FUNDAÇÃO UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DOS MATERIAIS

MATHEUS CICERO DA SILVA RIBEIRO

DIFERENCIAÇÃO ENTRE GRÃOS DE MILHO TRANSGÊNICOS E CONVENCIONAIS UTILIZANDO ESPECTROSCOPIA ÓPTICA

CAMPO GRANDE, MS

2019

MATHEUS CICERO DA SILVA RIBEIRO

DIFERENCIAÇÃO ENTRE GRÃOS DE MILHO TRANSGÊNICOS E CONVENCIONAIS UTILIZANDO ESPECTROSCOPIA ÓPTICA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência dos Materiais da Universidade Federal de Mato Grosso do Sul, como requisito á obtenção do título de Mestre em Ciência dos Materiais.

Orientador: Prof. Dr. Bruno Spolon Marangoni

Coorientador: Prof. Dr. Gustavo Nicolodelli

CAMPO GRANDE, MS

2019

DEDICATÓRIA

A Deus, que em sua infinita sabedoria guia meus caminhos me proporcionando saúde, serenidade e disposição para enfrentar todas as etapas desta árdua caminhada.

A minha mãe, Célia, meu pai, Sirval, meus irmãos Jarbas, Sirval Filho e Rian, minha tia Vilma Aparecida, meu tio Jair Camilo e toda a minha família pelo apoio, incentivo e carinho, que sempre foi essencial para minha caminhada e que não mediram esforços para me ajudar a chegar até aqui.

A Keila por todo o apoio psicológico e emocional que me proporcionou, me incentivando, me dando forças e permanecendo ao meu lado para que fosse possível a realização desse trabalho.

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Bruno Spolon Marangoni, por ter me aceitado a fazer parte do desenvolvimento dessa pesquisa e pela confiança depositada em mim. Agradeço pelos conhecimentos adquiridos, pela atenção e os incentivos fornecidos, pelas discussões e direcionamento á respeito de qual o próximo passo a ser feito, por sempre estar disposto a ensinar e também a aprender, e pela grande amizade adquirida.

Ao meu coorientador, Prof. Dr. Gustavo Nicolodelli, pelos conhecimentos á respeito da técnica LIBS ensinados, pelo auxílio no desenvolvimento da pesquisa e pelos conselhos dados á respeito do melhoramento da pesquisa.

Ao Prof. Dr. Jader de Souza Cabral do Instituto de Física da Universidade Federal de Uberlândia pelo fornecimento das amostras que foram primordiais para o desenvolvimento dessa pesquisa.

Ao grupo de pesquisa do Laboratório de Óptica e Fotônica do Instituto de Física da Universidade Federal de Mato Grosso do Sul, e também pela Embrapa Instrumentação, pela realização e obtenção das medidas espectroscópicas ópticas das amostras que foram essenciais para o desenvolvimento da pesquisa.

Ao Instituto de Física da Universidade Federal de Mato Grosso do Sul, e a todo o seu corpo docente, pela união e troca de conhecimentos, pelas orientações e conselhos, pela disponibilização dos laboratórios e de toda a infraestrutura fornecida.

Ao CNPq pelo apoio e incentivo (Processo 461743/2014-0).

"As vezes são as pessoas que ninguém consegue imaginar nada delas, que são aquelas que fazem as coisas que ninguém consegue imaginar."

Alan Turing

RESUMO

O milho tem um papel de grande importância no aumento da produtividade agrícola vivenciadas nas últimas décadas. Esse aumento se deve, principalmente, ao uso de grãos geneticamente modificados, ou também conhecidos como grãos transgênicos. Organismos Geneticamente Modificados (OGM's) são organismos vivos, sejam plantas, animais ou microorganismos, na qual sofreram alterações nas sequências de seus genes de *DNA*, com a finalidade do melhoramento de suas características biológicas, visando o aumento de sua qualidade e também da sua produtividade, tornando a planta mais resistente à herbicidas e a incidência de pragas. O cumprimento da legislação que regulamenta a comercialização de alimentos e ingredientes contendo OGM é totalmente dependente da sensibilidade e confiabilidade dos métodos de detecção e quantificação. Esse processo de análise demanda tempo e alto custo, pois exigem dados confidenciais á respeito da sequência da modificação genética e necessidade de material de referência certificado, cuja disponibilidade é limitado devido à confidencialidade por parte da empresa fornecedora das sementes. Neste trabalho, a combinação de espectroscopia óptica com os métodos computacionais de aprendizado supervisionado, levou a um método de diferenciação entre amostras transgênicas e não transgênicas, em grãos de milho, com o intuito de diminuir o tempo do processo de medida e reduzir os custos por análise. As técnicas utilizadas foram a Emissão Óptica por Plasma Induzido por Laser (LIBS) e Infravermelho por Transformada de Fourier (FT-IR). Os métodos computacionais de aprendizado supervisionado de reconhecimento de padrões foram utilizados, com os classificadores k-NN (*k-nearest neighbour*) e SVM (*Support Vector Machine*). Os resultados obtidos apresentaram que a precisão de classificação, junto com a técnica FT-IR foi de 64,7%, e para a técnica LIBS foi de 83,7%. Isso mostra que as técnicas de espectroscopia ópticas juntos com os métodos computacionais de aprendizado supervisionado possuem potencial na diferenciação entre as amostra, abrindo caminho para o desenvolvimento de equipamentos espectroscópicos portáteis para aplicação comercial dessa atividade em específico.

Palavras-Chave: ESPECTROSCOPIA, MILHO, LIBS e FT-IR.

ABSTRACT

The Corn worked like a major role in increasing agricultural productivity in recent decades. The increase in maize productivity is mainly due to the use of genetically modified grains, or also known as transgenic grains. Genetically Modified Organisms (GMOs) are living organisms, be they plants, animals or microorganisms, in which they have undergone alterations in the sequences of their DNA genes, with the purpose of improving their biological characteristics, aiming to increase their quality and also their productivity , making the plant more resistant to herbicides and the incidence of pests. Compliance with legislation regulating the marketing of GMO-containing foods and ingredients is totally dependent on the sensitivity and reliability of detection and quantification methods. This process requires time and cost because it requires confidential data regarding the sequence of the genetic modification and the need for certified reference material, the availability of which is limited due to the confidentiality of the seed supplier. In this work, the combination of optical spectroscopy and computational methods of supervised learning, led to a method of differentiation between transgenic and non-transgenic samples, in corn kernels, in order to reduce the time of the measurement process and reduce costs by analyze. The techniques used were Laser Induced Breakdown Spectroscopy (LIBS) and Fourier Transform Infrared (FT-IR). The computational methods of supervised learning of pattern recognition were used with k-NN (k-nearest neighbor) and SVM (Support Vector Machine) classifiers. The results obtained showed that the classification accuracy, together with the FT-IR technique was 64.7%, and for the LIBS technique it was 83.7%. This shows that the optical spectroscopy techniques together with the supervised learning computational methods have potential in the differentiation between the samples, paving the way for the development of portable spectroscopic equipment for commercial application of this specific activity.

Keywords: SPECTROSCOPY, CORN, LIBS and FT-IR.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O grande aumento que a produtividade agrícola vem sofrendo nos últimos anos se deve, principalmente, ao uso de grãos geneticamente modificados (GM), ou grãos transgênicos. Em 1996, a área plantada com grãos de origem transgênica era de 1,7 milhões de hectares, enquanto em 2017 foi de 189,8 milhões de hectares. (ISAAA, 2017).

Organismos Geneticamente Modificados (OGMs) são organismos vivos, sejam plantas, animais ou microorganismos, na qual sofreram alterações nas sequências de seus genes de *DNA*, com a finalidade do melhoramento de suas características biológicas. Nas plantas, são utilizadas técnicas do DNA recombinante que consiste na inserção, no genoma da planta, um ou mais genes de forma a garantir a expressão das características de interesse. As principais razões que envolvem o desenvolvimento de OGMs são a melhoria da qualidade e o aumento da produtividade das plantas, tornando-as mais nutritivas e mais tolerantes á herbicidas e resistentes á pragas e doenças. (DINON, 2007).

A espécie de milho transgênica que é mais cultivada no Brasil é o milho *Bt*. *Bacillus thuringiensis* (*Bt*) é um microorganismo encontrado no solo de várias regiões do Brasil e que tem servido de inseticida biológico desde a década de 60, por meio de pulverização dos esporos sobre a lavoura. Diferentes genes *Bt* têm sido isolados e incorporados ao milho, dentre eles Cry1AB, Cry1F e Cry1Ac, que produzem proteínas capazes de controlar a população de lagartas, como a mais destrutiva praga do milho, a lagarta-do-cartucho. (ABIMILHO, 2017).

Devido ao grande aumento da produtividade utilizando OGMs e seu respectivo consumo pela população, tanto de maneira direta (consumo de derivados de milho), como indireta (consumo de carnes provenientes de animais alimentados com rações ou milho de origem transgênica), há a regulação e fiscalização da origem dos alimentos geneticamente modificados. O Protocolo de Cartagena, entrado em vigor em Janeiro de 2000 e na qual o Brasil faz parte desde novembro de 2003, tem como principal objetivo

assegurar um nível de proteção adequado em relação à transferência (comércio), manipulação e uso dos OGMs. (PROTOCOLO DE CARTAGENA)

O cumprimento da legislação que regulamenta a comercialização de alimentos e ingredientes contendo Organismos Geneticamente Modificados (OGMs) é totalmente dependente da sensibilidade e confiabilidade dos métodos de detecção e quantificação de OGMs. Para a identificação do grão transgênico, a análise mais comum é a prática de amostragem de determinada quantidade de grão e envio da amostra para um laboratório especializado, aonde a amostra é submetida a uma extração do DNA. A técnica mais usada para a extração do DNA é a PCR (*Polymerase Chain Reaction* – Reação em Cadeia da Polimerase), que consiste na amplificação seletiva de sequências específicas da molécula de DNA. Apesar de ser um método seguro capaz de detectar uma ampla série de eventos e variedades genéticas, a *PCR* é um procedimento complexo, pois exige dados confidenciais à respeito da sequência da modificação genética e necessidade de material de referência certificado, cuja disponibilidade é limitada devido á confidencialidade por parte da empresa fornecedora das sementes transgênicas. (CONCEIÇÃO, 2006). De modo geral, a extração de DNA e identificação de grão transgênico é um processo extremamente complexo, que envolve várias etapas de diferentes reações químicas, além de ser um processo lento, de alto custo e ainda requer uma complexa preparação da amostra.

O uso de técnicas ópticas para detecção e quantificação dos compostos presentes em amostras vem se mostrando como uma técnica viável e rápida, que não requer ou necessita pouco preparo das amostras. O uso de técnicas ópticas para análise de alimentos e derivados vem sendo utilizados com o intuito de diminuir o tempo de processo de medida e reduzir os custos por análise. Em trabalhos recentes utilizando a técnica LIBS, foram demonstrados que, por Liu et al. (2018), LIBS e análises multivariadas foram utilizadas para detecção da presença de cobre em arroz, na qual se mostrou uma técnica eficiente na quantificação de cobre presente nas amostras. Sezer et al. (2017), utilizou LIBS para avaliar a eficiência da técnica na detecção de Lítio (Li) em amostras de carne, pois a ingestão de Lítio em grandes quantidades pode afetar o sistema nervoso central dos seres humanos . Os resultados mostraram

que a técnica LIBS possui grande potencial na determinação de Li em carnes. Em outro trabalho publicado por Sezer et al. (2017), a técnica LIBS foi utilizada para detectar traços de presença de adulterações em leite, pois a detecção de adulteração garante a proteção para o consumidor contra a compra de produtos que possam ter origem ilegal. A detecção do Li elementar no leite informa a presença de adulterantes na composição do alimento. Por Liu et. al. (2019), a técnica LIBS foi utilizada para identificar e diferenciar amostras transgênicas de milho das amostras não-transgênicas. Os resultados mostraram que a técnica LIBS junto com analises multivariadas conseguiram diferenciar as amostras entre transgênicas e não-transgênicas em uma precisão de 100%. No trabalho, os pesquisadores utilizaram apenas 1 (uma) classe de grão transgênico para diferenciar de 1 (uma) classe de seu parente não transgênico, o que facilita a diferenciação entre as duas classes. Já na pesquisa desenvolvida nessa dissertação, 4 (quatro) diferentes classes de grãos transgênicos são utilizadas para serem diferenciadas de duas classes diferentes de grãos não transgênicos, sendo essas as classes mais utilizadas comercialmente, indicando a potencial aplicação comercial que esta pesquisa possui.

Este trabalho tem como objetivo, utilizar técnicas espectroscópicas ópticas juntamente com métodos estatísticos multivariados computacionais de aprendizado supervisionado para desenvolver uma rotina de diferenciação entre amostras transgênicas e não-transgênicas em grãos de milho. Os métodos computacionais de aprendizado supervisionado de reconhecimento de padrões foram utilizados, juntos com classificadores k-NN (*k-nearest neighbour*) e SVM (*Support Vector Machine*), associados com a análise estatística multivariada não supervisionada de componentes principais (*Principal Component Analysis* – PCA). Com isso, será apresentado o potencial da técnica na diferenciação entre uma amostra transgênica e não-transgênica, o que se pode avaliar a possibilidade de portabilização e aplicação comercial das técnicas espectroscópicas ópticas para essa atividade em específico.

2. OBJETIVOS

2.1. Objetivo Geral

Este trabalho possui como objetivo principal a aplicação de técnicas espectroscópicas ópticas, sendo essas LIBS e FT-IR, associadas á métodos supervisionados de aprendizagem computacional, para a obtenção de uma rotina que diferencie amostras de grãos de milho transgênico de não transgênico.

2.2. Objetivos Específicos

- Obtenção do espectro óptico das amostras pela técnica Absorção no Infravermelho por Transformada de Fourier (FT-IR), e pela técnica Emissão Óptica por Plasma Induzido por Laser (LIBS)

- Análise das Componentes Principais (*Principal Component Analysis* – PCA), utilizando como base de dados o espectro óptico obtido em cada técnica, para obter informações das variáveis que são importantes para a diferenciação entre as amostras;

- Identificação dos elementos presentes nas amostras ou regiões espectrais importantes, que possam influenciar na diferenciação entre as amostras transgênicas e não transgênicas, através da análise do espectro óptico obtido através das técnicas FT-IR e LIBS;

- Utilização de métodos supervisionados de reconhecimento de padrões, com classificadores do tipo k-vizinhos próximos (k-NN) e Máquinas de Vetores de Suporte (SVM), para obtenção de rotina computacional que diferencie uma amostra transgênica da amostra não transgênica, com base nas informações espectrais obtidas.

- Análise quantitativa do poder de separação das técnicas através da análise da matriz de confusão. Comparar e verificar qual técnica espectroscópica óptica aplicada com métodos supervisionados de reconhecimento de padrões obteve a melhor precisão de diferenciação entre amostras transgênicas e não transgênicas.

3. REVISÃO DA LITERATURA

3.1. A Importância do Milho

A mais antiga espiga de milho conhecida é datada de 7.000 a.C., que foi encontrada no vale do Tehucan, região onde hoje se localiza o México. O milho é uma espécie da família das gramíneas, que se originou por um processo de seleção artificial, a partir do seu ancestral conhecido como Teosinte (gramínea com várias espigas sem sabugo). Esse processo de seleção artificial ocorreu através de uma maneira inconsciente de seleção, devido a escolha das espigas mais fáceis de serem colhidas e armazenadas. Com o passar dos tempos, eram-se escolhidas às plantas mais vigorosas, produtivas e de maior qualidade, na qual contribuíram para o surgimento de variedades com capacidade de adaptação em altas e baixas altitudes. Com o passar dos anos, o alto nível de domesticação e o melhoramento genético tornaram a planta completamente dependente da ação do homem. (GUIA DO MILHO, 2006)

A cultura do milho é realizada em quase todos os continentes, sendo os Estados Unidos da América o maior produtor mundial, representando aproximadamente 15% da quantidade total de milho exportado no mundo, seguido de Argentina e China. Vários países como Brasil, Ucrânia, Rússia, Índia e África do Sul também tiveram significantes impactos nas exportações de grãos de milho no mercado internacional. (USDA, 2017)

O Brasil espera colher 237,2 milhões de toneladas de grãos, sendo que para o milho a estimativa de produção para a safra 2016/2017 foi de 96 milhões de toneladas. A região Centro-Oeste se destaca como principal produtora nacional de grãos de milho, com uma produção para a safra 2016/2017 de 43.877,1 mil toneladas, em uma área de plantio estimada de 7.504 mil hectares. (CONAB, 2017). O estado de Mato Grosso do Sul, para o milho primeira safra 2016/2017, a área de plantio é de 28 mil hectares, com uma produtividade média de 8.880 kg/há, produzindo 248,6 mil t de milho. Para o milho segunda safra 2016/2017, uma área de 1.749,9 mil hectares é esperada á ser utilizada, com uma produção média de 5.300 kg/há, produzindo 9.274,5

mil t de milho. A comercialização ocorreu basicamente no mercado local, tendo como destinação as granjas agrícolas e de suínos. (CONAB, 2017).

Por ser uma fonte barata de carboidratos, proteínas e óleo, com uma ampla distribuição geográfica, o milho não é somente utilizado de forma direta na dieta humana e de animais, como também tem valor industrial para a produção de diversos produtos. Os principais produtos obtidos a partir do milho e vendidos diretamente ao consumidor são: creme, farinha, farinha pré-cozida flocada, polenta, flocos de milho, fubá, canjica (branca e amarela), pipoca, salgadinhos, cuscuz, angu, óleo de milho refinado, amidos pré-gelatinizados empregados nos cereais matinais, alimentos infantis e sopas instantâneas.

Assim, considerada como uma importante cultura para as necessidades atuais da sociedade moderna, a demando de consumo e de mercado, a produção de milho vem sofrendo contínuo aumento, tanto em níveis nacionais como mundiais. A própria elevação do consumo de derivados de aves e suínos exige indiretamente aumento na disponibilidade de milho, devido a sua incorporação junto com outros nutrientes nas rações específicas para a dieta balanceada dos animais. Esse aumento na produtividade do milho se deve ao uso de grãos geneticamente modificados (OGMs), ou também conhecidos como grãos transgênicos.

3.2. Comércio Do Milho

O milho é de longe o maior componente do comércio de grãos global, totalizando quase três/quartos do volume total produzido nos recentes anos. A maioria desses grãos produzidos é utilizada para alimentação, tanto diretamente na alimentação humana quanto indiretamente, e uma pequena parte é utilizada na área industrial. (USDA, 2018)

Embora os Estados Unidos da América seja o maior exportador mundial de grãos de milho, a exportação é uma relativa pequena parte da demanda de grãos de milho dos EUA, contabilizando um valor inferior a 15%. Essa baixa demanda de exportação significa que os preços do milho são largamente

determinados pela relação de oferta e demanda do mercado do EUA, na qual o restante dos países produtores de milho devem se ajustar para prevalecer os preços estabelecidos. Devido á grande influência dos EUA, o comércio e os preços dos grãos de milho mundiais dependem diretamente do *Corn Belt* norte-americano. Os outros países produtores de grãos de milho, na qual a maior parte está no hemisfério sul, plantam seus grãos de milho depois de descobrir o tamanho da plantação de milho no *Corn Belt* norte-americano, assim proporcionando um rápido e orientado suplemento no mercado mundial. Vários países, como Brasil, Ucrânia, Rússia, Índia e África do Sul tiveram significativas exportações de milho quando a produção era grande e os preços internacionais atrativos. (USDA, 2018)

A China tem sido uma significante fonte de incerteza no mercado mundial de milho, alternando de segundo maior exportador mundial em alguns anos, para um grande importador de grãos de milho em outros anos. A exportação de milho da China é largamente influenciada pelas taxas e subsídios de exportação do governo, pois os preços do milho na China são maiores do que os do mercado mundial. Grandes estoques de milho custam caro para manter, e a política de comércio de milho chinês flutuam com a pouca relação com a produção de grãos de milho no país, fazendo o mercado de milho chinês difícil de ser previsto. (USDA, 2018)

O comércio mundial de milho alcançou 78 milhões de toneladas nos anos de 1980/1981, com grandes importações da União Soviética e da Europa. Desde então, a importação de milho dos países que fazem parte da União Europeia declinou constantemente devido a Política de Agricultura Comum, limitando a importação de grãos na União Europeia e focando na produção local. A Hungria, por exemplo, depois que se juntou a União Europeia teve suas exportações de grãos de milho focalizadas para países da União Europeia, tirando o foco do comércio mundial. Durante o mesmo período, países como Japão, Coréia do Sul a Taiwan continuaram a aumentar suas importações de milho para suprir a crescente produção de carne, como alimento para os animais, e também para o próprio consumo direto humano. Países em desenvolvimento continuaram a aumentar as importações de milho

constantemente desde 1980, o que fez o comércio de milho ter uma produção de 130 milhões de toneladas na temporada 2013/2014. (USDA, 2018)

O Japão tornou-se o maior importador de grãos de milho nos recentes anos. A importação de grãos para alimentação humana no Japão tem se estagnado nos últimos anos, enquanto a importação para uso industrial e manufaturamento do amido tem crescido constantemente. A importação de milho pela União Europeia tem sido variável nos recentes anos, variando de 3 milhões de toneladas no ano de 2009/2010 para 16 milhões toneladas em 2013/2014. O preço do milho comparado com o preço do trigo da União Europeia e com as políticas de decisões sobre importações ajudam a entender essa variação de importação. O México é um importador crescente de grãos de milho. Mesmo sendo um grande produtor de milho, o México processa a maioria de seus grãos de milho branco para produtos de alimentação humana, e tem focado no comércio de importação de milho amarelo para alimentação de gado para suprir a crescente produção de carne bovina. (USDA, 2018)

3.3. Milhos Geneticamente Modificados

A produção de milho GM utiliza a tecnologia de DNA recombinante e baseia-se na transformação genética através da inserção no genoma da planta de uma ou mais sequências, geralmente isoladas de espécies diferentes, de forma a garantir a expressão do gene de interesse. (DINON, 2007)

A construção de OGMs para expressar determinada proteína normalmente utiliza três elementos básicos para sua expressão: o promotor, que controla a expressão da proteína recombinante no organismo; a região codificadora, que codifica a proteína recombinante de interesse; e a região terminadora, que determina o final do processo de transcrição do gene. O elemento promotor mais utilizado é o CaMV 35S, que é derivado do vírus fitopatogênico do mosaico da couve-flor, e o elemento terminador mais utilizado é o NOS, derivado do gene da nopalina sintase do plasmídio Ti da bactéria *Agrobacterium tumefaciens.* (DINON, 2007) O promotor da região codificadora é o que garante as diferentes expressões genéticas do milho, sendo a principal

característica a resistência á insetos e pragas, de acordo com a toxina inserida no DNA da planta.

Um dos principais fatores que comprometem o rendimento e a qualidade da produção na cultura do milho é a incidência de pragas. Dentre as principais, podemos destacar a lagarta-do-cartucho, lagarta-da-espiga e a broca-da-cana-de-açúcar. (REVISTA CULTIVAR, 2011). Raças adaptadas da lagarta-do-cartucho (*Spodoptera frugiperda*) compõe uma das mais importantes pragas que afetam genótipos tropicais de milho, chegando a causar até 34% de redução na produção dessa cultura no Brasil. (LOGUERCIO et al., 2002). A lagarta-da-espiga (*Helicoverpa zea*) causa danos diretos e indiretos pela abertura da espiga, facilitando a entrada de outras pragas, umidade e fungos causadores de podridões.

Tradicionalmente, o controle de pragas é realizado com base em inseticidas químicos, que intrinsicamente podem trazer consequências colaterais negativas em termos de toxicidade ao homem, aos animais e ao meio ambiente em geral. O uso abusivo e impróprio desses produtos sintéticos causaram vários problemas ambientais e de saúde, ameaçando sobremaneira a sustentabilidade do sistema de produção agrícola convencional. (LOGUERCIO et al., 2002). A importância da utilização dos milhos geneticamente modificados melhora o controle das lagartas que atingem a espiga, e que com a utilização de inseticidas não seriam combatidas satisfatoriamente, mesmo quando a aplicação fosse realizada diretamente nas espigas. (REVISTA CULTIVAR, 2011)

A espécie bacteriana de solo *Bacillus thuringiensis* é de ocorrência cosmopolita, sendo encontrada nos mais diversos ecossistemas do planeta. O gênero *Bacillus* possui uma fase de esporulação característica no seu desenvolvimento, na qual o esporo bacteriano e cristais protéicos são simultaneamente formados, sendo estes últimos sob forma de inclusões parasporais. Tais cristais em *Bt*, também chamados de “δ-endotoxinas” ou *ICPS* (*Insecticidal Crystal Proteins)*, e codificados pelos chamados gene *cry*. Uma das principais e importantes características das proteínas inseticidas *cry* é sua alta especificidade em relação as espécies-alvo de insetos afetados.

