ROBERTA DE AMORIM FERREIRA

REDES NEURAIS ARTIFICIAIS COM COMPONENTES PRINCIPAIS PARA A CONSTRUÇÃO DE MODELOS DE PREDIÇÃO EM DADOS DE ESPECTROSCOPIA NIR

Tese 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 *Doctor Scientiae*.

Orientador: Luiz Alexandre Peternelli

VIÇOSA - MINAS GERAIS
2022

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

T
F383r
2022

Ferreira, Roberta, 1990-
Redes neurais artificiais com componentes principais para a construção de modelos de predição em dados de espectroscopia NIR / Roberta Ferreira. – Viçosa, MG, 2022.
1 tese eletrônica (72 f.): il. (algumas color.).

Orientador: Luiz Alexandre Peternelli.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Estatística, 2022.
Referências bibliográficas: f. 64-72.
DOI: https://doi.org/10.47328/ufvbbt.2022.557
Modo de acesso: World Wide Web.

1. Análise de regressão. 2. Análise de componentes principais. 3. Redes neurais (Computação). 4. Quimiometria. 5. Predição. 6. Espectroscopia de infravermelho. I. Peternelli, Luiz Alexandre, 1966-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 519.536

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB-6/2523

ROBERTA DE AMORIM FERREIRA

REDES NEURAIS ARTIFICIAIS COM COMPONENTES PRINCIPAIS PARA A CONSTRUÇÃO DE MODELOS DE PREDIÇÃO EM DADOS DE ESPECTROSCOPIA NIR

Tese 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 *Doctor Scientiae*.

APROVADA: 06 de julho de 2022.

Assentimento:

Roberta de Amorim Ferreira
Autora

Luiz Alexandre Peternelli
Orientador

Aos negros e negras deste país que engrossam as estatísticas de analfabetismo, desemprego e exclusão social, assim como àqueles que conseguiram vencer as barreiras do preconceito e garantir um lugar ao sol, sem esquecer o desafio de construir uma sociedade livre de preconceitos.

AGRADECIMENTOS

Agradeço a Deus por ter me dado forças e ter me guiado durante todo o caminho, tornando possível a conclusão de mais uma fase da minha vida.

A minha mãe, Inês, pelo amor incondicional, dedicação e confiança. Sem você, a realização de mais esse sonho não seria possível.

À Letícia, que além de ser o amor da minha vida é minha melhor amiga. Obrigada pelos conselhos, pelo colo, amizade e paciência.

A meu irmão Gustavo e a meu sobrinho Kevin, meu agradecimento especial.

A meus familiares que, mesmo distantes, sempre apoiaram minhas escolhas e sonhos.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, pela oportunidade de cursar um doutorado de excelência.

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria, por contribuírem para minha formação acadêmica.

Um especial agradecimento ao professor Peternelli, que além das grandes contribuições como orientador, me ensinou muito sobre cooperação, trabalho em grupo e amizade. Você é um exemplo para todos nós.

Aos membros da banca examinadora, pelas contribuições e críticas ao trabalho.

Aos amigos de Viçosa, principalmente as meninas do "Jaja mais que sincera", que juntamente comigo enfrentaram dificuldades ao longo do curso. Jaqui, Leísa, Gabi França, Gaby Lazzarini e Carol... Amo vocês!

À República Malagueta (minha segunda casa em Viçosa) e aos demais amigos, pelo carinho e amizade.

Ao Instituto Federal de Minas Gerais (IFMG), por todo apoio e suporte na fase final do doutorado.

À CAPES, pelo suporte financeiro para o desenvolvimento deste trabalho.

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

Enfim, a todos que direta ou indiretamente colaboraram de alguma forma, fosse com um abraço ou com uma palavra de incentivo, para o sucesso deste trabalho.

MEU MUITO OBRIGADA!

BIOGRAFIA

Roberta de Amorim Ferreira, filha de Maria Inês de Amorim Ferreira e Roberto Divino Ferreira, nasceu em João Monlevade, Minas Gerais, em 06 de dezembro de 1990.

Em julho de 2015, graduou-se em Licenciatura em Matemática pela Universidade Federal de Viçosa, Viçosa-MG.

Em fevereiro de 2016, concluiu o curso de Mestrado em Estatística Aplicada e Biometria pela Universidade Federal de Viçosa, Viçosa-MG.

Em março de 2018 iniciou o curso de Doutorado em Estatística Aplicada e Biometria pela Universidade Federal de Viçosa, Viçosa-MG.

RESUMO

FERREIRA, Roberta de Amorim, D.Sc., Universidade Federal de Viçosa, julho de 2022. **Redes Neurais Artificiais com componentes principais para a construção de modelos de predição em dados de espectroscopia NIR**. Orientador: Luiz Alexandre Peternelli.

A espectroscopia no infravermelho próximo (NIR), associada a métodos estatísticos multivariados, vem sendo utilizada para a predição de indivíduos com maior produtividade. O método *Partial Least Squares* (PLS) é comumente empregado para ajuste de modelos de dados NIR. Entretanto, este método considera que a relação espectro/propriedade seja linear, o que não é sempre garantido em dados dessa natureza e o que pode, consequentemente, influenciar na acurácia do modelo. Alternativamente, a rede neural artificial (ANN) associada à análise de componentes principais (PCA), denominada PCA-ANN, possui a vantagem de ser eficiente em lidar com dados não lineares, incompletos e com ruídos, captando assim algumas complexidades presentes nos mesmos. Além disso, tal abordagem permite o não uso de pré-tratamentos, o que pode aumentar a capacidade preditiva dos modelos, além de diminuir o custo e o tempo das análises, quando comparada ao PLS associado aos pré-tratamentos usuais. O objetivo deste estudo foi construir e validar modelos de predição e processos de classificação, além de verificar se existe diferença significativa entre o método PLS, associado à matriz de espectros pré-tratados, e o método PCA-ANN, aplicado aos dados brutos. Para tanto, tais métodos foram aplicados a oito conjuntos de dados NIR, na forma bruta (sem pré-tratamentos) e com pré-tratamentos. A avaliação da capacidade preditiva dos modelos obtidos ocorreu por meio da correlação entre os valores preditos e os valores originais, e da raiz quadrada do erro quadrático médio de predição. Já a avaliação dos processos de classificação ocorreu através da taxa de erro aparente (TEA) e da taxa de verdadeiros positivos (TVP). Os resultados alcançados indicam que, na maioria dos conjuntos de dados analisados, o método PCA-ANN não difere estatisticamente do PLS para a predição dos modelos e para os processos de classificação, por meio da aplicação dos testes t e de Wilcoxon (valor-p > 0,01). O PCA-ANN deve ser escolhido para a realização de futuras análises, pois apresenta tempo computacional inferior àquele referente ao PLS.

Palavras-chave: PLS. PCA-ANN. Pré-tratamentos. Quimiometria. Predição.

ABSTRACT

FERREIRA, Roberta de Amorim, D.Sc., Universidade Federal de Viçosa, July, 2022**. Artificial Neural Networks with principal components for constructing prediction models in NIR spectroscopy data.** Adviser: Luiz Alexandre Peternelli.

