CRISTINA SILVA DIAS

COMPARAÇÃO DE MÉTODOS DE CLASSIFICAÇÃO EM DADOS DE ESPECTROSCOPIA NIR

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae.*

Orientador: Luiz Alexandre Peternelli

VIÇOSA - MINAS GERAIS
2020

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
D541c
2020

Dias, Cristina Silva, 1991-
Comparação de métodos de classificação em dados de espectroscopia NIR / Cristina Silva Dias. – Viçosa, MG, 2020.
44 f. : il. (algumas color.) ; 29 cm.

Orientador: Luiz Alexandre Peternelli.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.41-44.

1. Cana-de-açúcar - Métodos estatísticos. 2. Máquinas de vetor suporte. 3. Espectroscopia de infravermelho.
I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519

CRISTINA SILVA DIAS

COMPARAÇÃO DE MÉTODOS DE CLASSIFICAÇÃO EM DADOS DE ESPECTROSCOPIA NIR

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae.*

APROVADA: 25 de junho de 2020.

Assentimento:

Cristina Silva Dias
Cristina Silva Dias
Autora

Luiz Alexandre Peternelli
Orientador

AGRADECIMENTOS

Primeiramente agradeço a Deus e a Nossa Senhora por abençoarem o meu caminho e me proporcionarem força e coragem para vencer os obstáculos.

Aos meus pais, pelo amor incondicional e por não medirem esforços para que eu possa realizar os meus sonhos. Aos meus irmãos pelo carinho, companheirismo, paciência e incentivo.

Às minhas avós pelos exemplos de vida e orações. A toda a minha família, tios, tias, primos, primas, pela torcida e por toda a ajuda durante minha trajetória.

Aos amigos que encontrei na estatística e que levarei por toda minha vida. Em especial: Pedro, Fred e Taiana, por sempre estarem ao meu lado, pelas risadas, carinho, amizade e por proporcionarem momentos inesquecíveis.

Aos amigos Samuel, Josiane, Wanessa, Geise, Bárbara, Gabriela e Poliana pelo apoio e incentivo.

Ao professor Luiz Alexandre Peternellli, pela orientação, confiança, paciência, incentivo, amizade e por todos os ensinamentos desde a iniciação científica.

Aos pesquisadores D. Sc.s: Édimo Fernando Alves Moreira, Jussara Valente Roque, Reinaldo Francisco Teófilo e José Ivo Ribeiro Júnior por aceitarem participar da minha banca de defesa e por todas as contribuições nesse trabalho.

Ao Felipe Guzzo pelo suporte nos scripts e ao Mateus Teles por ter cedido os dados.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, pela oportunidade e pelo ensino de qualidade.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

RESUMO

DIAS, Cristina Silva, M.Sc., Universidade Federal de Viçosa, junho de 2020. **Comparação de métodos de classificação em dados de espectroscopia NIR.** Orientador: Luiz Alexandre Peternelli.

A crescente demanda de biomassa para produção de energia e etanol de segunda geração tem impulsionado a seleção de cultivares de cana-de-açúcar com maiores teores de fibra e sacarose aparente. Nesse sentido, torna-se importante buscar métodos de classificação aliados a técnica de espectroscopia do infravermelho próximo (NIR) para facilitar a seleção desses indivíduos. O objetivo desse trabalho é comparar os métodos de classificação: Análise Discriminante por Quadrados Mínimos Parciais (PLS-DA), Máquinas de Vetores de Suporte (SVM) e Florestas Aleatórias (RF) para verificar qual deles apresenta um melhor desempenho para a classificação dessas propriedades a partir de dados de espectroscopia NIR. Foi utilizado um conjunto de dados NIR composto por 460 amostras para classificação de fibra (FIB) e sacarose aparente (PC). A análise foi realizada em duas etapas. Na primeira etapa o conjunto de dados foi separado em conjunto treino e conjunto teste via algoritmo Kernard-Stone para a escolha dos pré-tratamentos. Na segunda etapa foi utilizado o pré-tratamento selecionado para cada método, separando novamente o conjunto original (460 amostras) em conjunto de treino e conjunto de teste, de forma aleatória com 10 repetições. Após todos os procedimentos, os resultados obtidos na comparação dos métodos mostraram que o PLS-DA e o SVM não apresentam diferença significativa ($p \geq 0,05$) e ambos diferiram do RF para a classificação das propriedades %PC e %FIB ($p < 0,05$). Foram avaliados os parâmetros de erro de classificação, sensibilidade e especificidade. Para todos esses parâmetros o PLS-DA e o SVM foram mais satisfatórios que o RF, pois apresentaram menores valores de erro de classificação e maiores valores para sensibilidade e especificidade podendo, assim, serem considerados métodos eficazes para classificação do dado de espectroscopia NIR utilizados nesse trabalho.

Palavras-chave: Cana-de-açúcar. Máquinas de Vetor Suporte. Floresta Aleatória. Análise Discriminante por Quadrados Mínimos Parciais. Infravermelho Próximo.

ABSTRACT

DIAS, Cristina Silva, M.Sc., Universidade Federal de Viçosa, June, 2020. **Comparison of classification methods in NIR spectroscopy data.** Adviser: Luiz Alexandre Peternelli.

The growing demand for biomass for power generation and second-generation ethanol has driven the selection of sugarcane cultivars with higher fiber and apparent sucrose levels. In this sense, it is crucial to seek classification methods combined with near-infrared spectroscopy (NIR) to facilitate the desired selection. The objective of this work is to compare the classification methods: Discriminant Analysis by Partial Least Squares (PLS-DA), Support Vector Machines (SVM), and Random Forests (RF) to see which one performs better for the classification of these properties from NIR spectroscopy data. A set of NIR data composed of 460 samples was used, classified as fiber (FIB) and apparent sucrose (PC). We performed the analysis in two stages. In the first stage, the data set was separated into training and test sets via the Kernard-Stone algorithm to choose pre-treatments. The pre-treatment selected for each method was used in the second stage, separating the original set (460 samples) again into a training set and test set, randomly with ten repetitions. After all the procedures, the results obtained in the comparison of the methods indicated that PLS-DA and SVM do not present significant differences ($p \geq 0.05$) and both differed from RF for the classification of properties %PC and %FIB ($p < 0.05$). We evaluated the classification error, sensitivity, and specificity among these methods. PLS-DA and SVM were more satisfactory for all these parameters than RF since the former presented lower classification errors and higher values of sensitivity and specificity. Therefore, these methods can be considered useful for classifying the NIR spectroscopy data used in this work.

Keywords: Sugarcane. Support Vector Machines. Random Forest. Partial Least Squares Discriminant Analysis. Near Infrared.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

A biomassa da cana-de-açúcar, tem se tornado uma alternativa sustentável e segura para a produção energética do país (TROMBETA e FILHO, 2017). Por meio dela, pode-se obter energia elétrica e biocombustíveis como biodiesel e etanol que são usados em substituição a derivados do petróleo como óleo diesel e gasolina (ANEEL, 2008). Consequentemente, o desenvolvimento de novas variedades de cana-de-açúcar com características de biomassa favoráveis, é de extrema importância no melhoramento da cana-de-açúcar (BARBOSA et al.,2012; SANTANA e JOSÉ, 2017). A composição da matéria-prima de biomassa é um aspecto primordial a ser considerado no processo de conversão de energia, informações sobre o teor de açúcar e fibra na biomassa em uma determinada população auxiliam na estimação do rendimento teórico de etanol e, portanto, maximiza o potencial de produção de biocombustível (HOANG et al., 2017). Nesse cenário, cultivares de cana-de-açúcar com maiores teores de fibra e com maiores quantidades de sacarose são altamente desejáveis e podem trazer mais sustentabilidade econômica e ambiental para a produção enérgica do país.

Para a determinação da porcentagem do teor de sacarose aparente e teor de fibra na cana-de-açúcar, as técnicas tradicionais demandam tempo e muitas das vezes tornam-se inviáveis. Dessa forma, o uso da espectroscopia no infravermelho próximo (NIR) associada a métodos estatísticos multivariados são excelentes alternativas para determinação das propriedades de interesse e para seleção de novas variedades de cana-de-açúcar. O NIR apresenta grande possibilidade de aplicação devido a sua facilidade de utilização, rapidez e exatidão (PASQUINI, 2018).

No melhoramento da cana-de-açúcar, tem-se grande número de indivíduos a serem avaliados no processo de seleção, processo este que demanda tempo de mão de obra no campo (PETERNELLI et al., 2017). Visando a seleção de indivíduos com mais biomassa, é de suma importância buscar métodos mais eficazes e de fácil implementação na classificação desses indivíduos como “selecionados” ou “não selecionados. Isso é possível por meio dos métodos de classificação, no qual cada uma das amostras é descrita por um conjunto de medidas experimentais, e a classificação das mesmas é dada de acordo com as propriedades de interesse (FERREIRA, 2015).

Segundo Ferreira (2015) para esse processo de classificação (análise supervisionada) é selecionado uma série de amostras representativas de cada classe. Esse conjunto de amostras determina o “conjunto de treino”. Posteriormente, as informações do conjunto de treino são

usadas para a construção do modelo ou de uma regra de classificação. Para a verificação da predição (ou classificação) do modelo construído, utiliza-se um conjunto de amostras externas, denominadas “conjunto de teste”. Baseado no resultado obtido, o modelo ajustado poderá ser então utilizado para classificar novas amostras.

Um método de classificação que apresenta grande eficiência é o método de Análise Discriminante por Quadrados Mínimos Parciais (PLS-DA) baseado na regressão PLS (*Partial Least Squares*). Este é um método atraente devido a sua capacidade de analisar, com sucesso, dados altamente colineares e ruidosos (BARKER e RAYENS, 2003). O método Máquinas de Vetores de Suporte (SVM) também vem apresentando destaque para problemas de classificação, pois apresenta robustez diante de dados com alta dimensão e uma boa capacidade de generalização (LORENA e CARVALHO, 2007; TORRES e REVERÓN, 2014). Uma alternativa que também está sendo amplamente utilizada em diversas áreas para classificação é o método Florestas Aleatórias (RF), por apresentar grande flexibilidade e capacidade de lidar com dados com características não lineares e com alta dimensionalidade (MARTINS et al., 2012; CHEMURA; MUTANGA; DUBE, 2017).

A avaliação constante de novas e modernas alternativas para predições contribuirá para a eficiência do melhoramento genético da cana de açúcar, tornando-a ainda mais promissora para o mercado. Diante disso, objetiva-se com esse trabalho, realizar análise comparativa dos métodos PLS-DA, SVM e RF utilizados para classificação de amostras provenientes de dado de espectroscopia NIR.

2. REFERENCIAL TEÓRICO

2.1 Cana-de-açúcar

A cana-de-açúcar desempenha um grande papel para a economia brasileira, pois além da produção do açúcar apresenta alta potencialidade na produção de energia renovável (CONAB, 2019).

Atualmente o Brasil apresenta destaque no ranking mundial na produção de cana-de-açúcar. O Brasil é o primeiro também na produção e na exportação de açúcar e o segundo maior produtor e exportador de etanol (AGRIMEC, 2019).

O bagaço da cana-de-açúcar é um resíduo utilizado em grande escala pela agroindústria brasileira. Uma tonelada de cana moída gera em torno 250 kg de bagaço, que é equivalente a

560.000 kcal de energia. Essa mesma quantidade de cana produz, aproximadamente, 70 litros de etanol, o que proporciona 392.000 kcal de energia (BONASSA et al., 2015).

As fibras do bagaço da cana-de-açúcar contêm, como principais componentes, 32% a 48% de celulose, 19% a 24% de hemicelulose, 23% a 32% de lignina e uma pequena quantidade de cinzas e extrativos (ROCHA et al., 2012; SOUZA et al., 2013).

2.2 Espectroscopia por infravermelho próximo (NIR)

A espectroscopia no infravermelho próximo é uma técnica espectroscópica que utiliza a região do infravermelho próximo do espectro eletromagnético com comprimento de onda de 780 a 2500nm (PASQUINI, 2018).

Os método de espectroscopia NIR são métodos rápidos, não destrutivos, não invasivos, com alta penetração do feixe de radiação de sondagem, adequado para uso em linha, aplicação quase universal (qualquer molécula contendo ligações CH, NH, SH ou OH), com exigências mínimas de preparação de amostras (PASQUINI, 2018).

Esta é uma técnica que está sendo bastante utilizada para resoluções de problemas nas indústrias alimentícias, nos setores farmacêuticos, nos ramos agrícolas e em diversas outras áreas (ALVES e POPPI, 2013).

2.3 Pré-tratamentos

Consideremos os dados organizados em forma de uma matriz **X** com I linhas (cada linha representa uma amostra i) e J colunas (cada coluna corresponde a uma variável j). Sobre essa matriz **X** são realizadas as aplicações das técnicas de pré-tratamentos, cujo objetivo é reduzir as variações sistemáticas ou aleatórias que podem mascarar informações relevantes e interferir nos resultados finais apresentados na matriz (FERREIRA, 2015).