(LOGUERCIO et al., 2002) Na membrana das células epiteliais do intestino, a interação toxina-receptor leva á formação de poros na membrana celular, o que altera o balanço osmótico das células epiteliais, que incham e sofrem rupturas, levando o inseto à morte por dificuldade de alimentação e infecção generalizada. Também logo após a ingestão da toxina pela lagarta, ocorre à inibição da ingestão de alimentos, levando á morte do inseto. (EMBRAPA, 2011). Esse efeito tóxico seletivo não se estende a outros organismos que não tenham tais receptores compatíveis, tomando as *ICPs* inertes a seres humanos, peixes, animais selvagens e outros insetos benéticos que podem auxiliar no controle das praga-alvo. (LOGUERCIO et al., 2002)

Com o advento da biotecnologia, foi desenvolvida uma nova tática de controle de pragas, que consiste nas plantas geneticamente modificadas resistentes a insetos. Através de técnicas de laboratórios, os genes *Bt* foram introduzidos em plantas de milho, dando origem ao milho geneticamente modificado, o milho *Bt*. (REVISTA CULTIVAR, 2011) Os níveis de localização da expressão dos genes *Bt* na planta geneticamente modificada podem ser regulados, permitindo a presença contínua da toxina em todo o corpo da planta ou somente nas partes relevantes, dependendo dos hábitos de ataque dos insetos-alvo. (LOGUERCIO et al., 2002) O milho geneticamente modificado *Bt,* é uma espécie de milho na qual foram introduzidos genes específicos da bactéria de solo, *Bacillus thuringiensis,* que promovem na planta a produção de uma proteína tóxica específica para determinados grupos de insetos. Assim, o milho *Bt* é uma cultiva de milho resistente a determinadas espécies de insetos sensíveis a essa toxina. (EMBRAPA, 2011) Atualmente, conforme a tabela 01 abaixo, os eventos liberados comercialmente no Brasil que expressam a toxina *Bt* em plantas de milho são:

Tabela 01 – Eventos presentes em milhos *Bt*. Modificado Revista Cultivar, 2011.

Empresa	Evento	Marca (sigla)	Toxina
Monsanto	MON810	YieldGard (YG, Y)	Cry 1Ab
Dow AgroSc.	TC1 507	Herculex (HX, H)	Cry 1F
Syngenta	BT11	Agrisure TL (TL)	Cry 1Ab
Monsanto	MON89034	YieldGard VTPRO (PRO)	Cry 1A105(1Ab, 1Ac, 1F) + Cry2Ab2
Syngenta	MIR 162	Viptera (VIP)	VIP3Aa20
Syngenta	176	Knockout	Cry1Ab

Os eventos que expressam as toxinas Cry 1A(b) e Cry 1F, com atividades sobre os lepidópteros, que incluem as lagartas-do-cartucho e a largarta-da-espiga. O evento contento os genes Cry1A.105 e Cry2Ab2, que representam uma segunda geração de milho transgênico resistente a insetos, produz simultaneamente duas proteínas ativas contra lagartas-praga. (EMBRAPA, 2011)

Recentemente, uma nova classe de proteínas entomocidas foi identificada como sendo secretada no sobrenadante de culturas de certas cepas de *Bt* em fase logarítmica de crescimento. Essas proteotoxinas são denominadas *VIPs* (*Vegetative Insecticidal Proteins*), tendo demonstrado ação sobre um espectro maior de espécies de insetos-praga quando comparadas a muitas “δ-endotoxinas” *cry*. As proteotoxinas *VIPs* são produzidas em etapas iniciais do processo de crescimento das bactérias em cultura, antecipando assim a sua obtenção. A forma de ação da *VIP* é idêntica a *cry,* destruindo a função digestiva dos insetos-alvo, mesmo que os receptores de membrana das células do intestino médio possam ser de naturezas distintas. (LOGUERCIO et al., 2002).

3.4. Legislação da Rotulagem em OGM’s

A grande quantidade de OGMs que vem sendo aprovada no mundo nos últimos anos e a suspeita de que os mesmos não sejam seguros para o consumo, levaram ao desenvolvimento de legislações para sua

comercialização. (CONCEIÇÃO, 2006) Em 29 de Janeiro de 2000, a Conferência das Partes da Convenção sobre Diversidade Biológica (CDB) adotou o seu primeiro protocolo suplementar conhecido como Protocolo de Cartagena sobre Biossegurança. Em vigor desde 11 de setembro de 2003, o Protocolo de Cartagena visa assegurar um nível adequado de proteção no campo da transferência, da manipulação e do uso seguro dos organismos vivos modificados, resultantes da biotecnologia moderna.. O protocolo cria uma instância internacional para discutir os procedimentos que deverão nortear a introdução de organismos vivos modificados em seus territórios. Também estabelece procedimento para um acordo de aviso prévio para assegurar que os países tenham as informações necessárias para tomar decisões conscientes antes de aceitarem a importação de organismos geneticamente modificados (OGMs) para seu território. Trata-se, portanto, de um instrumento de direito internacional que tem por objetivo proteger os direitos humanos fundamentais, tais como a saúde humana, a biodiversidade e o equilíbrio ecológico do meio ambiente, sem os quais ficam prejudicados os direitos a dignidade, a qualidade de vida e a própria vida. (PROTOCOLO DE CARTAGENA.)

A simples detecção desses organismos geneticamente modificados não garante a segurança de seu uso, mas a detecção é necessária por algumas razões: a Lei nº 8.078, de 11 de setembro de 1990, dispõe sobre a proteção do consumidor e os direitos básicos, segundo quais todos os cidadãos têm direito á informação adequada sobre produtos e serviços. (BRASIL, Lei n 8.078) A lei federal nº11. 105, de 14 de março de 2005, que estabelece normas de segurança e mecanismos de fiscalização de atividades que envolvam organismos geneticamente modificados e seus derivados, no art.40 estabelece que os alimentos e ingredientes alimentares destinados ao consumo humano ou animal que contenham, ou seja, produzidos a partir de OGMs ou derivados, deverão conter informação nesse sentido em seus rótulos. (BRASIL, Lei n 11.105) O decreto nº4.680, de 24 de abril de 2003, assegura quanto aos alimentos e ingredientes alimentares destinados ao consumo humano ou animal que contenham ou sejam produzidos a partir de organismos geneticamente modificados. Estabelece no art.2 que na comercialização de

alimentos e ingredientes alimentares destinados ao consumo humano ou animal que, contenham ou sejam produzidos a partir de OGMs com presença acima de um por cento do produto, o consumidor deverá ser informado da natureza transgênica desse produto. O decreto estabelece que a rotulagem deva ser feita tanto em produtos a granel ou *in natura.* (BRASIL, Decreto n 4.680)

3.5. Técnicas de Detecção de OGM's em Alimentos

A análise de rotina de produtos alimentícios contendo OGMs compreende três etapas: 1 – detecção; 2 – identificação do OGM presente na amostra, para determinar se este é autorizado e; 3 – quantificação do OGM no produto, para checar a necessidade de rotulagem ou não, conforme a legislação. Os OGM's são caracterizados pela presença de um ou mais segmentos de DNA exógenos, que podem ou não proporcionar a expressão de novas proteínas. Assim, a detecção de OGMs é focalizada na sequência de DNA exógena ou na proteína transgênica. (CONCEIÇÃO, 2006)

3.5.1. Detecção Baseada Na Presença De Proteína

Os métodos baseados na análise de proteínas detectam a presença de proteínas recombinantes que podem ser produzidas durante certos estágios de desenvolvimento ou em apenas algumas partes da planta. Bioensaios e Imunoensaios são os mais utilizados para detectar a presença de proteína. (DINON, 2007)

Os Bioensaios são ensaios baseados na presença de proteínas que detectam somente a presença de proteínas em OGMs que são resistentes á herbicidas. É um método na qual consiste em germinar as sementes alvo em uma solução diluída de herbicida. Se a semente for resistente ao herbicida contido na solução, ocorrerá a germinação e o desenvolvimento normal da planta. As principais limitações desse método é o longo tempo para a obtenção

do resultado, que é de aproximadamente uma semana, e a utilização restrita em sementes OGMs resistentes á herbicidas, e não á pragas e fungos. Bioensaios são utilizados geralmente por companhias exportadoras de sementes e grãos. (CONCEIÇÃO, 2006)

Imunoensaios são ideais para a detecção qualitativa e quantitativa de proteínas complexas. Mas a detecção de OGMs através de imunoensaios, nem sempre é possível, pois quando o nível de expressão da proteína transgênica nas partes das plantas que são utilizados na produção de alimentos é baixo, a sua detecção torna-se difícil. A concentração de proteína transgênicas nos tecidos de plantas varia em função da idade, variedade e condições ambientais. O processamento de alimento é outra situação que desfavorece os imunoensaios, devido a promover a remoção das proteínas em determinados produtos, alterando a sua conformação e impedindo o seu reconhecimento pelo anticorpo. Imunoensaios também não podem ser utilizados na detecção de OGMs cuja modificação genética não resulta em nova proteína, pois quando a proteína transgênica é muito similar á proteína nativa, não é possível produzir anticorpos específicos que reconheçam apenas a proteína transgênica, tendo em vista a semelhança de epitomo. Outra limitação dos imunoensaios é a incapacidade de detectar alimentos contendo OGMs cuja modificação genética resulta no aumento da expressão de uma proteína nativa, e também a incapacidade de distinguir variedades GM que apresentam diferentes eventos, porém expressam a mesma proteína transgênica. (CONCEIÇÃO, 2006)

3.5.2. Detecção Baseada Na Presença De DNA

Os métodos baseados na detecção de DNA são eficientes até mesmo para amostras altamente processadas, quando o DNA está fragmentado, mas não seriamente degradado, como geralmente ocorre durante o processamento de alimentos. O DNA pode apresentar-se fragmentado no caso de alimentos cujo processamento inclui alterações no pH e o uso de alta temperaturas (DINON, 2007), como derivados de amido e óleos refinados. O processamento de milho com uso de calor em meio ácido favorece a degradação do seu DNA.

A análise de DNA baseia-se na capacidade de detectar sequências únicas de DNA recombinante ou endógeno da planta e gera resultados evento-específicos para plantas diferentes que expressam a mesma proteína recombinante. Além disso, permite a detecção e quantificação de plantas GM que não expressam nenhuma nova proteína devido ao silenciamento da expressão do gene. Nas amostras em que existe DNA recombinante, todo o DNA exógeno é, em princípio, suscetível a detecção: sequências de promotores, genes de interesse introduzidos, sinais de terminação e genes marcadores usados para seleção das plantas modificadas em laboratório. (DINON, 2007)

O protocolo para análise de OGMs em alimentos, baseado na presença de DNA, segue algumas linhas gerais que incluem: extração e purificação do DNA da amostra, determinação da concentração de DNA extraído, amplificação do DNA por Reação em Cadeia da Polimerase (PCR) com iniciadores específicos para sequências presentes no OGM e eletroforese do DNA amplificado. (DINON, 2007) A PCR, que consiste na amplificação seletiva de sequências específicas da molécula de DNA, é o principal método utilizado na detecção e quantificação de alimentos contendo OGMs. A PCR é utilizada por ser um método sensível, específico, seguro e capaz de detectar uma ampla série de eventos, e de distinguir as variedades GM que apresentam diferentes construções gênicas, porém expressam a mesma proteína. (CONCEIÇÃO, 2006) Um pré-requisito essencial para a detecção de OGMs em alimentos compreende o conhecimento do tipo de modificação genética, incluindo a construção genética do inserto e elementos regulatórios (promotores e reguladores) que o flanqueiam. Para análise, é necessária uma quantidade mínima de amostra contendo DNA amplificável compreendendo a sequência de DNA alvo. (DINON, 2007)

Um passo crítico para a análise do DNA é a sua correta extração, seguindo um método que assegure a pureza e a qualidade do DNA extraído. (DINON, 2007) Outro problema na reação PCR é que o DNA polimerase utilizada pode ser inibida por substâncias contidas no alimento, como proteínas, gorduras e polissacarídeos. Por isso é fundamental a utilização de

controles de qualidade para evitar resultados falso-negativos. (CONCEIÇÃO, 2006)

3.6. Espectroscopia Óptica

Em termos genéricos, a espectroscopia é definida como a técnica de análise da interação de qualquer radiação eletromagnética com a matéria, que pode ser reflexão, refração, espalhamento, interferência, difração e absorção. Os métodos espectroscópicos, que são classificados de acordo com a região do espectro eletromagnético, quando relacionados aos níveis energéticos de moléculas ou átomos, são baseados na medida da quantidade de radiação emitida ou absorvida pelas moléculas ou pelas espécies atômicas. São empregadas para a elucidação de estruturas moleculares, bem como na determinação qualitativa e quantitativa de compostos orgânicos ou inorgânicos. (SKOOG, 2006)

3.6.1. Radiação Eletromagnética

A radiação eletromagnética é uma forma de energia que pode ser descrita com propriedades como comprimento de onda, frequência, velocidade e amplitude. As radiações eletromagnéticas, em seu aspecto ondulatório, consistem de uma campo elétrico e uma campo magnético, que oscilam senoidalmente e são perpendiculares entre si e perpendiculares em relação a direção de propagação da onda. A figura 1 abaixo ilustra o comportamento de propagação da onda. (SKOOG, 2006)

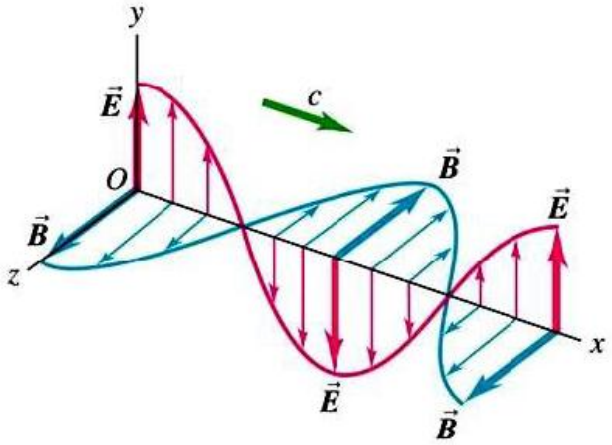

Figura 1 - Propagação de uma onda eletromagnética, mostrando que o campo elétrico (E) é perpendicular ao campo magnético (B), e perpendiculares á direção de propagação da onda (eixo x). Fonte: Imagens Google. Acesso em setembro de 2018.

A amplitude de uma onda eletromagnética é uma quantidade vetorial que fornece a medida da intensidade do campo elétrico ou magnético no ponto máximo da onda. O período (*p*) é o tempo, em segundos, necessário para a passagem de dois pontos da mesma fase por um ponto fixo no espaço. A frequência (*f*) é o número de oscilações que ocorrem em um segundo. A frequência da onda de qualquer radiação eletromagnética é determinada pela fonte que a emite e permanece constante independentemente do meio que esta atravessa. O comprimento de onda (*λ*) é a distância linear entre dois máximos ou mínimos sucessivos de uma onda. A velocidade da onda (*v*) é dada pela multiplicação da frequência (em onda por unidade de tempo) pelo comprimento de onda (em distância por onda) ($v = f.\lambda$), resultando na sua unidade em distância por unidade de tempo (m/s). A velocidade da onda e o comprimento da onda dependem do meio em que estão. O número de onda ($\bar{\nu}$), que também pode ser usado para descrever a radiação eletromagnética, é definido como o número de ondas por centímetro, e é igual a $\frac{1}{\lambda}$, com unidade de cm^{-1}. A intensidade da onda eletromagnética é energia de um feixe que atinge uma determinada área por unidade de tempo, por unidade de ângulo sólido. (SKOOG, 2006)

A radiação eletromagnética, quando tratada como partícula, é considerada como sendo constituída de pacotes de energia chamados fótons.

Fótons são partículas de radiação eletromagnética que possuem massa zero e a sua energia é relacionada com o seu comprimento de onda, frequência e número de onda. A quantidade de energia contida em um feixe luminoso pode ser dado pela expressão abaixo:

$$\mathrm{E} = h.f = \frac{hc}{\lambda} = hc\bar{\nu} \qquad (1)$$

, em que *h* é a constante de Planck (6,63 X 10^{-34} J.s) e *c* é a velocidade da luz no vácuo.

Os tipos específicos de interações da radiação com a matéria dependem fortemente da energia da radiação empregada e o modo de detecção. O espectro eletromagnético é um espectro cujo, no eixo das ordenadas tem-se a intensidade da radiação absorvida ou transmitida pela molécula ou átomo, e no eixo das abscissas tem-se a frequência (*f*), comprimento de onda (*λ*) ou o número de onda ($\bar{\nu}$). Essa observação é entendida somente se as energias dos átomos ou das moléculas estiverem contidas em quantidades discretas, na qual quando um átomo ou molécula perde ou ganha energia *E*. Essa energia é captada como uma radiação de frequência *f* proporcional a *E*, e então, um pico de emissão ou de absorção aparece nessa frequência, qual a sequência de emissão ou absorção se constrói o espectro eletromagnético. (ATKINS, 2006) A imagem abaixo demonstra um espectro de emissão de átomos de ferro excitados.

Para moléculas, o espectro eletromagnético também pode ser chamado de espectro de transição. Quando o analito (substância de interesse na análise) se encontra no seu estado de mais baixa energia, este é dito estar no estado fundamental. Quando a energia absorvida da luz é suficiente para fazer a molécula sair do estado fundamental e ir para um estado de maior energia, é dito que a molécula está no seu estado excitado. (SKOOG, 2006) Uma molécula que absorve energia em um determinado comprimento de onda, na qual promove a transição entre o estado fundamental e o estado excitado, é chamada de cromóforo. Uma molécula excitada possuirá uma das possíveis quantidades discretas de energia descritas pelas leis da mecânica quântica, os níveis de energia da molécula. Os níveis de energia, em maioria, são

determinados pelas possíveis distribuições espaciais dos elétrons e são chamados níveis eletrônicos de energia. Sobre estes níveis de energia, existem os níveis vibracionais, que indicam os vários modos de vibração da molécula. Há subdivisões menores entre os níveis vibracionais que são chamados de níveis rotacionais, que está relacionada com a rotação da molécula em torno do seu centro de gravidade. (PUC-RIO 0321127/CA) A figura 2 abaixo ilustra os níveis de energia de uma molécula diatômica.

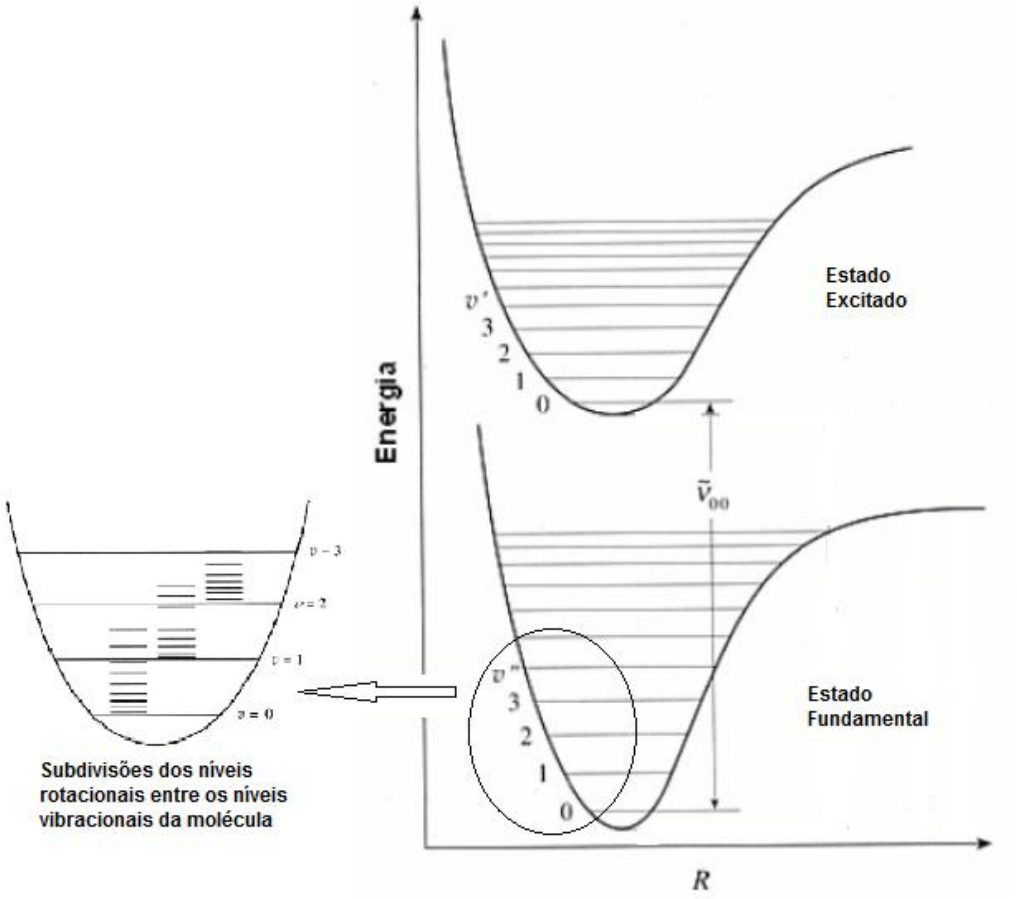

Figura 2 - Níveis de energia de uma molécula diatômica, mostrando seus níveis vibracionais e rotacionais. Adaptado de BASSI, 2001.

A energia total, *E*, associada com uma molécula é então dada por:

$$E = E_{eletrônica} + E_{vibracional} + E_{rotacional} \qquad (2)$$

, onde $E_{eletrônica}$ é a energia associada com os elétrons nos vários orbitais externos da molécula, $E_{vibracional}$ é a energia da molécula como um todo devido

ás vibrações interatômicas, e $E_{rotacional}$ é a energia associada com a rotação da molécula em torno do seu centro de gravidade. (ATKINS, 2006)

Com base no espectro eletromagnético, podem-se obter informações sobre uma amostra. Quando a amostra é estimulada pela aplicação de uma fonte de radiação eletromagnética externa, muitos processos de interações podem ocorrer, como por exemplo, espalhamento ou reflexão. Mas a radiação incidente também pode ser absorvida pela amostra, e fornecer informações qualitativas e quantitativas. Na espectroscopia de absorção, a quantidade de luz absorvida em função do comprimento de onda é mensurada, enquanto na espectroscopia de fotoluminescência, como fosforescência e fluorescência, é mensurada a intensidade da radiação emitida pela amostra. (SKOOG, 2006) A figura 3 abaixo demonstra e classifica o campo eletromagnético de acordo com a sua frequência e comprimento de onda. A luz branca é uma mistura de radiação eletromagnética com comprimento de onda que variam de 380nm até aproximadamente 700nm (1nm = 10^{-9}m). (ATKINS, 2006)

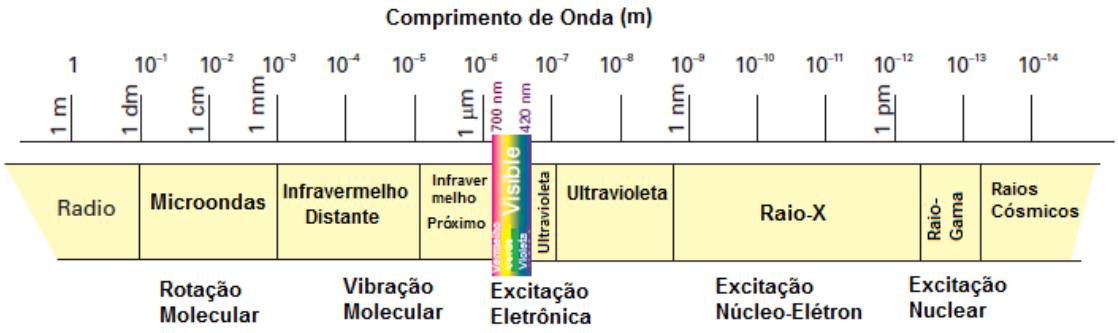

Figura 3 - O espectro eletromagnético e a classificação das regiões espectrais. Adaptado do livro Atkins. Traduzido pelo autor. Fonte: Atkins.

3.6.2. Espectroscopia de Absorção

Na espectroscopia de absorção, a quantidade de luz absorvida por uma amostra é medida em função do comprimento de onda da fonte de emissão de radiação eletromagnética. Cada espécie molecular é capaz de absorver suas próprias frequências características da radiação eletromagnética, num processo que transfere energia para a molécula e resulta em um decréscimo da intensidade da radiação eletromagnética incidente. A medida que a luz

atravessa um meio contendo um analito que absorve, um decréscimo da intensidade ocorre na proporção que o analito é excitado. Quanto mais longo for o comprimento do caminho do meio através do qual a luz passa, o chamado caminho óptico, ou quanto maior for a concentração de absorventes no caminho óptico, maior será a atenuação (diminuição do número de fótons por segundos presentes no feixe) da radiação incidente. (SKOOG, 2006)

Quando radiação eletromagnética entra em contato com um composto químico qualquer, e essa radiação tiver comprimento de onda na faixa do visível ou do ultravioleta, muito provavelmente essa radiação vai permitir que um elétron seja promovido para um nível energético eletrônico superior. Para que a molécula absorva um fóton para excitar um elétron, esse fóton deverá ter energia exatamente igual á diferença entre os orbitais dos níveis de energia. Se essa energia não for exatamente igual, a transição não ocorrerá. Algumas outras restrições também existem nesse processo, por exemplo, não será qualquer elétron que vai sofrer a excitação, e sim apenas um dos elétrons do orbital molecular mais alto ocupado, chamado de HOMO (*Highest Occupied Molecular Orbital*), na qual também não pode ir para qualquer orbital. Geralmente o elétron excitado irá para um orbital molecular de menor energia que não esteja ocupado com outros elétrons, o chamado orbital LUMO (*Lowest Unoccuppied Molecular Orbital*), respeitando as regras de transição eletrônica. (OLIVEIRA, 2001) Transições de baixa energia também são possíveis entre níveis vibracionais no interior de um mesmo nível eletrônico, utilizando radiação eletromagnética na faixa do infravermelho. (PUC RIO 0321127/CA). Essa interação da radiação eletromagnética com a matéria também depende do movimento vibracional das moléculas que compõe a amostra, e também da eletronegatividade da molécula.