Near infrared (NIR) spectroscopy, associated with multivariate statistical methods, has been used to predict individuals with higher productivity. The Partial Least Squares (PLS) method is commonly used to fit NIR data models. However, this method considers that the spectrum/property relationship is linear, which is not always guaranteed in data of this nature and can, consequently, influence the accuracy of the model. Alternatively, the artificial neural network (ANN) associated with principal component analysis (PCA), called PCA-ANN, has the advantage of efficiently dealing with non-linear, incomplete, and noisy data. PCA-ANN can, thus, capture some of the complexities of this kind of data. Also, the PCA-ANN approach allows for the non-use of pretreatments, in addition to reducing the cost and time of the analysis compared to the PLS associated with the usual pretreatments. The objective of this study was to build and validate prediction models and classification processes under PLS and PCA-ANN approaches, besides considering pre-treated and untrated sprectra. These methods were applied to eight NIR datasets, in raw form (without pretreatments) and with pretreatments. The model's predictive capacity was evaluated through the correlation between the predicted values and the original values, and the square root of the mean squared error of prediction. The evaluation of the classification processes was carried out through the apparent error rate (AER) and the true positive rate (TPR). In most of the analyzed datasets, the PCA-ANN method does not differ statistically from the PLS for the prediction and classification purposes through the application of t and Wilcoxon tests (p-value > 0.01). PCA-ANN should be chosen for future analysis as it has a shorter computational time than PLS.

Keywords: PLS. PCA-ANN. Pre-treatments. Chemometrics. Prediction.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

A espectroscopia no infravermelho próximo (NIR) associada a métodos estatísticos vêm sendo utilizada em diversas áreas da pesquisa para previsões de propriedades químicas e físicas de amostras. Dentre tais áreas, pode-se citar: a indústria farmacêutica, na predição de ingredientes ativos (ZHAO et al., 2021); a área florestal, com o objetivo de acelerar o processo de identificação de árvores (OLIVEIRA et al., 2015); o setor agrícola, na detecção rápida da poluição da água (CHEN et al., 2020); além de outras áreas, como a agricultura, a medicina, as indústrias de alimentos, petróleo, carvão e química (ARMENTA et al., 2010; DALE et al., 2013; GRASSI; ALAMPRESE, 2018; ZONTOV et al., 2016). A espectroscopia NIR é uma técnica, simples, rápida, exata e não gera resíduos no ambiente (MORGANO et al., 2008; VALDERRAMA et al., 2007).

Em quimiometria, constantemente utiliza-se o método *Partial Least Squares* (PLS), para ajuste de modelos de dados de espectroscopia NIR (TEÓFILO; MARTINS; FERREIRA, 2009). Entretanto, esse método considera que a relação espectro/propriedade seja linear, o que não é sempre garantido em dados dessa natureza, podendo, consequentemente, influenciar na exatidão do modelo (BALABIN; LOMAKINA, 2011).

A escolha do número de variáveis latentes (nVL) a ser utilizado no PLS muitas vezes é um impasse enfrentado pelo pesquisador. Além disso, sabe-se que dados não lineares podem ser modelados pelo PLS, porém de maneira limitada, adotando-se algumas estratégias como a aplicação de funções de pré-tratamentos de dados ou a utilização de mais variáveis latentes (JANIK; FORRESTER; RAWSON, 2009). No entanto, neste último caso pode ocorrer um super ajuste do modelo (BALABIN; LOMAKINA, 2011; JANIK; FORRESTER; RAWSON, 2009; MANLY, 2004).

Segundo Pudełko e Chodak (2020), o uso das componentes principais (CPs), geradas a partir da análise de componentes principais (PCA), como dados de entrada da rede neural artificial (ANN) pode ser mais eficiente na modelagem de dados espectrais não lineares do que a utilização do PLS, proporcionando melhor capacidade de predição dos modelos. Assim como nos trabalhos de Oliveira-Esquerre, Mori e Bruns (2002) e Yoplac et al. (2019), a associação do PCA à ANN foi denominada neste estudo como PCA-ANN.

Nesse contexto, a PCA-ANN constitui-se um novo paradigma, que pode ser empregado para predição de dados NIR em concomitância ou em substituição às metodologias tradicionais utilizadas. Essa abordagem permite o não uso de pré-tratamentos sobre a matriz de espectros, e ainda assim aumentar a capacidade preditiva, além de diminuir o custo e o tempo das análises,

quando comparada ao PLS associado aos pré-tratamentos comumente aplicados em dados NIR. O PCA-ANN vem sendo aplicado com sucesso na solução de diversos problemas envolvendo dados NIR (MIREEI; SADEGHI, 2013; MOZAFFARI; SADEGHI; ASEFI, 2022; PUDEŁKO; CHODAK, 2020).

Desta forma, este trabalho teve como objetivo construir e validar modelos de predição e processos de classificação, empregando-se os métodos PLS e PCA-ANN aplicados a diversos conjuntos de dados NIR, na forma bruta (sem pré-tratamentos) e com pré-tratamentos, além de verificar se existe diferença significativa entre o PLS associado a pré-tratamentos e o PCA-ANN aplicado aos dados brutos.

2. REVISÃO DE LITERATURA

2.1.Quimiometria

A quimiometria é uma disciplina da área de química, que estuda e analisa o ajuste de métodos estatísticos e matemáticos a dados de origem química, e tem como objetivo a obtenção de informações de um conjunto de dados complexo (FERREIRA et al., 1999; BRERETON, 2003).

O grande avanço computacional dos últimos anos possibilitou a manipulação e processamento de grandes conjuntos de dados. Tal fato, associado à evolução das técnicas instrumentais, permitiu que a quimiometria se difundisse ainda mais na comunidade científica, nas indústrias químicas, farmacêuticas, de alimentos para controle de qualidade, dentre outras (FERREIRA, 2015).

Uma maneira de se obter dados em quimiometria é através de instrumentos de espectroscopia. Este trabalho abordará, especialmente, a espectroscopia NIR associada ao método estatístico tradicional PLS, e também ao método não usual PCA-ANN. Tais procedimentos, serão apresentados brevemente a seguir.

2.2.Espectroscopia no Infravermelho Próximo (NIR)

A espectroscopia é o estudo das interações da radiação eletromagnética com a matéria. A partir da utilização de instrumentos de espectroscopia em amostras coletadas pelo pesquisador, obtém-se o espectro, que contém informações importantes sobre as suas propriedades químicas (BLANCO; VILLARROYA, 2002).

Quando os comprimentos de onda na região do infravermelho incidem sobre os átomos de uma certa molécula, parte da radiação é absorvida e parte é transmitida, sendo esta última responsável por gerar o espectro NIR (VERHOEVEN, 2008). A espectroscopia NIR é uma técnica simples, rápida, precisa, não destrutiva, que não gera resíduos no ambiente, além de não exigir grandes preparos na amostra (FERNANDEZ et al., 2020; MORGANO et al., 2008; PASQUINI, 2003; VALDERRAMA; BRAGA; POPPI, 2007).