Existem dois tipos de pré-tratamentos: Transformação, quando o pré-tratamento é aplicado nos espectros de cada amostra (nas linhas da matriz **X**) e o Pré-processamento, no qual os pré-tratamentos são feitos nas variáveis (nas colunas da matriz **X**) (FERREIRA, 2015).

O pré-processamento utilizado nesse trabalho foi:

Centrar na média: Calcula-se o valor médio de cada coluna da matriz de dados e, posteriormente, subtrai-se esse valor médio de cada um dos valores da respectiva coluna, de

acordo com a expressão:$x_{ij(cm)} = x_{ij} - \overline{x_j}$. Após esta etapa, todas as médias serão iguais a zero e as variações são espalhadas em torno de zero (BURNS e CIURCZAK, 2007).

As transformações aplicadas no conjunto de dados nesse trabalho foram:

Correção Multiplicativa de Sinal (*Multiplicative Scatter Correction* - MSC): Esse tratamento é aplicado para a correção da linha de base, corrige os efeitos dos fenômenos físicos, como a dispersão da luz, causado pela falta de homogeneidade das amostras e as variações causadas pelas diferenças no percurso óptico das amostras (FERREIRA, 2015). Para realizar essa correção, utiliza-se a linearização do espectro médio $\mathbf{x_m}$, em que cada espectro pode ser escrito como função linear do espectro médio, da seguinte forma: $x_i = a_i\mathbf{1} + b_i\mathbf{x_m}$. Os coeficientes a_i e b_i são estimados pelo método dos mínimos quadrados, fazendo-se a regressão de cada espectro i no espectro médio $\boldsymbol{x_m}$. Assim, a correção é expressa da seguinte forma: $x_{i\ (msc)} = \frac{x_{i-a_i}}{b_i}$.

Primeira Derivada: A primeira derivada tem como finalidade corrigir um deslocamento da linha de base (no eixo das ordenadas) causado por um problema instrumental ou de amostragem, em que pode deslocar um espectro como um todo, de um valor positivo ou negativo em relação ao zero de absorbância (FERREIRA, 2015). O algoritmo mais utilizado é o de Savitsky-Golay. A primeira derivada pode ser obtida pela equação: $2\delta\frac{dA}{d\lambda}(\lambda_j) \cong \Delta A(\lambda_j) = A_{\lambda_j+\delta} - A_{\lambda_j-\delta}$.

Segunda Derivada: A segunda derivada, nos informa sobre a curvatura ou a concavidade de uma curva. Por meio desse pré-tratamento é possível corrigir a inclinação da linha de base (FERREIRA, 2015). A segunda derivada pode ser obtida pela equação: $(2\delta)^2\frac{d^2A}{d\lambda^2}(\lambda_j) \cong \Delta^2 A(\lambda_j) = A_{\lambda_j+\delta} + A_{\lambda_j-\delta} - 2A_{\lambda_j}$.

Padrão Normal de Variação (*Standard Normal Variate*- SNV): O SNV, assim como o MSC, tem como finalidade corrigir problemas como espalhamento de luz. Esse método autoescala cada linha da matriz original de dados. O valor médio de cada linha é subtraído de todos os valores da respectiva linha, depois são divididos pelos respectivos desvios-padrão (FERREIRA, 2015). O SNV é feito por meio da equação: $\mathbf{x}_{i(SNV)} = \frac{(\mathbf{x}_i - \overline{\mathbf{x}_i})}{s_i}$, onde $\overline{x_i} = \frac{1}{J}\sum_{j=1}^{J} x_{ij}$ e $s_i = \sqrt{\sum_{j=1}^{J}(x_{ij} - \overline{x_i})^2}$.

2.4 Análise Discriminante por Quadrados Mínimos Parciais (*Partial Least Squares Discriminant Analysis* – PLS-DA)

O método PLS-DA está fundamentado no método de regressão por Quadrados Mínimos Parciais (PLS) e pode ser considerado uma extensão da análise linear discriminante (LDA) (BARKER; RAYENS, 2003). Assim como no PLS, no método PLS-DA as variáveis da matriz $\mathbf{X}_{IJ}$ fazem relação com a matriz $\mathbf{Y}_{IQ}$ ou com um vetor $\mathbf{y}_{I1}$. No entanto, a matriz **Y,** ou vetor **y,** apresenta variáveis categóricas (discreta) (WOLD et al., 2001).

Nesse trabalho utilizamos o algoritmo SIMPLS proposto por Jong (1993). Esse algoritmo é realizado da seguinte forma:

Primeiramente compute o valor de **s:**

$$\mathbf{s} = \mathbf{X^t y}$$

Para $i = 1, 2, \cdots, a$ variáveis latentes, calcule **r** (fator-peso para **X**), **t** (escores para X), **q** (pesos para **y**) e **p** (pesos para **X**):

$\mathbf{r}_i = \mathbf{s}$

$\mathbf{t_i} = \mathbf{X}\mathbf{r}_i$

$\mathbf{t_i} = \mathbf{t_i}/\|\mathbf{t_i}\|$

$\mathbf{r_i} = \mathbf{r_i}/\|\mathbf{r_i}\|$

$\mathbf{P_i} = \mathbf{X^t t_i}$

$q_i = \mathbf{y^t t_i}$

Armazene $\mathbf{R} = (\mathbf{r_1}, \cdots, \mathbf{r_a}), \mathbf{T} = (\mathbf{t_1}, \cdots, \mathbf{t_a}), \mathbf{P} = (\mathbf{p_1}, \cdots, \mathbf{p_a})$ e $\mathbf{q} = (q_1, \cdots, q_a)$ e projete em um espaço ortogonal a **P**:

$$\mathbf{s} = \mathbf{s} - \mathbf{P}(\mathbf{P^t P})^{-1}\mathbf{s}$$

Após a etapa de projeção **r, t, P** e q serão calculados para a próxima variável latente até atingir $i = a$. A última etapa calcula o vetor de regressão:

$$\mathbf{b} = \mathbf{R}q$$

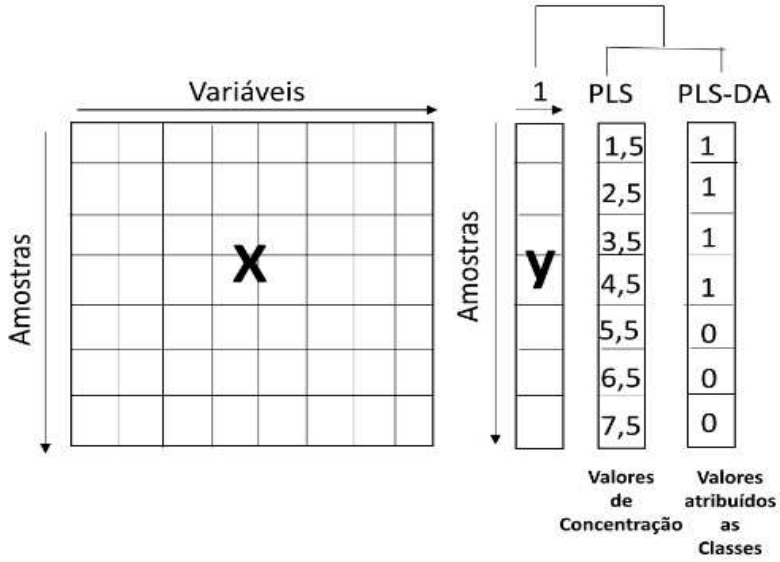
Como o PLS maximiza a relação entre as variáveis dependentes e os escores, temos que as variáveis latentes (VL) representam as direções que melhor separam as classes.

A matriz **Y** assume valores binários, os quais indicam se a amostra pertence ou não à classe de interesse (BARKER e RAYENS, 2003), ou seja, se temos H classes envolvidas ($h = 1,2, \cdots, H$). Para a primeira classe, correspondente a h =1, o número 1 é atribuído às amostras dessa classe, e 0 é atribuído a todas as outras amostras, não pertencentes a essa classe. Para $h =$ 2, são as amostras dessa classe que são atribuídas o número 1 e as demais por 0, e assim

sucessivamente até que cada uma das classes tenha sido representada pelo número 1 (FERREIRA, 2015).

Os valores preditos pelo modelo PLS-DA serão, idealmente, os valores 0 e 1. No entanto, na maioria das vezes ele não preverá 0 ou 1 perfeitamente, acarretando o uso de uma aproximação para esses valores. Para o processo de classificação é calculado um valor limite (*threshold*) entre os valores preditos. Dessa forma, valores preditos acima deste *threshold* indicam que a amostra pertence à classe de interesse, enquanto valores preditos abaixo deste *threshold* indicam que a amostra não pertence à classe (PATACA, 2006).

Figura 1 - Representação esquemática dos métodos PLS e PLS-DA e suas diferenças.



Fonte: Adaptado de LOPES (2015).

2.5 Máquinas de vetor suporte (*Suporte Vector Machines* - SVM)

Para facilitar a abordagem das SVMs, consideraremos o caso das SVMs lineares que consiste em separar duas classes. As SVMs lineares com margens rígidas definem fronteiras lineares a partir de dados linearmente separáveis por meio de um hiperplano de separação. Seja T um conjunto de treinamento com *N* dados, $x_i \in$ X e $y_i \in$ Y em que X constitui o espaço dos dados e **Y** = {-1,1}. T é linearmente separável se é possível separar os dados por um hiperplano. Um hiperplano de separação ideal separa as duas classes e maximiza a distância até o ponto mais próximo de qualquer classe (VAPNIK, 1995). Geometricamente, a margem corresponde

à menor distância entre os pontos de dados (amostras) mais próximos a um ponto no hiperplano (JAMES et al., 2013).

A equação de um hiperplano é dada por:

$$f(\mathrm{x}) = \beta_0 + \beta^T \mathrm{x} = 0$$

Em que $\beta^T \mathrm{x}$ é o produto escalar entre os vetores **β** e **x**, **β** ∈ X é o vetor normal ao hiperplano descrito e $\frac{\beta_0}{\|\beta\|}$ é a distância entre o hiperplano e a origem com $\beta_0 \in \Re$.

Essa equação divide o espaço X em duas regiões: $\beta_0 + \beta^T \mathrm{x} \geq 0$ se $y_i = 1$ e $\beta_0 + \beta^T \mathrm{x} \leq 0$ se $y_i = -1$. A obtenção das classificações pode ser dada por meio de uma função sinal $g(\mathrm{x}) = sng\big(f(\mathrm{x})\big)$ conforme apresentada abaixo:

$$g(\mathrm{x}) = sng\big(f(\mathrm{x})\big) = \begin{cases} +1 \; se \; \beta_0 + \beta^T \mathrm{x} > 0 \\ -1 \; se \; \beta_0 + \beta^T \mathrm{x} < 0 \end{cases}$$

Dessa forma um hiperplano de separação tem a seguinte propriedade:

$$f(x) = y_i(\beta_0 + \beta^T \mathrm{x}) \geq 0$$

A partir de $f(\mathrm{x})$, é possível obter infinitos hiperplanos equivalentes, pela multiplicação de β^Te β_0 por uma mesma constante. O hiperplano canônico é definido ao conjunto *T* como aquele em que β^T e β_0 são escalados de forma que os exemplos mais próximos ao hiperplano $(\beta_0 + \beta^T \mathrm{x})$ sejam dados por:

$$|(\beta_0 + \beta^T x_i)| = 1$$

que resulta nas inequações: $\begin{cases} y_i(\beta_0 + \beta^T x_i) \geq 1, se \; y_i = 1 \\ y_i(\beta_0 + \beta^T x_i) \leq -1, se \; y_i = 1 \end{cases}$

Estas são reduzidas em:

$y_i(\beta_0 + \beta^T x_i) - 1 \geq 0, \; \forall \; x_i \text{ e } y_i \in T$

Algumas propriedades de álgebra linear serão usadas a fim de simplificar alguns cálculos. Considere dois pontos do hiperplano x_1 e x_2, como eles pertencem ao plano, logo devem satisfazer a equação do plano $f(\mathrm{x}) = \beta_0 + \beta^T \mathrm{x}$. Dessa forma temos:

$f(\mathrm{x}) = \beta_0 + \beta^T \mathrm{x}_1 = 0$ (*) e $f(\mathrm{x}) = \beta_0 + \beta^T \mathrm{x}_2 = 0$ (**)

Subtraindo (*) de (**) segue que:

$$\beta_0 + \beta^T \mathrm{x}_1 - \beta_0 - \beta^T \mathrm{x}_2 = 0 \Rightarrow \beta^T (\mathrm{x}_1 - \mathrm{x}_2) = 0$$

Logo $\beta^* = \frac{\beta}{\|\beta\|}$ é um vetor ortogonal ao plano.