A representação da probabilidade de absorção em função do comprimento de onda é chamada de espectro de absorção. (PUC RIO 0321127/CA) O espectro é uma espécie de impressão digital de um composto químico, uma vez que cada composto difere de outro em função da composição química, ou seja, de diferentes átomos que o formam e também da geometria molecular. Dessa forma, a análise do espectro permite dizer qual é a molécula em questão. (OLIVEIRA, 2001)

A lei da absorção, conhecida como a lei de Beer-Lambert, relaciona quantitativamente como a grandeza da atenuação depende da concentração das moléculas absorventes e da extensão do caminho sobre o qual ocorre a absorção. Se luz de intensidade I_0 passa através de uma substância, que pode estar em uma solução, de espessura *d* e concentração molar *c*, a intensidade *I* da luz transmitida obedece a lei de Beer-Lambert da seguinte maneira: (PUC RIO 0321127/CA)

$$I = I_0 * 10^{-\varepsilon dc} \quad \text{ou} \quad \log(\frac{I}{I0}) = -\varepsilon dc \qquad (3)$$

O coeficiente ε é chamado de coeficiente de absorção molar. O coeficiente de concentração molar depende da frequência da radiação incidente e é maior onde a absorção é mais intensa. O coeficiente de absorção molar indica a seção transversal molar da absorção, e quanto maior for a área da seção transversal da molécula para absorção, maior é a habilidade de bloquear a passagem da radiação incidente. Para simplificar a equação de Beer-Lambert, é utilizado a absorbância, *A,* da amostra para um dado número de onda como:

$$A = \log(\frac{Io}{I}) \quad \text{ou } A = -\log T \qquad (4)$$

, onde T é a transmitância da amostra, que relaciona a potência radiante do feixe de luz, P_0, e a potência do feixe de luz após passar pela amostra, *P*.

Assim a lei de Beer-Lambert fica da seguinte forma:

$$A = \varepsilon dc = \log(\frac{P0}{P}) \qquad (5)$$

, onde o produto ε*dc* é conhecido como densidade óptica da amostra. (ATCKINS, 2006)

3.6.3. Absorção Molecular no Infravermelho

A região do infravermelho do espectro eletromagnético, em termos de número de onda, vai de 4000 a 400 cm^{-1}. O infravermelho fornece evidências da presença de vários compostos que estão na amostra, devido á interação entre momento de dipolo elétrico da molécula com a radiação incidente. Quase todos os compostos que tenham ligações covalentes, apresentam absorção na região do infravermelho, estes podendo ser compostos orgânicos ou inorgânicos. A radiação infravermelha geralmente não é energética suficiente para causar transições eletrônicas, porém podem induzir transições nos estados vibracionais associados com o estado eletrônico fundamental da molécula. Os modos de como as moléculas vibram podem variar. Nas vibrações de estiramento, por exemplo, os átomos das moléculas se aproximam e depois se afastam um do outro. A energia potencial desse sistema a qualquer instante depende da energia relacionada com a ligação química entre esses átomos, quando “comprimem” ou “esticam” essa ligação. Outros tipos de vibrações moleculares na qual há a deformação angular entre as ligações dos átomos podem acontecer, como as vibrações balanço no plano (*rocking*), tesoura no plano (*scissoring*), oscilação fora do plano (*wagging*) e torção fora do plano (*twisting*). (SKOOG, 2006)

A absorção molecular no infravermelho também pode ser realizada como Absorção no Infravermelho por Transformada de Fourier (FTIR), quando os aspectos construtivos do equipamento utilizado possui um interferômetro ao invés de uma grade de difração, como apresenta a figura 4 abaixo. Na FTIR, a radiação infravermelha é dividida em dois sinais iguais, sendo que um deles percorre um caminho fixo enquanto o outro percorre um caminho óptico variável, e depois os dois sinais são recombinados antes de ser incidido na amostra. O fato de existir um caminho óptico variável faz com que haja interferências construtivas e destrutivas na onda infravermelha formada, o que da origem ao interferograma, que contém toda a energia radiativa que veio da fonte, além de uma grande faixa de comprimentos de onda. Quando o feixe incide na amostra, essa absorve de forma simultânea todos os comprimentos de onda que promovem suas transições vibracionais, e que quando chega ao

detector, contém informações sobre a quantidade de energia absorvida em cada comprimento de onda. O sinal obtido é comparado com um sinal de referência, e o sinal de absorção da amostra pelo interferograma é construído. Mas o sinal final do interferograma contém sinais no domínio temporal, na qual necessita que o procedimento matemático de Transformada de Fourier seja realizado pelo computador, para que o sinal seja observado no domínio das frequências, ou mais comumente usado, no domínio dos números de onda. A Transformada de Fourier extrai as frequências individuais que foram absorvidas e reconstrói o gráfico de absorção para a correta visualização do espectro a partir da soma dos vários sinais obtidos percorridos pelo feixe dentre os vários caminhos ópticos percorridos. (APOSTILA ESPECTROSCOPIA; SKOOG, 2006)

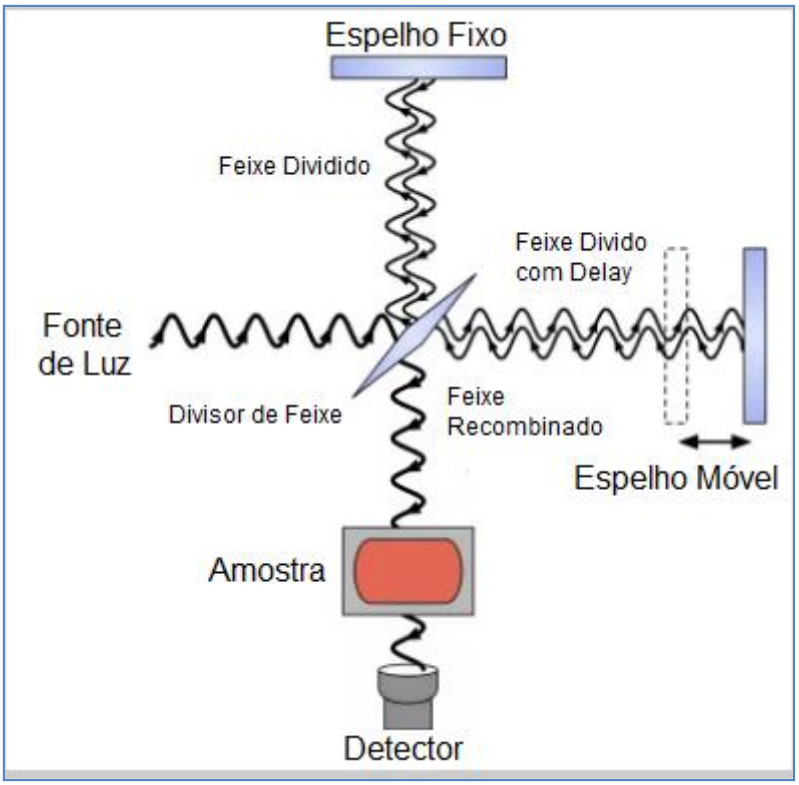

Figura 4 - Esquema do funcionamento do interferômetro de Michelson. Fonte: fciencias.com. Adaptado pelo Autor.

3.6.4. Espectroscopia De Emissão Óptica com Plasma Induzido por Laser – LIBS

A espectroscopia de emissão óptica com plasma induzido por laser, acrônimo para *Laser Induced Breakdown Spectroscopy* – LIBS, é um tipo de técnica óptica, de caráter multi-elementar, que busca identificar os elementos químicos presentes na amostra através da radiação eletromagnética emitida por eles, quando excitados a partir de um laser de alta intensidade.

Na técnica LIBS, um laser pulsado de alta energia é direcionado e focalizado na superfície da amostra que se deseja analisar. Quando um laser de alta intensidade, cuja irradiância é da ordem de GW/cm^2, atinge a superfície da amostra, este absorve radiação num curto intervalo de tempo, da ordem de 10^{-13}s, fazendo com que a temperatura superficial da amostra atinja as ordens de, aproximadamente, 10^4 *K*. (RANULFI, 2019) Essa alta temperatura e o gradiente de campo elétrico ali formados promovem a quebra das ligações químicas do material, resultando na produção inicial de íons e elétrons livres. Tal fenômeno de interação é conhecido como ruptura induzida por laser. A ruptura leva á produção de íons através da absorção de fótons pelos átomos do material-alvo, gerando aquecimento superficial e pares elétron-íons, os quais serão os componentes da pluma do plasma (região externa que envolve o plasma) e do plasma. Neste ponto, o elétron livre absorve radiação e, devido a sua interação com o campo elétrico das espécies existentes na região, este é acelerado (ganha mais energia cinética) e pode colidir com outro átomo, ionizando-o. Os dois elétrons livres absorvem radiação e, sucessivamente, ionizam outros átomos, formando um efeito de ionização em cascata. Tal efeito em cascata, juntamente com a absorção intrínseca da radiação por parte dos elementos da amostra, causará um rápido aquecimento da sua superfície, o que leva a evaporação de material, dando origem á ablação (retirada de pequena quantidade da amostra) e formação do plasma. (FRANCO, 2017) Após o fim do pulso de laser e conforme o plasma se expande e resfria, este perde energia por emissão de radiação devido á recombinação entre elétrons livres e, assim decai. Durante o processo de decaimento do átomo, os elétrons emitem um fóton, onde é possível medir os comprimentos de onda das linhas

de emissão atômica/iônica, e identificar os elementos presentes. (RANULFI, 2019) A emissão de plasmas induzidos por laser consiste de linhas espectrais atômicas e iônicas características das espécies constituintes. A figura 5 abaixo apresenta de forma simplificada os processos ocorridos durante as medidas LIBS.

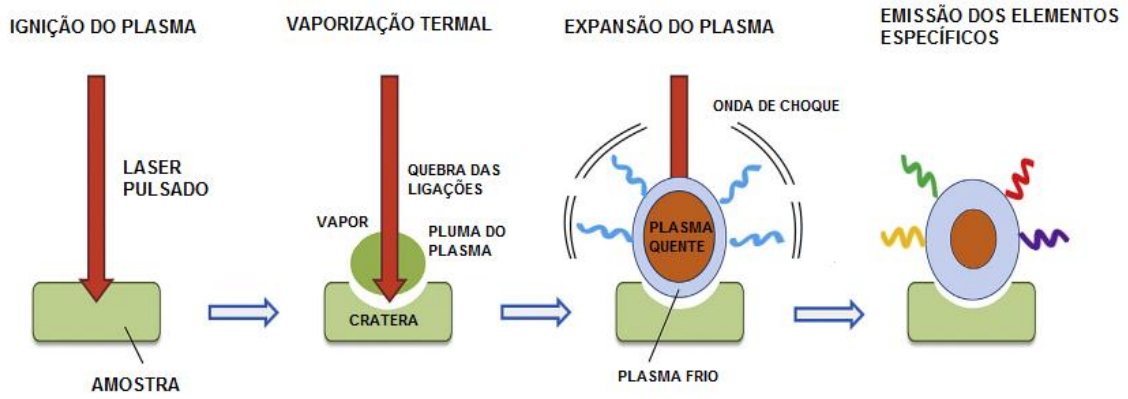

Figura 5 - Esquema da interação plasma-amostra para LIBS. Fonte: Markiewicz-Keszycka et al, 2017. Adaptado pelo autor.

LIBS é considerada uma técnica que causa pouco dano a amostra, podendo ser considerada não destrutiva, que não necessita ou necessita de pouco preparo da amostra, que possui rapidez nas análises, podem ser aplicadas *in situ*, em materiais líquidos, sólidos ou gases, e avaliações multielementares simultâneas, tornam a sua aplicação viável nas mais variadas análises. (PASQUINI et al., 2007.) O espectro registrado contém informação qualitativa e quantitativa que pode ser correlacionada com a identidade da amostra ou empregada na determinação da quantidade de seus constituintes. O espectro *LIBS* possui um caráter multi-elementar, na qual contém a informação da presença de todos os elementos constituintes da amostra, de uma só vez. É importante notar que, devido ao fato de que a técnica LIBS analisa apenas uma pequena parte quantidade de material, a acurácia e precisão das medidas são extremamente dependentes da homogeneidade da amostra. (RANULFI, 2019)

LIBS é uma técnica relativamente simples do ponto de vista da instrumentação. Os equipamentos consistem de um laser pulsado de alta energia focalizado na superfície da amostra. Ocorre o processo de ablação, vaporização e geração do plasma. Um conjunto de lentes é utilizado para coletar as radiações características que são emitidas pelos átomos quando esses começam a decair, devido à nuvem de plasma começar a se resfriar. A luz que é coletada é então conduzida a um espectrômetro através de fibras ópticas, que tem a função de difratar a luz e enviar ao detector. Os sinais são armazenados e processados, podendo ser visualizados por meio de um software dedicado. A figura 6 abaixo apresenta um esquema geral do equipamento usado para medidas LIBS.

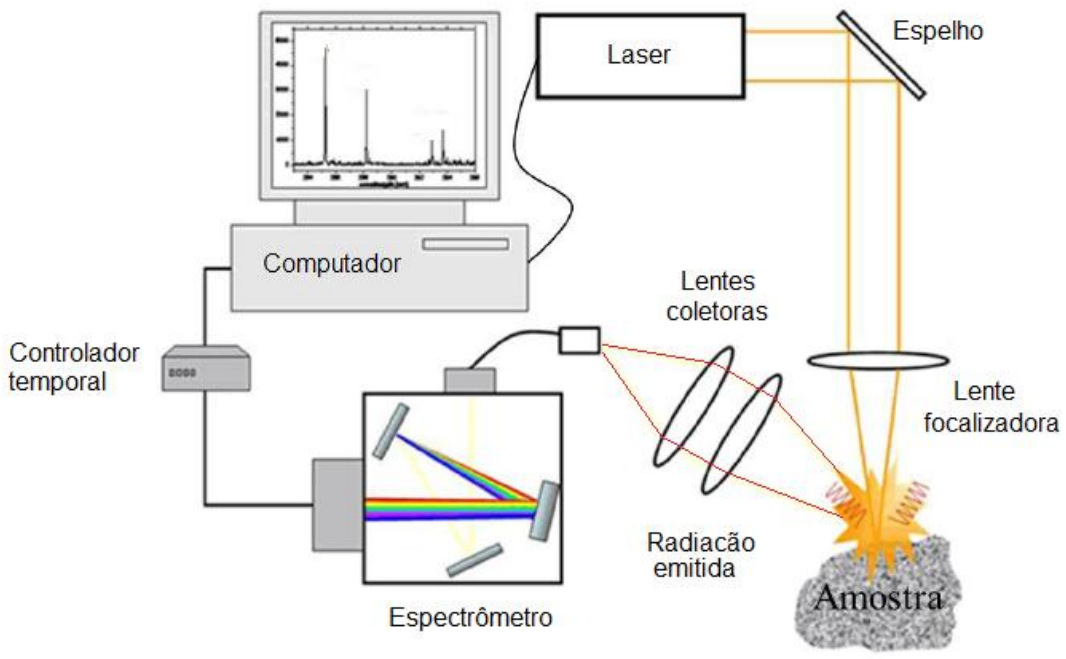

Figura 6 - Configuração geral de um equipamento LIBS mostrando os principais componentes. Fonte: Adaptado de Ranulfi 2019.

3.7. Análise das Componentes Principais – PCA

A análise das componentes principais, PCA (acrônimo do inglês Principal Component Analysis), é um método estatístico multivariado, que é utilizado para a compressão de dados sem perda de informações relevantes. A transformação é desenvolvida de maneira que os conjuntos de dados originais possam ser representados por um número reduzido de novas variáveis, chamadas componentes principais, que são combinações lineares das variáveis originais. A PCA decompõe a matriz de dados originais em uma soma de matrizes, produtos de vetores chamados scores (relacionam as componentes principais com as amostras (observações)) e loadings (relacionam as componentes principais com as variáveis), na qual através dos gráficos de scores e loadings permitem avaliar a influência de cada variável em cada amostra, encontrando similaridades ou diferenças nos dados. A primeira componente principal, chamada de PC1, é definida na direção da máxima variância do conjunto de dados. A segunda componente principal, PC2, é definida na direção que descreve a máxima variância no espaço da PC1, de forma que cada componente principal (PC1, PC2, PC3, PC4,...) é responsável pela fração sucessiva de variância de dados, consistindo em um sistema de coordenadas ortogonais entre si e, portanto, não correlacionadas. (FERRARINI, 2004)

A Análise de Componentes Principais é utilizada para identificar relações entre características extraídas de dados, e usada para reduzir a dimensionalidade do parâmetro espaço, a partir de combinações lineares das variáveis originais de maneira que retenham as informações essenciais e possam ajudar significativamente na interpretação e análise dos dados. (VASCONCELOS).

3.7.1. Procedimento Matemático para Obtenção das Componentes Principais

Para a obtenção dos valores das componentes principais, primeiramente se deve obter a matriz de dados originais, formada a partir dos valores das amostras. A matriz de dados originais terá dimensão m x n, onde m e n representam, respectivamente, o número de linhas (observações) e o número de colunas (variáveis). Cada linha da matriz de dados originais representa uma observação dos dados amostrais, enquanto cada coluna representa uma variável dos dados amostrais.

Após a obtenção da matriz de dados originais, se calcula a matriz covariância desses dados originais, a partir das eq. 6 ou eq. 7, na qual tem a função de relacionar a interdependência numérica entre duas variáveis aleatórias, informando quais variáveis são redundantes para a redução da dimensão do parâmetro espaço no calculo da PCA. A matriz covariância é uma matriz simétrica e positiva, com dimensão n x n, sendo n o número de colunas (variáveis) da matriz de dados originais.

A covariância é calculada sempre entre duas dimensões (variáveis), sendo que calcular a covariância entre uma dimensão e ela mesma resulta na sua variância. Para o cálculo da matriz covariância, entre duas dimensões (X e Y), de uma dada matriz de dados originais (A), a covariância pode ser calculada como:

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x) * (y_i - \mu_y) \qquad (6)$$

, onde n representa o número de linhas (observações) da matriz de dados originais, x_i e y_i representam cada um dos elementos da variável (coluna), na i-ésima posição, μ_x e μ_y representam a média aritmética da variável (média aritmética por coluna).

$$\mathrm{Cov(A)} = \frac{1}{n-1} * ([a]^{\mathrm{T}} * [a]) \qquad (7)$$

$$\mathrm{Cov(A)} = \frac{1}{n-1} * \left(\left[\mathrm{A_{ij}} - \mu_{\mathrm{j}}\right]^{\mathrm{T}} * \left[\mathrm{A_{ij}} - \mu_{\mathrm{j}}\right]\right) \qquad (8)$$

, onde n representa o número de linhas (observações) da matriz de dados originais, $a = \mathrm{A_{ij}} - \mu_{\mathrm{j}}$, sendo $\mathrm{A_{ij}}$ um elemento da variável (coluna) na linha i coluna j, e μ_{j} é a média aritmética da variável (média por coluna), e o índica T indica matriz transposta.

A matriz covariância (Cx) para dados com mais de duas dimensões (colunas), será formada pela covariância entre cada par de dimensões (colunas). Se a matriz de dados originais possuir 3 dimensões (x, y e z), terá o seguinte formato:

$$\mathrm{C_X} = \begin{pmatrix} cov(\mathrm{x},\mathrm{x}) & cov(\mathrm{x},\mathrm{y}) & cov(\mathrm{x},\mathrm{z}) \\ cov(\mathrm{y},\mathrm{x}) & cov(\mathrm{y},\mathrm{y}) & cov(\mathrm{y},\mathrm{z}) \\ cov(\mathrm{z},\mathrm{x}) & cov(\mathrm{z},\mathrm{y}) & cov(\mathrm{z},\mathrm{z}) \end{pmatrix} \qquad [1]$$

A diagonal principal da matriz contém as variâncias, e as demais posições as correlações entre as direções, de modo que é sempre possível encontrar um conjunto de autovalores ortonormais. (VASCONCELOS) A matriz covariância é uma matriz simétrica e positiva, com dimensão n x n, sendo n o número de colunas (variáveis) da matriz de dados originais.

Após calculada a matriz covariância dos dados originais, se deve calcular os autovalores e autovetores da matriz covariância, e depois arranjar esses dados em uma matriz cuja as colunas são formadas a partir dos autovetores da matriz de covariância.

Um autovetor v é um autovetor de uma matriz quadrada M, quando o produto de Mv resulta em um múltiplo de v, chamado de autovalor, representado pela letra λ. Uma das propriedades dos autovetores é que eles são perpendiculares entre si, na qual torna possível expressar os dados em

termos de autovetores em vez de termos dos eixos x, y e z. Para a obtenção dos autovalores a seguinte expressão é usada:

$$\det(M - \lambda I) = 0 \qquad (9)$$

, onde I é a matriz identidades, M é a matriz de dados a ser utilizada (matriz covariância) e λ será os valores dos autovalores calculados.

Os autovetores associados aos autovalores serão dados pela expressão:

$$(M - \lambda \mathrm{I})\, v = 0 \qquad (10)$$

Se uma matriz *n* x *n* tem *n* autovalores linearmente independentes, então ela é diagonalizável. Se uma matriz é diagonalizável então ela tem *n* autovalores linearmente independentes que serão os seus elementos da diagonal principal. Com os autovetores obtidos, se monta uma matriz coluna com esses autovetores ($P = [v_1\ v_2\ v_3]$). Com a seguinte expressão abaixo, a matriz diagonal da matriz de covariância ($C_{xDiagonal}$) é obtida:

$$\mathrm{C_{xDiagonal}} = \mathrm{P^{-1}.M.P} \qquad (11)$$

Onde P^{-1} é a matriz inversa da matriz formada pelos autovetores, M é a matriz de dados utilizadas para se obter os autovalores e os autovetores e P é a matriz formada pelos autovetores.

No processo de diagonalização, não existem uma ordem preferencial na qual se arranjam os autovalores. Portanto a matriz obtida através da eq. 11 é uma matriz diagonal, na qual seus autovetores não estão ordenados de modo crescente ou decrescente, estão aleatórios. A transformada de *Hotelling* é utilizada para se obter uma matriz diagonal na qual os valores dos autovetores estarão ordenados de forma decrescente, o que auxilia no processo de interpretação dos dados da PCA. A expressão abaixo define a transformada de *Hotelling*

$$y = A(x - \mu_x) \qquad (12)$$

, onde *A* é uma matriz cujas colunas são os autovetores da matriz de covariância C_x, mas ordenados de forma decrescente, x são os valores da variável de *A* e μ_x é a média aritmética por variável de *A.*

Com a expressão abaixo, a matriz C_y diagonalizada é obtida através de *A* e C_x, na qual os autovetores estarão ordenados de forma decrescente e os valores fora da diagonal principal terão valor nulo.

$$\mathrm{C_y} = A.\mathrm{C_x}.\mathrm{A^T} \qquad (13)$$

Em reconhecimento de padrões, é desejável dispor uma representação compacta e com bom poder de discriminação de classes de padrões, na qual não haja redundância entre as diferentes características dos padrões, ou seja, que não haja covariância entre os vetores da base do espaço de característica. Realizando a mudança de base e diagonalizando a matriz de covariância, as variáveis padrões podem ser representadas em termos dessa nova base do espaço de características que não possuem correlação entre sí. E como na PCA os autovalores da matriz de covariância são iguais á variância das características transformadas, quando um autovetor possuir um autovalor grande, significa que esse fica em uma direção em que há uma grande variância dos padrões. A importância disso está no fato de que, em geral, é mais fácil distinguir padrões usando uma base em que seus vetores apontam para a direção da maior variância dos dados, além de não serem correlacionados entre si. (VASCONCELOS) Em resumo, há uma redução da dimensionalidade do conjunto, facilitando trabalhar com os dados.

3.8. Análise Multivariada de Dados

A análise multivariada de dados é definida e caracterizada pela análise simultânea de múltiplas variáveis presentes em uma única relação ou em um conjunto de relações, na qual os “indivíduos” possuem características diferentes. Auxilia na compreensão de comportamentos complexos de grupos de dados em ambientes afins e pode, dependendo da aplicação, acrescentar informações potencialmente úteis a eles, além de permitir e preservar as

correlações naturais entre as múltiplas influências de comportamento, sem isolar qualquer indivíduo ou variável. (SILVA, 2015)

Uma das mais importantes áreas da análise multivariada de dados é denominada análise discriminante de dados, que pode ser descrita como uma técnica multivariada de interdependência entre objetos. Essa técnica é baseada na associação entre os dados e um conjunto de características descritivas, ou atributos especificados pelo manuseador dos dados, que utiliza a análise discriminante para prover a classificação. Um dos objetivos da análise discriminante também é a redução dimensional da classificação dos objetos ou dados, em um conjunto de atributos representativos, na qual passam a pertencer a determinada classe, com a qual os outros participantes partilham características em comum. (SILVA, 2015)

3.9. Métodos de Classificação ou Reconhecimento Supervisionado de Padrões

As técnicas de reconhecimento de padrões são usadas para identificar as semelhanças e diferenças contidas em amostras, comparando-as entre si. Os métodos supervisionados são métodos na qual são fornecidas para um modelo quais amostras são semelhantes e quais são diferentes, para que possam ser encontrados os critérios de classificação. O modelo utiliza amostras pré-classificadas em determinadas categorias, chamadas de amostras de treino, para construir os critérios de classificação, e a performance do modelo é posteriormente avaliada com base nas amostras de teste.