A região espectral do NIR de bancada abrange, em geral, uma faixa de radiação que varia de 700 a 2500 nm, em comprimento de onda (BOKOBZA, 1998). Este instrumento apresenta maior acurácia, ou seja, maior poder preditivo. Contudo, além do mesmo estar

associado a um custo elevado, existe um outro fator limitante, uma vez que as amostras precisam ser levadas ao laboratório onde o instrumento se encontra (PASQUINI, 2003).

Já o NIR portátil, geralmente, engloba uma faixa de radiação que varia de 900 a 1700 nm, em comprimento de onda. Tal instrumento possui as vantagens de ter um menor custo, além de uma maior praticidade, visto que pode ser levado até o local onde se encontram as amostras (SANTOS et al., 2020).

Os dados oriundos do NIR são dispostos em uma matriz **X** (N x p), em que cada linha i (i = 1 a N) se refere a uma amostra (espectro) e, cada coluna j (j = 1 a p) se refere aos comprimentos de ondas (variáveis). Nota-se que apenas uma única amostra origina muitas variáveis, de maneira que o número de comprimentos de onda lidos no NIR é superior ao de amostras coletadas ($p >> N$).

Percebe-se que a alta dimensionalidade e a multicolinearidade são características encontradas em dados NIR (FREUND; WILSON; SA, 2006), o que implica na necessidade de utilização de métodos específicos para análise e modelagem desses dados (TEÓFILO, 2007). Contudo, na quimiometria, é comum a aplicação de alguns pré-tratamentos a fim de reduzir ruídos experimentais que podem prejudicar a interpretação dos resultados, antes do uso de tais métodos (FERREIRA, 2015).

2.3.Pré-tratamentos

A aplicação de alguns pré-tratamentos na matriz de dados NIR permite, frequentemente, que o pesquisador tenha uma melhor interpretação qualitativa e quantitativa dos espectros (PASQUINI, 2018; STWART et al., 1995). A aplicação de pré-tratamentos promove a remoção e/ou redução de variações indesejáveis (ruídos experimentais), por meio do uso de operações matemáticas nos conjuntos de dados (DE SOUZA; POPPI, 2012; XU et al., 2008).

Os espectros das amostras são constituídos do sinal verdadeiro, que possui informações relevantes ou não sobre as propriedades de interesse, e do sinal aleatório, que contém os ruídos experimentais (CHEN et al., 2004). Tais ruídos podem ocorrer em virtude do espalhamento da luz, da resolução do instrumento, da superfície da amostra, dentre outros fatores (ENGEL et al., 2013). Deve-se ter cautela ao escolher os pré-tratamentos pois, caso contrário, informações relevantes podem ser eliminadas, o que poderá prejudicar a qualidade do modelo a ser construído (FERREIRA, 2015; PASQUINI, 2018).

Os pré-tratamentos são divididos em dois tipos, denominados transformações e pré-processamentos. As transformações são aplicadas às amostras (linhas da matriz **X** de dados); os pré-processamentos são aplicados às variáveis (colunas da matriz **X** de dados) (FERREIRA, 2015; PASQUINI, 2018). Ferreira (2015) descreve algumas técnicas de pré-tratamentos que podem ser aplicadas isoladamente ou combinadas:

- **Centrar na média**, que faz com que o centro das coordenadas do sistema seja movido para a média multivariada dos dados;
- **Alisamento**, que tem como finalidade suavizar o ruído que acompanha o sinal analítico, reduzindo a componente aleatória e aumentando a razão sinal/ruído. O método mais usado para esse fim é o alisamento Savitzky-Golay, proposto por Savitzky e Golay (1964);
- **Correção Multiplicativa de Espalhamento ou *Multiplicative Scatter Correction* (MSC)**, que tenta remover variações nos espectros ocasionadas pelo espalhamento de luz pelas amostras;
- **Primeira e segunda derivada**, que geralmente acentuam as informações presentes nas variáveis, e que devem ser aplicadas com cuidado, pois além de acentuarem o sinal verdadeiro, podem também acentuar o sinal aleatório.

Após a escolha dos pré-tratamentos a serem aplicados, deve-se empregar métodos estatísticos para a construção do modelo matemático de calibração multivariada, que tem como objetivo estabelecer uma relação funcional entre os comprimentos de ondas, obtidos a partir de cada amostra analisada, e a propriedade relevante ao estudo (FERREIRA, 2015).

A seguir, são descritos os métodos estatísticos utilizados neste trabalho para a análise de espectros NIR.

2.4.Análise de Componentes Principais (PCA)

A Análise de Componentes Principais (PCA) é um dos métodos estatísticos mais utilizados na quimiometria para modelagem de dados NIR. Isso ocorre devido à alta correlação presente entre as variáveis em dados desta natureza, que muitas vezes apresentam informações redundantes (FERREIRA, 2002; PASQUINI, 2018).

Ao realizar-se a aplicação do PCA no conjunto de dados, as informações presentes nas variáveis originais são substituídas por novas informações mais significativas (MARTENS; NAES, 1989). As variáveis desse novo conjunto de dados são denominadas componentes

principais (CPs). Assim, tem-se um novo conjunto de dados com menor número de variáveis, informações menos redundantes e mais significativas quando comparado ao conjunto original, o que pode resolver os problemas de dimensionalidade e multicolinearidade presentes em dados NIR (SABIN; FERRÃO; FURTADO, 2004).

Para obter as CPs da matriz de dados, pode-se utilizar o algoritmo diagonal que é baseado na decomposição por valores singulares (SVD) (GOLUB; VAN LOAN, 1989). Tal algoritmo afirma que toda matriz pode ser escrita de acordo com a Equação 1:

$$\mathbf{X} = \mathbf{URV^t} \tag{1}$$

Na bibliografia, **X** é descrita como a matriz inicial de dados e **UR** é denominada matriz de escores. A matriz **V**, chamada de matriz de *loadings*, armazena a informação sobre a relação entre as variáveis originais e as CPs. Já a matriz diagonal **R** contém informações sobre a quantidade de variância em cada componente.

As matrizes **U** e **V** produzem um novo sistema de coordenadas formado pelas n CPs, sendo que cada uma delas é construída a partir da combinação das variáveis originais. A primeira componente principal (CP1) e a segunda componente principal (CP2) são definidas na direção (eixo) de maior variância e de segunda maior variância do conjunto de variáveis originais, respectivamente, e assim sucessivamente, até a n-ésima componente principal (CPn) definida. Ressalta-se que todas as CPs são sempre ortogonais entre si (SABIN; FERRÃO; FURTADO, 2004).

2.5.Regressão por Quadrados Mínimos Parciais (*Partial Least Squares* - PLS)

Na construção de modelos estatísticos para predição de dados NIR, o método PLS (WOLD, 1982) é atualmente o mais utilizado. Além de ser eficiente em lidar com ruídos experimentais, a regressão PLS também é capaz de lidar com a multicolinearidade e a alta dimensionalidade, que são características fortemente encontradas em dados espectrais (PASQUINI, 2018; TEÓFILO; MARTINS; FERREIRA, 2009).