Seja x_1 um ponto no hiperplano H_1 e x um ponto no hiperplano separador. Para obtermos o tamanho da margem M, basta projetarmos $(\mathrm{x}_1 - \mathrm{x})$ ao vetor normal $\beta^* = \frac{\beta}{\|\beta\|}$. Essa projeção é dada por:

$$M = (\mathrm{x}_1 - \mathrm{x}) \left(\frac{\beta}{\|\beta\|} \cdot \frac{(\mathrm{x}_1 - \mathrm{x})}{\|\mathrm{x}_1 - \mathrm{x}\|} \right) \Rightarrow M = (\mathrm{x}_1 - \mathrm{x}) \frac{1}{\|\beta\|} \left(\frac{\beta(\mathrm{x}_1 - \mathrm{x})}{\|\mathrm{x}_1 - \mathrm{x}\|} \right)$$

$$\Rightarrow M = (\mathrm{x}_1 - \mathrm{x}) \frac{1}{\|\beta\|} \left(\frac{\beta \mathrm{x}_1 + \beta_0 - (\beta \mathrm{x} + \beta_0)}{\|\mathrm{x}_1 - \mathrm{x}\|} \right)$$

Note que $\beta_0 + \beta^T \mathrm{x}_1 = 1$ e $\beta_0 + \beta^T \mathrm{x} = 0$. Substituindo na equação anterior,

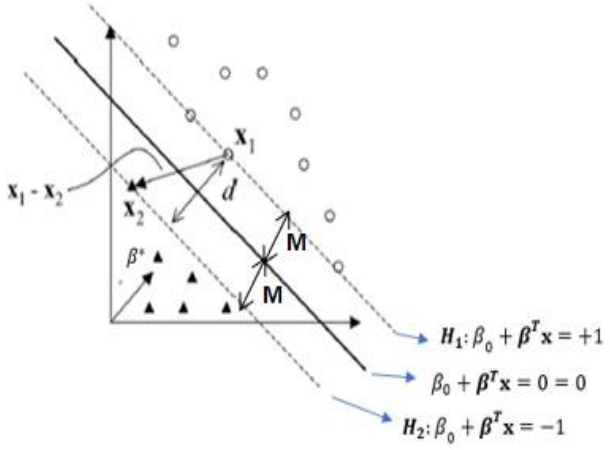
$$M = (\mathrm{x}_1 - \mathrm{x}) \frac{1}{\|\beta\|} \left(\frac{1 - 0}{\|\mathrm{x}_1 - \mathrm{x}\|} \right) = \frac{1}{\|\beta\|} \cdot \frac{(\mathrm{x}_1 - \mathrm{x})}{\|\mathrm{x}_1 - \mathrm{x}\|}$$

Como o interesse é calcular o comprimento do vetor projetado, temos que $\frac{(\mathrm{x}_1 - \mathrm{x})}{\|\mathrm{x}_1 - \mathrm{x}\|} = 1$. Contudo, temos que $M = \frac{1}{\|\beta\|}$ é a distância mínima entre o hiperplano separador e os dados de treinamento.

O objetivo do método é maximizar a margem de separação M em relação ao hiperplano separador dado por $f(x) = \beta_0 + \beta^T \mathrm{x} = 0$, ou seja, $\underset{\beta, \beta_0}{\text{maximizar}} \frac{1}{\|\beta\|}$ o que é equivalente a $\underset{\beta, \beta_0}{\text{minimizar}} \frac{1}{2} \|\beta\|^2$, com a seguinte restrição: $y_i(\beta_0 + \beta^T x_i) - 1 \geq 0, \ \forall\, i = 1,2, \cdots, n$

Essas restrições garantem que os dados de treinamento não estão entre as margens de separação das classes. Por essa razão, a SVM obtida também é chamada de SVM com margens rígidas (LORENA e CARVALHO, 2007).

Figura 2 – Cálculo do tamanho da margem



Fonte: Adaptado de LORENA e CARVALHO (2007).

Esse é um problema de otimização convexo e, portanto, possui apenas um mínimo global. A solução é dada usando uma função Lagrangiana, que conglomera as restrições à função objetivo, relacionadas aos parâmetros α_i, que são designados multiplicadores de Lagrange.

$$L(\beta, \beta_0, \alpha) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N} \alpha_i (y_i(\beta x_i + \beta_0) - 1)$$

Minimizar a função Lagrangiana implica em minimizar β_0 e β e maximizar as variáveis α_i. Definindo as derivadas parciais iguais a zero, obtemos:

$$\frac{\partial L}{\partial\, \beta} = 0 \Rightarrow \beta = \sum_{i=1}^{N} \alpha_i\, y_i x_i$$

$$\frac{\partial L}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i\, y_i = 0$$

Substituindo na equação de Lagrange temos como resultado:

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i\, \alpha_j y_i y_j \left(x_i x_j\right)$$

Com as restrições: $\begin{cases} \alpha_i \geq 0 \\ \sum_{1=1}^{N} \alpha_i y_i = 0 \end{cases}$

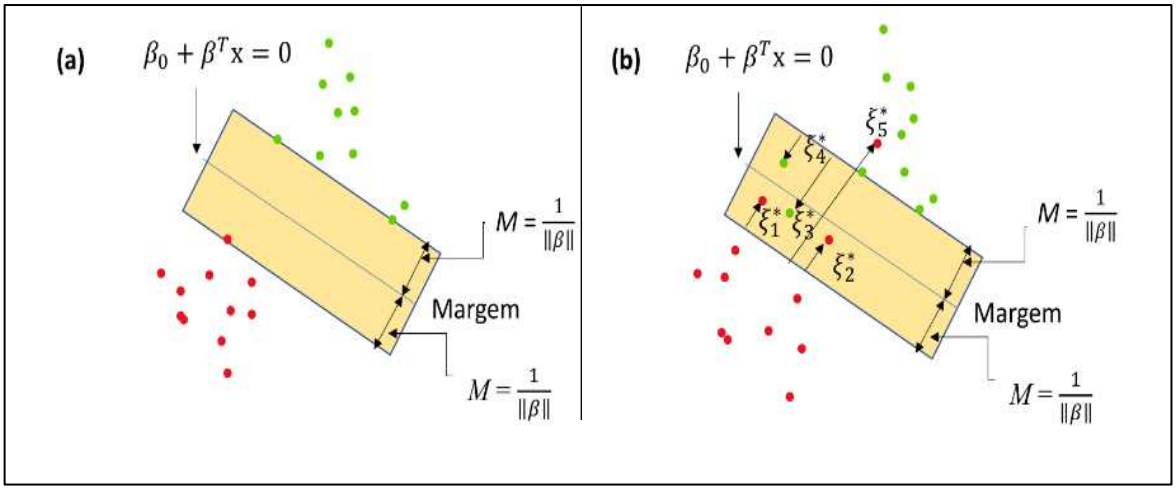
A solução deve atender às condições de Karush–Kuhn–Tucker (KKT), satisfazendo as condições: $\beta = \sum_{i=1}^{N} \alpha_i \, y_i x_i$, $\sum_{i=1}^{N} \alpha_i \, y_i = 0$, $\begin{cases} \alpha_i \geq 0 \\ \sum_{1=1}^{N} \alpha_i y_i = 0 \end{cases}$ e $\alpha_i(y_i(x_i + \beta_0) - 1) = 0 \; \forall i.$

Podemos observar pela condição $\alpha_i(y_i(x_i + \beta_0) - 1) = 0$, que $y_i(x_i + \beta_0) - 1 = 0$ ou $\alpha_i = 0$.

Dessa forma, se $y_i(x_i + \beta_0) - 1 = 0$ e $\alpha_i > 0$, os pontos $x_i$estão sobre a margem, caso contrário, os pontos estarão fora da margem.

Nem sempre os dados são linearmente separáveis, suponhamos que as classes se sobrepõem no espaço de recurso. Uma maneira de lidar com a sobreposição ainda é maximizar a margem, mas permitindo que alguns pontos estejam do lado errado da margem. Esse é o caso das SVMs com margens suaves. Conforme pode ser visto na Figura 3.

Figura 3 – Tipos de margens. Margens rígidas: caso linearmente separável (a); Margens Suaves: Caso inseparável que permite erro (b).



Fonte: Adaptado de HASTIE et al. (2009).

Dessa forma será definida uma variável de folga ξ_i, em que as restrições passam a ser modificadas para todo conjunto de treinamento $i = 1 \cdots N$.

$$y_i(\beta_0 + \beta^T x) \geq 1 - \xi_i, \; \xi_i \geq 0, \forall i = 1, \cdots, N$$

Os erros de treinamento ocorrem quando $\xi_i > 1$. Logo, $\sum \xi_i$, representa o limite dos erros de classificação nos dados de treinamento (HASTIE et al., 2009).

Levando em consideração esse termo, e minimizando os erros sobre os dados de treinamento, a equação $y_i(\beta_0 + \beta^T x) \geq 1 - \xi_i$ torna-se equivalente a:

$$\underset{\beta,\beta_0,\xi}{\text{minimizar}}\frac{1}{2}\|\beta\|^2 + \mathrm{C}\left(\sum_{i=1}^{N}\xi_i\right)$$

$$\text{Sujeito a } \xi_i \geq 0, y_i(\beta_0 + \beta^T x) \geq 1 - \xi_i \; \forall i = 1, \cdots, N$$

Onde o parâmetro C (custo), consiste no parâmetro de regularização, que atribui um peso à minimização dos erros no conjunto de treinamento correspondente à minimização da complexidade do modelo.

Novamente, temos um problema quadrático com restrições lineares de desigualdade e, portanto, trata-se de um problema de otimização convexa (HASTIE et al., 2009). A solução é obtida usando multiplicadores Lagrange:

$L(\beta, \beta_0, \alpha, \mu, \xi) = \frac{1}{2}\|\beta\|^2 + \mathrm{C}(\sum_{i=1}^{N}\xi_i) - \sum_{i=1}^{N}\alpha_i\big(y_i(\beta x_i + \beta_0) - (1 - \xi_i)\big) - \sum_{i=1}^{N}\mu_i\xi_i$

Definindo as derivadas em relação a β, β_0 e ξ iguais a zero, obtemos como resultado:

$$\beta = \sum_{i=1}^{N}\alpha_i y_i x_i$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0$$

$$\alpha_i = C - \mu_i$$

Tomando as contantes $\alpha_i, \mu_i, \xi_i \geq 0 \; \forall i$, substituindo os resultados na aquação lagrangiana, temos:

$$L(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\,\alpha_j y_i y_j\big(x_i x_j\big)$$

Com as restrições:

$$\begin{cases} 0 \leq \alpha_i \leq C, \forall i = 1, \cdots, N \\ \displaystyle\sum_{i=1}^{N}\alpha_i y_i = 0 \end{cases}$$

Adicionalmente, as condições de Karush–Kuhn–Tucker (KKT) incluem as restrições:

$$\alpha_i\big(y_i(x_i + \beta_0) - (1 - \xi_i)\big) = 0$$

$$\mu_i y_i = 0$$

$$\big(y_i(x_i + \beta_0) - (1 - \xi_i)\big) \geq 0$$

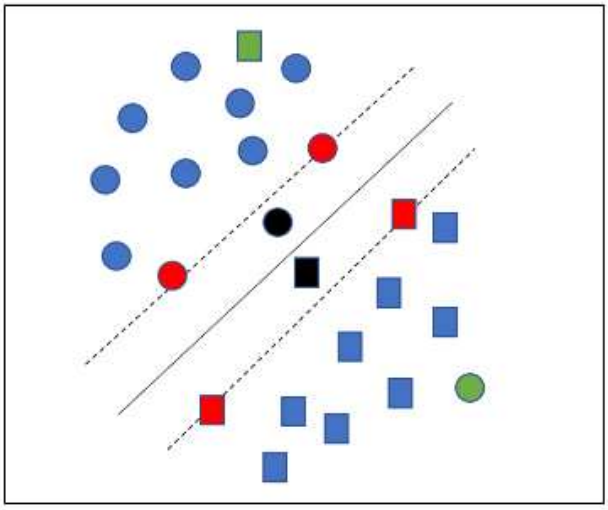
A solução para β continua sendo dado por $\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i \, y_i x_i$. Em que os pontos x_i para os quais $\hat{\alpha}_i > 0$ são considerados vetores de suportes (SVs), são os dados responsáveis pela formação do hiperplano separador. Pode-se apontar diferentes tipos de SVs, se $\hat{\alpha}_i < \mathrm{C}$ então ($\hat{\xi} = 0$), os pontos são corretamente classificados situando-se sobre as margens. Quando $\hat{\alpha}_i =$ C podemos representar três casos:

i. Se $\hat{\xi} > 1$, há erros de classificação
ii. Se $0 \leq \hat{\xi} \leq 1$, os pontos são classificados corretamente, porém situam-se dentro das margens
iii. Se $\hat{\xi} = 0$, significa que os pontos estão sobre a margem.

Para $\hat{\alpha}_i = 0$ e $\hat{\xi} = 0$ os pontos são classificados corretamente localizando-se ao lado opostos das margens. Esses pontos não são utilizados para função de decisão (PASSERINE, 2004).

Na Figura 4 é possível obter uma ilustração desses casos. Os pontos na cor verde estão incorretamente classificados (caso i); os pontos na cor preta estão classificados corretamente e se situam dentro das margens (caso ii); os pontos na cor vermelha estão localizados sobre a margem (caso iii); e quando $\hat{\alpha}_i = 0$ e $\hat{\xi} = 0$ determina os pontos na cor azul .