Na classificação de padrões, o algoritmo responsável pela classificação deve primeiramente “aprender” como classificar os padrões do problema ao qual se deseja obter resposta, treinando o algoritmo de forma a torná-lo capaz de, após o treinamento, classificar um padrão desconhecido dentre uma das classes existentes. A fase de treinamento possui um peso muito grande no desempenho do algoritmo como um todo, pois durante essa fase, o algoritmo deve armazenar protótipos, que podem ser padrões de treinamento, que sejam capazes de generalizar ao máximo a classificação feita para os padrões

desconhecidos. Após a fase de treinamento, o algoritmo deve passar pela fase de teste. Essa fase é responsável por medir o desempenho da classificação feita pelo algoritmo, na qual ao final da fase de teste, devem ser apresentadas as porcentagens de padrões incorretamente classificados para se analisar o desempenho do classificador para o problema em questão. (BEZERRA, 2006)

As técnicas de classificação são categorizadas das seguintes formas:

- Paramétricas: Técnicas paramétricas consideram que as variáveis tenham uma distribuição normal, devendo satisfazer os requisitos para o número de graus de liberdade (amostra/variável), e a homogeneidade da matriz de variância – covariância.

- Não Paramétricas: Técnicas não paramétricas deixam que os próprios dados definam sua própria estrutura e serem capazes de encontrar esta estrutura explícita ou implicitamente.

- Discriminantes: As amostras pertencem a uma e somente uma classe, na qual existe uma fronteira entre as classes.

- Modelativas: As amostras podem pertencer a nenhuma classe, uma classe ou várias classes ao mesmo tempo, na qual as fronteiras das classes podem se superpor com fronteiras de outras classes.

- Probabilísticas: Estimam um grau de confiança da classificação.

- Determinísticas: Não estimam um grau de certeza de uma classificação.

3.9.1. Métodos de Validação dos Dados

Para determinar a precisão do método supervisionado de reconhecimento de padrões, alguns métodos de avaliação são normalmente utilizados para avaliar o seu desempenho, sendo os mais conhecidos:

- *Holdout Validation:* Uma porcentagem dos dados é selecionada para ser usada como dados de teste. O modelo é treinado com o restante dos dados

não utilizados para teste e a performance do modelo é avaliada com os dados de teste. Então, esse tipo de avaliação consiste em um modelo que utiliza somente uma porção dos dados para determinar a sua acurácia. (SILVA, 2015)

- *Cross-Validation:* Informados os dados a serem usados para a construção do modelo, esses são divididos em *k-fold* divisões iguais, na qual 1/*k* dos dados é utilizado para teste, chamados de *in-fold*, enquanto (*k* – 1)/*k* é utilizado para treinamento, chamados de *out-of-fold.* Após fazer esse treinamento, o modelo é testado com o *in-fold* e sua precisão armazenada. Em seguida, um outro conjunto de dados do *out-of-fold* é usado para teste, e os dados do *in-fold* agora é usado para treinamento. A precisão desse segundo teste é armazenada. Isso se repete até que todos os *out-of-fold* sejam usados um a um com *in-fold*, e a acurácia total do modelo é feita através da média desses *k* testes. (SILVA, 2015)

As utilizações dos métodos de avaliação são importantes para o desenvolvimento dos classificadores, pois eles evitam o problema de sobreajuste (*overfitting*). O sobreajuste ocorre quando um classificador se adapta aos documentos de treino, e fica muito eficaz em classificar e distinguir os documentos usados no treinamento mas não consegue distinguir dados novos externos, o que pode reduzir a taxa de acerto na classificação dos novos dados. (BEZERRA, 2006)

3.10. Classificadores

3.10.1. Método k-nearest neighbor (k-NN)

O método *k* vizinhos próximos, acrônimo do inglês *k*-nearest neighbor (k-NN), é um método não paramétrico usados para classificação e regressão. Em ambos os casos, a entrada consiste no conjunto de treino e de testes, além da determinação do *k* mais próximo em um espaço característico. O conjunto de dados de treinamentos e testes é utilizado para a construção do algoritmo *k*-NN para aquela determinada situação. Na classificação, a amostra é classificada

baseado em seus *k* vizinhos mais próximos mais frequentes, e também baseados na função de distância escolhida para calcular os *k*-vizinhos mais próximos. Assim, os dados que a partir de uma determinada restrição sejam os mais similares a ele, será atribuída à classe do elemento desconhecido. *K* é um número inteiro positivo, normalmente pequeno, que define o número de vizinhos próximos que o algoritmo pode usar para classificar uma amostra. Se o valor de *k* é igual a 1, a amostra é simplesmente associada á classe daquele único vizinho próximo. A saída do método depende se ele será empregado em regressão ou em classificação. (SILVA, 2015) A figura 7 abaixo demonstra o método *kNN.*

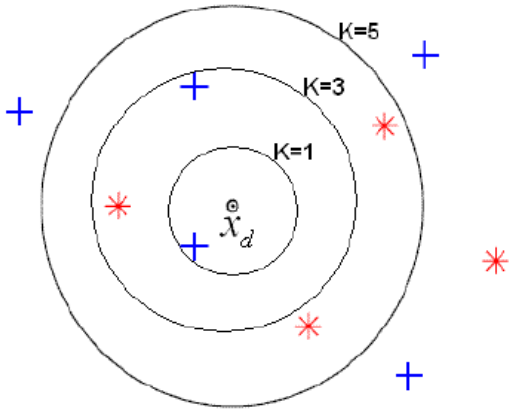

Figura 7 - Representação do método kNN. Fonte: Modificado de Bezerra, 2006.

Para estimar a classe de um novo padrão de amostras *X* que não pertençam ao conjunto de treinamento, o algoritmo *k*-NN calcula a distância para os *k*-vizinhos mais próximos a *X*, e classifica-o como sendo da classe que aparece com maior frequência dentre os seus vizinhos. Durante a fase de classificação do *k*-NN, algumas vezes ocorre um problema, onde, dado um padrão de teste *X*, os seus *k*-vizinhos mais próximos são de uma mesma classe e o algoritmo não consegue decidir com qual classes dos *k*-vizinhos ele deve comparar o padrão *X*. Para resolver essa situação, o padrão que teve esse problema, será rodado de forma recursiva pelo algoritmo, o qual agora usará apenas o *k-1* vizinhos para o cálculo, até que uma das classes dos *k*-vizinhos apareça com maior frequência em relação ás demais. (BEZERRA, 2006) O desempenho dos algoritmos de treinamento baseados no *k*-NN podem

ser medidos por três variáveis que influenciam o seu desempenho, que são o número de protótipos armazenados durante a fase de treinamento, o tempo computacional necessário para classificar um padrão desconhecido e a taxa de erro do conjunto de teste.

O método *k*-NN é uma método *lazy learning*, ou seja, um "aprendiz preguiçoso", que simplesmente armazena as amostras de treino e realiza uma única etapa para classificar essas amostras.

Para calcular a distância entre uma amostra de teste e os seus k vizinhos mais próximos, normalmente é necessário calcular todas as distâncias entre as amostras de teste e de treino. Considere X = (x_1, x_2, x_3, ..) e Y = (y_1, y_2, y_3, ...) como dois pontos, as funções de distâncias mais comuns no cálculo entre a distância entre dois pontos usando o *k*-NN são:

- Distância métrica Euclidiana entre X e Y é dada por

$$\mathrm{d}(\mathrm{x},\mathrm{y}) = \sqrt{(\mathrm{x}_1 - \mathrm{y}_1)^2 + (\mathrm{x}_2 - \mathrm{y}_2)^2 + \cdots + (\mathrm{x}_\mathrm{n} - \mathrm{y}_\mathrm{n})^2} \qquad (14)$$

- Distância métrica dos cossenos entre dois vetores **a** e **b** é dada por

$$cos\Theta = \frac{\vec{a}.\vec{b}}{\|\vec{a}\| \left\|\vec{b}\right\|} \qquad (15)$$

A medida da distância métrica dos cossenos é usada para calcular a similaridade entre dois vetores medindo o cosseno do ângulo entre eles. Esse tipo de medição de distância métrica é a medida da orientação e não da magnitude, mostrando como os dois vetores estão relacionados olhando os ângulos entre eles ao invés da magnitude. Esse tipo de distância com o valor pequeno de ângulo indica que os dois estão na mesma direção, e mesmo que possa haver uma grande distância Euclidiana entre pontos de dois vetores, se esse vetores estão na mesma direção, o ângulo entre eles será menor, indicando que possuem alta similaridade entre eles.

- Distância métrica baseada no peso w_i entre a amostra a ser classificada x_q e as amostras de treinamento x_i, é dada por

$$w_i = \frac{1}{d(x_q, x_i)^2} \qquad (16)$$

Um refinamento da classificação do método *k*-NN é dar pesos de contribuição de cada *k* vizinhos de acordo com a distância entre o ponto a ser classificado x_q e os pontos de treinamento x_i, dando pesos maiores para os vizinhos mais próximos.

- Distância métrica cúbica entre 2 vetores *u* e v, de *n*-dimensões é dado por

$$\sqrt[3]{\sum_{i=1}^{n} |u_i - v_i|^3} \qquad (17)$$

3.10.2. Método Support Vector Machine (SVM)

Máquinas de Vetores de Suporte, acrônimo do inglês *Support Vector Machine* (SVM), é um classificador que se baseia na análise estatísticas dos dados para construir um algoritmo que diferencie as várias classes a serem classificadas. Os classificadores SVM classificam dados encontrando o melhor hiperplano que separam os dados limites que separam uma classe da outra. O melhor hiperplano para um classificador SVM é aquele que possui a margem mais larga entre dados limites que separam duas classes. Margem é definida como a máxima largura entre duas "linhas" paralelas ao hiperplano que não possui dados em seu interior. Os vetores de suporte são pontos dos dados que estão mais próximos ao hiperplano que separam as classes, na qual esses pontos estão na fronteira das "linhas" paralelas ao hiperplano, chamadas de linha *gutter*. A figura 8 abaixo ilustra as definições para um SVM, na qual separa dados positivos (+) de dados negativos (-) através de um hiperplano.

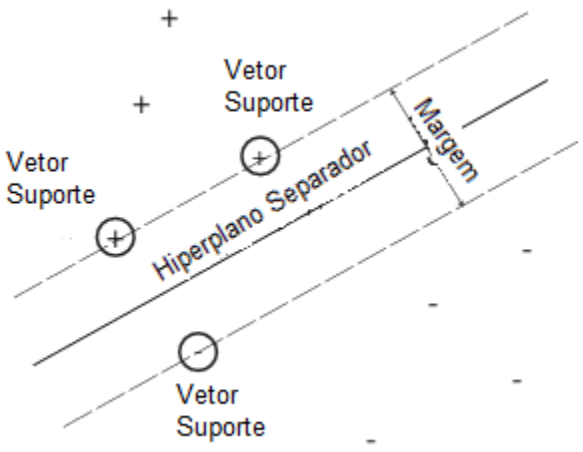

Figura 8 - Hiperplano separando duas classes através do método SVM. Fonte: Mathworks.com

3.10.2.1. SVM Linear

Support Vector Machine lineares com margens rígidas definem fronteiras lineares para dados linearmente separáveis, que possuem apenas duas classes de diferenciação, separando os dados por meio de um hiperplano.

Seja **w** um vetor perpendicular ao hiperplano que separa as duas classes, na qual o comprimento desse vetor não é importante nesse momento, e **u** um vetor desconhecido, localizado entre as duas “linhas” que separam as classes positivas das classes negativas, chamada de linha *gutter*. A equação do hiperplano separador é dada pelo produto escalar entre o vetor **w** e o vetor **u**, na qual o resultado deve ser maior ou igual a uma constante, chamada de constante C.

$$\vec{w}.\vec{u} \geq C \qquad (18)$$

Igualando a zero a equação acima, e introduzindo uma constante *b*, que por enquanto é definida como *b* = *-C,* tem:

$$\vec{w}.\vec{u} + b \geq 0 \qquad (19)$$

A equação acima é uma regra de decisão para construir o algoritmo de máquinas de vetores de suporte, que divide o espaço dos dados em duas

regiões, uma maior que zero e outra menor que zero, na qual quando for maior que zero indica que o vetor desconhecido **u** está entre a linha do hiperplano separador e a "linha" *gutter* da classe positiva, e na qual quando for menor que zero indica que o vetor **u** está entre a linha do hiperplano separador e a "linha" *gutter* da classe negativa. Da equação 19 acima, **w** e *b* ainda não possuem valores específicos, pois mesmo sabendo que **w** deve ser perpendicular ao hiperplano separador, existem vários vetores que podem ser perpendiculares ao hiperplano separador dependendo do seu comprimento. Para os dados positivos, introduzimos o vetor $\mathbf{x}_+$, e para os dados negativos introduzimos o vetor **x.,** na qual a eq. 19 deve satisfazer as seguintes condições:

$$\vec{w} \cdot \overrightarrow{x_+} + b \geq 1 \qquad (20)$$

$$\vec{w} \cdot \overrightarrow{x_-} + b \leq -1 \qquad (21)$$

Para facilitar a manipulação das equações, uma variável Y_i é introduzida, na qual Y_i = +1 para dados positivos, e Y_i = -1 para dados negativos, na qual essa variável vai indicar se um dado é positivo ou negativo. Então, multiplicando o lado esquerdo das eq. 20 e 21 por Y_i, e a equação 21 por -1, tem:

$$Y_i(\vec{w} \cdot \overrightarrow{x_i} + b) \geq 1 \qquad (22)$$

$$Y_i(\vec{w} \cdot \overrightarrow{x_i} + b) - 1 \geq 0 \qquad (23)$$

A eq. 23 implica que, para os dados que estão na linha *gutter*, tanto sendo a linha *gutter* das classes positivas quanto das classes negativas, a equação 23 será igual a zero.

$$Y_i(\vec{w} \cdot \overrightarrow{x_i} + b) - 1 = 0 \qquad (24)$$

Então, para os outros dados que não estão nessa linha *gutter*, terão valor maior que 1 para dados positivos e valor menor que 1 para dados negativos.

Para o calculo da distância entre as linhas *gutter* que separam as duas classes, é primeiramente calculado a diferença entre os vetores $\mathbf{x}_+$ dos dados positivos, e o vetor $\mathbf{x}_-$ dos dados negativos. Com esse resultado, é feito o

produto escalar com um vetor unitário que seja perpendicular as linhas *gutter* paralelas ao hiperplano separador, e como o vetor **w** é um vetor perpendicular ao hiperplano separador, para obter seu valor unitário divide-se o vetor **w** pela sua magnitude, da forma que a eq. Abaixo é obtida:

$$\text{margem} = (\mathbf{x}_+ - \mathbf{x}_-) \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|} \qquad (25)$$

Da eq. 24, para dados positivos, Y_i é igual a +1, de forma que

$$+1 \cdot (\vec{w} \cdot \overrightarrow{x_+} + b) - 1 = 0$$

$$(\vec{w} \cdot \overrightarrow{x_+}) = 1 - b \qquad (26)$$

Da eq. 24 para dados negativos, Y_i é igual a -1, de forma que

$$-1 \cdot (\vec{w} \cdot \overrightarrow{x_-} + b) - 1 = 0$$

$$(-\vec{w} \cdot \overrightarrow{x_-}) - b) - 1 = 0$$

$$-(\vec{w} \cdot \overrightarrow{x_-}) = 1 + b \qquad (27)$$

Substituindo na eq. 25 tem-se

$$\text{margem} = \big((1 - b) + (1 + b)\big) \cdot \frac{1}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|} \qquad (28)$$

O valor obtido na eq. 28 é a distância entre os hiperplanos *gutter* paralelos ao hiperplano separador.

Para se obter a maior distância entre os hiperplanos que separam os dados das duas diferentes classes, a maximização da eq. 28 é necessária, o que significa que se é necessário minimizar a magnitude de **w**. A minimização de **w** é dado pela seguinte equação:

$$\frac{1}{2} \cdot \|\mathbf{w}\|^2 \qquad (29)$$

Com as restrições de $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) - 1 >= 0$, i = 1, ... , *n.*

As restrições são impostas de maneira a assegurar que não haja dados de treinamento entre as margens de separação das classes. Por esse motivo, a SVM obtida possui também a nomenclatura de SVM com margens rígidas. O problema de otimização obtido é quadrático, cuja solução possui uma ampla e estabelecida teoria matemática. Como a função objetivo sendo minimizada é convexa e os pontos que satisfazem as restrições formam um conjunto convexo, esse problema possui um único mínimo global. Problemas desse tipo podem ser solucionados com a introdução de uma função Lagrangiana, que engloba as restrições á função objetivo, associadas a parâmetros denominados multiplicadores de Lagrange α_i. (LORENA et al., 2007) A eq. 30 abaixo demonstra essa função Lagrangiana:

$$\mathrm{L} = \frac{1}{2} . \|\mathbf{w}\|^2 \; - \; \sum_{\mathrm{i}=1}^{n} \propto_{\mathrm{i}} \; [\mathrm{y_i} . (\mathbf{w} . \mathbf{x_i} + b) - 1] \qquad (30)$$

A função Lagrangiana deve minimizar **w** e *b*, o que implica:

$$\frac{\partial L}{\partial b} = 0 \quad e \quad \frac{\partial L}{\partial \mathbf{w}} = 0 \qquad (31)$$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{\mathrm{i}=1}^{n} \propto_i . \; \mathrm{y_i} . \mathbf{x_i} = 0 \; \therefore \mathbf{w} = \sum_{\mathrm{i}=1}^{n} \propto_{\mathrm{i}} . \mathrm{y_i} . \mathbf{x_i} \qquad (32)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \propto_{\mathrm{i}} . \mathrm{y_i} = 0 \qquad (33)$$

Substituindo as eq. 32 e 33 na eq. 30, obtém

$$\mathrm{L} = \frac{1}{2} . \left(\sum_{\mathrm{i}=1}^{n} \propto_{\mathrm{i}} . \mathrm{y_i} . \mathbf{x_i} \right) . \left(\sum_{\mathrm{j}=1}^{n} \propto_{\mathrm{j}} . \mathrm{y_j} . \mathbf{x}_j \right) - \sum_{\mathrm{i}=1}^{n} \propto_{\mathrm{i}} . \mathrm{y_i} . \mathbf{x_i} . \left(\sum_{\mathrm{j}=1}^{n} \propto_{\mathrm{j}} . \mathrm{y_j} . \mathbf{x_j} \right) - \sum_{\mathrm{i}=1}^{n} \propto_{\mathrm{i}} . \mathrm{y_i} . b + \sum_{\mathrm{i}=1}^{n} \propto_{\mathrm{i}} \quad (34)$$

$$\mathrm{L} = \sum_{\mathrm{i}=1}^{n} \propto_{\mathrm{i}} \; - \; \frac{1}{2} \sum_{\mathrm{i,j}=1}^{n} \propto_{\mathrm{i}} . \propto_{\mathrm{j}} . \mathrm{y_i} . \mathrm{y_j} . \mathbf{x_i} . \mathbf{x_j} \qquad (35)$$

Substituindo a eq. 32 na eq. 19, que representa a regra de decisão, tem:

$$\sum_{i=1}^{n} \propto_i . y_i . \mathbf{x}_i . \mathbf{u} + b \geq 0 \qquad (36)$$

A eq. acima indica que a regra de decisão é apenas em função do produto entre o vetor dos dados e o vetor desconhecido, e é a função linear que representa o hiperplano que separa os dados com maior margem. O vetor dos dados que participam da regra de decisão são chamados de vetores de suporte, pois somente eles participam na determinação da equação do hiperplano separador e podem ser considerados os dados mais informativos do conjunto de treinamento. (LORENA et al., 2007)

3.10.2.2. SVMs Não Lineares

Alguns dados de conjuntos de classes podem possuir características que o tornam não linearmente separáveis por um hiperplano, como mostra a figura 9 abaixo.

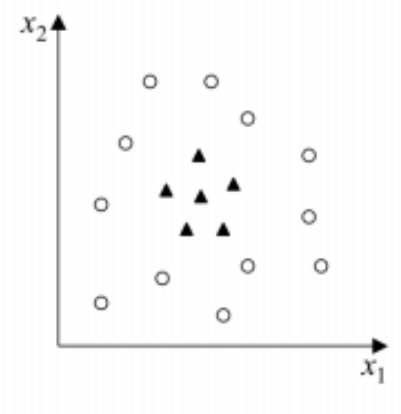

Figura 9 - Exemplo de dados não linearmente separáveis. Fonte: Internet.

As SVMs lidam com problemas não lineares mapeando o conjunto de treinamento de seu espaço original, referenciado como entradas, para um novo espaço de maior dimensão, denominado espaço de características. Dado um conjunto de dados não linear no espaço de entradas *X*, esse conjunto de dados

é transformado em um espaço de características na qual, com alta probabilidade, os dados são linearmente separáveis. Para isso, condições de que a transformação deva ser não linear e que a dimensão do espaço característico seja suficientemente alta devem ser satisfeitas.

Assim, mapea-se inicialmente os dados para um espaço de maior dimensão, e depois aplica-se a SVM linear sobre este espaço para encontrar um hiperplano capaz de realizar a separação. Como a dimensão desses espaços de características podem ser muito altas, a utilização de funções denominadas Kernel (*K*) são de grande importância, pois essas funções levam apenas em consideração o produto escalar entre dois pontos, do espaço de entradas, no espaço de característica. É comum empregar o uso de funções Kernel sem conhecer o mapeamento que é gerado implicitamente, o que mostra a simplicidade de seu uso e sua capacidade de representar espaços abstratos. Alguns dos Kernels mais utilizados na prática são os polinomiais, os Gaussianos ou RBF (*Radial-Basis Function*). (LORENA et al., 2007)

Em códigos de classificação utilizando o algoritmo *Support Vector Machine*, algumas funções avançadas para classificação de dados multi-classes, além das funções Kernel, podem ser usados. Essas funções, ou técnicas, são utilizadas nos dados de treinamento para otimizar os parâmetros de classificação. Uma dessas técnicas é chamada *one-against-one*, na qual para um problema de multi-classes, o algoritmo realiza o treinamento entre todas as combinações binárias (classes 2 a 2), de forma que cada combinação entre elas irá retornar padrões de treinamento que serão armazenados pelo algoritmo e que depois irá “construir” o código de treinamento com base na combinação de todos esses padrões de treinamento. (BEZERRA, 2006) Esse tipo de técnicas aprende a distinguir uma classe da outra. A figura 10 abaixo ilustra o funcionamento do *one-against-one.*

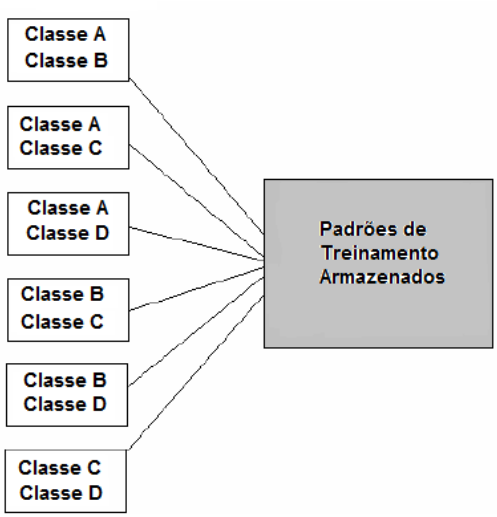

Figura 10 - Técnica one-against-one para 4 classes distintas. Fonte: Modificado de Bezerra, 2006.

A outra técnica utilizada é chamada *one-against-all*, na qual nessa técnica os dados multi-classe serão treinados um de cada vez, na qual a classe que está sendo treinada é dada um valor positivo e o restante das classes é dado um valor negativo. É feito o treinamento com essas classes e seus padrões armazenados. Em seguida, outra classe que antes estava classificada como negativa passa a ser classificada como positiva, enquanto a classe que era positiva agora é classificada como negativa. Isso ocorre até que todos os dados sejam treinados dessa forma e seus padrões armazenados. Em seguida, um padrão de treinamento é montado para classificar novos dados. Esse tipo de técnica aprende a distinguir uma classe de todas as outras.

4. METODOLOGIA

A metodologia proposta é dividida em duas partes principais. A primeira etapa consiste no preparo das amostras de grãos de milho, e obtenção dos espectros ópticos utilizando as técnicas ópticas FTIR e LIBS. A segunda etapa consiste na análise qualitativa dos dados e, obtenção de rotina computacional para diferenciação de amostras transgênicas de não transgênicas, utilizando um algoritmo de reconhecimento de padrões supervisionado.

Seis amostras diferentes de grãos de milho foram utilizadas como material de pesquisa, sendo 2 (duas) amostras não transgênicas e 4 (quatro) amostras transgênicas. As amostras de milho foram nomeadas de acordo com a nomenclatura comercialmente definida, como apresenta a tabela 02 abaixo.

