. “Técnicas de caracterização de polímeros”. *Artliber*, São Paulo, v. 430, n. 2004, 2004.
- [45] BUIJS, H. “Infrared Spectroscopy”. *Springer Handbooks*, p. 607–613, 2006. ISSN 25228706.
- [46] RENCHER, A.C. “Methods of Multivariate Analysis”. 2. ed., Wiley-Interscience, 2002. 738 p. ISBN 978-0-471-46172-2.
- [47] GELADI, Paul. “Chemometrics in spectroscopy”. Part 1. Classical chemometrics. *Spectrochimica Acta Part B: Atomic Spectroscopy*, v. 58, n. 5, p. 767-782, 2003.
- [48] PIERCE, Karisa M.; MOHLER, Rachel E. “A review of chemometrics applied to comprehensive two-dimensional separations from 2008–2010”. *Separation & Purification Reviews*, v. 41, n. 2, p. 143-168, 2012. <https://doi/10.1080/15422119.2011.591868>.
- [49] Supervised vs. Unsupervised Learning. IBM, 2021 Disponível em: <https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning>
- [50] BRERETON, Richard G. et al. “Chemometrics in analytical chemistry—part II: modeling, validation, and applications”. *Analytical and bioanalytical chemistry*, v. 410, n. 26, p. 6691- 6704, 2018. <https://doi/10.1007/s00216-018-1283-4>.
- [51] JOLLIFFE, I.T.; CADIMA, J. “Principal Component Analysis: a review and recent developments”. *Phil. Trans. R.Soc. A* 374: 20150202, 2016. <https://doi/10.1098/rsta.2015.0202>.
- [52] HÄRDLE, W. K.; SIMAR, L. “Applied Multivariate Statistical Analysis”. 4. ed., Springer, 2015.
- [53] Abdullah, D. M., & Abdulazeez, A. M. “Machine Learning Applications based on SVM Classification A Review”. *Qubahan Academic Journal*, 1(2), 81-90, 2021.
- [54] Chauhan, V. K., Dahiya, K., & Sharma, A. “Problem formulations and solvers in linear SVM: a review”. *Artificial Intelligence Review*, 2018. <https://doi/10.1007/s10462-018-9614-6>.
- [55] Abu Alfeilat, Haneen Arafat, et al. “Effects of distance measure choice on k-nearest neighbor classifier performance: a review”. *Big data* 7.4 (2019): 221-248.
- [56] Dhanabal, S., and S. J. I. J. C. A. Chandramathi. “A review of various k-nearest neighbor query processing techniques”. *International Journal of Computer Applications* 31.7 (2011): 14-22.

- [57] KUBAT, M. “An Introduction to Machine Learning”. 2. ed., Springer, 2017. 348 p.
- [58] Syed, Ali R. “A review of cross validation and adaptive model selection”. 2011.
- [59] Saccenti, Edoardo, and José Camacho. “Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods”. *Chemometrics and Intelligent Laboratory Systems*, 149 (2015): 99-116.
- [60] ENGEL, JASPER ET AL. “Breaking with trends in pre-processing”. *TrAC Trends in Analytical Chemistry*. V.50, p.96-106, 2013.
- [61] RUDAKIYA, D.M.; GUPTE, A. “Assessment of White rot fungus mediated hardwood degradation by FTIR spectroscopy and multivariate Analysis”. *Journal of Microbiological Methods*. Vol. 157, p. 123-130, 2019.
- [62] MÜLLER, G., SCHÖPPER, C., VOS, H., KHARAZIPOUR, A., & POLLE, A. “FTIR-ATR spectroscopic analysis of changes in wood properties during particle- and fibreboard production of hard- and softwood trees”. *Bioresources*. Vol. 4, 49–71, 2009.
- [63] Poletto, M., Zattera, A. J., & Santana, R. M. C. “Structural differences between wood species: Evidence from chemical composition, FTIR spectroscopy, and thermogravimetric analysis”. *Journal of Applied Polymer Science*, 126(S1), E337–E344, 2012. <https://doi/10.1002/app.36991>.
- [64] Pandey, K. K., & Pitman, A. “FTIR studies of the changes in wood chemistry following decay by brown-rot and white-rot fungi”. *International Biodeterioration & Biodegradation*, 52(3), 151–160, 2003. [https://doi/10.1016/s0964-8305\(03\)00052-0](https://doi/10.1016/s0964-8305(03)00052-0).
- [65] Popescu, Carmen-Mihaela, et al. “Vibrational spectroscopy and X-ray diffraction methods to establish the differences between hardwood and softwood”. *Carbohydrate Polymers* 77.4 (2009): 851-857. <https://doi/10.1016/j.carbpol.2009.03.011>.
- [66] Yang, Z., Peng, H., Wang, W., & Liu, T. “Crystallization behavior of poly(ε-caprolactone)/layered double hydroxide nanocomposites”. *Journal of Applied Polymer Science*, NA–NA, 2010. <https://doi/10.1002/app.31787>.
- [67] Rodrigues, J., Faix, O., & Pereira, H. “Determination of Lignin Content of Eucalyptus globulus Wood Using FTIR Spectroscopy”. *Holzforschung*, 52(1), 46–50, 1998. <https://doi/10.1515/hfsg.1998.52.1.46>.