Figura 4 - Tipos de vetores de suporte (SVs)



Fonte : Adaptado de PASSERINE (2004).

O parâmetro β_0 é definido por α e é calculado a partir dos SV (*Support Vector*), denotamos *n*SV por número de vetores de suporte. Assim temos que:

$$\hat{\beta}_0 = \frac{1}{nSV} \sum_{x_j \in SV} \frac{1}{y_i} - \hat{\beta}.x_j$$

Dadas as soluções para $\beta \; e \;\; \beta_0$ a função de decisão é dada por :

$$g(x) = sng\big(f(x)\big) = \left(\sum_{x_i \in SV} y_i \hat{\alpha}_i x_i \cdot \mathrm{x} + \hat{\beta}_0 \right)$$

Com as restrições : $\begin{cases} 0 \leq \alpha_i \leq \mathrm{C}, \forall i = 1, \cdots, N \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases}$

O parâmetro de ajuste deste procedimento é o parâmetro de custo C.

2.6 Floresta Aleatória (*Random Forest* - RF)

Floresta Aleatória (*Random Forest* - RF (BREIMAN, 2001)) faz parte do contexto da classificação por árvores. RF é uma adaptação do método *bagging*, o qual cria uma grande coleção de árvores não correlacionadas e calcula a média delas (HASTIE et al., 2009). O principal objetivo dos *bagging* é reduzir a variação de um método de aprendizado estatístico, ou seja, dado um conjunto de I observações independentes $X_1, \cdots, X_I$, com variância σ^2, a variância da média $\bar{X}$ das observações é dada por $\frac{\sigma^2}{I}$. Em outras palavras, calcular a média de um conjunto de observações reduz a variação e consequentemente aumenta a precisão (JAMES et al., 2013).

No método das RF para classificação construímos uma floresta numérica de árvores de decisão em amostras de treinamento inicializadas. Ao construir estas árvores de decisão, cada vez que é considerada uma divisão em uma árvore, uma amostra aleatória de m preditores é escolhida como candidatos do conjunto completo de J preditores. A divisão pode usar apenas um desses m preditores, uma nova amostra de m preditores é obtida a cada divisão. Podendo escolher $m \approx J$, $m = \frac{J}{2}$ ou $m = \sqrt{J}$. Na maioria das vezes escolhe-se $m \approx \sqrt{J}$, ou seja, o número de preditores considerados em cada divisão é aproximadamente igual à raiz quadrada do número total de preditores (JAMES et al., 2013).

Uma RF obtém um voto de classe de cada árvore e depois classifica a amostra usando voto majoritário. É recomendado o valor padrão para $m = \sqrt{J}$ e o valor mínimo do tamanho do nó igual a 1.

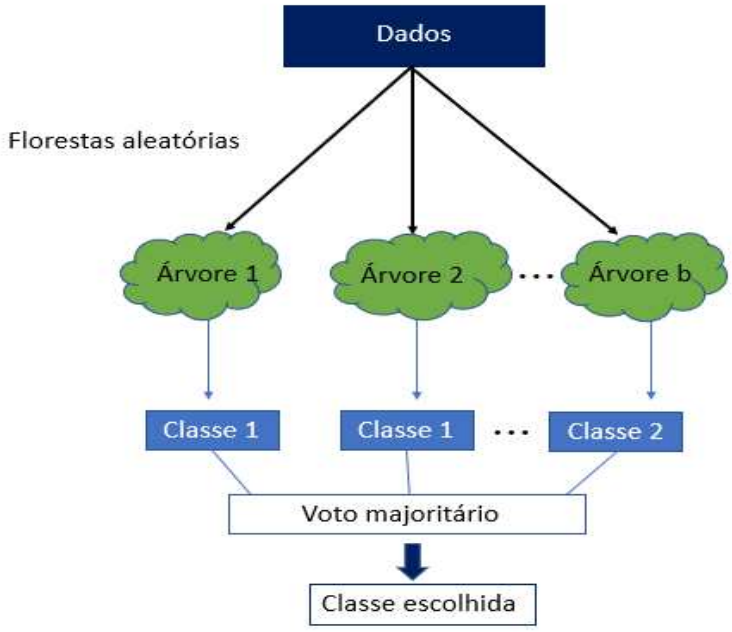
Considerando um conjunto de treinamento de tamanho B, segue o algoritmo de Floresta Aleatória para classificação (HASTIE et al., 2009):

1. Para b =1 até B
   (a) Inicie com uma amostra $X^*_{(1,N)}$ de acordo com os dados de treinamento
   (b) Aumente uma árvore de floresta aleatória A_b para os dados de inicialização, recursivamente repetindo as etapas a seguir para cada nó de cada árvore, até que o tamanho mínimo do nó, n_{min}, seja atingido.
      i. Selecione m variáveis aleatórias entre as J variáveis
      ii. Escolha a melhor variável/ponto de divisão entre as m
      iii. Divida o nó em dois nós filhos
2. Saída do conjunto de árvores:$\{A_b\}_1^B$

Para fazer uma previsão em um novo ponto x:

Seja $\widehat{D}_b(x)$ a previsão de classe da b-*ésima* árvore da floresta aleatória. Então $\widehat{D}_{rf}^B(x) =$ voto majoritário $\{D_b(x)\}_1^B$.

Figura 5 - Ilustração do método de florestas aleatórias.



FONTE: Adaptado de <https://en.wikipedia.org/wiki/Random_forest>Acesso em: 25 de ago. 2020.

3. MATERIAL E MÉTODOS

3.1 Material Vegetal

Foram avaliadas 460 amostras provenientes de uma população originalmente de mudas de 98 famílias de meio irmãos. As mudas da população, onde cada planta é um único genótipo, foi o resultado de cruzamentos realizados na Estação de Floração e Melhoramento da Serra do Ouro, município de Murici, estado de Alagoas, Brasil (09° 13 'S, 35° 50' W, 450 m de altitude).

Após o processamento, as sementes foram enviadas para a Estação de Pesquisa em Melhoramento Genético da Cana-de-Açúcar (CECA) da Universidade Federal de Viçosa, município de Oratórios, Minas Gerais, Brasil (20° 25 'S, 42° 48' W, 494 m de altitude) e colocadas para germinarem em uma casa de vegetação.

Posteriormente, mudas obtidas de cada família foram transplantadas para o campo e avaliadas na primeira (cana-planta) e segunda (soca) safras com base nas características desejáveis no primeiro (T1) e segundo (T2) ensaio clonal (Barbosa et al., 2012) .

3.2 Desenho Experimental

O experimento foi instalado no CECA, em blocos aumentados, em maio de 2016. O CECA está localizado no município de Oratórios, Estado de Minas Gerais, Brasil (20° 25’S, 42° 48’ W, 494 m de altitude). Os tratamentos comuns - RB867515, RB966928 e RB92579 - foram incluídos uma vez em cada bloco, e os tratamentos regulares foram dispostos em 21 blocos (FEDERER, 1961). No total, 18 blocos continham 24 parcelas; um bloco continha 16 parcelas e dois blocos continham 12 parcelas.

3.4 Dados fenotípicos e Análise de referência

Os clones foram avaliados na primeira safra (12 meses após o plantio) e na segunda safra (26 meses após o plantio). Neste estudo, foram usados apenas dados da segunda safra. A estimação da porcentagem de sacarose aparente na cana (%PC) e porcentagem do teor de fibras (%FIB) foi realizada de acordo com as recomendações do manual do CONSECANA (CONSECANA, 2006).

Para trabalhar com um conjunto representativo das amostras, foram cortados dez colmos (de dez diferentes touceiras) selecionados aleatoriamente de cada parcela de linha dupla. Os cortes foram feitos ao nível do solo com um facão. Topos verdes, folhas aderentes e as bainhas das folhas foram removidas antes dos caules serem agrupados e pesados usando um dinamômetro. Depois, os dez colmos selecionados aleatoriamente em cada parcela foram triturados usando uma picadora de forragem estacionária (modelo EN-6500, indústrias Nogueira & Brait,Itapira, Brasil). Uma subamostra de 500 g dos talos triturados foi coletada e prensado com uma prensa hidráulica (PL011 Model, Dedini, Inc., Piracicaba, Brasil) a 250 kgf / cm^2 (24,5 MPa) por 1 minuto.

Após prensagem, o suco e o restante do bolo de fibras foram recolhidos e levados para o laboratório. O suco foi analisado para %POL por polarimetria usando um sacarímetro (Modelo SDA2500, Acatec, Brasil) após clareamento da solução com acetato de chumbo $Pb(C_2H_3O_2)_2$. O restante do bolo de fibras foi pesado (WC) e utilizado para calcular o teor de fibra (CONSECANA, 2006):

$$\%FIB = 0{,}08 \cdot WC + 0{,}876$$

A %PC foi calculada da seguinte forma:

$$\%PC = \%POL \cdot (1 - 0{,}01 \cdot FIB\%) \cdot SC$$

onde SC é o coeficiente que converter a sacarose de suco em sacarose de cana, e é obtido por $SC = 1{,}0313 - 0{,}00575 \cdot \%FIB$. Os valores finais foram todos expressos em base total de biomassa fresca (500g de colmos triturados).

3.5 Preparação de amostras e obtenção de espectros NIR

Uma subamostra de 100g dos colmos triturados foi coletada e imediatamente seca em estufa de circulação de ar a 50 °C por 24h ou até que atingisse uma massa constante. Posteriormente, para obter um tamanho de partícula homogêneo, as amostras secas foram moídas usando um moinho com uma peneira de fundo de 0,4 mm, embalado em um saco plástico e armazenado. O processo de preparação teve duração de dois meses (o processamento de cada amostra foi de, aproximadamente, 5 minutos).

Os espectros NIR das amostras foram medidos em condições de laboratório à temperatura ambiente de 21 ºC. O instrumento utilizado foi um espectrômetro de infravermelho próximo (FT-NIR) com transformada de Fourier (modelo Antaris ™ II, Thermo Scientific Inc., EUA). As condições de operação do instrumento foram definidas da seguinte forma: 4

cm^{-1} resolução em uma faixa de número de onda investigada de 10000 a 4000 cm^{-1} e modo de refletância difusa com $log\ (1\ /\ R)$, onde R é a refletância medida. As amostras (em pó) foram colocadas em um acessório de vidro, no formato de copo, e dispostos na janela do instrumento.

A cada digitalização, o acessório era movido para cobrir diferentes posições da amostra, totalizando seis posições. Para cada amostra, um total de 192 varreduras foram feitas e depois calculadas a média, representando o espectro final.

3.6 Análise estatística e comparação dos métodos

Inicialmente foi feita a análise descritiva dos dados com a finalidade de separar o conjunto de dados em duas classes de acordo com as médias da porcentagem de fibra (%FIB) e porcentagem do teor sacarose aparente (%PC). Em cada caso, valores acima da média foram definidos como **classe 1** e valores abaixo da média como **classe 2.**

Adicionalmente, foram construídos gráficos para visualização dos espectros e para auxiliar nas escolhas dos pré-tratamentos que seriam aplicados. Os pré-tratamentos aplicados foram: Centragem na Média (CM), Correção Multiplicativa de Sinal (MSC), Padrão de Normal de Variação (SNV), Primeira Derivada (1D) e Segunda Derivada (2D), ambas usando o método Savitzky-Golay com janelas de tamanho 15 e polinômio de grau 2 (FERREIRA, 2015).

Para definição dos pré-tratamentos para cada método, o conjunto de dados foi estratificado pelo algoritmo de KENNARD e STONE (KENNARD e STONE, 1969) em dois conjuntos: o de treino contendo 368 amostras, e o de teste, contendo 92 amostras. A separação foi realizada individualmente para %PC e %FIB e feita na mesma proporção do conjunto completo. Este algoritmo tem como objetivo a seleção de um subconjunto de amostras que possam representar o máximo da variabilidade do conjunto total (KENNARD e STONE, 1969) e é amplamente utilizado em pesquisas com NIR (SOUSA et al., 2011; LI et al., 2018; LV et al., 2019; SIMEONE et al., 2019). O algoritmo Kernard-Stone, sempre seleciona as mesmas amostras para o conjunto de dados usando a mesma configuração do algoritmo, nesse sentido, ele foi utilizado para ajuste dos modelos em uma primeira etapa. Os pré-tratamentos que proporcionaram o menor erro de classificação na validação cruzada (erro CV) foram designados como mais apropriados para cada método.

Após a escolha dos pré-tratamentos para cada método (etapa 1), o conjunto de dados constituído por 460 amostras foi novamente particionado em conjuntos de treino (368 amostras) e teste (92 amostras) de forma aleatória com o objetivo de comparar os métodos de

classificação. Nessa segunda etapa, para cada repetição foi construído um modelo com os pré-tratamentos definidos para cada método. Foram realizadas dez repetições gerando dez modelos de classificação, o que possibilitou montar intervalos de confiança e aplicar testes estatísticos.

Em todos os métodos utilizados para a classificação nesse trabalho, deve ser feito a escolha dos valores de seus parâmetros. No PLS-DA é necessário a escolha do número de componentes ou variáveis latentes (nVL). Já o método SVM a escolha do valor do parâmetro de custo (C) e para o método RF o número de preditores (m).