Tabela 02 – Nomenclatura e classificação das espécies de milho usadas

Nomenclatura	Classificação
Impacto C	NÃO TRANSGÊNICO
Fórmula C	NÃO TRANSGÊNICO
Fórmula VIP3	TRANSGÊNICO
Impacto VIP3	TRANSGÊNICO
Fórmula VIP	TRANSGÊNICO
Fórmula TL	TRANSGÊNICO

4.1. Primeira Etapa

As amostras de milho foram fornecidas através de uma colaboração com a Universidade Federal de Uberlândia (UFU), localizada no estado de Minas Gerais. A coleta das amostras foi feita na fazenda experimental da UFU, englobando variedades transgênicas e não transgênicas, como apresenta a tabela 02 acima. Após a coleta, as amostras passaram por um processo de limpeza manual, secagem em estufa a 65ºC para a retirada de umidade e moagem. No processo de moagem manual dos grãos de milho, se utilizou almofariz e pistilo, na qual posteriormente as amostras foram peneiradas para

homogeneização. As amostras foram peneiradas, por peneiras padronizadas ABNT, de medida *mesh* 100, que possui abertura de malha de 150 µm. Após essa etapa, as amostras foram embaladas em recipientes contendo 20 g de cada amostra, e posteriormente levadas ao grupo de pesquisa para a realização das medidas espectroscópicas ópticas.

4.1.1. Medidas da Espectroscopia de Absorção no Infravermelho por Transformada de Fourirer – FTIR

O equipamento utilizado foi um espectrômetro comercial modelo Spectrum(R) *100 Series FT-IR,* da fabricante PerkinElmer(R), utilizando um intervalo de medida de 4000 – 600 cm^{-1}. Foram feitas medidas no modo de transmissão com acessório ATR, onde as amostras foram colocadas em cima de um cristal e pressionadas de tal modo que se obteve o máximo contato com o cristal, com resolução de 4 cm^{-1} e 15 varreduras. As medidas foram realizadas na Universidade Federal de Mato Grosso do Sul.

4.1.2. Medidas da Espectroscopia de Emissão Óptica com Plasma Induzido por Laser – LIBS

Para a caracterização multielementar e simultânea dos elementos presentes nas amostras, foi utilizado o sistema LIBS Pulso Duplo (PD) pertencente a Embrapa Instrumentação (Nicolodelli et al., 2015). O sistema é composto por dois lasers com diferentes comprimentos de onda, sendo um no infravermelho (1064 nm) e outro no verde (532 nm). O pulso de 1064 nm possui energia máxima de 50 mJ, largura de 6 ns e é gerado pelo laser Q-switch de Nd:YAG Ultra, da fabricante Quantel. Já o pulso de 532 nm possui energia máxima de 180 mJ, largura de 4 ns e é gerado pelo laser Q-switch de Nd:YAG Brillant, da fabricante Quantel acoplado a um módulo de geração de segundo harmônico. A geometria utilizada pode ser colinear ou ortogonal, ou seja, ambos os feixes foram alinhados e direcionados até a amostra, atingindo-a de forma sobreposta ou perpendicularmente, com um atraso entre eles. Ele

também pode operar em modo convencional (pulso único), que foi o modo adotado nesse trabalho. Para a detecção e seleção de comprimentos de onda foi utilizado um sistema ARYELLE 400-Butterfly. Esse espectrômetro opera em duas faixas espectrais de 175-330 nm (Faixa UV)/275-750 nm (Faixa UV-VIS), com resolução espectral de 13-24/29-80 pm, respectivamente, e possui uma CCD intensificada (ICCD) com 1024 x 1024 pixels.

Alguns componentes secundários foram necessários para aquisição dos espectros. Os feixes provenientes dos dois lasers foram direcionados ao alvo (amostra) através de espelhos dicroicos para os apropriados comprimentos de onda (532/1064 nm). Utilizamos lentes (distancia focal de 100 mm) com filmes antirreflexos (532/1064 nm), para o melhor aproveitamento óptico da energia do laser. Estas lentes são utilizadas para focalizar o feixe do laser sobre a superfície da amostra. Duas lentes de sílica fundida foram posicionadas entre a amostra e a ponta da fibra para uma eficiente coleta do plasma emitido. O suporte para amostra (pastilha) foi posicionado em uma mesa x-y micro-controlada para uma fácil e rápida varredura do feixe do laser sobre ela. Para a sincronização temporal do experimento (tempo de atraso entre pulsos e atraso da detecção) foi utilizado um gerador de atraso de pulso com oito canais da fabricante Quantum Composers, modelo 9618. Com esse gerador de atraso, ajustamos o tempo entre o pulso e o acionamento do espectrômetro para aquisição do espectro LIBS (tempo de atraso).

Neste trabalho, foi utilizado apenas o laser 532 nm com energia de 60 mJ para a excitação da amostra. O tempo de atraso entre os pulsos do laser e a coleta com espectrômetro foi 0,5 µs e o tempo de integração do sinal 10,0 µs. Para obtenção de cada espectro foram acumulados 3 pulsos de laser sobre um mesmo ponto. Foram obtidos um total 60 espectros por amostra.

As amostras de milho foram submetidas a um processo de moagem para redução da heterogeneidade. Após a moagem, as amostras foram submetidas a uma pressão de cinco toneladas para obtenção de pastilhas. Em seguida, ambas as amostras foram caracterizadas pela técnica espectroscópica LIBS. As medidas LIBS foram realizadas em parceria com a Embrapa Instrumentação de São Carlos.

4.2. Segunda Etapa

A segunda etapa consistiu na análise espectral dos resultados ópticos obtidos. Todos os resultados ópticos obtidos, sendo eles Absorção FT-IR e LIBS Faixa UV e LIBS Faixa UV-Vis, são compostos basicamente por uma matriz de dados, relacionando o comprimento de onda a uma intensidade óptica. Dessa maneira se foi feita a análise comparativa qualitativa para elementos específicos e, análise multivariada para diferenciação entre amostras transgênicas e não transgênicas.

4.2.1. Análise Comparativa

Regiões específicas dos espectros foram verificadas, com o objetivo de identificar qual é o fenômeno por trás da transição observada. Para os espectros de absorção FT-IR, os resultados obtidos foram comparados com a literatura existente, para identificar a origem das transições. Para o LIBS, o processo de identificação das transições foi feito com base na comparação das transições de interesse com as que estão catalogadas no site do NIST Database.

4.2.2. Análise Multivariada para Diferenciação

Para as análises multivariadas, se utilizou a Análise das Componentes Principais (PCA), para analisar o potencial de diferenciação entre as amostras, e métodos supervisionados de reconhecimento de padrões, para a obtenção de uma rotina que diferencie as espécies transgênicas de não transgênicas, através dos dados obtidos pelas medidas espectroscópicas ópticas.

Para o treinamento dos Métodos Supervisionados de Reconhecimento de Padrões, uma rotina foi utilizada, na qual os dados ópticos foram ordenados em forma de tabela, da seguinte maneira: as linhas correspondem às

observações, que são as espécies do milho. As colunas correspondem as variáveis, que são os comprimentos de onda. Em cada célula da tabela, foi armazenado o valor da intensidade óptica correspondente ao comprimento de onda incidente, na qual a última célula de cada linha, ou seja, a última coluna da tabela contém a classe na qual os dados são verdadeiramente classificados.

Depois de informado a tabela com os dados de intensidade óptica, foram treinados doze tipos de classificadores, sendo no geral seis classificadores do tipo *k*-vizinhos próximos (kNN) e seis classificadores do tipo Máquinas de Vetores de Suporte (SVM), para verificar qual possui a melhor performance através da observação do valor de sua precisão e de sua matriz de erro.

5. RESULTADOS E DISCUSSÕES

5.1. Absorção no Infravermelho por Transformada de Fourier – FTIR

O espectro da absorção na região do infravermelho compreendeu os números de onda entre 600 á 4000 cm^{-1}. A figura 11 abaixo apresenta o espectro óptico de 17 amostras, sendo 6 de espécies não transgênicas (3 de Impacto C e 3 de Fórmula C) e 11 de espécies transgênicas (3 de Fórmula VIP3, 3 de Impacto VIP3, 3 de Fórmula VIP e 2 de Fórmula TL).

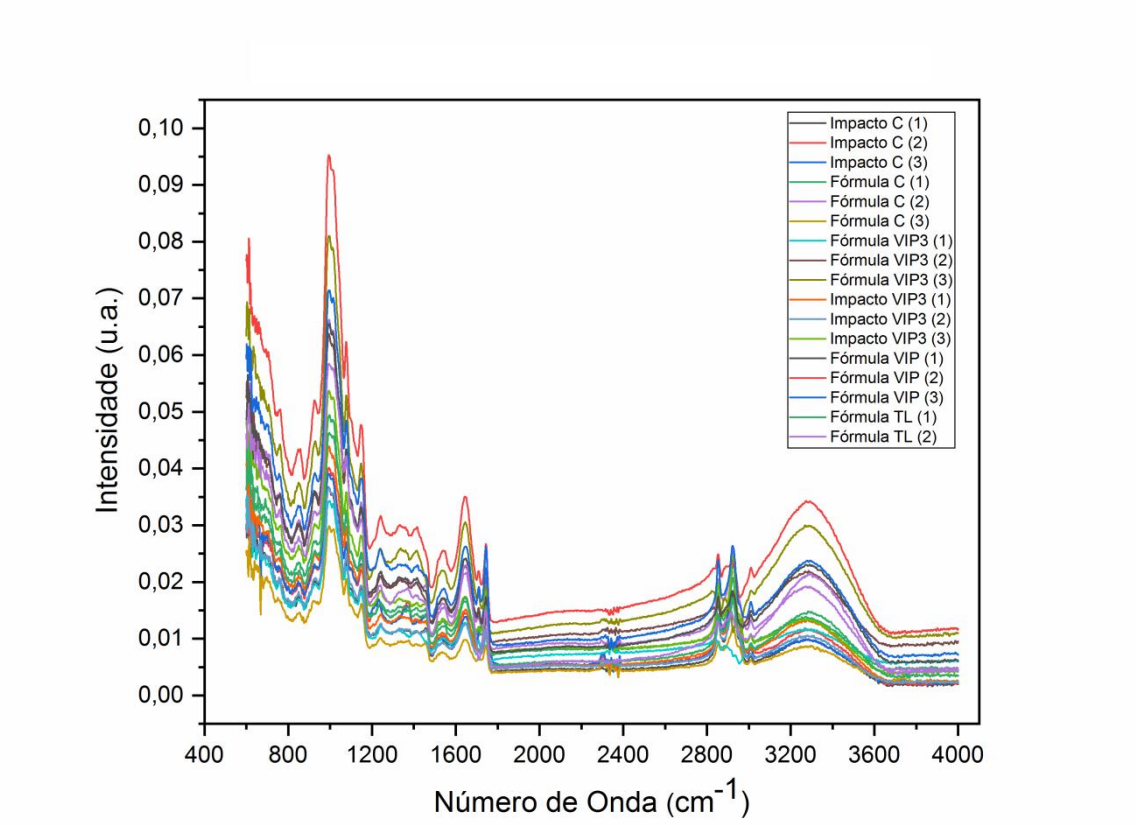

Figura 11 - Espectro óptico FTIR das espécies de milho transgênicas e não transgênicas. Fonte: Própria.

No espectro óptico da figura 11, se é visto que todas as espécies de milho possuem um espectro óptico bastante similar, mas que diferem para valores de intensidade óptica. O comportamento do espectro óptico similar torna difícil diferenciar uma espécie de milho da outra, o que se faz necessário

à utilização de ferramentas matemáticas, no intuito de se obter dados, ou variáveis, que tornem possível a diferenciação entre as amostras.

A análise das componentes principais foi utilizada, com o objetivo de se explorar a separabilidade entre os dados amostrais, e obter informações de quais variáveis possam influenciar melhor na diferenciação entre as espécies. A figura 12 apresenta a dispersão pelas 3 (três) primeiras componentes principais.

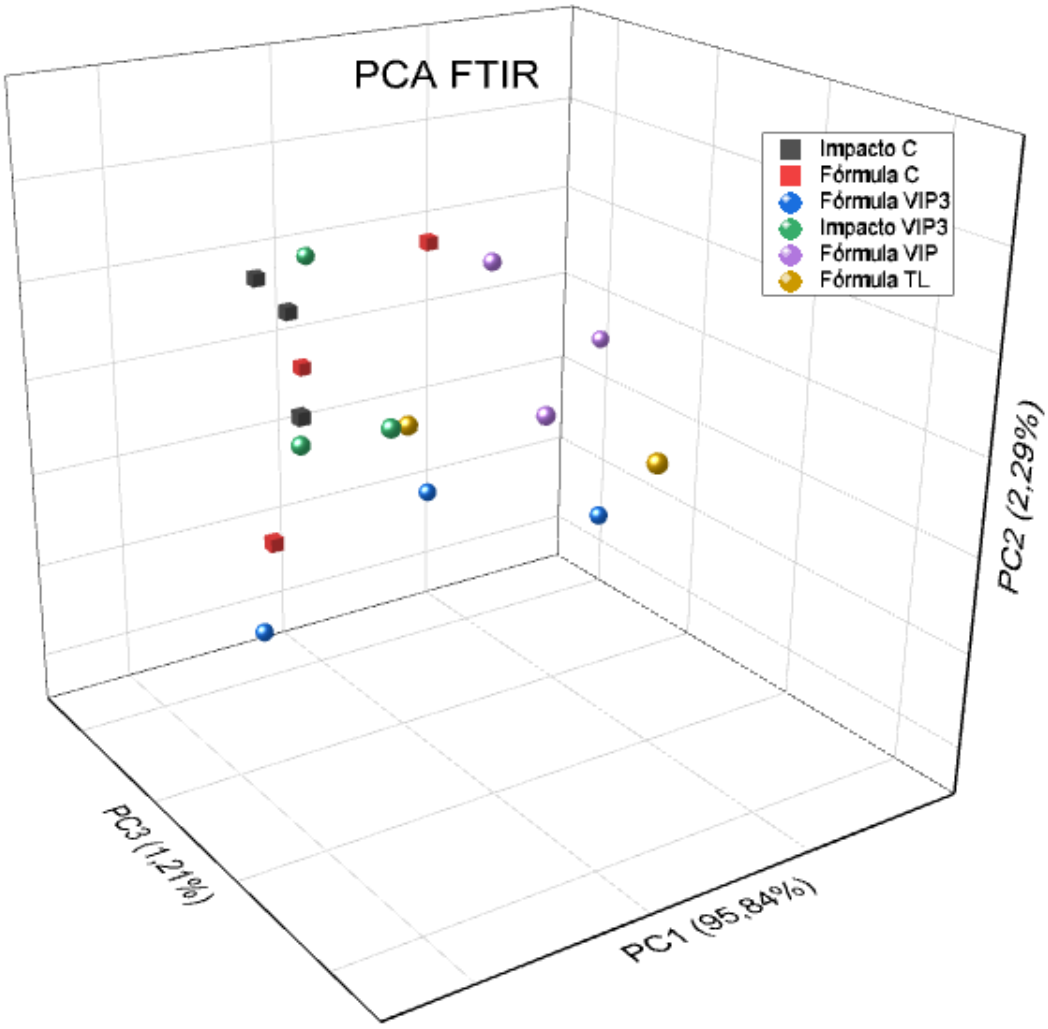

Figura 12 - Dispersão por Componentes Principais das amostras de milho transgênicas e não transgênicas – FTIR. O número de cada eixo indica o número da componente principal, e entre parêntesis a variância. Scores de formato quadrado são amostras convencionais, e de formato esféricos são amostras transgênicas. Fonte: Própria.

Na figura 12, a dispersão dos *scores* (pontuações referentes a cada amostra) realizada pela PCA é projetada, na qual não se conseguiu uma dispersão que fosse suficiente para separar amostras transgênicas das não transgênicas, pois os mesmos ficaram projetados de maneira que as classes transgênicas estavam muitos próximas das classes não transgênicas. Mesmo que a variância das componentes principais PC1, PC2 e PC3 foram de, respectivamente, 95,84%, 2,29% e 1,21%, totalizando 99,34% da dispersão total dos dados, se observa que a PCA possui dificuldades em obter informações que consiga diferenciar as amostras transgênicas das amostras não transgênicas.

Com a matriz de informações gerada pela Análise das Componentes Principais, os *loadings* das componentes principais foram plotados, com o objetivo de analisar qual foi a variável (elemento) que possa ter uma maior influência na diferenciação de uma amostra da outra. A figura 13 abaixo mostra os *loadings* das componentes principais PC1 e PC2.

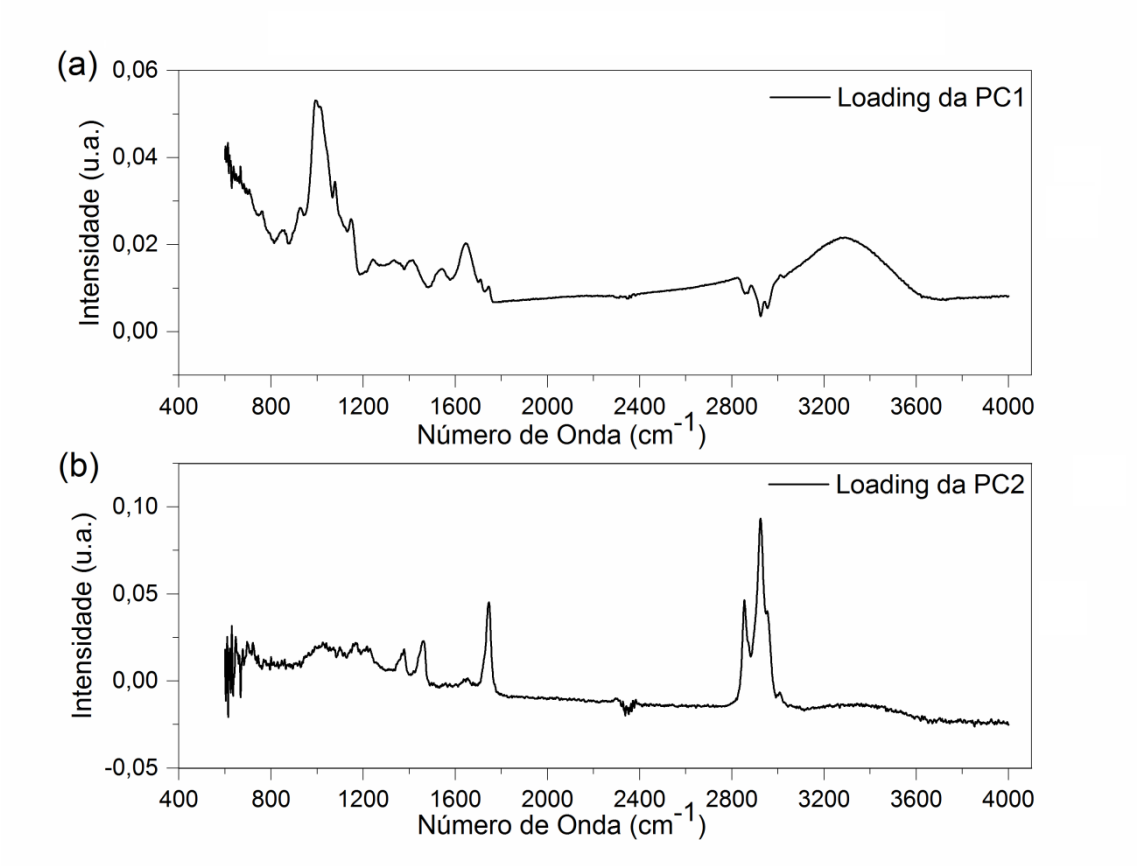

Figura 13 - Loadings das componentes principais das amostras transgênicas e não transgênicas – FTIR: (a) loadings da componente principal 1, (b) loadings da componente principal 2. Fonte: Própria.

Softwares de reconhecimento de padrão supervisionado foram utilizados, com o objetivo de, junto com as análises das componentes principais, obter um algoritmo que seja capaz de diferenciar as espécies de milho transgênicas de não transgênicas, somente utilizando informações de seu espectro óptico. Com os resultados da acurácia do treinamento dos classificadores de reconhecimento de padrão supervisionado, mostrado nas tabelas 03 e 04, 6 classificadores obtiveram as melhores *performances* durante seus treinamentos, que foi de 64,7%.

Tabela 03 – Dados do Treinamento para SVM para FTIR

Tipo do Classificador	Método de Validação	Função *Kernel*	Método Multiclasse	Acurácia (%)
SVM	5-fold Cross Validation	Linear	one-against-one	58,8
SVM	5-fold Cross Validation	Quadratica	one-against-one	58,8
SVM	5-fold Cross Validation	Cúbica	one-against-one	41,2
SVM	5-fold Cross Validation	*Fine* Gaussiana	one-against-one	64,7
SVM	5-fold Cross Validation	*Medium* Gaussiana	one-against-one	64,7
SVM	5-fold Cross Validation	*Coarse* Gaussiana	one-against-one	64,7

Tabela 04 – Dados do Treinamento para *k*-NN para FTIR

Tipo do Classificador	Método de Validação	Número de Vizinhos Próximos	Tipo de Distância	Acurácia (%)
k-NN	5-fold Cross Validation	1	Métrica Euclidiana	52,9
k-NN	5-fold Cross Validation	10	Métrica Euclidiana	64,7
k-NN	5-fold Cross Validation	100	Métrica Euclidiana	64,7
k-NN	5-fold Cross Validation	10	Métrica Cosseno	52,9
k-NN	5-fold Cross Validation	10	Cúbica (Mnkowski)	64,7
k-NN	5-fold Cross Validation	10	Inverso do Quadrado do Peso	52,9

Com base na acurácia, os classificadores SVM com função *Kernel Fine* Gaussiana, *Medium* Gaussiana, *Coarse* Gaussiana, k-NN com distância Euclidiana com 10 e 100 vizinhos mais próximos, e k-NN com distância cúbica, foram os que obtiveram as melhores precisões, de 64,7%. As figuras 14 a 19 abaixo apresentam a matriz de confusão desses classificadores, que é uma forma de visualizar a *performance* de um algoritmo de aprendizado supervisionado.

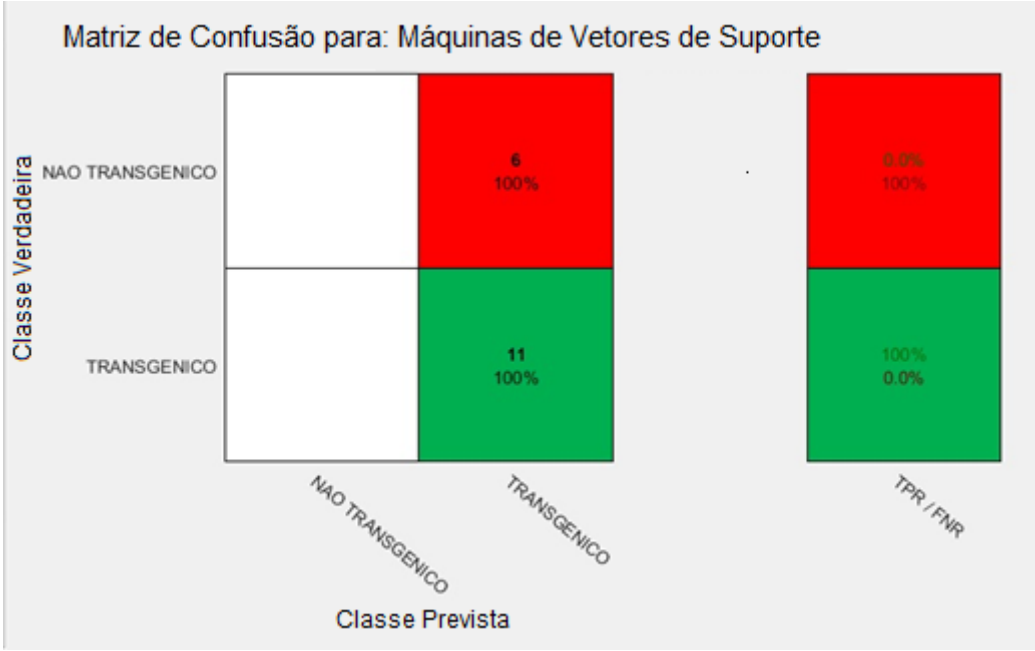

Figura 14 - Matriz de Erro do classificador SVM com função kernel Fine Gaussiana – FTIR. Fonte: Própria.