Assim como o PCA, o PLS é um método fundamentado na compressão dos dados originais. Para executá-lo, também pode-se utilizar o algoritmo bidiagonal que é baseado na decomposição de valores singulares (SVD), conforme apresentado na Equação 2 (BARLOW; BOSNER; DRMAČ, 2005):

$$\mathbf{y} \rightarrow \mathbf{X} = \mathbf{URV^t} \tag{2}$$

Conforme descrito na literatura, as colunas da matriz **U** e as linhas da matriz $\mathbf{V^t}$ criam os novos subespaços conhecidos como variáveis latentes, que possuem informações presentes na matriz de dados **X** e no vetor **y**, que contém valores de alguma propriedade específica das amostras (WOLD; SJÖSTRÖM; ERIKSSON, 2001)**.** Geralmente, as primeiras variáveis latentes (2 a 10) fornecem quase toda (aproximadamente 100%) informação da matriz de dados **X** original (ROQUE, 2015).

O algoritmo para execução do PLS é apresentado, resumidamente, a seguir (MARTINS; TEOFILO; FERREIRA, 2010):

- Inicialize o algoritmo para primeira componente:

$$\mathbf{X} = \mathbf{y}\mathbf{v_1^t}$$
$$\mathbf{X^t} = \mathbf{v_1}\mathbf{y^t}$$
$$\mathbf{X^t}\mathbf{y} = \mathbf{v_1}\mathbf{y^t}\mathbf{y}$$
$$\mathbf{X^t}\mathbf{y}(\mathbf{y^t}\mathbf{y})^{-1} = \mathbf{v_1}(\mathbf{y^t}\mathbf{y})(\mathbf{y^t}\mathbf{y})^{-1}$$
$$\mathbf{v_1} = \mathbf{X^t}\mathbf{y}(\mathbf{y^t}\mathbf{y})^{-1} \tag{3}$$

Para normalizar o vetor, basta dividi-lo por sua norma euclidiana, ou seja, $\|v_1\| = \sqrt{v_1^t v_1}$. O vetor normalizado da Equação 3 é:

$$\mathbf{v_1} = \frac{\mathbf{X^t}\mathbf{y}}{\|\mathbf{X^t}\mathbf{y}\|} \; ; \; \alpha_1 \mu_1 = \mathbf{X}\mathbf{v_1}$$

- Para i = 2, ..., h componentes:

$$y_{i-1}\mathbf{v_1} = \mathbf{X^t}\mu_{i-1} - \alpha_{i-1}\mathbf{v_{i-1}}$$
$$\alpha_i \mu_1 = \mathbf{X}\mathbf{v_i} - y_{i-1}\mu_{i-1}$$

$$\mathbf{V_h} = (\mathbf{v_1}, \dots, \mathbf{v_h}) \quad \mathbf{U_h} = (\mu_1, \dots, \mu_i) \quad \text{e} \quad \mathbf{R_h} = \begin{pmatrix} \alpha_1 & y_1 & & \\ & \ddots & & \\ & & \alpha_{k-1} & y_{k-1} \\ & & & \alpha_k \end{pmatrix}$$

Sendo h o nVL escolhido, prova-se que $\mathbf{U_h R_h = XV_h}$ e, então, $\mathbf{R_{h=} U_h^t XV_h}$. Portanto, a partir do cálculo das matrizes **U**, **V** e **S**, pode-se estimar a pseudo-inversa de Moore-Penrose de **X** e construir o modelo da seguinte maneira:

$$\mathbf{y = X\beta \rightarrow X = U_h R_h V_h^t \ \rightarrow y = U_h R_h V_h^t \beta \rightarrow \ \widehat{\beta} = V_h R_h^{-1} U_h^t y}$$

O método *leave-one-out* de validação cruzada pode ser empregado para a escolha do nVL (GOURVÉNEC et al., 2003). Nesta tese, o parâmetro utilizado para escolha do nVL foi a raiz do erro quadrático médio de validação cruzada (RMSECV), apresentada na Equação 4.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \qquad (4)$$

Na Equação 4, N e y_i referem-se, respectivamente, ao número de amostras da validação e ao i-ésimo valor referente ao vetor de referência **y**. Já $\hat{y}_i$ diz respeito ao i-ésimo valor previsto pelo modelo.

Diversos pesquisadores mostram a potencialidade do PLS na modelagem de dados (BOGREKCI; LEE, 2005; MCCARTY et al., 2002; VASQUES; GRUNWALD; SICKMAN, 2008). Entretanto, há pesquisadores que apontam outros métodos com maior capacidade preditiva do que o PLS, como por exemplo, a ANN (KODAIRA; SHIBUSAWA, 2013; ROSSEL; BEHRENS, 2010).

Considerando o grande potencial do uso da abordagem ANN nos dados NIR, o tópico a seguir apresenta uma revisão de literatura sobre este método.

2.6.Rede Neural Artificial (*Artificial Neural Network* - ANN)

A inteligência humana é uma das mais desenvolvidas no meio biológico, o que desperta grande interesse na reprodução das funções cognitivas da mesma. O cérebro humano é capaz de reconhecer padrões e relacioná-los, além de adquirir conhecimentos através da experiência. É composto de neurônios, que são a base da ANN (BRAGA; CARVALHO; LUDERMIR, 2007).

A ANN é uma das abordagens utilizadas na área de Inteligência Computacional que fundamenta-se na simulação do cérebro humano e no seu comportamento (HAYKIN, 2001).

Conforme Braga, Carvalho e Ludermir (2007), estudos envolvendo ANN têm instigado pesquisadores devido à possibilidade de obtenção de modelos ainda mais eficientes do que aqueles usados comumente nas suas respectivas áreas de pesquisa.

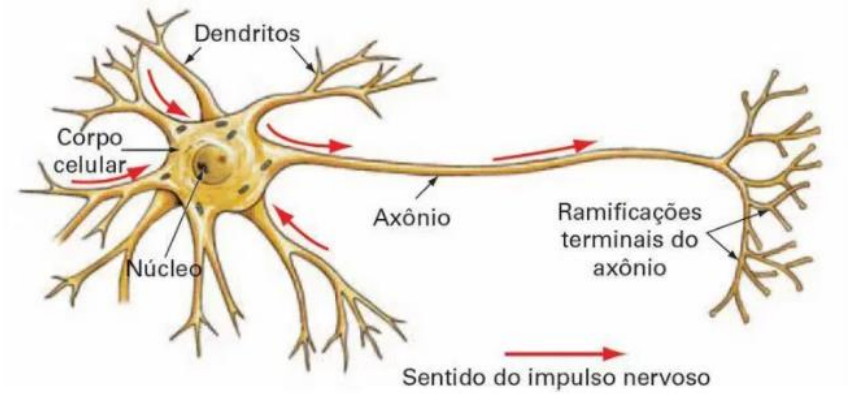
Constantemente, observa-se na literatura a utilização das técnicas de ANN aplicadas a diversas áreas de pesquisa (DASGAONKAR, 2018; GIANOLA et al., 2011; GOYAL, 2014; PETERNELLI et al., 2017; SILVA et al., 2016; SIQUEIRA-BATISTA et al., 2014). Vale ressaltar que o emprego desse método no melhoramento genético, animal e vegetal, representa uma boa ferramenta na tomada de decisões (NASCIMENTO et al., 2013; VENTURA et al., 2012).