Nesse sentido, foi empregado a validação cruzada *k-fold*, com $k = 10$ nas amostras do conjunto de calibração para auxiliar na escolha dos valores desses parâmetros. A validação *k-fold*, divide o conjunto de treinamentos em *k* partes iguais. Utiliza-se uma parte para teste e as outras $k - 1$ partes restantes são utilizados para estimação dos parâmetros e validação do modelo (HASTIE et al., 2009). Os parâmetros escolhidos foram aqueles que apresentaram uma menor estimativa do erro de classificação na validação cruzada.

Finalmente, foi feita a validação externa utilizando cada conjunto teste correspondente a cada modelo construído e então foi possível obter a matriz de confusão (Tabela 1) para cada modelo e, assim, calcular os respectivos erros de classificação, a sensibilidade e a especificidade.

Tabela 1 - Matriz de confusão entre a classificação verdadeira e a classificação obtida a partir dos valores preditos pelos modelos.

	Classe 1 (Real)	Classe 2 (Real)
Classe 1 (Predita)	TP	FP
Classe 2 (Predita)	FN	TN

TP = número de casos verdadeiro positivo; TN= número de casos verdadeiro negativo; FP = número de casos falso positivo e FN = número de casos falso negativo.

Os parâmetros de classificação que foram usados para avaliar o desempenho de cada método são definidos como:

$$Sensibilidade = \frac{TP}{TP+FN} \qquad Especificidade = \frac{TN}{TN+FP} \qquad Erro = \frac{FP+FN}{TP+TN+FP+FN}$$

A sensibilidade mede a taxa de verdadeiros positivos, ou seja, da observação pertencer à classe h ($h = 1$ ou 2) e ser classificada na classe h, e a especificidade mede a taxa de verdadeiro negativo, ou seja, da observação não pertencer à classe *h* e ser classificada como não sendo da classe *h*. O erro corresponde ao erro de classificação, ou seja, o quanto o método classificou incorretamente.

Como foram realizadas dez repetições, foram gerados dez modelos para cada método de classificação. Para fazer a comparação dos métodos e verificar qual apresenta um melhor desempenho para classificação de amostras de dados NIR, foi calculado a média e o desvio padrão dos resultados de sensibilidade, especificidade e erro de classificação. Dessa forma, foram empregados testes estatísticos para comparação das médias e avaliação dos métodos de classificação.

Foi aplicado uma análise de variância sob o delineamento em blocos casualisados, em que as variáveis respostas foram os erros, a sensibilidade e especificidade, obtida por cada método, e os blocos foram as 10 realizações independentes realizadas. As hipóteses testadas sobre as médias obtidas para cada método, para um nível de 5% de significância, foram:

$H_0: \mu_1 = \mu_2 = \mu_3$

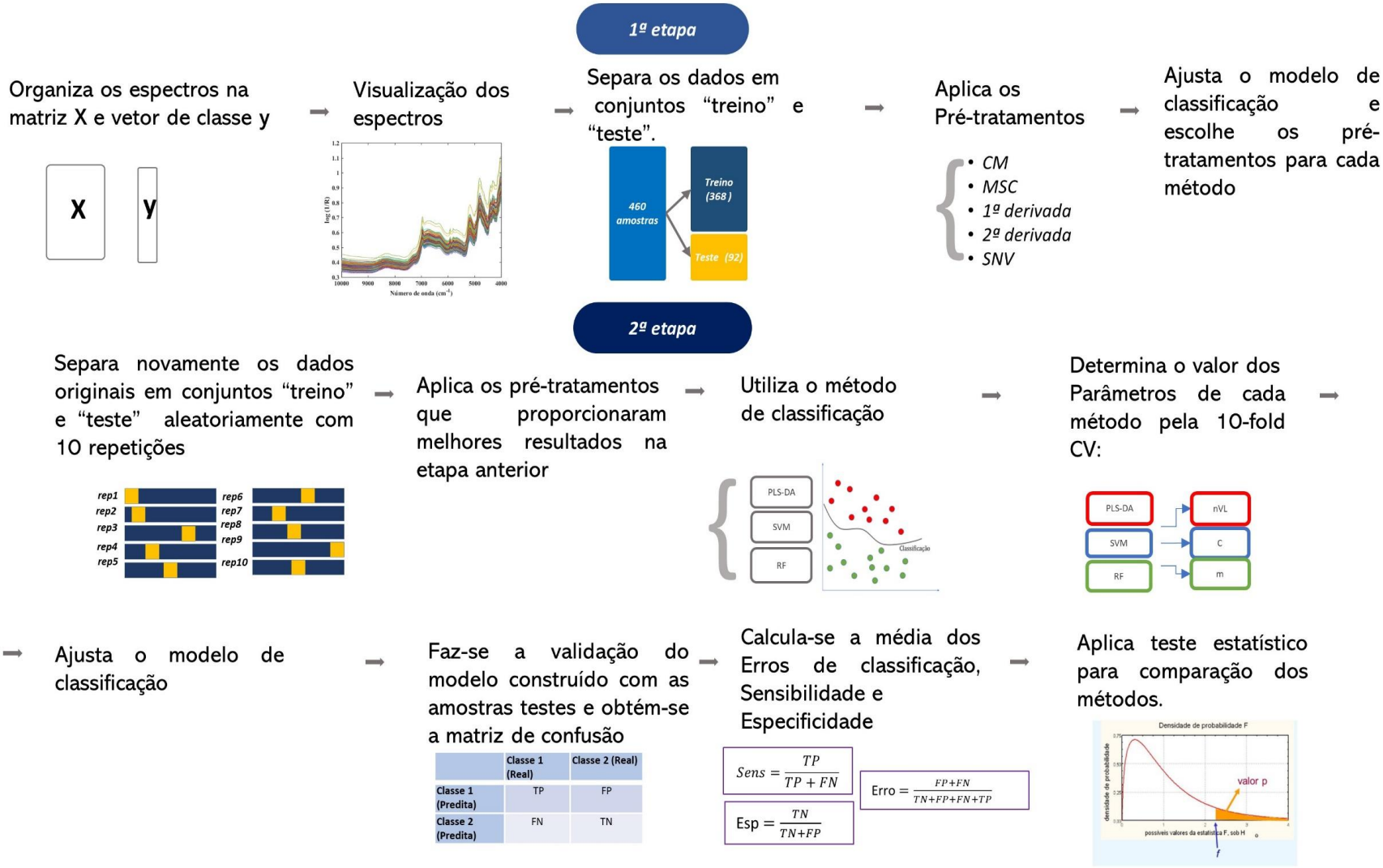
H_a: Pelo menos um contraste entre médias é estatisticamente diferente de zero.

Para comparações múltiplas das médias dos tratamentos foi aplicado o teste *t* de *student* para um nível de significância de 5%.

As análises foram realizadas por meio do software Matlab (Matlab R2016a, 9.0, The MathWorks Inc., Natick, EUA), PLS-Toolbox 8.2 (Eigenvector Research, Inc. Wenatchee, EUA) e pelo software R (R Core Team, 2019).

O diagrama na Figura 6 ilustra todo o procedimento realizado.

Figura 6 - Diagrama referente às etapas realizadas.



4. RESULTADOS E DISCUSSÃO

4.1 Análise descritiva dos dados

A propriedade %PC apresentou média (15,20) e desvio padrão (1,83), maiores do que aqueles para %FIB. Os valores de teor de fibra variaram de 10,53-17,76% e de sacarose aparente 9,31-20,69% (Tabela 2).

De acordo com o criterio estabelecido para definir as classes, após a separação por meio do algoritmo Kenard-Stone foram designadas um total de 219 amostras na classe 1 e 241 amostras na classe 2 para %FIB, e 232 amostras na classe 1 e 228 amostras na classe 2 para %PC. Os resultados do teste qui-quadrado para homogeneidade ($p > 0{,}05$), mostram que houve uma proporção razoável em relação ao número de amostras na classe 1 e classe 2 para os conjuntos de treino e teste, tanto para %PC, quanto para %FIB (Tabela 2).

Tabela 2 - Análise descritiva dos valores de porcentagem do teor de fibra (%FIB) e porcentagem do teor de sacarose aparente (%PC).

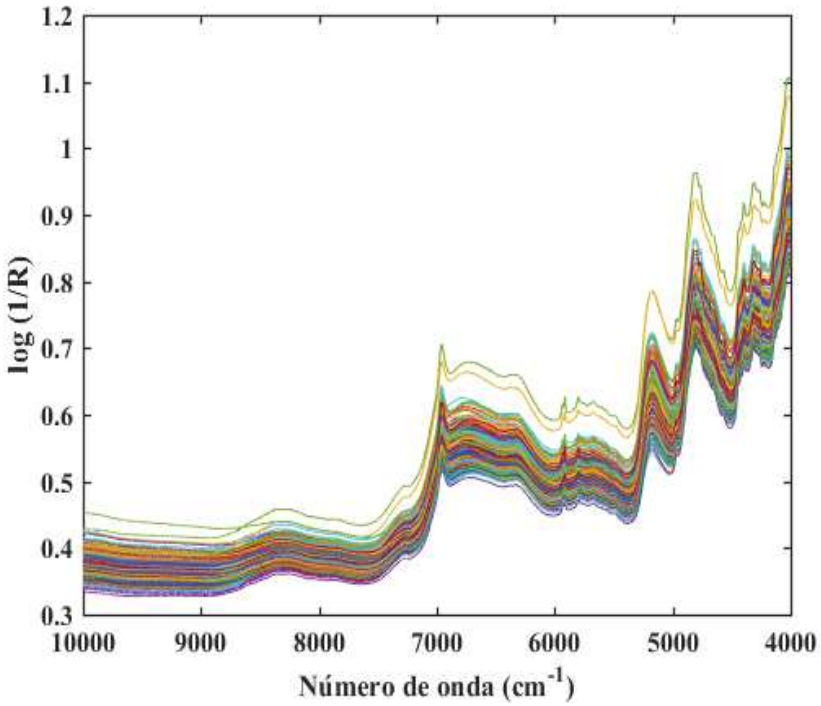
Propriedade	Média	Desvio Padrão	Mínimo	Máximo	*p*-valor (qui-quadrado)
%FIB	13,66	1,20	10,53	17,76	0,305
%PC	15,20	1,83	9,31	20,69	0,852

4.2 Análise dos espectros

O conjunto de dados dos espectros NIR da cana-de-açúcar utilizados para obter a porcentagem de teor sacarose aparente e de teor de fibra da cana de açúcar é composto por 460 amostras e 3112 variáveis.

Ao visualizar os espectros (Figura 7) percebe-se uma pequeno deslocamento e inclinação na linha de base. Devido a isso foram utilizados os pré-tratamentos: Correção Multiplicativa de Sinal (MSC), Padrão de Normal de Variação (SNV), Primeira e Segunda Derivada (Savitzky-Golay, janela =15 e polinômio de grau = 2).

Figura 7 - Espectros NIR obtidos da cana de açúcar para predição do teor de sacarose e fibra



4.3. Definição dos pré-tratamentos

Visando encontrar um melhor resultado na classificação, inicialmente foram testados diferentes pré-tratamentos de dados no conjunto de treino obtido por meio do algoritmo Kennard-Stone. Na Tabela 3 encontram-se os resultados de cada pré-tratamento aplicado, e os respectivos erros CV obtido para os métodos PLS-DA, SVM e RF, para a classificação em relação a porcentagem de fibra (%FIB) e porcentagem de sacarose aparente (%PC).

Nota-se que, para a propriedade %PC, o melhor modelo PLS-DA foi obtido com a escolha do pré-tratamento Centrar na Média juntamente com o MSC e Segunda Derivada. Em relação ao método SVM, os pré-tratamentos mais adequados foram Centrar na Média e MSC. O método RF apresentou menor erro com os pré-tratamentos centrar na média e SNV.

Referente à propriedade %FIB, o melhor modelo PLS-DA foi obtido com os pré-tratamentos Centrar na Média, juntamente com o MSC e a Primeira Derivada. O método SVM obteve um menor erro sem aplicação de pré-tratamentos. O método RF apresentou resultados mais satisfatórios com os pré-tratamentos Centrar na Média e Segunda derivada (Tabela 3).

Tabela 3 - Valores do erro de validação cruzada (Erro CV) para diferentes tratamentos em dados de cana-de-açúcar para classificação da porcentagem de fibra (%FIB) e porcentagem de sacarose aparente (%PC) utilizando o modelo PLS-DA, SVM e RF.