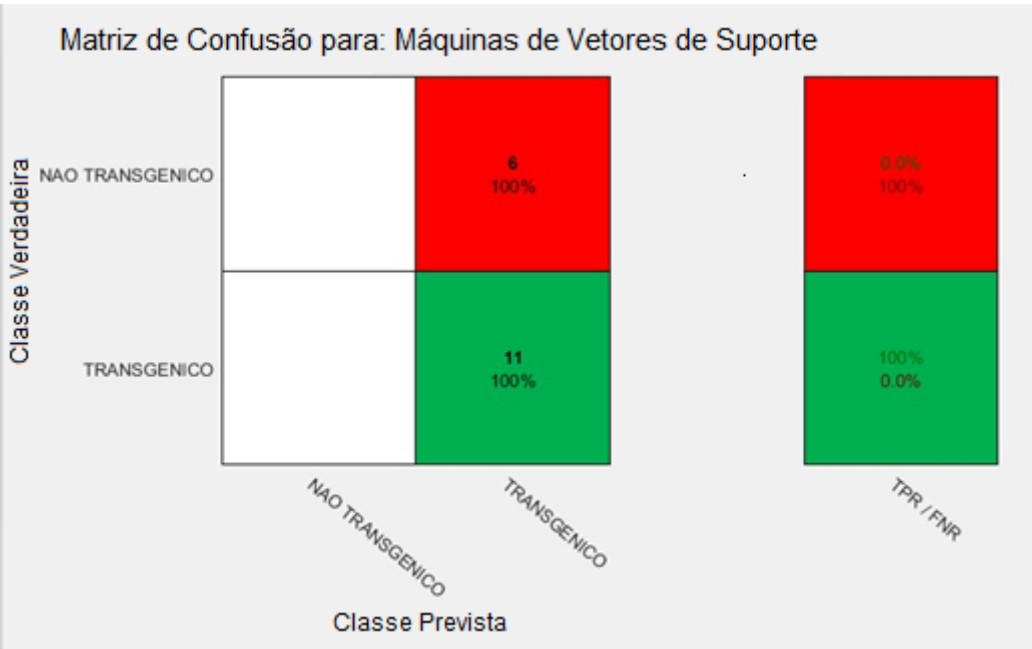

Figura 15 - Matriz de Erro do classificador SVM com função kernel Medium Gaussiana – FTIR. Fonte: Própria.

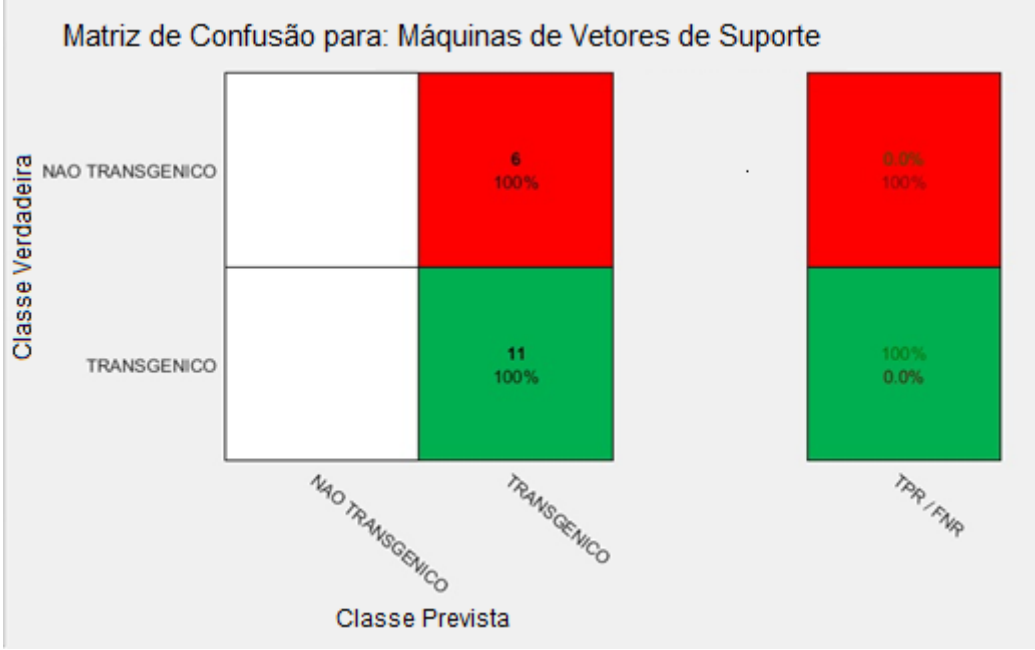

Figura 16 - - Matriz de Erro do classificador SVM com função kernel Coarse Gaussiana -- FTIR. Fonte: Própria.

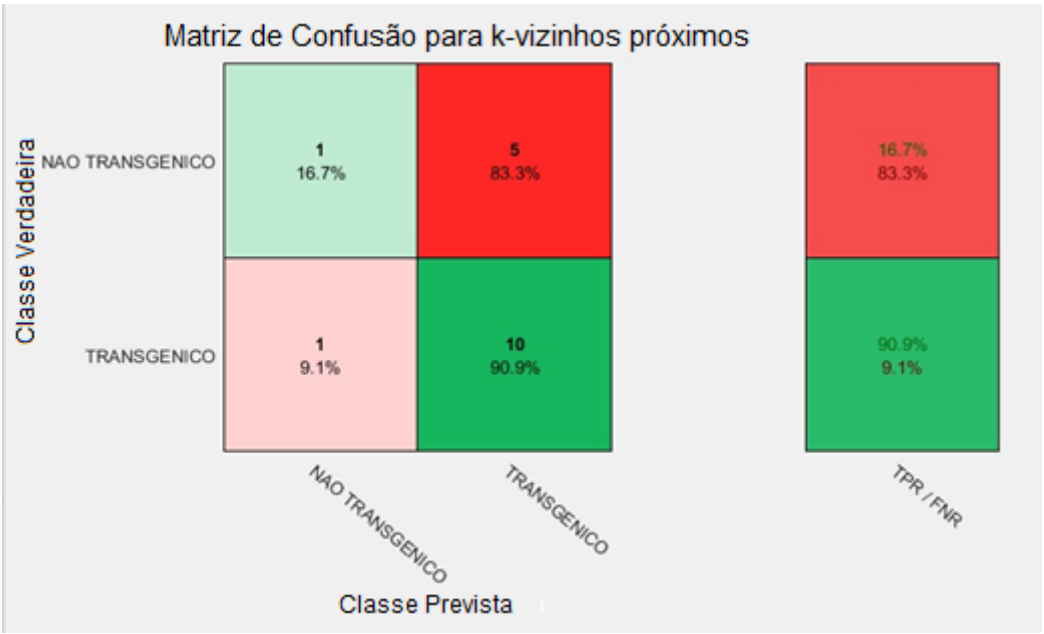

Figura 17 - Matriz de Erro do classificador k-NN com função distância cúbica – FTIR. Fonte: Própria.

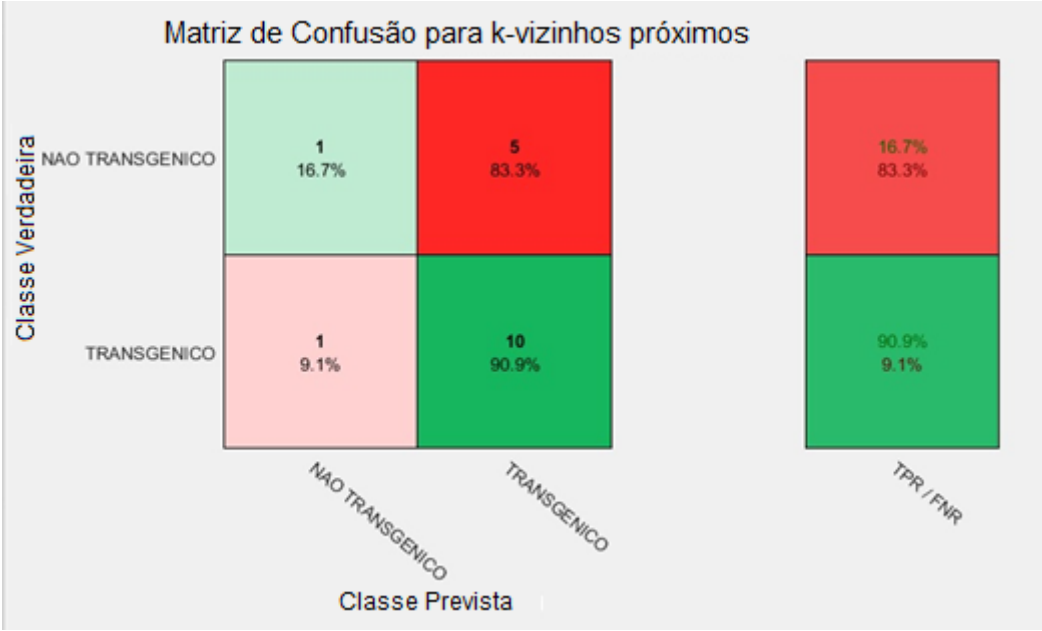

Figura 18 - Matriz de Erro do classificador k-NN com função distância métrica Euclidiana – FTIR. Fonte: Própria.

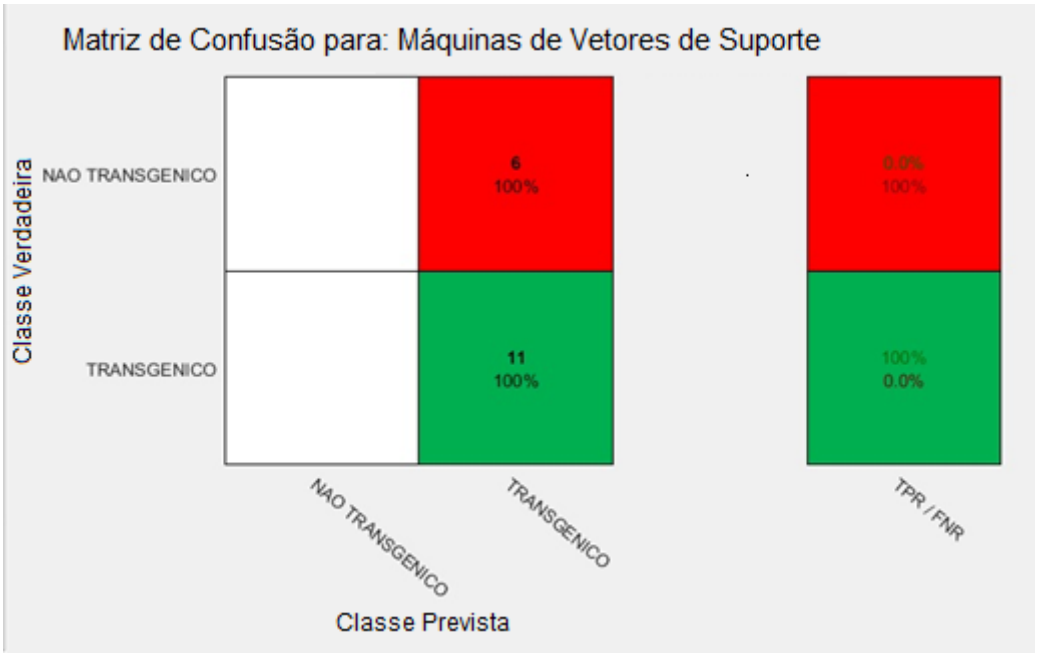

Figura 19 - Matriz de Erro do classificador k-NN com função distância Euclidiana – FTIR. Fonte: Própria.

Mesmo que esses classificadores possuam a mesma precisão, se pode notar que os melhores foram o *k-NN* com função distância cúbica, e com função distância métrica Euclidiana, pois as suas matrizes de erro são as únicas que possuem a taxa de verdadeiro positivo (TPR – *True Positive Rate*), tanto para as espécies transgênicas quanto para as espécies não transgênicas. Essa taxa TPR indica que o classificador classificou o espectro de uma classe prevista na sua verdadeira classe, o que em outras palavras diz que o classificador classificou “não transgênico” em “não transgênico” ao invés de “não transgênico” em “transgênico”, ou vice-versa.

5.2. Laser Induced Breakdown Spectroscopy – Faixa UV

Os espectros de LIBS, na região do Ultravioleta, compreenderam os comprimentos de onda entre 175 á 341 nm. A figura 20 abaixo demonstra o espectro óptico das amostras, para 89 espécies de milho não transgênicas (44 de Impacto C e 45 de Fórmula C) e 174 de espécies transgênicas (47 de Fórmula VIP3, 43 de Impacto VIP3, 41 de Fórmula VIP e 43 de Fórmula TL).

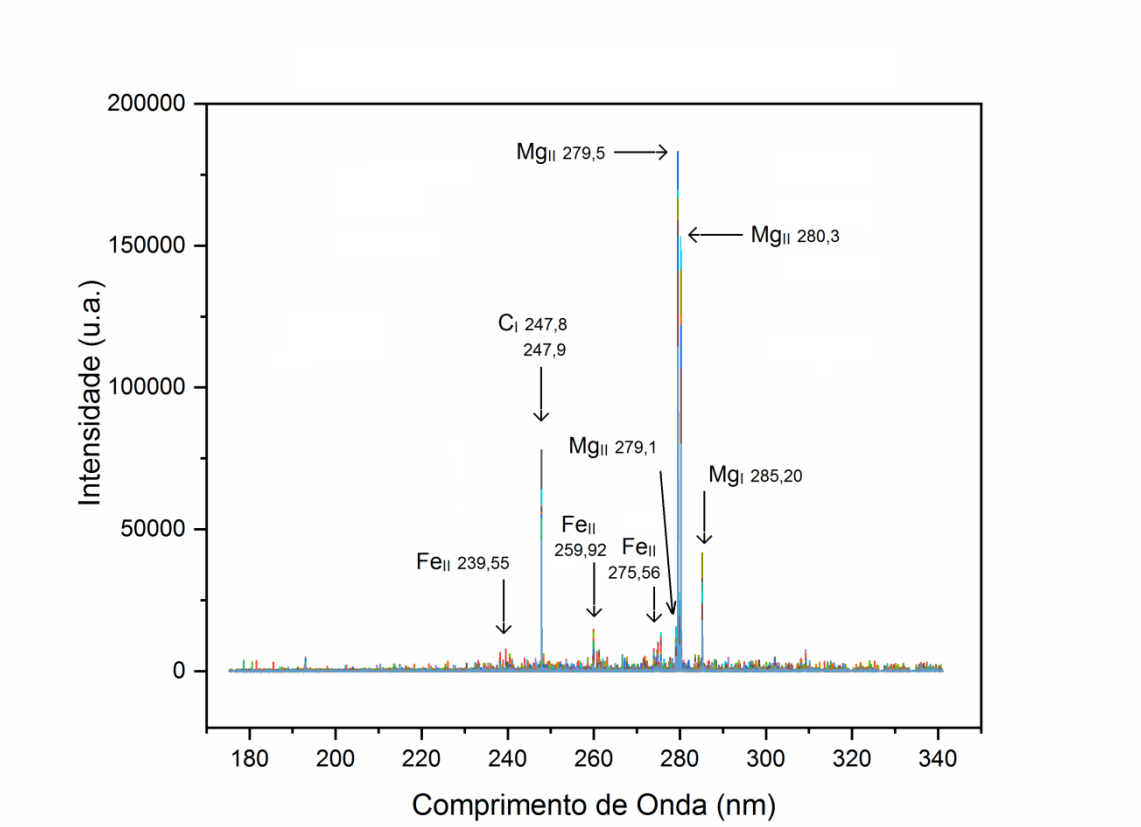

Figura 20 - Espectro LIBS na região do Ultravioleta para as espécies de milho. Fonte: Própria.

Analisando o espectro óptico de intensidade LIBS, na figura 20, se pode notar que os 263 espectros ópticos, plotados sobrepostos, possuem comportamento similar nas suas linhas de emissão, mas que diferem para valores de intensidade. Essa similaridade espectral dificulta diferenciar uma espécie de milho da outra, o que se faz necessário à utilização de ferramentas matemáticas, no intuito de se obter dados que tornem possível a diferenciação entre as amostras. De acordo com o Instituto Nacional de Padrões e Tecnologia (*National Institute of Standards and Technology – NIST database*),

e relacionando com as intensidades do espectro óptico, os elementos mais presentes nas amostras são: Fe_{II} (239,55 nm, 259,92 nm, 275, 56 nm), C_I (247,8 nm, 247,9 nm), Mg_{II} (279,1 nm, 279,5 nm, 280,3 nm) e Mg_I (285,20 nm).

A análise das componentes principais foi utilizada, com o objetivo de se explorar a separabilidade entre os dados amostrais, e obter informações de quais variáveis possam influenciar melhor na diferenciação entre as espécies. Através da dispersão dos *scores* da PCA mostrado na figura 21, visualmente não é possível separar em *clusters* os scores das espécies transgênicas e das não transgênicas, pois ficaram “misturados” entre as espécies. A figura 21 apresenta a dispersão pelas 3 (três) primeiras componentes principais.

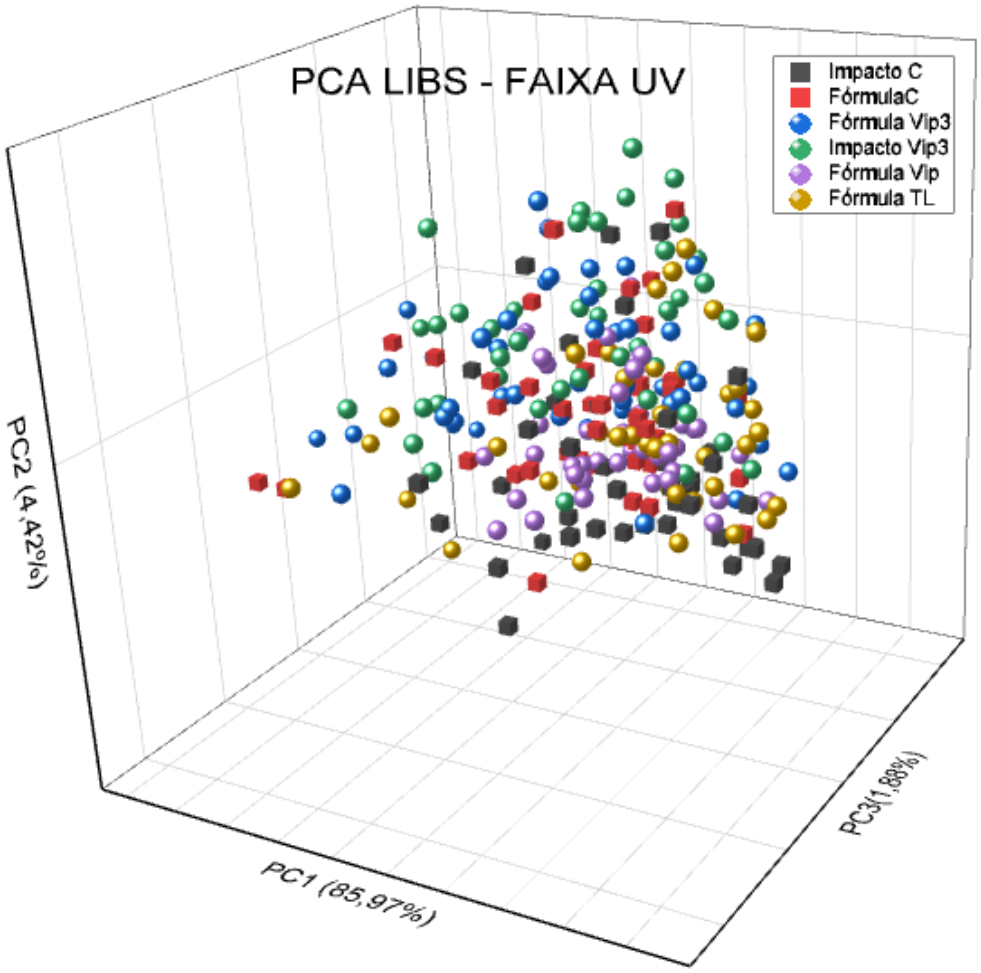

Figura 21 - Dispersão dos scores por PCA das espécies de milho transgênicas e não transgênicas – LIBS Faixa UV. O número de cada eixo indica o número da componente principal, e entre parêntesis a variância. Scores de formato quadrado são amostras convencionais, e de formato esféricos são amostras transgênicas. Fonte: Própria.

Mesmo havendo uma grande variância da PC1, PC2 e PC3, que foram, respectivamente, de 85,97%, 4,42% e 1,88%, com a projeção da dispersão dos *scores* não se é possível observar diferenciação entre as espécies. Quando a PCA não consegue separar as espécies, é um indicativo de que as espécies transgênicas e não transgênicas possuem um grande grau de similaridade em sua composição, o que mostra a dificuldade de diferenciar as espécies de milho somente através de uma análise espectral.

Com a matriz de informações gerada pela Análise das Componentes Principais, os *loadings* das componentes principais foram plotados, com o objetivo de analisar qual foi a variável (elemento) que possa ter uma maior influência na diferenciação de uma amostra da outra. A figura 22 abaixo mostra os *loadings* das componentes principais PC1 e PC2.

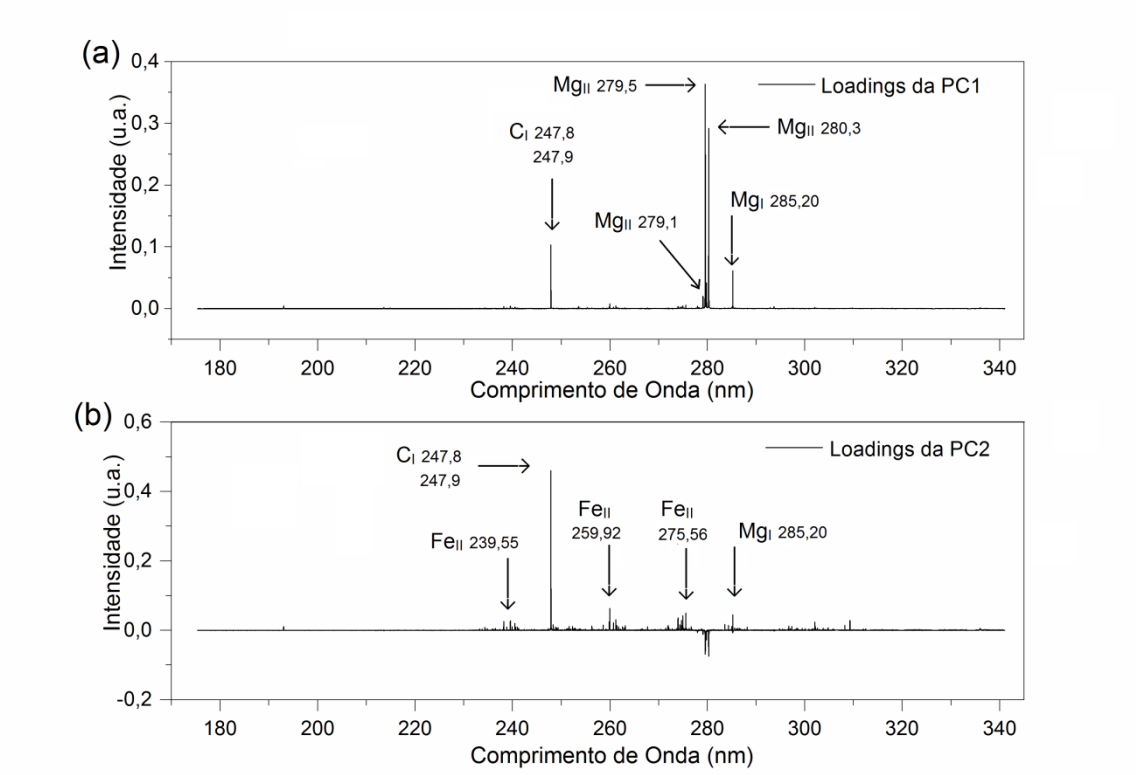

Figura 22 - Loadings das Componentes Principais para amostras de milho no LIBS UV: (a) loadings da componente principal 1, (b) loadings da componente principal 2. Fonte: Própria.

Os picos de maiores intensidade dos *loadings* indicam que, naquele correspondente comprimento de onda, a variável possui uma grande importância na diferenciação ou dispersão dos dados. De acordo com o Instituto Nacional de Padrões e Tecnologia (*National Institute of Standards and Technology – NIST)* database, os picos de maiores intensidades dos *loadings* indicam a presença dos seguintes elementos que mais possam influenciar na diferenciação entre as amostars: Fe_{II} (239,55 nm, 259,92 nm, 275,56 nm), C_{I} (247,8 nm), Mg_{II} (279,1 nm, 279,5 nm, 279,8 nm, 280,3 nm) e Mg_{I} (285,20 nm).

S*oftwares* de reconhecimento de padrão supervisionado foram utilizados, com o objetivo de, junto com as análises das componentes principais, obter um algoritmo que seja capaz de diferenciar as espécies de milho transgênicas das não transgênicas, utilizando informações de seu espectro óptico. Através das análises das tabelas 05 e 06, se nota que o melhor classificador foi o SVM com função *kernel* cúbica, alcançando uma precisão de 83,7%. Através da observação da matriz de erro desse classificador, figura 23, se pode ver que 73% das espécies convencionais (classe verdadeira), ou não transgênicas, foram classificadas corretamente em suas classes previstas, e 89,1% das espécies transgênicas (classe verdadeira) foram classificadas corretamente em suas classes previstas. Essas precisões indicam que o classificador possui uma capacidade de classificar os espectros ópticos entre as espécies que se deseja, mas que pode cometer erros de classificação, já que não possui uma precisão de classificação de 100%.

Tabela 05 – Dados do Treinamento para SVM para LIBS Faixa UV.

Tipo do Classificador	Método de Validação	Função *Kernel*	Método Multiclasse	Acurácia (%)
SVM	5-fold Cross Validation	Linear	one-against-one	83,3
SVM	5-fold Cross Validation	Quadratica	one-against-one	82,1
SVM	5-fold Cross Validation	Cúbica	one-against-one	83,7
SVM	5-fold Cross Validation	*Fine* Gaussiana	one-against-one	66,2
SVM	5-fold Cross Validation	*Medium* Gaussiana	one-against-one	66,2
SVM	5-fold Cross Validation	*Coarse* Gaussiana	one-against-one	66,2

Tabela 06 – Dados do Treinamento para *k*-NN para LIBS Faixa UV.