A ANN possui a vantagem de ser eficiente em lidar com dados não lineares, incompletos e com ruídos (BISHOP, 2006; NAGY et al., 2019), captando assim algumas complexidades presentes em tais dados, o que nem sempre poderia ser possível com a aplicação dos métodos tradicionais (GALVÃO et al., 1999). Além disso, a rede é capaz de aprender a partir de exemplos, identificar padrões escondidos em dados anteriores e usá-los para previsão (MAURO; CAMPOS, 2020).

2.6.1.Neurônios biológicos e neurônios artificiais

Conforme apresentado na Figura 1, os neurônios biológicos são constituídos por três elementos principais: corpo celular, dendritos e axônio. Os dendritos são responsáveis por conduzir os sinais (impulsos nervosos) das extremidades ao corpo celular, onde são processados, formando um sinal excitante ou inibitório. Logo após, esses sinais são levados até os dendritos de outros neurônios através do axônio. A conexão entre o axônio de um neurônio e o dendrito posterior é chamada de sinapse. As sinapses são responsáveis pela relação funcional entre os neurônios, formando as redes neurais biológicas (GUYTON, 1988).

Figura 1- Modelo de neurônio Biológico

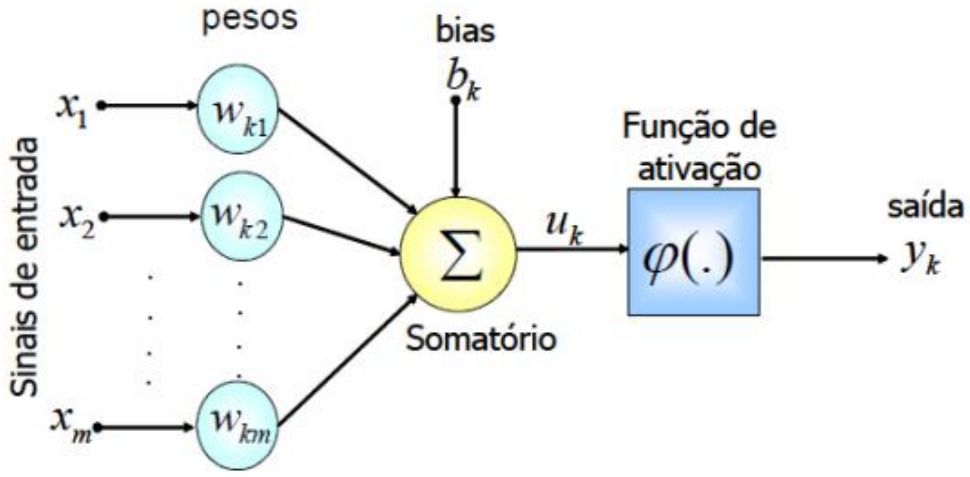


Fonte: Lopes (2000).

O neurônio artificial é baseado no funcionamento do neurônio biológico. Além disso, é formado por: m entradas (x_1, x_2, ..., x_m), pesos ponderados (w_{k1}, w_{k2}, ..., w_{km}), uma porta do limiar (somatório), uma função de ativação φ(x), e um terminal de saída do neurônio (HAYKIN, 2001).

O primeiro modelo artificial de um neurônio biológico foi desenvolvido por Warren McCulloch e Walter Pitts, em 1943 (MCCULLOCH; PITTS, 1943) (Figura 2). Nesse modelo, é possível observar todos os elementos do neurônio artificial.

Figura 2 - Modelo não linear de um neurônio artificial



Onde: x_1, x_2, ..., x_m são as entradas da rede; w_{k1}, w_{k2}, ..., w_{km} são os pesos ou pesos sinápticos, associados a cada entrada; b_k é o termo bias (porta do limiar); u_k é a combinação linear dos sinais de entrada; φ(.) é a função de ativação; e y_k é a saída do neurônio.
Fonte: Vendruscolo et al. (2015).

No processo de treinamento de uma ANN, inicialmente tem-se o ajuste dos km pesos às m entradas da rede. Estes pesos são parâmetros que podem ser valores positivos ou

negativos, dependendo do sinal sináptico (inibitório ou excitatório), e variam à medida que um novo conjunto de treinamento é apresentado à rede (CRUZ; NASCIMENTO, 2018). Desta forma, os pesos são responsáveis por todo o conhecimento adquirido pela rede.

Treinamento, aprendizado e validação são etapas importantes para a construção de modelos envolvendo ANN, assim como a escolha adequada de funções de ativação, associada a uma arquitetura apropriada de rede e a um bom ajuste de seus pesos. A seguir, são detalhados alguns componentes relevantes do modelo neural.

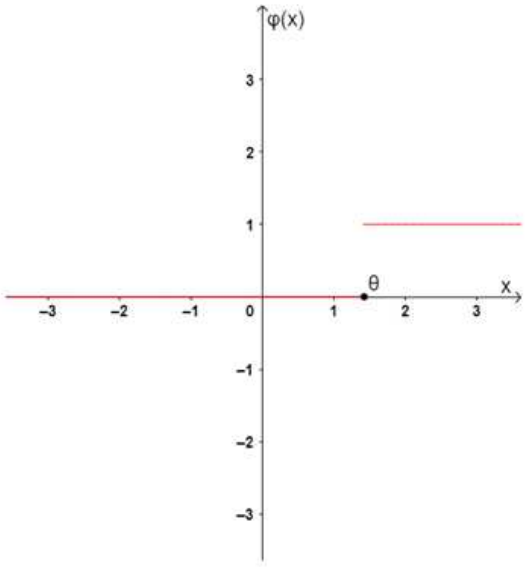
2.6.2.Funções de ativação

No processo da construção de modelos envolvendo ANN, é realizada uma soma ponderada de cada sinal de entrada com um determinado peso ajustável. Cada neurônio possui uma porta do limiar, que tem a função de comparar cada soma ponderada com um determinado valor limite, chamado de limiar. Caso o valor da soma seja maior ou igual ao limiar, a saída é ativada; caso contrário, é desativada (BRAGA; CARVALHO; LUDERMIR, 2007). Diante disso, será possível saber a influência de um determinado peso na saída da unidade.

As funções de ativação são baseadas nas somas ponderadas e fornecem o valor das saídas y dos neurônios, dentro de um intervalo de valores, conforme a imagem da função escolhida. Geralmente, os intervalos são [0, 1] ou [-1, 1]. A função de ativação será escolhida, conforme o problema em estudo (HAYKIN, 2001).

A função limiar ou degrau é utilizada no modelo de McCulloch e Pits (MCCULLOCH; PITTS, 1943). A Figura 3 exemplifica o gráfico dessa função.