		PLS-DA	SVM	RF
	Pré-tratamento	**Erro CV**	**Erro CV**	**Erro CV**
% PC	Nenhum	0,3160	0,3262	0,3851
	CM	0,2912	0,3097	0,3831
	CM + 1D	0,2939	0,3110	0,3306
	CM + 2D	0,2883	0,3426	0,3397
	CM + MSC	0,3027	**0,2961**	0,3614
	CM + MSC + 1D	0,3033	0,3290	0,3337
	CM + MSC + 2D	**0,2783**	0,3318	0,3367
	CM + SNV	0,3047	0,3290	**0,3257**
	CM + SNV + 1D	0,3048	0,3452	0,3348
	CM + SNV + 2D	0,2858	0,3071	0,3380
% FIB	Nenhum	0,2856	**0,2904**	0,3823
	CM	0,2824	0,3110	0,3750
	CM + 1D	0,2530	0,3259	0,3415
	CM + 2D	0,3043	0,3588	**0,3364**
	CM + MSC	0,2667	0,2964	0,3617
	CM + MSC + 1D	**0,2476**	0,3211	0,3516
	CM + MSC + 2D	0,3104	0,3648	0,3555
	CM + SNV	0,2773	0,2961	0,3515
	CM + SNV + 1D	0,2558	0,3567	0,3564
	CM + SNV + 2D	0,3017	0, 3644	0,3530

CM: Centrar na média; MSC: Correção multiplicativo de sinal; 1D: Primeira derivada; 2D: Segunda derivada; SNV: Padrão normal de variação.

Na propriedade %PC, o uso de pré-tramento para o método SVM melhora apenas 1,96% em relação ao erro CVquando não utiliza pré-tratamento. Já o PLS-DA ao utilizar pré-tratamentos apresenta melhoria de 13,54% e 13,21% para as propriedades %PC e %FIB, respectivamente. O RF também se mostra favorável ao uso de pré-tramentos, em que os erros CV diminuem 18,23% para a propriedade %PC e 14,48% para a propriedade %FIB.

Os métodos PLS-DA e RF apresentam melhorias significativas quando são submetidos a mais etapas de pré-tratamentos. O mesmo não ocorre com o SVM em que, para a %PC, foram utilizados apenas dois pré-tratamentos, e para a %FIB os dados não precisaram ser pré-tratados.

Devos (2014) avaliou se os pré-tratamentos eram tão úteis para os modelos SVM quanto demonstram ser para outros métodos de classificação. Este autor concluiu, utilizando conjuntos de dados NIR, que o ganho de precisão do modelo ajustado pelo SVM, comparado ao PLS-DA, deve-se principalmente pela característica do método ter ajuste não linear, visto que apresentou pouca melhoria em relação ao uso de pré-tratamentos.

4.4. Ajuste dos modelos de classificação

Após a escolha dos pré-tratamentos adequados para cada método, os dados originais constituído por 460 amostras foram novamente particionados em conjuntos de treino (368 amostras) e teste (92 amostras) de forma aleatória, esse processo foi repetido 10 vezes, gerando 10 modelos de classificação para cada método. A Tabela 4 mostra o resultado dos valores dos parâmetros de cada método e seus respectivos erros na validação cruzada (Erro CV).

Tabela 4 - Valores dos parâmetros e erro CV para os métodos PLS-DA, SVM e RF nos modelos ajustados.

		PLS-DA		SVM		RF	
	Modelo	nVL	**Erro CV**	C	**Erro CV**	m	**Erro CV**
% PC	Modelo 1	6	0,311	1	0,303	3112	0,359
	Modelo 2	6	0,278	1	0,307	3112	0,364
	Modelo 3	6	0,293	1	0,364	3112	0,363
	Modelo 4	5	0,335	1	0,353	3112	0,342
	Modelo 5	7	0,296	1	0,295	2	0,383
	Modelo 6	6	0,284	1	0,308	78	0,378
	Modelo 7	8	0,215	1	0,317	78	0,320
	Modelo 8	6	0,269	1	0,320	78	0,339
	Modelo 9	5	0,311	1	0,364	78	0,345
	Modelo 10	6	0,288	1	0,318	3112	0,375
% FIB	Modelo 1	7	0,282	1	0,290	78	0,347
	Modelo 2	7	0,278	1	0,290	3112	0,329
	Modelo 3	8	0,246	1	0,274	3112	0,352
	Modelo 4	7	0,258	1	0,269	78	0,315
	Modelo 5	8	0,244	1	0,255	3112	0,314
	Modelo 6	5	0,267	1	0,306	78	0,359
	Modelo 7	7	0,278	1	0,288	78	0,334
	Modelo 8	8	0,257	1	0,290	78	0,377
	Modelo 9	7	0,264	1	0,255	3112	0,356
	Modelo 10	7	0,290	1	0,310	3112	0,370

nVL: número de variáveis latentes; *C*: parâmetro de penalização; *m*: número de preditores;

Em seguida, foi realizada a validação de cada modelo obtido em cada partição para a classificação referente a %PC e a %FIB utilizando o conjunto de treino composto por 92 amostras. Assim, foi possível obter o erro de classificação, a sensibilidade e a especificidade de cada modelo. Como o objetivo é selecionar indivíduos com maiores teores de fibra e sacarose aparente, dessa forma, serão apresentados os parâmetros de classificação em relação a classe 1, uma vez que, como há apenas duas classes, os valores dos erros de classificação são

equivalentes e os valores de sensibilidade referentes a classe 2 são complementares aos valores de especificidade referentes a classe1 e vice e versa.

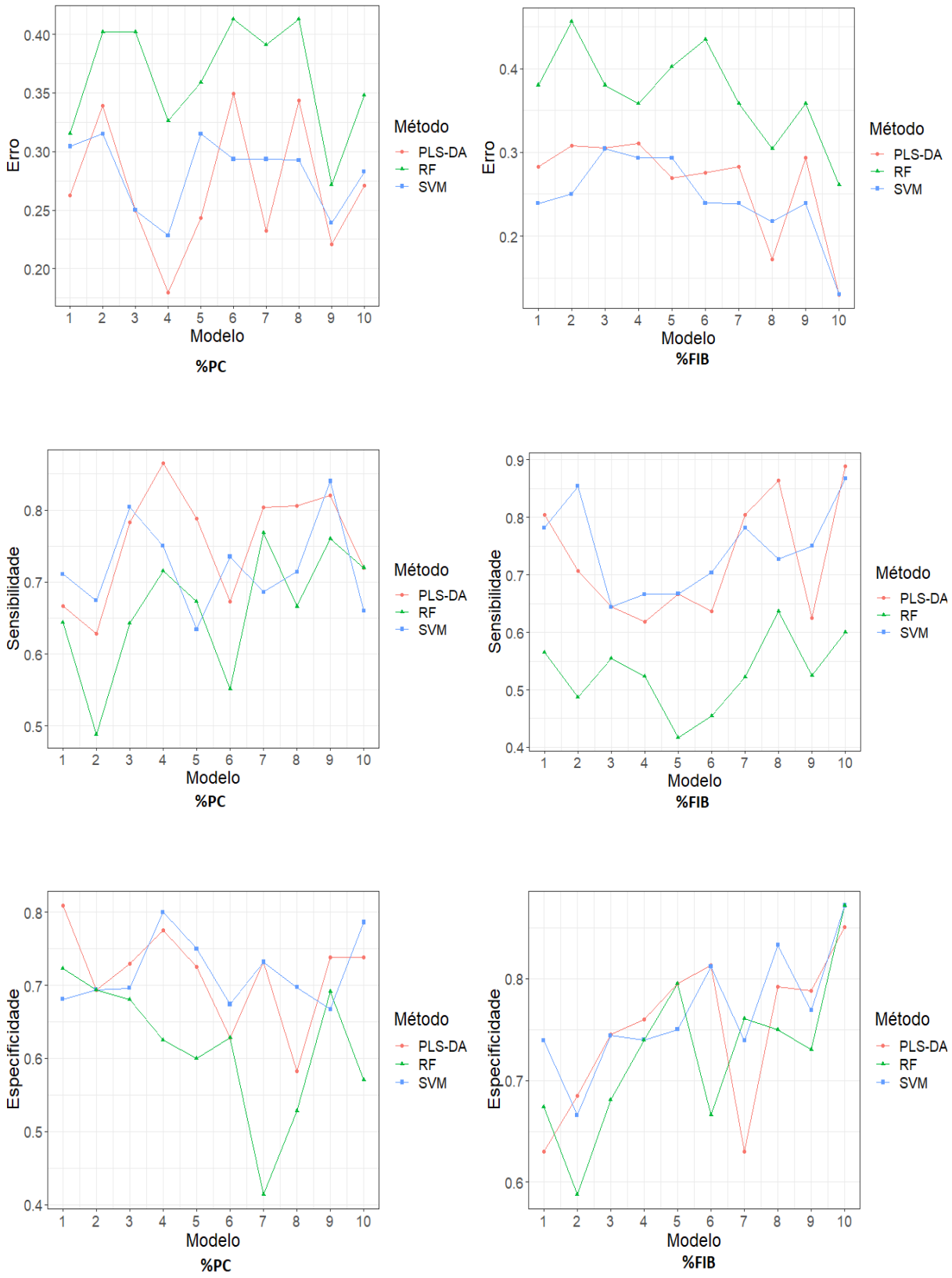
O método RF apresenta maiores valores de erros e menores valores de sensibilidade para os dois conjuntos de dados em comparação com o PLS-DA e SVM. O PLS-DA e o SVM, apresentam valores análogos para o erro, sensibilidade e especificidade na maioria dos modelos. Além disso, é possível observar que houve uma variabilidade nos valores dos parâmetros de classificação dentro de cada conjunto de dados usados para a construção dos modelos (Tabela 5 e Figura 8).

Tabela 5 - Valores de sensibilidade, especificidade e erros de classificação obtidos na predição para os métodos PLS-DA, SVM e RF em cada modelo para classificação de % PC e % FIB.

		PLS-DA			**SVM**			**RF**		
	Modelo	**Erro**	**Sens**	**Esp**	**Erro**	**Sens**	**Esp**	**Erro**	**Sens**	**Esp**
%PC	1	0,262	0,667	0,809	0,304	0,711	0,681	0,315	0,644	0,723
	2	0,339	0,628	0,694	0,315	0,674	0,694	0,402	0,488	0,694
	3	0,250	0,783	0,729	0,250	0,804	0,696	0,402	0,643	0,680
	4	0,179	0,865	0,775	0,228	0,750	0,800	0,326	0,715	0,625
	5	0,243	0,788	0,725	0,315	0,634	0,750	0,358	0,673	0,600
	6	0,349	0,673	0,628	0,293	0,735	0,674	0,413	0,551	0,628
	7	0,232	0,804	0,732	0,293	0,686	0,731	0,391	0,768	0,414
	8	0,343	0,806	0,582	0,292	0,714	0,697	0,413	0,666	0,528
	9	0,221	0,820	0,738	0,239	0,840	0,667	0,271	0,760	0,691
	10	0,271	0,720	0,738	0,282	0,660	0,786	0,347	0,720	0,571
%FIB	1	0,282	0,804	0,630	0,239	0,782	0,739	0,380	0,565	0,674
	2	0,308	0,707	0,685	0,250	0,854	0,666	0,456	0,487	0,588
	3	0,305	0,644	0,745	0,304	0,644	0,744	0,380	0,555	0,681
	4	0,310	0,619	0,760	0,293	0,666	0,740	0,358	0,524	0,740
	5	0,268	0,667	0,795	0,293	0,667	0,750	0,402	0,417	0,795
	6	0,275	0,636	0,813	0,239	0,704	0,812	0,434	0,454	0,666
	7	0,282	0,804	0,630	0,239	0,782	0,739	0,358	0,522	0,761
	8	0,172	0,864	0,792	0,217	0,727	0,833	0,304	0,636	0,750
	9	0,293	0,625	0,788	0,239	0,750	0,769	0,358	0,525	0,730
	10	0,130	0,889	0,851	0,130	0,867	0,872	0,261	0,600	0,872

Sens: Sensibilidade; Esp: Especificidade; Erro: Erro de classificação.

Figura 8 – Erro de classificação (Erro), sensibilidade e especificidade para os métodos PLS-DA, SVM e RF, referente à classe 1 das propriedades %PC e %FIB nos dez modelos obtidos.



Foi calculada a média dos erros de classificação, sensibilidade e especificidade e seus respectivos desvios padrões (Tabela 6). Para uma análise exploratória desses valores, foram construídos Boxplots (Figura 9) os quais mostram as dispersões das medidas avaliadas em cada método.