Tipo do Classificador	Método de Validação	Número de Vizinhos Próximos	Tipo de Distância	Acurácia (%)
k-NN	5-fold Cross Validation	1	Métrica Euclidiana	72,2
k-NN	5-fold Cross Validation	10	Métrica Euclidiana	82,1
k-NN	5-fold Cross Validation	100	Métrica Euclidiana	70,3
k-NN	5-fold Cross Validation	10	Métrica Cosseno	81
k-NN	5-fold Cross Validation	10	Cúbica (Mnkowski)	81,7
k-NN	5-fold Cross Validation	10	Inverso do Quadrado do Peso	81,4

A figura 23 abaixo mostra a matriz de confusão do classificador SVM de melhor precisão.

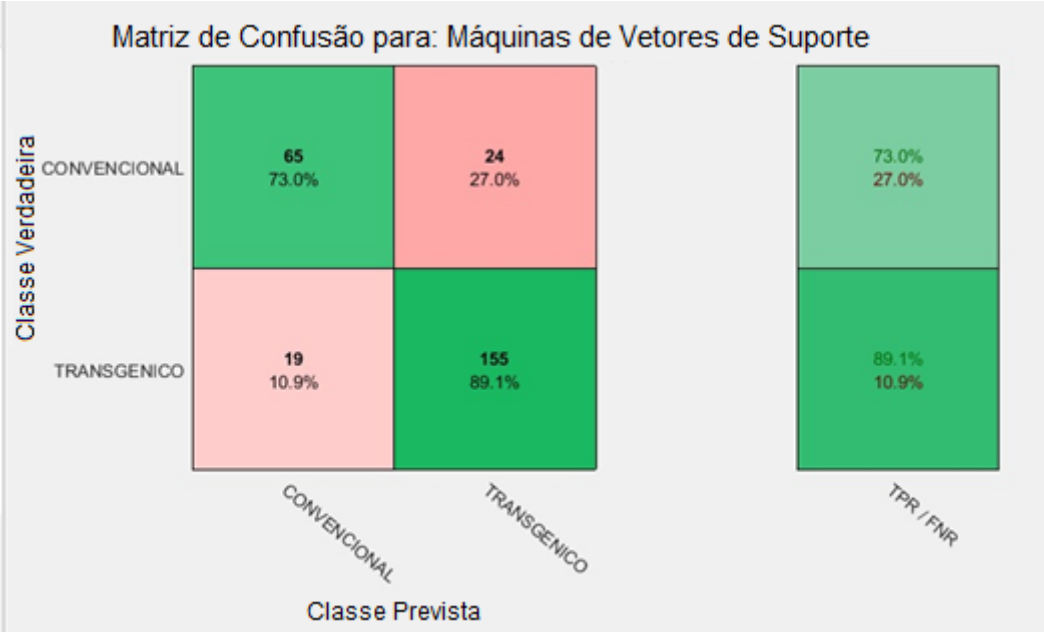

Figura 23 - Matriz de Erro do classificador SVM com função kernel distância cúbica – LIBS Faixa UV. Fonte: Própria.

5.3. Laser Induced Breakdown Spectroscopy – Faixa UV-Vis

Os espectros de LIBS na região Ultravioleta e do Visível compreenderam os comprimentos de onda entre 273,76 á 771,52 nm. A figura 24 abaixo apresenta o espectro óptico das amostras, para 88 espécies de milho não transgênicas (44 de Impacto C e 44 de Fórmula C) e 218 de espécies transgênicas (53 de Fórmula VIP3, 61 de Impacto VIP3, 54 de Fórmula VIP e 50 de Fórmula TL).

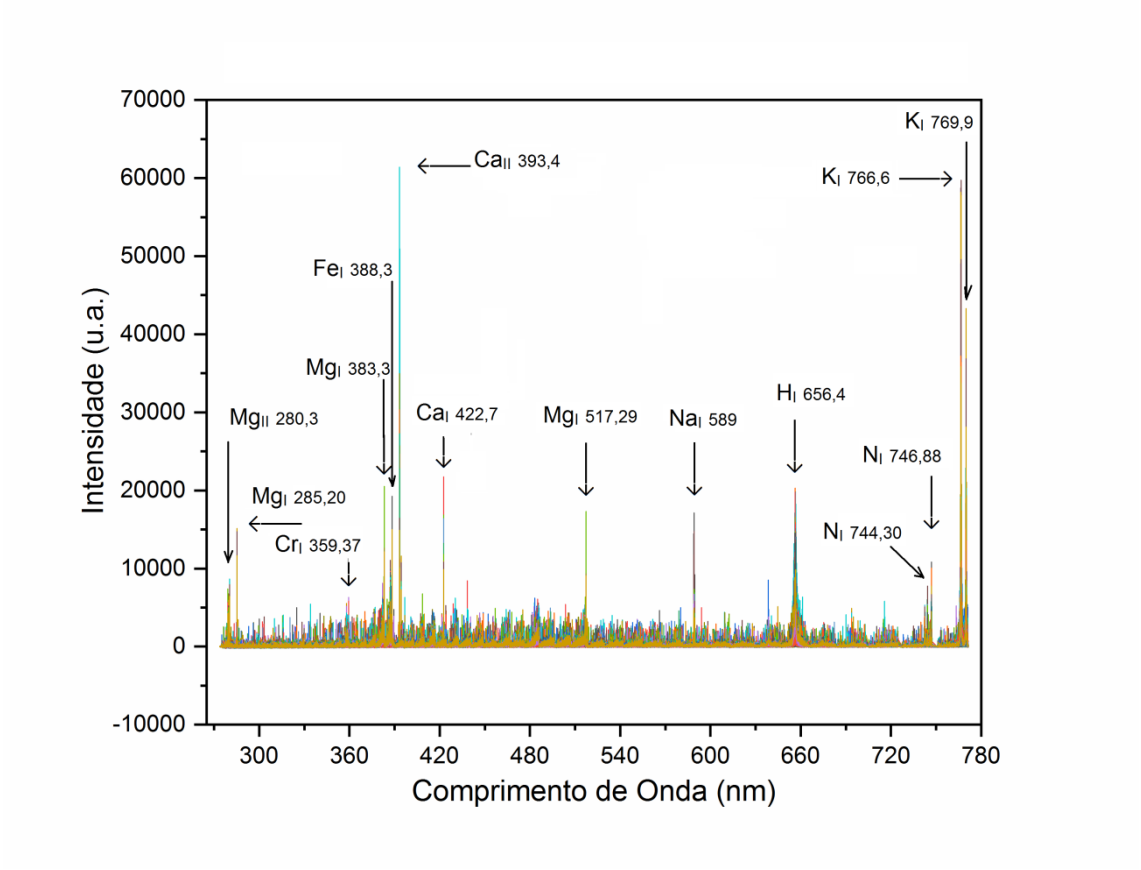

Figura 24 - Espectro LIBS na faixa UV-Vis para as espécies de milho. Fonte: Própria.

A figura 24 mostra o espectro óptico LIBS Vis. Visualmente, o comportamento espectral possui muitas similaridades entre as espécies de milho transgênicas e não transgênicas, mas que diferenças nas intensidades ópticas podem ser notadas. Similaridade no espectro óptico torna difícil a diferenciação entre uma espécie da outra, somente pela análise de seu comportamento espectral. De acordo com o Instituto Nacional de Padrões e

Tecnologia (*National Institute of Standards and Technology – NIST database*), e relacionando com as intensidades do espectro óptico, os elementos mais presentes nas amostras são: Mg_{II} (280,3 nm) e Mg_{I} (285,20 nm, 383,3 nm, 517, 29 nm), Cr_{I} (359,37 nm), Fe_{I} (388,3 nm), Ca_{II} (393,4 nm), Ca_{I} (422,7 nm), Na_{I} (589 nm), H_{I} (656,4 nm), N_{I} (744,30 nm, 746,88 nm) e K_{I} (766,6 nm, 769,9 nm).

A análise das componentes principais foi utilizada, com o objetivo de se explorar a separabilidade entre os dados amostrais, e obter informações de quais variáveis possam influenciar melhor na diferenciação entre as espécies. Projetando a dispersão dos *scores* das 3 (três) primeiras componentes principais, figura 25, visualmente também não é possível diferenciar e nem separar as espécies transgênicas das não transgênicas, pois os *scores* das espécies foram projetados “misturados” uns aos outros. A variância da PC1 foi de 39,61%, e que se comparado com as variâncias das PC’s1 da técnica óptica FTIR e LIBS Faixa UV, é a menor variância obtida. Isso indica que a PCA no LIBS Faixa UV-Vis tem uma dificuldade maior para dispersar os dados, o que pode indicar que a técnica óptica LIBS Faixa UV-Vis não seja a mais indicada para diferenciar espécies de milho transgênicas de não transgênicas.

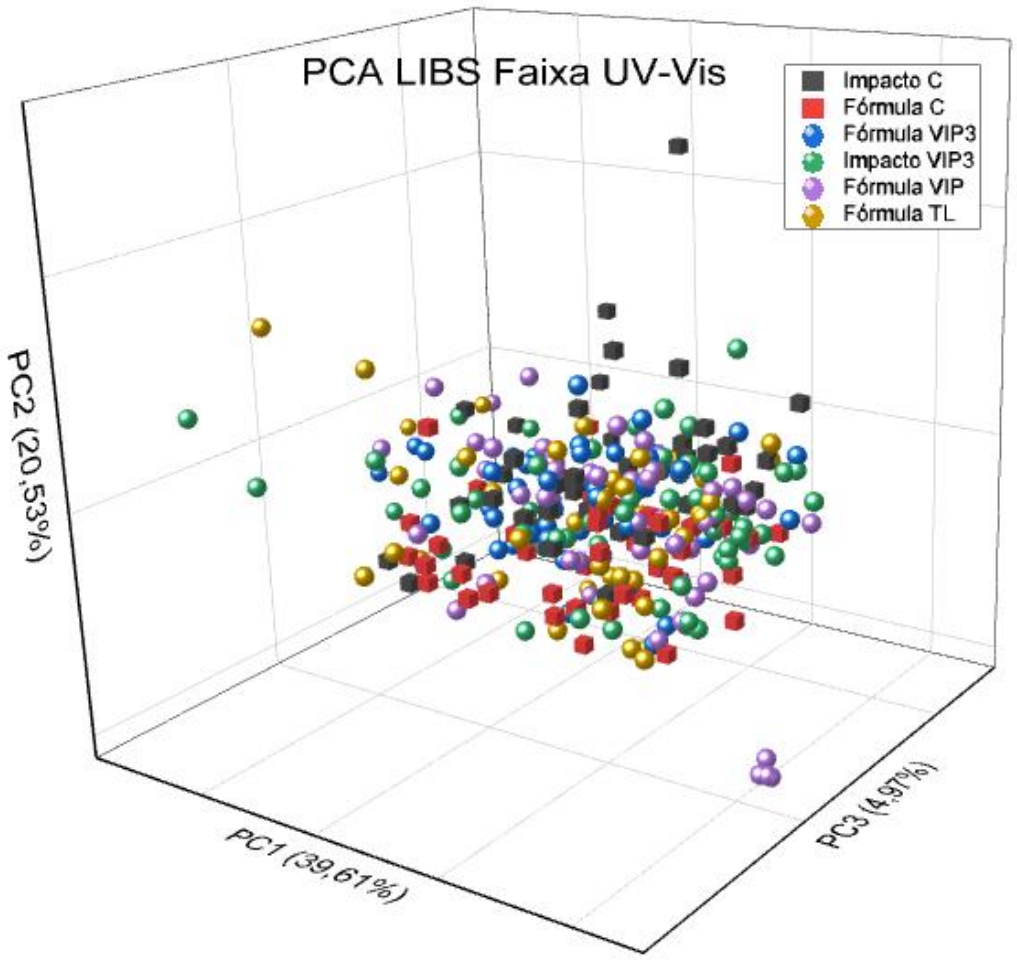

Figura 25 - Dispersão dos scores por PCA das espécies de milho transgênicas e não transgênicas – LIBS Faixa UV- Vis. O número de cada eixo indica o número da componente principal, e entre parêntesis a variância. Scores de formato quadrado são amostras convencionais, e de formato esféricos são amostras transgênicas. Fonte: Própria.

Com a matriz de informações gerada pela Análise das Componentes Principais, os *loadings* das componentes principais foram plotados, com o objetivo de analisar qual foi a variável (elemento) que possa ter uma maior influência na diferenciação de uma amostra da outra. A figura 26 abaixo mostra os *loadings* das componentes principais PC1 e PC2.

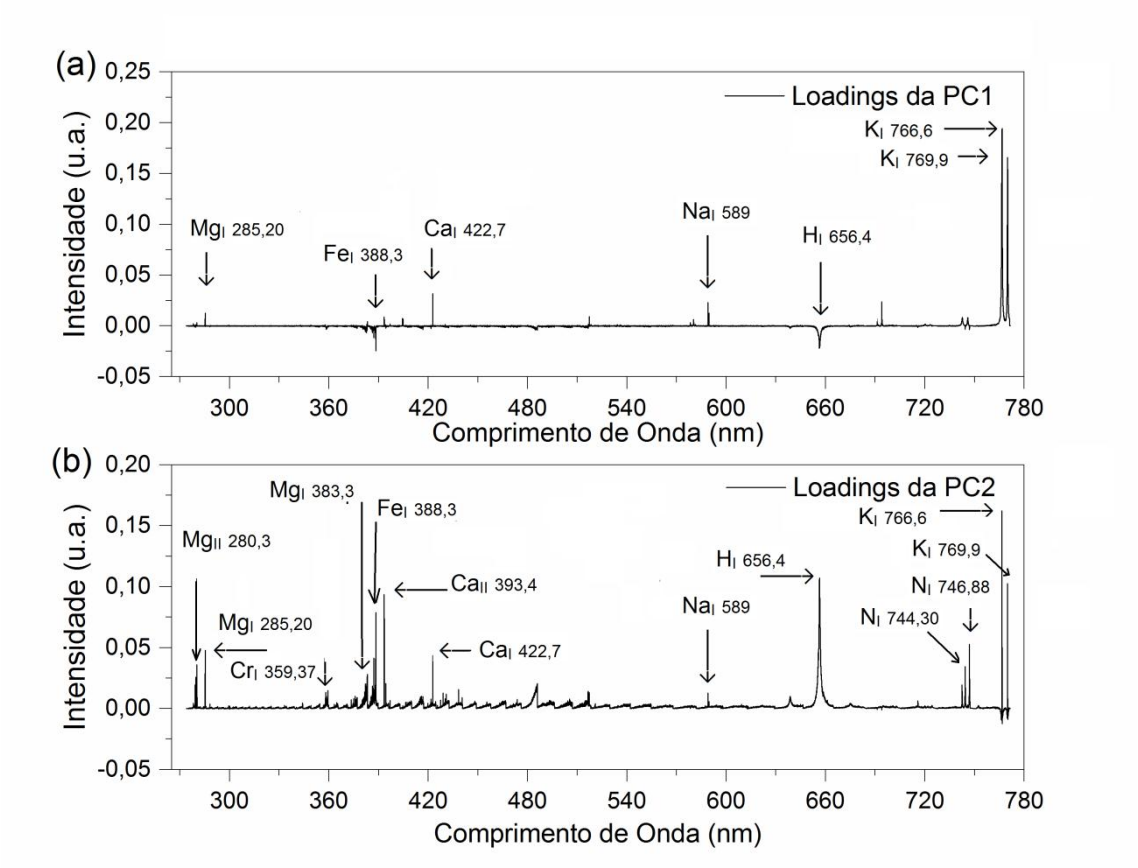

Figura 26 - Loadings das Componentes Principais para amostras de milho no LIBS Faixa UV-Vis: (a) loadings da componente principal 1, (b) loadings da componente principal 2. Fonte: Própria.

A figura 26 mostra os *loadings* das componentes principais PC1 e PC2. Os *loadings* carregam informações que relacionam as componentes principais com as variáveis utilizadas durante a obtenção das PC's. Os picos de maiores intensidade dos *loadings* indicam que, naquele correspondente comprimento de onda, a variável possui uma grande importância na diferenciação ou dispersão dos dados. De acordo com o Instituto Nacional de Padrões e Tecnologia (*National Institute of Standards and Technology – NIST)* database, os picos de maiores intensidades dos *loadings* indicam a presença dos seguintes elementos que mais influenciam na diferenciação dos dados: Mg_{II} (280,3 nm), Mg_{I} (285,20 nm, 383,3 nm), Cr_{I} (359,4 nm), Fe_{I} (388,3 nm), Ca_{II} (393,4 nm), Ca_{I} (422,7 nm), Na_{I} (589 nm, 589,6 nm), H_{I} (656,3 nm, 656,4 nm), N_{I} (744,3 nm, 746,9 nm) e K_{I} (766,6 nm, 766,9 nm e 770 nm).

Softwares de reconhecimento de padrão supervisionado foram utilizados, com o objetivo de, junto com as análises das componentes principais, obter um algoritmo que seja capaz de diferenciar as espécies de milho transgênicas de não transgênicas, somente utilizando o seu espectro óptico. Pela análise das tabelas 07 e 08, se pode ver que o classificador SVM com função *kernel* cúbica foi o que atingiu a melhor performance, que foi de 81% de acurácia. Essa performance é melhor que as obtidas nos treinamentos para FTIR, mas é menor do que a obtida nos treinamentos de LIBS Faixa UV. Na análise da matriz de erro da figura 27, se verifica que somente 39,8% das espécies da classe verdadeira convencional foram classificadas na classe prevista convencional, o que é uma TPR menor que a obtida no LIBS UV. A TPR para as classes transgênicas foi de 97,7%, que é uma taxa maior que a do LIBS UV, mas que, em uma performance geral, fez esse classificador atingir somente 81% de performance

Tabela 07 – Dados do Treinamento para SVM para LIBS faixa UV-Vis.

Tipo do Classificador	Método de Validação	Função *Kernel*	Método Multiclasse	Acurácia (%)
SVM	5-fold Cross Validation	Linear	one-against-one	72,9
SVM	5-fold Cross Validation	Quadratica	one-against-one	80,7
SVM	5-fold Cross Validation	Cúbica	one-against-one	81
SVM	5-fold Cross Validation	*Fine* Gaussiana	one-against-one	71,2
SVM	5-fold Cross Validation	*Medium* Gaussiana	one-against-one	71,2
SVM	5-fold Cross Validation	*Coarse* Gaussiana	one-against-one	71,2

Tabela 08 – Dados do Treinamento para kNN para LIBS faixa UV-Vis.

Tipo do Classificador	Método de Validação	Número de Vizinhos Próximos	Tipo de Distância	Acurácia (%)
k-NN	5-fold Cross Validation	1	Métrica Euclidiana	65
k-NN	5-fold Cross Validation	10	Métrica Euclidiana	71,6
k-NN	5-fold Cross Validation	100	Métrica Euclidiana	71,2
k-NN	5-fold Cross Validation	10	Métrica Cosseno	77,8
k-NN	5-fold Cross Validation	10	Cúbica (Mnkowski)	74,2
k-NN	5-fold Cross Validation	10	Inverso do Quadrado do Peso	71,6

A figura 27 abaixo mostra a matriz de confusão do classificador com a melhor precisão, que é uma forma de visualizar a *performance* de um algoritmo de aprendizado supervisionado.

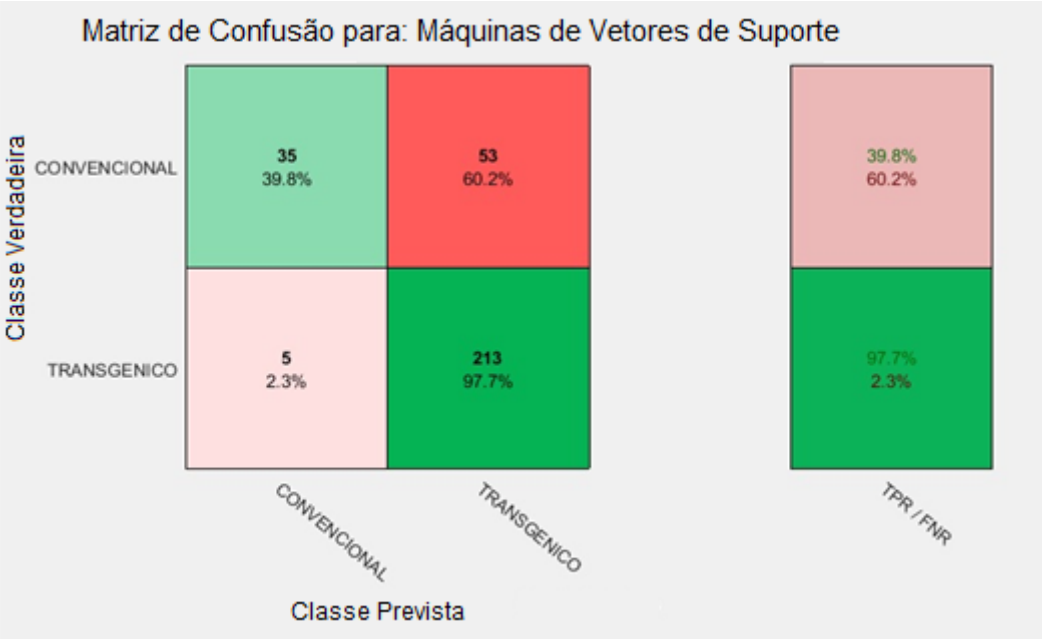

Figura 27 - Matriz de Erro do SVM com distância cúbica – LIBS faixa UV-Vis. Fonte: Própria.

6. CONCLUSÃO

Com base nos resultados obtidos e as discussões realizadas, para as 6 espécies de milho utilizadas como amostras, a técnica LIBS demonstrou ter um grande potencial para a diferenciação de amostras transgênicas e não transgênicas. A técnica FT-IR mostrou potencial limitado, sendo necessário um estudo mais aprofundado. Quando a PCA foi utilizada em associação com a técnica FT-IR, essa foi a técnica que adquiriu a maior variância das 3 primeiras componentes principais, com o valor acumulado de 99,34%. Na mesma condição, a técnica óptica LIBS Faixa Ultravioleta adquiriu 92,27% de variância, enquanto o LIBS na Faixa do Ultravioleta e Visível adquiriu 65,11%. Os gráficos dos *loadings* das componentes principais foram gerados, a fim de descobrir quais variáveis podem melhor influenciar na diferenciação entre as amostras. Para a técnica LIBS, tanto na faixa do UV quanto na faixa do UV-Vis, vários elementos foram identificados, podendo esses serem os mais influentes na diferenciação das amostras, sendo Carbono (C), Potássio (K), Cálcio (Ca), Magnésio (Mg) e Ferro (Fe). Métodos Supervisionados de Reconhecimento de Padrões foram aplicados, juntamente com ferramentas de classificação, para obtenção de rotina que diferencie classe transgênica da não transgênica. Através das acurácias obtidas durante os treinamentos desses classificadores, a melhor performance foi obtida com o classificador SVM com função *kernel* cúbica, quando associada com a técnica LIBS UV, com acurácia de 83,7%, Para a técnica LIBS UV-Vis, o classificador SVM com função *kernel* cúbica também foi o que atingiu a melhor acurácia, de 81%, e para a técnica FT-IR, o classificador k-NN com distância cúbica e o k-NN com distância métrica Euclidiana atingiram a melhor performance de 64,1%. A baixa precisão de classificação associada à técnica FT-IR e, taxas de classificação maiores associadas a técnicas LIBS na faixa UV e UV-Vis, se deve ao fato de que a técnica LIBS é capaz de informar um número maior de dados ópticos de cada amostra, que são utilizadas para verificar similaridades ou diferenças entre as classes, aumentando o poder de diferenciar as amostras. O fato dos gráficos de dispersão dos *scores* da PCA indicar que algumas amostras transgênicas possuem similaridades com amostras não transgênicas, pois foram plotados “misturados” uns aos outros, indica que a técnica óptica que conseguir obter

um número maior de informações das amostras, será aquela na qual o classificador alcançara a maior precisão de classificação. A utilização de técnicas ópticas junto com métodos supervisionados de reconhecimento de padrões se mostraram bastantes promissoras para a diferenciação entre amostras transgênicas e não transgênicas, em grãos de milho, o que também se pode avaliar a possibilidade de aplicação comercial para essa atividade em específico.

7. REFERÊNCIAS BIBLIOGRÁFICAS

- ABIMILHO. Associação Brasileira das Indústrias de Milho. Disponível em: <http://www.abimilho.com.br>. Acesso em: 05/08/2017.

- ATKINS, Peter; DE PAULA, Julio. *Physical Chemistry.* 8 ed. W.H. Freeman and Company. New York. 2006. ISBN: 0-7167-8759-8.

- BASSI, Adalberto B. M. S. Conceitos Fundamentais em Espectroscopia. *Revista Chemkeys.* Universidade Estadual de Campinas. São Paulo. 2001.

- BEZERRA, Miguel Eugênio Ramalho. *Métodos baseados na regra do vizinho mais próximo para reconhecimento de imagens.* 2006. 90 f. Trabalho de Conclusão de Curso (Bacharelado em Engenharia da Computação) – Escola Politécnica de Pernambuco – Universidade de Pernambuco, Recife.

- BRASIL. Decreto n. 4.680, 24 de abril de 2003. Regulamenta o direito á informação. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto/2003/D4680.htm>. Acesso em : 27/08/2018.

- BRASIL. Lei n. 11.105, 14 de março de 2005. Normas de segurança e fiscalização de OGM's. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2004-006/2005/Lei/L11105.htm>. Acesso em : 27/08/2018.