Figura 3 - Gráfico da função limiar ou degrau



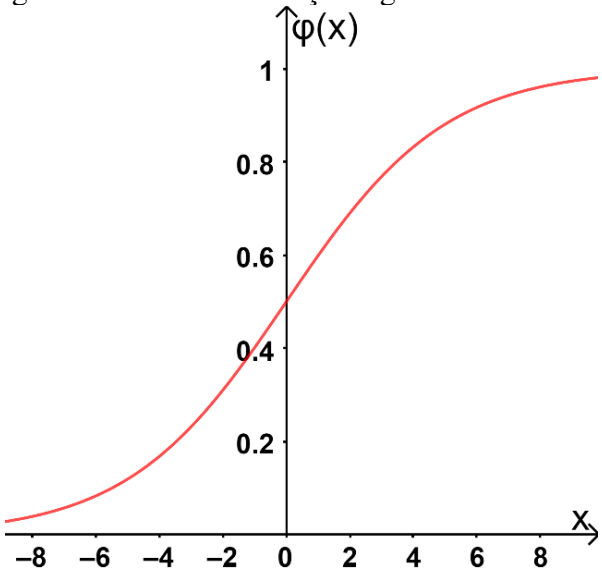
Fonte: Adaptado de Haykin (2001).

A lei da função limiar é representada na Equação 5, na qual observa-se que a função limiar assume apenas os valores 0 ou 1:

$$\varphi(x) = \begin{cases} 0, & se\ x < \theta \\ 1, & se\ x \geq \theta \end{cases} \tag{5}$$

A função de ativação sigmóide logística (logsig) é uma das mais utilizadas em ANN (HAYKIN, 2001), e seu gráfico pode ser visto na Figura 4.

Figura 4 - Gráfico da função sigmoidal



Fonte: Adaptado de Haykin (2001).

A função logística é um exemplo de função sigmoidal, definida pela Equação 6:

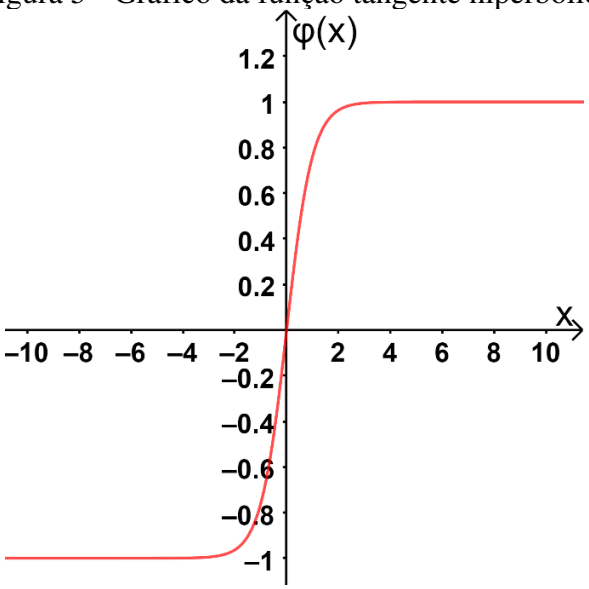
$$\varphi(x) = \frac{1}{1 + e^{-ax}} \tag{6}$$

A variação do parâmetro "a" produzirá diferentes inclinações para a função sigmoidal. Quando este parâmetro tender ao infinito, esta função se tornará uma função limiar, porém, com valores da imagem pertencentes a um intervalo contínuo entre 0 e 1. As principais características da função sigmoidal são a suavidade e a continuidade. Tal função é frequentemente utilizada em redes multicamadas ou em redes com sinais contínuos (HAYKIN, 2001).

Outra função interessante é a tangente hiperbólica (tansig), que tem a forma antissimétrica em relação à origem, com valores da imagem variando num intervalo entre -1 e 1 (HAYKIN, 2001). Essa função é definida pela Equação 7 e seu gráfico pode ser observado na Figura 5:

$$\varphi(x) = tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{7}$$

Figura 5 - Gráfico da função tangente hiperbólica



Fonte: Adaptado de Haykin (2001).

Geralmente, a função tansig é utilizada em redes com sinais contínuos, e seu grande diferencial é a possibilidade de os resultados também serem números negativos (HAYKIN, 2001).

2.6.3.Arquitetura da rede

Segundo Haykin (2001), o número de camadas, as conexões existentes entre camadas, o número de neurônios em cada camada e, por fim, o algoritmo de aprendizado são os elementos que compõem a arquitetura de uma rede neural.

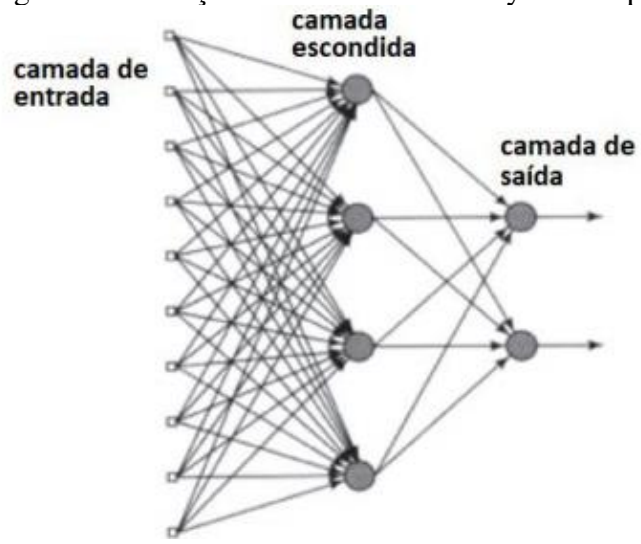
O número de neurônios em cada camada influencia diretamente na exatidão e na capacidade da rede em reconhecer padrões. Esse número depende do problema a ser resolvido, da quantidade de ruídos nos dados, do número de exemplos de treinamento, da função a ser aprendida pela rede e da distribuição dos dados de treinamento (TAFNER; XEREZ; FILHO, 1995). Um número alto de neurônios em cada camada pode gerar problemas de *overfitting*, tornando a rede incapaz de reconhecer padrões. Por outro lado, um número inferior ao necessário fará com que a rede não encontre uma solução apropriada. Portanto, é fundamental a escolha adequada desse número (HAYKIN, 2001).

A escolha do número de camadas de uma rede neural é de grande relevância, pois esse valor influencia diretamente na capacidade da rede de solucionar problemas. A rede de camada única possui uma estrutura mais simples, constituída por uma camada de saída e uma camada de entrada, com uma direção bem definida. Em uma arquitetura de rede com múltiplas camadas, geralmente, são identificados três tipos de camadas (HAYKIN, 2001):

- **Camada de Entrada:** onde os padrões são apresentados à rede;
- **Camadas Intermediárias ou Ocultas ou Escondidas:** onde os pesos são atribuídos às entradas da rede, mediante conexões; são responsáveis por boa parte do processamento dos dados;
- **Camadas de Saída:** onde tem-se a saída de rede e a saída desejada, que fornecem os valores preditos para as variáveis de interesse.

A Figura 6 ilustra uma arquitetura simples de rede, na qual podem ser observadas as camadas de entrada, as camadas escondidas (intermediárias ou ocultas) e as camadas de saída. Esse modelo neural é conhecido como Perceptron múltiplas camadas ou *Multilayer Perceptron* (MPL), e é capaz de lidar com problemas não linearmente separáveis, além de ser o mais utilizado em ANN (BRAGA; CARVALHO; LUDERMIR, 2007).