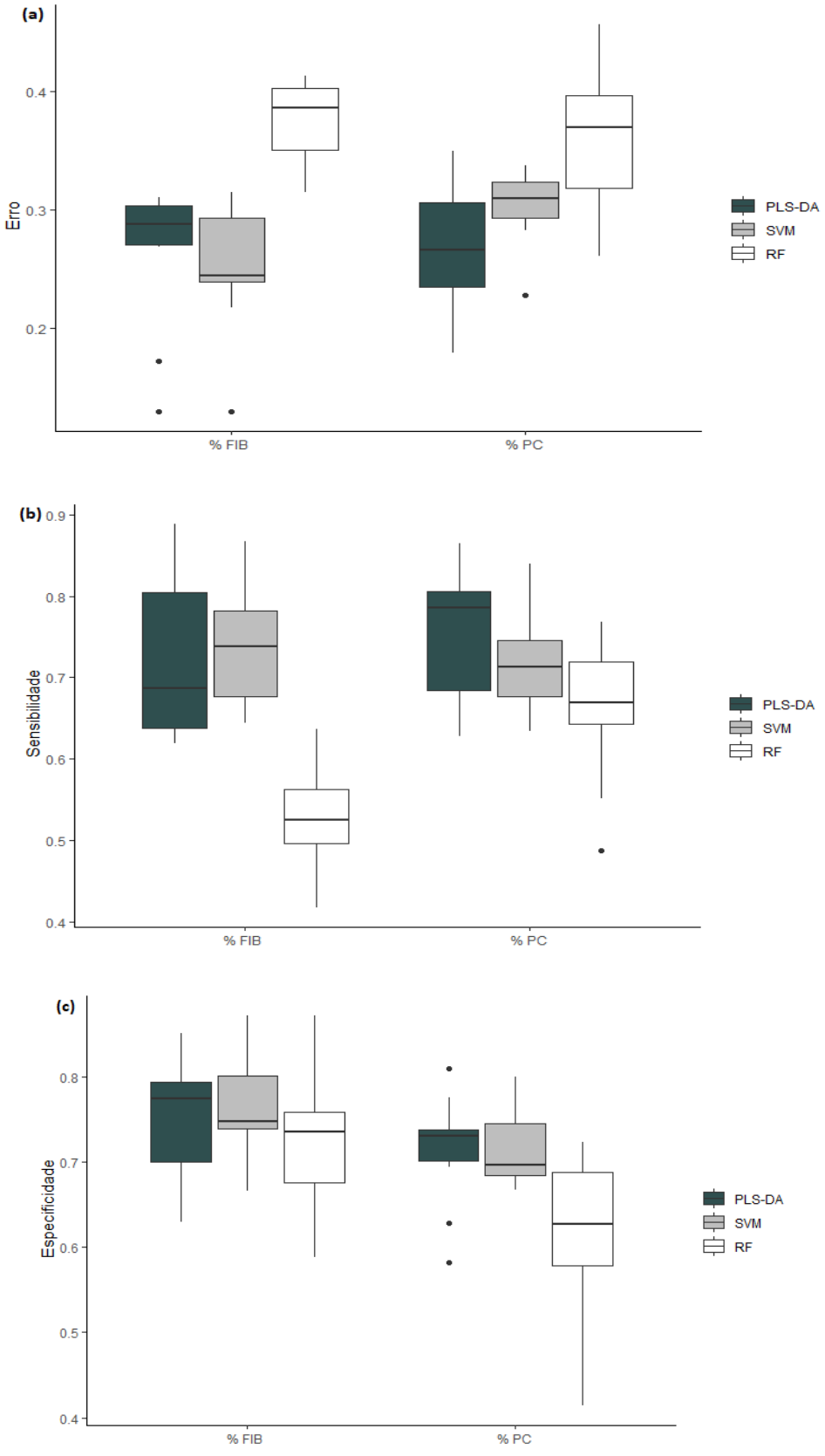
Nota-se que o método RF se difere dos demais, apresentando resultado inferiores em relação aos parâmetros de classificação. Os métodos PLS-DA e SVM aparentemente não apresentam diferenças entre si, tanto para a classificação em relação a porcentagem do teor de sacarose aparente (%PC) quanto para a classificação em função da porcentagem de fibra (%FIB).

Tabela 6 - Valores da média e do desvio padrão dos parâmetros sensibilidade, especificidade e erro de classificação obtido em cada método para a classificação de %PC e %FIB.

		PLS-DA			**SVM**			**RF**		
		Erro	**Sens**	**Esp**	**Erro**	**Sens**	**Esp**	**Erro**	**Sens**	**Esp**
% PC	Média	0,269	0,755	0,715	0,281	0,720	0,717	0,364	0,662	0,597
	DP	0,057	0,078	0,061	0,031	0,064	0,067	0,036	0,084	0,094
% FIB	Média	0,262	0,725	0,758	0,244	0,744	0,763	0,369	0,500	0,722
	DP	0,062	0,101	0,065	0,054	0,080	0,061	0,067	0,098	0,077

Sens: Sensibilidade; Esp: Especificidade; DP: Desvio padrão

Figura 9 - Boxplot em função do Erro de classificação (Erro) (a); Sensibilidade (b) e Especificidade (c) referente a classe 1 para cada método em relação as propriedades %PC e %FIB.



4.5 Comparação dos métodos de classificação

Foram verificadas as pressuposições de normalidade e homocedasticidade para realizar a ANOVA. Em todos os conjuntos de dados (Tabela 7) de acordo com o teste de Shapiro e Wilk (1965) os resíduos podem ser considerados normais ($p > 0,05$) e, de acordo com o teste de Oneill e Mathews (2000), as variâncias podem ser consideradas homogêneas ($p > 0,05$).

Tabela 7 - Resultados dos testes para verificação da normalidade dos resíduos e homogeneidade das variâncias dos resíduos.

Conjuntos	Shapiro -Wilk	Oneill-Mathews
Erro	*p*-valor = 0,105	*p*-valor = 0,912
Sensibilidade	*p*-valor = 0,831	*p*- valor = 0,942
Especificidade	*p*-valor = 0,146	*p*-valor = 0,060

A ANOVA (Tabela 8) nos mostra que pelo menos uma média dos tratamentos (PLS-DA, SVM e RF) se diferem das demais para as variáveis respostas: erro de classificação, sensibilidade e especificidade referente a discriminação quanto a % PC.

Ao analisar a ANOVA (Tabela 8) referente a %FIB para as variáveis repostas: erro de classificação e sensibilidade concluímos que pelo menos uma média se difere das demais, quanto a especificidade concluímos que as médias são estatisticamente iguais. No entanto, maior ênfase deve ser dada ao erro de classificação e à sensibilidade, já que a ocorrência de falsos positivos não representa um grande problema para a seleção de plantas, uma vez que as plantas selecionadas indevidamente poderão ser descartadas no futuro (PETERNELLI et al., 2017, 2018) dentro do programa de melhoramento.

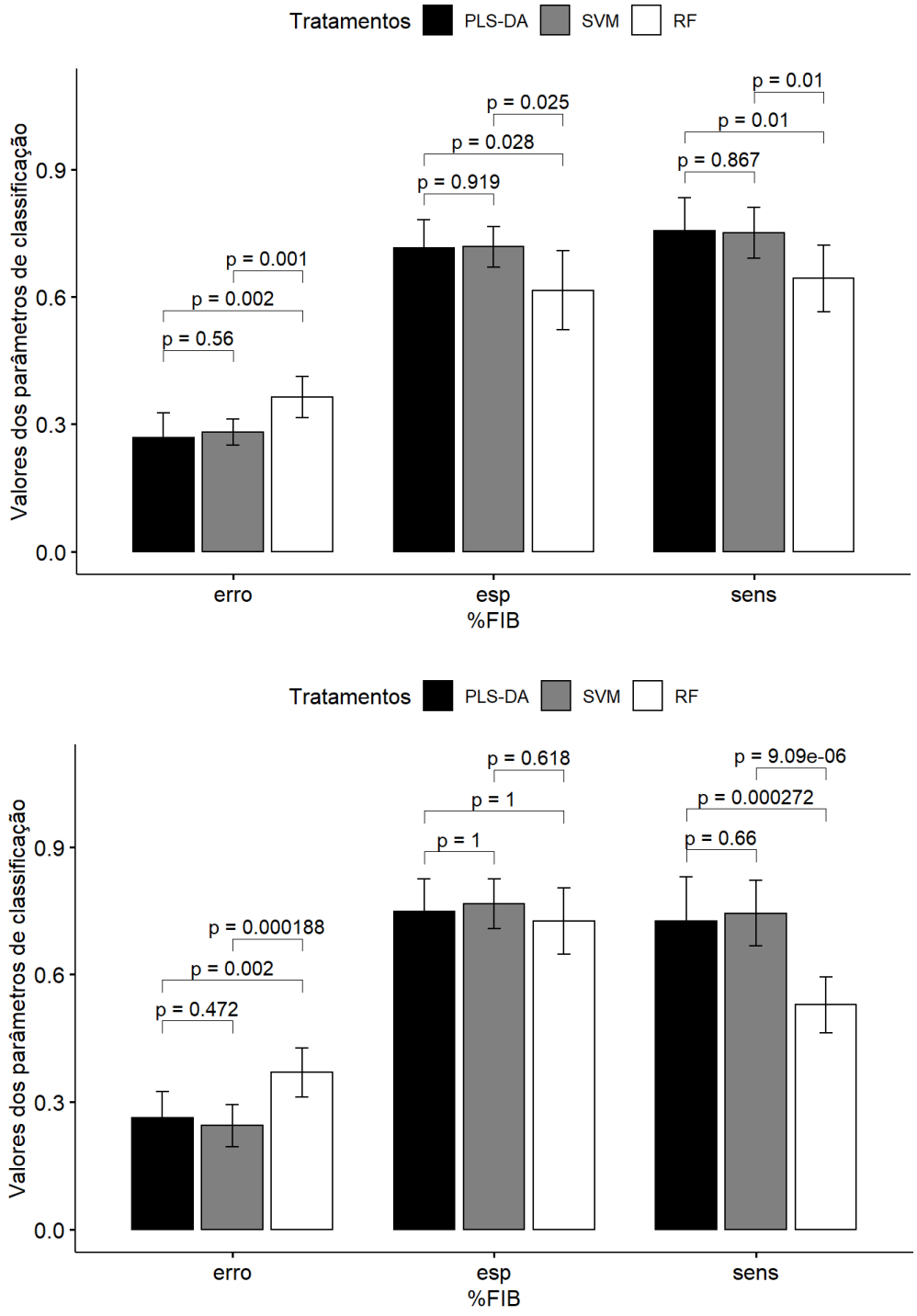
Tabela 8 - ANOVA em função das variáveis respostas: Erro de classificação, sensibilidade e especificidade

			Erro		Sensibilidade		Especificidade	
	FV	GL	QM	*p*-valor	QM	*p*-valor	QM	*p*-valor
%PC	Tratamentos	2	0,0266	0,008	0,0219	0,0065	0,0339	0,0049
	Blocos	9	0,0046	0,002	0,0115	0,0107	0,0059	0,3162
	Resíduos	18	0,0008		0,0032		0,0046	
%FIB	Tratamentos	2	0,0455	0,0006	0,1431	0,0048	0,0041	0,1599
	Blocos	9	0,0078	0,0199	0,0144	0,013	0,0113	0,0010
	Resíduos	18	0,0007		0,0033		0,0020	

Ao fazer a comparação múltiplas das médias de erro de classificação e sensibilidade (Figura 10), nota-se que não há diferença significativa entre os métodos PLS-DA e SVM, e que

ambos diferem do RF. Além disso, é possível observar que o PLS-DA e SVM apresentam maiores valores médios para os parâmetros de sensibilidade e especificidade e menores valores do erro de classificação quando comparados com RF (Figura 10).

Figura 10 - Gráfico de barras das médias e intervalo de confiança de cada tratamento de acordo com o teste *t* de *student* com 95% de confiança referente as variáveis resposta: erro de classificação (erro), especificidade (esp) e sensibilidade (sens) para as propriedades %PC e %FIB.



Um diferencial desse trabalho foi o procedimento que usamos para comparar os classificadores. Fizemos a divisão em duas etapas: Na primeira etapa o conjunto original foi particionado em conjunto treino (368 amostras) e teste (92 amostras) usando o algoritmo Kennard-Stone (KENNARD e STONE, 1969) para definir os pré-tratamentos mais adequados para cada método. Na segunda etapa o conjunto original foi novamente particionado em conjunto de treino (368 amostras) e teste (92 amostras) de forma aleatória com 10 repetições e empregado os pré-tratamentos definidos na etapa anterior. A divisão em duas etapas possibilitou que os mesmos procedimentos fossem utilizados para os diferentes métodos de classificação, facilitando a comparação, e tornando-a mais justa.

Os resultados obtidos para a classificação de %PC e %FIB mostram que o PLS-DA e o SVM não diferem entre si ($p > 0,05$) e apresentam melhor desempenho para classificação do dado NIR utilizado nesse trabalho comparados ao RF. Esse resultado é compatível com as conclusões de Riccioli et al. (2018) e por Shao et al. (2015), que utilizaram dados NIR e de nariz eletrônico respectivamente, e concluíram em seus estudos que o modelo obtido pelo método RF apresentou desempenho inferior em comparação ao PLS-DA e SVM.

Embora o SVM e PLS-DA não tenham apresentado diferenças significativas ($p > 0,05$) para o conjunto de dados utilizados nesse trabalho, muitos autores obtiveram em suas pesquisas resultados que mostram a distinção desses classificadores. Em uma análise comparativa dos três métodos PLS-DA, SVM e RF utilizando dados NIR. Xu et al. (2016) concluiu que o SVM apresentou maiores valores de sensibilidade que os demais. Sampaio et al. (2020) obteve resultados satisfatórios para identificação dos tipos de farinha de arroz com espectroscopia no infravermelho próximo utilizando os métodos PLS-DA e SVM, porém reforça que o SVM supera o PLS-DA na robustez. Lu et al. (2014) também compararam os métodos PLS-DA e SVM utilizando dados NIR e ressalta que o SVM apresenta valores menores de erros de classificação que o PLS-DA.