- BRASIL. Lei n. 8.078, 11 de setembro de 1990. Código de defesa do consumidor. Disponível em: <http://www.planalto.gov.br/ccivil_03/LEIS/L8078.htm>. Acesso em : 27/08/2018.

- CONAB – Companhia Nacional de Abastecimento. *Acompanhamento da Safra Brasileira grãos.* v.4, safra 2016/17 – n.10. Brasília, 2017. 170 p. ISSN: 2318-6852.

- CONCEIÇÃO, Fabricio R; MOREIRA, Ângela N; BINSFIELD, Pedro C. Detecção e quantificação de organismos geneticamente modificados em alimentos e ingredientes alimentares. *Revista Ciência Rural*, Santa Maria, v. 36, n. 1, p.315-324, jan-fev. 2006. ISSN: 0103-8478.

- CONSELHO DE INFORMAÇÕES SOBRE BIOTECNOLOGIA. *Guia do milho: Tecnologia do campo á mesa.* São Paulo, 2006.

- DINON, Andréia Zilio. *Detecção por PCR de milho geneticamente modificado (MON 810) em farinha de milho, fubá, biju e polenta.*2007. 81 f. Dissertação (Mestrado em Ciência dos Alimentos) – Universidade Federal de Santa Catarina, Florianópolis.

- EISBERG, Robert; RESNICK, Robert. *Física Quântica:* átomos, moléculas, sólidos, núcleos e partículas. 461 p. 23ª tiragem, ed. Elsevier.

- EMBRAPA. *Folheto de perguntas e respostas sobre milho.* Brasília, 2011. 15 p

- ESPECTROSCOPIA no infravermelho. 45 p. cap. 2.

- FERRARINI, Hair. *Determinação de teores nutricionais do milho por espectroscopia no infravermelho e calibração multivariada.* 2004. 125 f. Dissertação (Mestrado em Química) – Universidade Federal do Paraná, Curitiba.

- FRANCO, Marco Aurélio de Menezes. *Efeitos de matriz nas propriedades do plasma LIBS para quantificação de carbono.* 2017. 124 f. Dissertação (Mestrado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Paulo.

- ISAAA. 2017. Global Status of Commercialized Biotech/GM Crops in 2017: Biotech Crop Adoption Surges as Economic Benefits Accumulate in 22 Years. *ISAAA Brief* No. 53. ISAAA: Ithaca, NY. Disponível em: <ISAAA.org/resources/publications/briefs/53>. Acesso em: 16/08/2018.

- LIU, Fei; YE, Lanhan; PENG, Jiyu; SONG, Kunlin; SHEN, Tingting; ZHANG, Chu; HE, Yong. *Fast Detection of Copper Content in Rice by Laser-Induced Breakdown Spectroscopy with Uni-and Multivariate Analysis.* Sensor. V.18, 15p, 2018. DOI: 10.3390/s18030705.

- LIU, Xiaodan; FENG, Xuping; LIU, Fei; PENG, Jiyu; HE, Young. *Rapid Identification of Genetically Modified Maize Using Laser-Induced Breakdown Spectroscopy.* Food and Bioprocess Technology. V.12, p.347-357, 2019. DOI: 10.1007/s11947-018-2216-0.

- LOGUERCIO, Leandro L; CARNEIRO, Newton P; CARNEIRO, Andréa A. *Milho Bt:* alternativa biotecnológica para controle biológico de insetos-praga. Biotecnologia Ciência & Desenvolvimento. n 24 – 2002.

- LORENA, Ana Carolina; CARVALHO, André C. P. L. F. de. Uma introdução ás *Support Vector Machines*. *Revista de Informática Teórica e Aplicada.* Vol. 14. n. 2. P. 43-67. 2007.

- MARANGONI, B.S.; SILVA, K.S.G.; NICOLODELLI, G.; SENESI, G.S.; CABRAL, J.S.; VILLAS-BOAS, P.R.; SILVA, C.S.; TEIXEIRA, P.C.; NOGUEIRA, A.R.A.; BENITES, V.M.; MILORI, D.M.B.P. *Phosphorus quantification in fertilizers using laser induced breakdown spectroscopy (LIBS): a methodology of analysis to correct physical matrix effects***.** Royal Society of Chemistry. Anal. Methods, v.8, p.78-82, 2016. DOI: 10.1039/c5ay01615k

- MARKIEWICZ-KESZYCKA, M.; CAMA-MONCUNILL, X.; P. CASADO-GAVALDA, M.; DIXIT, Yash.; CAMA-MONCUNILL, R.; CULLEN, P.J.; SULLIVAN, Carl. *Laser-induced breakdown spectroscopy (LIBS) for food analysis: A review.* Trends in Food Science & Technology. V.65, p. 80 – 93, 2017. DOI: 10.1016/j.tifs.2017.05.005 0924-2244.

- NICOLODELLI, Gustavo; SENESI, Giorgio S.; ROMANO, Renan A.; PERAZZOLI, Ivan L. O.; MILORI, Débora M. B. P. Signal enhancement in collinear double-pulse laser-induced breakdown spectroscopy applied to different soils. *Spectrochimica Acta Part B: Atomic Spectroscopy*, v. 111, p. 23-29, 2015.

- OLIVEIRA, Luiz F. C. *Espectroscopia Molecular.* Minas Gerias: Cadernos temáticos de Química Nova na Escola. n. 4. p. 24 – 30. 2001.

- PASQUINI, Celio; CORTEZ, Juliana; SILVA, Lucas M. C.; GONZAGA, Fabiano B. *Laer Induced Breakdown Spectroscopy.* J. Braz. Chem. Soc. Vol.18, n. 3. P. 463-512. 2007.

- PONTIFICE UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO – PUC RIO. *Folheto de técnicas espectroscópicas.* Certificação digital nº 0321127/CA. Rio de Janeiro, 16 p.

- PONTIFICE UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO – PUC RIO. *Folheto de espectrometria de fluorescência molecular.* Certificação digital nº 0212136/CA. Rio de Janeiro, 10 p.

- PONTIFICE UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO – PUC RIO. *Folheto de introdução á fluorescência.* Certificação digital nº 0510942/CA. Rio de Janeiro, 18 p.

- PONTIFICE UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO – PUC RIO. *Folheto de técnicas experimentais.* Certificação digital nº 0521270/CA. Rio de Janeiro, 28 p.

- Protocolo de Cartagena. Protocolo de Cartagena sobre Biossegurança. Disponível em: http://www.mma.gov.br/biodiversidade/conven%C3%A7%C3%A3o-da-diversidade-biol%C3%B3gica/protocolo-de-cartagena-sobre-biosseguranca. Acesso em: 27/08/2018.

- RANULFI, Anielle Coelho. *LIBS como ferramenta diagnóstica em plantas: um estudo nutricional de folhas de soja na busca pelos efeitos da infestação por Aphelenchoides besseyi.* 2019. 151 f. Tese (Doutorado em Ciências) – Instituto de Física de São Carlos, Universidade de São Paulo, São Paulo.

- REVISTA CULTIVAR. *Tecnologia Protetora.* p. 36 – 38. 2011

- SEZER, Banu; DURNA, Sahin; BILGE, Gonca; BERKKAN, Aysel; YETISEMIYEN, Atilla; BOYACI, Ismail H. *Identification of milk fraud using laser-induced breakdown spectroscopy (LIBS).* International Dairy Journal. V.81, p.1-7, 2017. DOI: 10.1016/j.idairyj.2017.12.005.

- SEZER, Banu; VELIOGLU, Hasan M.; BILGE, Gonca; BERKKAN, Aysel; OZDIC, Nese; TAMER, Ugur; BOYACI, Ismail H. *Meat Science.* V.135, p.123-128, 2017. DOI: 10.1016/j.meatsci.2017.09.010.

- SILVA, Renato Rodrigues da. *Desenvolvimento de toolbox de análise multivariada para o Matlab.* 2015. 62 f. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) – Universidade Federal de Uberlândia, Uberlândia.

- SKOOG, Douglas A.; WEST, Donald M.; HOLLER, F. J.; CROUCH, Stanley R. *Fundamentos da Química Analítica.* Tradução da 8ª edição norte americana. Editora Thomson. 2006.

- USDA – United States Department of Agriculture. World Agricultural Supply and Demand Estimates, Washington, DC. Disponível em: <https://www.ers.usda.gov/topics/crops/corn/trade.aspx#world> . Acesso em: 20/08/2018.

- USDA – United States Department of Agriculture. World Agricultural Supply and Demand Estimates, Washington, DC. Disponível em: <https://www.ers.usda.gov/topics/crops/corn/trade.aspx#world> . Acesso em: 05/08/2017.

- VASCONCELOS, Simone. *Análise de Componentes Principais (PCA).* 17 f.

APÊNDICE A

O apêndice tem por objetivo o detalhamento do passo-a-passo realizado para a execução do trabalho.

A tabela A1 abaixo informa as espécies de milho utilizadas no FTIR, LIBS faixa UV e LIBS faixa UV-Vis, e suas respectivas classes.

Tabela A1 – Dados das espécies de milho para FTIR, LIBS UV e LIBS UV-Vis

Amostra	Espécies	Classes
1	Impacto C	Convencional
2	Fórmula C	Convencional
3	Fórmula VIP3	Transgênico
4	Impacto VIP3	Transgênico
5	Fórmula VIP	Transgênico
6	Fórmula TL	Transgênico

• A1 - ESPECTROSCOPIA NO INFRAVERMELHO POR TRANSFORMADA DE FOURIER – FTIR

As espécies de milho da tabela A1 sofreram análise óptica por FTIR, obtendo os dados numéricos do número de onda incidente na amostra, e a sua respectiva intensidade.

Utilizando o software MatLab® para a extração desses dados numéricos, e também para o tratamento e manipulação dos dados, obteve uma primeira tabela, na qual as linhas eram representadas pelas variáveis (comprimentos de onda), e as colunas pelas observações (espécies de milho). Os dados ópticos de cada espécie de milho foram armazenados, cada um em uma célula diferente da tabela nomeada de “data.Struct”, na qual cada grupo de colunas representam uma espécie de milho diferente. Concatenou os dados horizontalmente de cada coluna da tabela “data.Struct”, obtendo uma segunda tabela de dimensões 3.401 linhas x 17 colunas, nomeada de “AbsorcaoDados”.

A figura A1.1 abaixo demonstra a parte do código responsável pela execução do procedimento descrito acima.

```
n = 17;

for i=1:n
    filename = ['dadosmilho' num2str(i)];
    dataStruct.(filename) = load([filename '.txt']);
end



AbsorcaoDados = [dataStruct.dadosmilho1(:,2),dataStruct.dadosmilho2(:,2),dataStruct.dado
ComprimentoOnda = [dataStruct.dadosmilho1(:,1)];
Subtypes = importdata('subtypes.txt',' ');
Tipos = importdata('genes.txt',' ');
```

Figura A1.1 – Código para execução da organização dos dados ópticos por FTIR. Fonte: Própria.

Para a análise das componentes principais - PCA, utilizou o comando “pca”. No comando “pca”, a tabela que contenha os dados ópticos deve estar ordenada com as observações (espécies de milho) representadas pelas linhas da tabela, e as variáveis (comprimentos de onda) representadas pelas colunas da tabela, na qual não deve conter dados com caracteres que não sejam numéricos. A figura A1.2 abaixo informa o código utilizado para obtenção da PCA.

```
[coeff, score, latent, tsquared, explained] = pca(AbsorcaoDados');
x=score(:,1);
y=score(:,2);
figure(2)
gscatter(x,y,Subtypes)
xlabel('PC1 (ValorExplainedPrimeiraCélula)')
ylabel ('PC2 (ValorExplainedSegundaCélula)')
title('Principal Component Analysis')
```

Figura A1.2 – Comando para execução do PCA com os dados ópticos FTIR. Fonte: Própria.

Utilizando o App *Classification Learner* presente no Matlab®, se foi feito o treinamento das ferramentas matemáticas de classificação Máquinas de Vetores de Suporte – SVM, e *k*-vizinhos próximos – kNN. A figura A1.3 abaixo mostra o código utilizado para montar e informar a tabela de dados que o classificador necessita para seu funcionamento.

```
AbsorcaoTransposta = AbsorcaoDados';
AbsorcaoTransTable = array2table(AbsorcaoTransposta);
AbsorcaoTransTable(:,end+1) = Tipos;
AbsorcaoTransTable.Properties.VariableNames(end) = {'Caracteristica_Genetica'};
```

Figura A1.3 – Comando para execução das ferramentas de classificação SVM e kNN para FTIR. Fonte: Própria.

Após o treinamento dos classificadores SVM e kNN com os dados ópticos por FTIR, a acurácia e tempo de treinamento de cada classificador foi obtida, conforme informam as tabelas A1.1 e A1.2 abaixo.

Tabela A1.1 – Dados do Treinamento para SVM para Absorção FTIR

Tipo do Classificador	Método de Validação	Função *Kernel*	Método Multiclasse	Tempo de Treinamento (min:seg)	Acurácia (%)
SVM	5-fold Cross Validation	Linear	one-against-one	01:17	58,8
SVM	5-fold Cross Validation	Quadratica	one-against-one	00:22	58,8
SVM	5-fold Cross Validation	Cúbica	one-against-one	00:22	41,2
SVM	5-fold Cross Validation	*Fine* Gaussiana	one-against-one	00:22	64,7
SVM	5-fold Cross Validation	*Medium* Gaussiana	one-against-one	00:22	64,7
SVM	5-fold Cross Validation	*Coarse* Gaussiana	one-against-one	00:22	64,7

Tabela A1.2 – Dados do Treinamento para kNN para Absorção FTIR

Tipo do Classificador	Método de Validação	Número de Vizinhos Próximos	Tipo de Distância	Tempo de Treinamento (min:seg)	Acurácia (%)
k-NN	5-fold Cross Validation	1	Métrica Euclidiana	00:23	52,9
k-NN	5-fold Cross Validation	10	Métrica Euclidiana	00:23	64,7
k-NN	5-fold Cross Validation	100	Métrica Euclidiana	00:23	64,7
k-NN	5-fold Cross Validation	10	Métrica Cosseno	00:23	52,9
k-NN	5-fold Cross Validation	10	Cúbica (Mnkowski)	00:23	64,7
k-NN	5-fold Cross Validation	10	Inverso do Quadrado do Peso	00:23	52,9

- **A2 – EMISSÃO POR PLASMA INDUZIDO POR LASER – LIBS faixa UV**

As espécies de milho da tabela A1 sofreram análise óptica por LIBS na faixa Ultravioleta, obtendo os dados numéricos do comprimento de onda incidente na amostra, e a sua respectiva intensidade.

Utilizando o software MatLab® para a extração desses dados numéricos, e também para o tratamento e manipulação dos dados, obteve uma primeira tabela, na qual as linhas eram representadas pelas variáveis (comprimentos de onda), e as colunas pelas observações (espécies de milho). Os dados ópticos de cada espécie de milho foram armazenados, cada um em uma célula diferente da matriz nomeada de “MatrizAbsorbancia{X}”, na qual cada “X” “MatrizAbsorbancia” representa os dados ópticos de várias medidas de cada espécie de milho. Concatenou os dados horizontalmente de cada matriz, obtendo uma segunda tabela de dimensões 35.597 linhas x 263 colunas, nomeada de “AbsorcaoDados”.

A figura A2.1 abaixo demonstra a parte do código responsável pela execução do procedimento descrito acima.

```
for m=1:(length(files)-2);
    cd(files(m+2).name);
    arquivospasta = dir('*.esf');
    for n=1:(length(arquivospasta));
        n
        clear b
        data = importdata(arquivospasta(n).name,';',40);
        b = sortrows(data.data,1);
        MatrizAbsorbancia{m}(:,n) = b(:,3);
    end
    MatrizAbsorbanciaMedia{m} = mean(MatrizAbsorbancia{m}(:,n),2);
    cd ..
end
cd ..

Subtypes = importdata('Subtypes.txt',' ');
Tipos = importdata('Tipos.txt',' ');
ComprimentoOnda = b(:,1);
AbsorcaoDados = horzcat(MatrizAbsorbancia{1},MatrizAbsorbancia{2},MatrizAbsorbancia{3},
MatrizAbsorbancia{4},MatrizAbsorbancia{5},MatrizAbsorbancia{6});
```

Figura A2.1 – Código para execução da organização dos dados ópticos por LIBS faixa UV. Fonte: Própria.

Para a análise das componentes principais - PCA, utilizou o comando “pca”. No comando “pca”, a tabela que contenha os dados ópticos deve estar ordenada com as observações (espécies de milho) representadas pelas linhas da tabela, e as variáveis (comprimentos de onda) representadas pelas colunas da tabela, na qual não deve conter dados com caracteres que não sejam

numéricos. A figura A2.2 abaixo informa o código utilizado para obtenção da PCA.

```
[coeff, score, latent, tsquared, explained] = pca(AbsorcaoDados');
x=score(:,1);
y=score(:,2);
figure(2)
gscatter(x,y,Subtypes)
xlabel('PC1 (ValorExplainedPrimeiraCélula)')
ylabel ('PC2 (ValorExplainedSegundaCélula)')
title('Principal Component Analysis')
```

Figura A2.2 – Comando para execução do PCA com os dados ópticos LIBS faixa UV. Fonte: Própria.

Utilizando o App *Classification Learner* presente no Matlab®, se foi feito o treinamento das ferramentas matemáticas de classificação Máquinas de Vetores de Suporte – SVM, e *k*-vizinhos próximos – kNN. A figura A2.3 abaixo mostra o código utilizado para montar e informar a tabela de dados que o classificador necessita para seu funcionamento.

```
AbsorcaoTransposta = AbsorcaoDados';
AbsorcaoTransTable = array2table(AbsorcaoTransposta);
AbsorcaoTransTable(:,end+1) = Tipos;
AbsorcaoTransTable.Properties.VariableNames(end) = {'Caracteristica_Genetica'};
```

Figura A2.3 – Comando para execução das ferramentas de classificação SVM e kNN para LIBS faixa UV. Fonte: Própria.

Após o treinamento dos classificadores SVM e kNN com os dados ópticos por LIBS faixa UV, a acurácia e tempo de treinamento de cada classificador foi obtida, conforme informam as tabelas A2.1 e A2.2 abaixo.

Tabela A2.1 – Dados do Treinamento para SVM para LIBS faixa UV

Tipo do Classificador	Método de Validação	Função *Kernel*	Método Multiclasse	Tempo de Treinamento (min:seg)	Acurácia (%)
SVM	5-fold Cross Validation	Linear	one-against-one	30:00	83,3
SVM	5-fold Cross Validation	Quadratica	one-against-one	30:28	82,1
SVM	5-fold Cross Validation	Cúbica	one-against-one	30:15	83,7
SVM	5-fold Cross Validation	*Fine* Gaussiana	one-against-one	31:55	66,2
SVM	5-fold Cross Validation	*Medium* Gaussiana	one-against-one	31:36	66,2
SVM	5-fold Cross Validation	*Coarse* Gaussiana	one-against-one	31:25	66,2

Tabela A2.2 – Dados do Treinamento para kNN para LIBS faixa UV

Tipo do Classificador	Método de Validação	Número de Vizinhos Próximos	Tipo de Distância	Tempo de Treinamento (min:seg)	Acurácia (%)
k-NN	5-fold Cross Validation	1	Métrica Euclidiana	31:01	72,2
k-NN	5-fold Cross Validation	10	Métrica Euclidiana	30:53	82,1
k-NN	5-fold Cross Validation	100	Métrica Euclidiana	31:57	70,3
k-NN	5-fold Cross Validation	10	Métrica Cosseno	30:30	81
k-NN	5-fold Cross Validation	10	Cúbica (Mnkowski)	30:45	81,7
k-NN	5-fold Cross Validation	10	Inverso do Quadrado do Peso	30:27	81,4

- **A3 – EMISSÃO POR PLASMA INDUZIDO POR LASER – LIBS faixa UV-Vis**

As espécies de milho da tabela A1 sofreram análise óptica por LIBS na faixa Ultravioleta e Visível, obtendo os dados numéricos do comprimento de onda incidente na amostra, e a sua respectiva intensidade.

Utilizando o software MatLab® para a extração desses dados numéricos, e também para o tratamento e manipulação dos dados, obteve uma primeira tabela, na qual as linhas eram representadas pelas variáveis (comprimentos de onda), e as colunas pelas observações (espécies de milho). Os dados ópticos de cada espécie de milho foram armazenados, cada um em uma célula diferente da matriz nomeada de “MatrizAbsorbancia{X}”, na qual cada “X” “MatrizAbsorbancia” representa os dados ópticos de várias medidas de cada espécie de milho. Concatenou os dados horizontalmente de cada matriz, obtendo uma segunda tabela de dimensões 38.468 linhas x 306 colunas, nomeada de “AbsorcaoDados”.

A figura A3.1 abaixo demonstra a parte do código responsável pela execução do procedimento descrito acima.

```
for m=1:(length(files)-2);
    cd(files(m+2).name);
    arquivospasta = dir('*.esf');
    for n=1:(length(arquivospasta));
        n
        clear b
        data = importdata(arquivospasta(n).name,';',40);
        b = sortrows(data.data,1);
        MatrizAbsorbancia{m}(:,n) = b(:,3);
    end
    MatrizAbsorbanciaMedia{m} = mean(MatrizAbsorbancia{m}(:,n),2);
    cd ..
end
cd ..

Subtypes = importdata('Subtypes.txt',' ');
Tipos = importdata('Tipos.txt',' ');
ComprimentoOnda = b(:,1);

AbsorcaoDados = horzcat(MatrizAbsorbancia{1},MatrizAbsorbancia{2},MatrizAbsorbancia{3},
MatrizAbsorbancia{4},MatrizAbsorbancia{5},MatrizAbsorbancia{6});
```

Figura A3.1 – Código para execução da organização dos dados ópticos por LIBS faixa UV-Vis. Fonte: Própria.

Para a análise das componentes principais - PCA, utilizou o comando “pca”. No comando “pca”, a tabela que contenha os dados ópticos deve estar ordenada com as observações (espécies de milho) representadas pelas linhas da tabela, e as variáveis (comprimentos de onda) representadas pelas colunas da tabela, na qual não deve conter dados com caracteres que não sejam

numéricos. A figura A3.2 abaixo informa o código utilizado para obtenção da PCA.

```
[coeff, score, latent, tsquared, explained] = pca(AbsorcaoDados');
x=score(:,1);
y=score(:,2);
figure(2)
gscatter(x,y,Subtypes)
xlabel('PC1 (ValorExplainedPrimeiraCélula)')
ylabel ('PC2 (ValorExplainedSegundaCélula)')
title('Principal Component Analysis')
```

Figura A3.2 – Comando para execução do PCA com os dados ópticos LIBS faixa UV-Vis. Fonte: Própria.

Utilizando o App *Classification Learner* presente no Matlab®, se foi feito o treinamento das ferramentas matemáticas de classificação Máquinas de Vetores de Suporte – SVM, e *k*-vizinhos próximos – kNN. A figura A3.3 abaixo mostra o código utilizado para montar e informar a tabela de dados que o classificador necessita para seu funcionamento.

```
AbsorcaoTransposta = AbsorcaoDados';
AbsorcaoTransTable = array2table(AbsorcaoTransposta);
AbsorcaoTransTable(:,end+1) = Tipos;
AbsorcaoTransTable.Properties.VariableNames(end) = {'Caracteristica_Genetica'};
```

Figura A3.3 – Comando para execução das ferramentas de classificação SVM e kNN para LIBS faixa UV. Fonte: Própria.

Após o treinamento dos classificadores SVM e kNN com os dados ópticos por LIBS faixa UV, a acurácia e tempo de treinamento de cada classificador foi obtida, conforme informam as tabelas A3.1 e A3.2 abaixo.

Tabela A3.1 – Dados do Treinamento para SVM para LIBS faixa UV

Tipo do Classificador	Método de Validação	Função *Kernel*	Método Multiclasse	Tempo de Treinamento (min:seg)	Acurácia (%)
SVM	5-fold Cross Validation	Linear	one-against-one	40:08	72,9
SVM	5-fold Cross Validation	Quadratica	one-against-one	37:38	80,7
SVM	5-fold Cross Validation	Cúbica	one-against-one	36:52	81
SVM	5-fold Cross Validation	*Fine* Gaussiana	one-against-one	36:55	71,2
SVM	5-fold Cross Validation	*Medium* Gaussiana	one-against-one	36:43	71,2
SVM	5-fold Cross Validation	*Coarse* Gaussiana	one-against-one	36:50	71,2

Tabela A3.2 – Dados do Treinamento para kNN para LIBS faixa UV

Tipo do Classificador	Método de Validação	Número de Vizinhos Próximos	Tipo de Distância	Tempo de Treinamento (min:seg)	Acurácia (%)
k-NN	5-fold Cross Validation	1	Métrica Euclidiana	36:55	65
k-NN	5-fold Cross Validation	10	Métrica Euclidiana	37:00	71,6
k-NN	5-fold Cross Validation	100	Métrica Euclidiana	37:05	71,2
k-NN	5-fold Cross Validation	10	Métrica Cosseno	37:09	77,8
k-NN	5-fold Cross Validation	10	Cúbica (Mnkowski)	37:09	74,2
k-NN	5-fold Cross Validation	10	Inverso do Quadrado do Peso	37:04	71,6