Figura 6 -Ilustração de uma rede Multilayer Perceptron (MLP)



Fonte: Santos et al. (2005).

O modelo MPL se diferencia dos demais pela utilização de uma função de ativação não linear, de camadas de entrada e saída, de uma ou mais camadas de neurônios ocultas e, por fim, pelo alto grau de conectividade, determinado pelas sinapses. Tal modelo tem grande capacidade de resolver problemas mais complexos, treinando sua rede através do método de aprendizado supervisionado com o algoritmo conhecido como *Backpropagation* (HAYKIN, 2001).

O algoritmo *Backpropagation* é baseado no cálculo do erro obtido, quando se compara a resposta fornecida na camada de saída da rede neural com a saída esperada. Inicialmente, são atribuídos pesos aleatórios às redes e, durante o processo de aprendizagem, esses pesos são associados aos valores de entrada, gerando uma resposta na camada de saída. O valor do erro é, então, calculado. Esse valor é retro-propagado da camada de saída para a camada de entrada, e os pesos são ajustados de forma a minimizar o erro obtido. Esse procedimento é repetido até que se tenha a saída desejada (com erro mínimo) (RUMELHART; HINTON; WILLIAMS, 1986).

Diante dos conceitos estudados, a ANN é uma ferramenta potencial a ser utilizada no melhoramento genético. O tópico seguinte apresenta uma revisão de literatura que abrange trabalhos envolvendo a aplicação da ANN em dados NIR.

2.6.4.Potencialidade das redes neurais artificiais na predição de dados NIR

No estudo conduzido por Chen et al. (2001), objetivou-se prever o conteúdo de drogas e a dureza de comprimidos usando ANN em dados NIR. Esse trabalho demonstrou que um modelo de ANN bem treinado representa uma técnica alternativa poderosa para a análise de tais dados. Além disso, os autores enfatizam que a ANN pode ser muito útil nos casos em que a modelagem convencional de dados não funciona adequadamente.

Mutlu et al. (2011) utilizaram o NIR associado à ANN, com o objetivo de investigar a capacidade desse método em prever parâmetros de qualidade da farinha de trigo. Evidencia-se, neste estudo, o uso da ANN devido à natureza não linear dos dados. Um total de 79 amostras de farinha de diferentes variedades de trigo foram analisadas. Os resultados indicam que tal associação pode ser considerada como uma ferramenta valiosa para a predição da qualidade da farinha de trigo.

Para corrigir problemas relacionados à não linearidade dos dados, algumas estratégias podem ser úteis, como a realização de pré-tratamentos nos mesmos ou, ainda, o uso de métodos não lineares (BALABIN; SAFIEVA; LOMAKINA, 2007). Com o objetivo de acelerar o processo de identificação de árvores, Oliveira et al. (2015) utilizaram a ANN associada à espectroscopia NIR para a classificação de quatro espécies de madeira. Destaca-se, neste trabalho, que os espectros não foram submetidos aos pré-tratamentos comumente aplicados a dados NIR e, mesmo assim, os resultados obtidos foram promissores, evidenciando que a ANN é uma metodologia flexível aos ruídos e distorções presentes nos dados.

Balabin, Safieva e Lomakina (2007) compararam a exatidão, a complexidade computacional e a facilidade de aplicação de seis métodos populares para a predição de propriedades químicas da gasolina, por meio de espectros NIR. Os métodos avaliados foram: regressão linear múltipla (MLR), regressão por componentes principais (PCR), regressão por quadrados mínimos parciais polinomiais (Poly-PLS), regressão por quadrados mínimos parciais de *spline* (Spline-PLS), PLS e ANN. Concluiu-se, neste estudo, que os métodos não lineares (Poly-PLS, Spline-PLS e ANN) foram superiores aos lineares (MLR, PCR e PLS), com a ANN sendo mais precisa, porém, com alto tempo computacional e baixa facilidade de aplicação.

Zhao et al. (2021) combinaram o uso de dados NIR com abordagens de aprendizado de máquina, incluindo a ANN para predição de princípio ativo de três medicamentos. O PLS também foi empregado e seu desempenho de predição comparado aos outros métodos

propostos. Os resultados indicam que as abordagens de aprendizado de máquina são mais estáveis, previsíveis e adequadas, quando é necessária uma análise de alta exatidão.

No estudo conduzido por Balabin e Lomakina (2011), objetivou-se comparar a robustez de cinco métodos estatísticos multivariados, incluindo o PLS e a ANN, em 14 conjuntos de dados distintos, sendo alguns desses dados NIR. O método ANN mostrou-se mais preciso na grande maioria dos conjuntos de dados, com reduções pontuais na média do erro de predição. Isso pode se justificar, pois a maioria dos dados utilizados apresentaram uma tendência de não linearidade.

Apesar da ANN lidar de forma eficiente com dados não lineares, incompletos e com ruídos (BISHOP, 2006; JANIK; FORRESTER; RAWSON, 2009; NAGY et al., 2019), o que é uma grande vantagem em relação aos métodos tradicionais aplicados a dados NIR, alguns autores relatam o elevado custo computacional da aplicação desse método, principalmente em conjuntos de dados com alta dimensionalidade (BALABIN; SAFIEVA; LOMAKINA, 2007; JANIK; FORRESTER; RAWSON, 2009). Uma alternativa para enfrentar esses desafios seria a utilização do PCA associado à ANN, conforme proposto neste trabalho, com a denominação de PCA-ANN.

2.7.Análise de Componentes Principais associada às Redes Neurais (PCA-ANN)

Para executar o PCA-ANN, inicialmente aplica-se o PCA ao conjunto original de dados a fim de obter um novo conjunto de dados com poucas variáveis (CPs), com informações menos redundantes e com maior variabilidade, quando comparado ao conjunto original. Em seguida, as CPs são utilizadas como entrada da ANN (GUO et al., 2016), o que poderá ser vantajoso pela possibilidade de não se aplicar os pré-tratamentos comumente utilizados em quimiometria.

O PCA-ANN pode contribuir para o aumento da capacidade preditiva e de classificação dos modelos, além de reduzir o tempo computacional da rede. Assim, utilizar CPs como entrada da ANN pode ser um meio eficaz para reduzir a dimensionalidade dos dados, mantendo as principais informações espectrais (MIREEI; MOHTASEBI; SADEGHI, 2014).

Encontram-se, na literatura, trabalhos que utilizaram o PCA-ANN para predição de alguma característica de interesse do pesquisador (MIREEI; MOHTASEBI; SADEGHI, 2014; MOZAFFARI; SADEGHI; ASEFI, 2022; NOVAIS; VALVERDE, 2021). Outros estudos utilizaram este método para classificação das amostras pesquisadas (JILANI, 2011; MIREEI;

SADEGHI, 2013; SACHDEVA et al., 2013). Todos estes trabalhos apontam a eficiência do PCA-ANN, tanto para problemas de predição quanto para de classificação.