Em contrapartida, Richter (2016) testou os métodos PLS-DA, SVM e RF quanto à suas capacidades de distinguir espécies de árvores espectralmente semelhantes e concluiu, na validação, que o PLS-DA superou o RF e SVM apresentando uma maior acurácia em todos os conjuntos de dados.

As SVMs apresentam como características uma boa capacidade de generalização e de adaptação para casos não lineares e são robustas diante de dados de grande dimensão, ou seja, apresenta resistência a pequenas e deliberadas variações dos parâmetros (LORENA e CARVALHO, 2007; WANG; GU; WANG, 2017; ZIDI; MOULAHI; ALAYA, 2018).

O PLS-DA têm como vantagem reduzir os efeitos da multicolinearidade entre as variáveis (LIU et al., 2019) e é amplamente aplicável à modelagem de dados de alta dimensão (LEE; LIONG; JEMAIN, 2018).

A escolha do método depende do objetivo do usuário e do tipo de dados que ele apresenta. O PLS-DA e o SVM possuem uma série de alto desempenho em aplicações práticas para classificação de amostras de dados NIR em diversas áreas como podem ser verificados nos trabalhos de Porto et al., 2019, Santana et al., 2020, Jianqiang et al., 2019 e Gupta et al., 2018.

Para uma análise mais precisa, seria interessante testar esses métodos em outro conjuntos de dados NIR e fazer a comparação com os resultados obtidos neste trabalho.

5. CONCLUSÃO

Tanto para a classificação em relação a porcentagem do teor de sacarose aparente, quanto na classificação em relação a porcentagem de fibra, o método PLS-DA e SVM apresentaram resultados semelhantes e superiores ao RF, podendo ser, assim, considerados métodos mais eficientes que o RF para problemas de classificação quando se usa dados NIR.

6. REFERÊNCIAS

AGRIMEC: Agroindustrial e Mecânica LTDA. **Estudo inovação e cana-de-açúcar a agrimec na agrishow 2019**.Disponível em:<https://agrimec.com.br/blog/estudo-inovacao-e-cana-de-acucar-a-agrimec-na-agrishow-2019/> acesso em 30 de novembro 2019.

ALVES, J. C. L.; POPPI, R. J. Biodiesel content determination in diesel fuel blends using near infrared (NIR) spetroscopy and support vector machines (SVM). **Talanta**, v. 104, p. 155–161, 2013.

ANEEL. Parte II: Fontes Renováveis - Biomassa. **Atlas de Energia Elétrica do Brasil**, p.

63–74, 2008.

BARBOSA, M. H. P. et al. Genetic improvement of sugar cane for bioenergy: the brazilian experience in network research with RIDESA. **Crop Breeding and Applied Biotechnology**, v. 12, n. spe, p. 87–98, 2012.

BARKER, M.; RAYENS, W. Partial least squares for discrimination. **Journal of Chemometrics**, v. 17, n. 3, p. 166–173, 2003.

BONASSA, G. et al. Subprodutos Gerados na Produção de Bioetanol: Bagaço, Torta de Filtro, água de Lavagem e Palhagem. **Revista Brasileira de Energias Renováveis**, v. 4, n. 3, p. 144–166, 2015.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

BURNS, D. A.; CIURCZAK, E. W. (3ª ed.). **Handbook of near-infrared analysis**. CRC press, 2007.

CHEMURA, A.; MUTANGA, O.; DUBE, T. Separability of coffee leaf rust infection levels with machine learning methods at Sentinel-2 MSI spectral resolutions. **Precision Agriculture**, v. 18, n. 5, p. 859–881, 2017.

CONAB: Companhia Nacional de Abastecimento. Central de informações agropecuárias: safras – cana. 2019/20. Disponível em:< https://www.conab.gov.br/ > acesso em 15 de outubro de 2019.

CONSECANA. Manual CONSECANA-SP. p. 54, 2006.

DEVOS, O.; DOWNEY, G.; DUPONCHEL, L. Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils. **Food Chemistry**, v. 148, p. 124–130, 2014.

FEDERER, W. T. Augmented designs with one-way elimination of heterogeneity. **Biometrics**, v. 17, n. 3, p. 447-473, 1961.

FERREIRA, M. M. C**. Quimiometria – Conceitos, Métodos e Aplicações**. Campinas, SP: Editora Unicamp, 2015.

GUPTA, O. et al. Machine learning approaches for large scale classification of produce. **Scientific Reports**, v. 8, n. 1, p. 1–8, 2018.

HASTIE, T; TIBSHIRANI, R; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. Stanford, Usa: Springer, 745 p, 2009.

HEUMANN, B. W. An object-based classification of mangroves using a hybrid decision tree Support vector machine approach. **Remote Sensing**, 3.11: 2440-2460, 2011.

HOANG, N. V. et al. High-Throughput Profiling of the Fiber and Sugar Composition of Sugarcane Biomass. **Bioenergy Research**, v. 10, n. 2, p. 400–416, 2017.

JAMES, G.et al**. An introduction to statistical learning: with applications in R**. Springer, New York, 2013.

JIANQIANG, Z. et al. Rapid and automatic classification of tobacco leaves using a hand-held DLP-based NIR spectroscopy device. **Journal of the Brazilian Chemical Society**, v. 30, n. 9, p. 1927–1932, 2019.

KENNARD, R. W.; STONE, L. A. Computer aided design of experiments. **Technometrics**, v. 11, n. 1, p. 137–148, 1969.

LEE, L. C.; LIONG, C. Y.; JEMAIN, A. A. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps. **Analyst**, v. 143, n. 15, p. 3526–3539, 2018.

LI, Xue-Ying, et al. Calibration Transfer of Soil Total Carbon and Total Nitrogen between Two Different Types of Soils Based on Visible-Near-Infrared Reflectance Spectroscopy. **Journal of Spectroscopy**, 2018.

LIU, Y. DE et al. Visual discrimination of citrus HLB based on image features. **Vibrational Spectroscopy**, v. 102, n. December 2018, p. 103–111, 2019.

LOPES, R. E. C. **Discriminação de madeiras similares por NIRS e PLS-DA considerando variações de temperatura e umidade**. Dissertação (mestrado). Instituto de Química, Universidade de Brasília, Brasília, 2015.

LORENA, A. C.; CARVALHO, A. C. P. L. F. Uma Introdução às Support Vector Machines. **Revista de informática Teórica e Aplicada**, vol.14, n.2, p 46-67, 2007.

LU, Y. et al. Classifying rapeseed varieties using Fourier transform infrared photoacoustic spectroscopy (FTIR-PAS). **Computers and Electronics in Agriculture**, v. 107, p. 58–63, 2014.

LV, C. et al. Identification of True and False Aksu Apple Based on NIRS and PLS-DA. **IOP Conference Series: Earth and Environmental Science**, v. 310, n. 4, 2019.

MARTINS, B; GALHARDAS, H; GONÇALVES, N. Using Random Forest classifiers to detect duplicate gazetteer records. In: **7th Iberian Conference on Information Systems and Technologies (CISTI 2012).** IEEE, p. 1-4, 2012.

MARTINS, J. P. A; TEOFILO, R. F.; FERREIRA, M. M. C. Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets. **Journal of Chemometrics**, v. 24, n. 6, p. 320–332, 2010.

O'NEILL, M. E.; MATHEWS, K. Y. Theory & methods: A weighted least squares approach to Levene's test of homogeneity of variance. **Australian & New Zealand Journal of Statistics**, v. 42, n. 1, p. 81-100, 2000.

PASQUINI, C. Analytica Chimica Acta Near infrared spectroscopy: A mature analytical technique with new perspectives e A review. **Analytica Chimica Acta**, v. 1026, p. 8–36, 2018.

PASSERINI, A. Kernel Methods, Multiclass Classification and Applications to Computational Molecular Biology. **Molecular Biology,** p. 141, 2004.

PATACA, L.C.M. **Análises de mel e própolis utilizando métodos quimiométricos de Classificação e Calibração**. 97f. Tese (Doutorado) – Universidade Estadual de Campinas, Campinas, 2006

PEREZ, I. M. N et al. Classification of Chicken Parts Using a Portable Near-Infrared (NIR) Spectrophotometer and Machine Learning. **Applied Spectroscopy**, v. 72, n. 12, p. 1774–1780, 2018.

PETERNELLI, L. A. et al. Artificial neural networks and linear discriminant analysis in early selection among sugarcane families. **Crop Breeding and Applied Biotechnology**, v. 17, n. 4, p. 299–305, 2017.

PETERNELLI, L. A. et al. Decision Trees as a Tool to Select Sugarcane Families. **American Journal of Plant Sciences**, v. 09, n. 02, p. 216–230, 2018.

PORTO, N. DE A. et al. Early prediction of sugarcane genotypes susceptible and resistant to Diatraea saccharalis using spectroscopies and classification techniques. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, v. 218, p. 69–75, 2019.

RANDOM FOREST - Wikipedia. Disponível em: https://en.wikipedia.org/wiki/Random_forest>. Acesso em 25 de agosto de 2020.

RICCIOLI, C.; PÉREZ-MARÍN, D.; GARRIDO-VARO, A. Identifying animal species in NIR hyperspectral images of processed animal proteins (PAPs): Comparison of multivariate techniques. **Chemometrics and Intelligent Laboratory Systems**, v. 172, n. October 2017, p. 139–149, 2018.

RICHTER, R. et al. The use of airborne hyperspectral data for tree species classification in a species-rich Central European forest area. **International Journal of Applied Earth Observation and Geoinformation,** v. 52, p. 464–474, 2016.

S. DE JONG. SIMPLS: an alternative approach to partial least squares regression. **Chemometrics and intelligent laboratory systems**, v. 18, p. 251–263, 1993.

SAMPAIO, P.S., CASTANHO, A., ALMEIDA, A.S. et al. Identification of rice flour types with near-infrared spectroscopy associated with PLS-DA and SVM methods. **European Food Res Technol 246,** 527–537, 2020.

SANTANA, F. et al. Experimento Didático De Quimiometria Para Classificação De Óleos Vegetais Comestíveis Por Espectroscopia No Infravermelho Médio Combinado Com Análise Discriminante Por Mínimos Quadrados Parciais: Um Tutorial, Parte V. **Química Nova**, 2020.

SANTANA, P. N. DE; JOSÉ, A. Combining ability of sugarcane genotypes based on the selection rates of single cross families. p. 47–53, 2017.

SHAO, X. et al. Comparison of different classification methods for analyzing electronic nose data to characterize sesame oils and blends. **Sensors**, 2015, 15.10: 26726-26742.

SHAPIRO, S. S; WILK, M.B. An analysis of variance test for normality (complete samples). **Biometrika,** 52.3/4: 591-611, 1965.

SIMEONE, MLF; RIBEIRO, M. R.; TRINDADE, R. dos S. Espectroscopia no infravermelho próximo e análise discriminante por quadrados mínimos parciais como método alternativo para a seleção de sementes haploides de milho. **Embrapa Milho e Sorgo-Boletim de Pesquisa e Desenvolvimento (INFOTECA-E)**, 2019.

SOUSA, L. C. et al. Desenvolvimento de modelos de calibração NIRS para minimização das análises de madeiras de Eucalyptus spp. **Ciência Florestal**, 21.3: 591-599, 2011.

SOUZA, A. P., LEITE, D. C. C, Pattathil S, Hahn MG, Buckeridge MS Composition and structure of sugarcane cell wall polysaccharides: implications for second-generation bioethanol production. **BioEnergy Research**, 6: 64–579, 2013.

TORRES, A; REVERÓN, J. "Integration of rock physics, seismic inversion, and support vector machines for reservoir characterization in the Orinoco Oil Belt, Venezuela." **The Leading Edge** 33.7: 774-782, 2014.

TROMBETA, N. DE C.; FILHO, J. V. C. Potencial e disponibilidade de biomassa de cana-de-açúcar na Região Centro-Sul do Brasil: Indicadores agroindustriais. **Revista de Economia e Sociologia Rural,** v. 55, n. 3, p. 479–496, 2017.

VAPNIK.N. V. **The nature of Statistical learning theory**. Springer-Verlag, New York, 1995.

WANG, H.; GU, J.; WANG, S. An effective intrusion detection framework based on SVM with feature augmentation. **Knowledge-Based Systems**, v. 136, p. 130–139, 2017.

WOLD, S. et al. Some recent developments in PLS modeling. **Chemometrics and Intelligent Laboratory Systems**, v. 58, n. 2, p. 131–150, 2001.

XU, J.L; RICCIOLI, C, SUN, D.W. Comparison of Vis or NIR Hyperspectral Imaging and Computer Vision for Automatic Differentiation of Organically and Conventionally Farmed Salmon, **Journal of Food Engineering**, 2016.

ZIDI, S.; MOULAHI, T.; ALAYA, B. Fault detection in wireless sensor networks through SVM classifier. **IEEE Sensors Journal**, v. 18, n. 1, p. 340–347, 2018.