2.8.Métodos de divisão dos conjuntos de dados

Uma etapa importante na construção dos modelos estatísticos é a divisão do conjunto de dados em três subconjuntos complementares e disjuntos: treino, teste e validação, (FERREIRA; TEIXEIRA; PETERNELLI, 2022). Os subconjuntos de treino e teste são empregados para a construção do modelo, enquanto o subconjunto de validação é utilizado para testar a sua capacidade preditiva (GALVÃO et al., 2005; LEE; LIONG; JEMAIN, 2018).

A amostragem aleatória (RS) (RAJER-KANDUČ; ZUPAN; MAJCEN, 2003) pode ser empregada na obtenção do subconjunto de validação. Já o algoritmo Kennard Stone (KS), proposto por R.W. Kennard e L.A. Stone em 1969 (KENNARD; STONE, 1969), pode ser aplicado na divisão dos subconjuntos de treino e teste.

No estudo desenvolvido por Ferreira, Teixeira e Peternelli (2022), inicialmente utilizou-se a RS para a seleção das amostras do subconjunto de validação. As amostras restantes foram empregadas na comparação entre os métodos RS e KS, a fim de avaliar a influência da escolha do subconjunto de treino na construção dos modelos, bem como sua validação em dados NIR. Nos resultados apresentados por aqueles autores, os métodos KS e RS diferiram estatisticamente entre si, sendo que o primeiro apresentou melhor capacidade preditiva do que o segundo para a divisão dos dados em treino e teste.

2.8.1.Amostragem Aleatória (RS)

Assim como o KS, a RS é um método útil e simples, capaz de selecionar os elementos que formarão os subconjuntos amostrais. De modo geral, de uma lista com N unidades elementares, a RS seleciona com igual probabilidade, sem reposição e sequencialmente, n ($n < N$) elementos que formarão o conjunto amostral (BOLFARINE; BUSSAB, 2005). Isso garante que o subconjunto selecionado siga a distribuição estatística de todo o conjunto (FERREIRA; TEIXEIRA; PETERNELLI, 2022; GALVÃO et al., 2005).

2.8.2.Kennard- Stone (KS)

O KS é um método que pode ser utilizado para dividir o conjunto de dados em dois subconjuntos (treino e teste), de maneira a selecionar amostras que representem o máximo de sua variabilidade (DE SOUSA et al., 2011; FERREIRA; TEIXEIRA; PETERNELLI, 2022). Para selecionar as amostras que formarão o subconjunto de treino, o KS utiliza a distância euclidiana para cada par (p, q) de amostras, conforme apresentado na Equação 8, em que J representa o número de comprimento de onda e N corresponde ao tamanho da amostra:

$$d_x(p,q) = \sqrt{\sum_{j=1}^{J}\left[x_p(j) - x_q(j)\right]^2} \quad p, q \in [1, N] \tag{8}$$

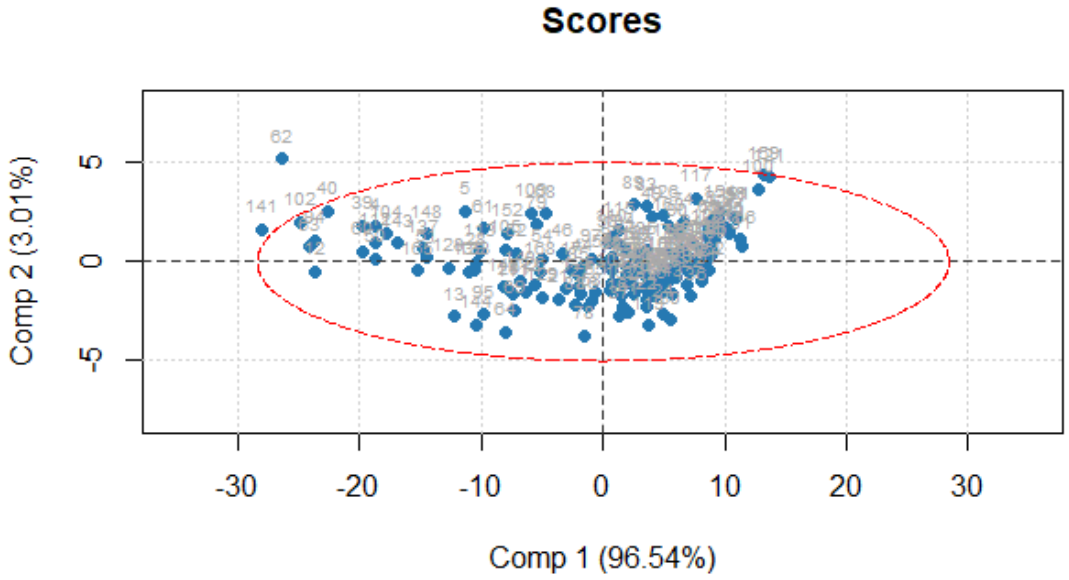
2.9.Análise de Componentes Principais (PCA) associada à Elipse de Hotelling T^2

Os *outliers* são valores discrepantes no conjunto de dados que, além de influenciarem na distribuição geral dos dados, podem afetar a capacidade de predição do modelo (WANG et al., 2018). Diante disso, deve-se utilizar métodos que possam identificar tais *outliers* e eliminá-los do conjunto de dados. Para esse fim, o PCA associado à Elipse de Hotelling T^2 tem se mostrado útil (ADNAN et al., 2017; AGUSSABTI et al., 2020; ASKE; KALLEVIK; SJÖBLOM, 2001; BROWNFIELD; KALIVAS, 2017).

Na execução do PCA associado à Elipse de Hotelling T^2, inicialmente aplica-se o PCA sobre os dados originais com o intuito de visualizar as principais tendências das amostras (ASKE; KALLEVIK; SJÖBLOM, 2001). Para tanto, utiliza-se a primeira e a segunda CPs, a fim de esboçar o gráfico bidimensional. Em seguida, constrói-se a Elipse de Hotelling T^2 (HOTELLING, 1947), sobreposta ao mesmo. As amostras localizadas fora da elipse (com nível de confiança de 95%) são consideradas fortes *outliers* e devem ser removidas do conjunto de dados (AGUSSABTI et al., 2020).

A Figura 7 ilustra um exemplo de aplicação do PCA associado a Elipse de Hotelling T^2, com um nível de confiança de 95%. Observa-se que as amostras 141 e 62 estão fora da elipse e, portanto, são consideradas potenciais *outliers*, devendo ser eliminadas do conjunto de dados.

Figura 7 - Gráfico do método PCA, associado a Elipse de Hotelling T^2, a um nível de confiança de 95%



Em que Comp 1 e Comp 2 correspondem à primeira e segunda componente principal, utilizadas para facilitar a visualização gráfica bidimensional. Os valores de porcentagem referem-se à proporção da variância explicada por cada componente. Dois pontos (141 e 62) se destacam como *outliers*.
Fonte: A autora.

2.10.Critérios para comparação dos métodos propostos

2.10.1.Parâmetros de comparação

A capacidade preditiva dos modelos pode ser averiguada através do coeficiente de correlação (r) e da raiz quadrada do erro quadrático médio (RMSE), dados pelas Equações 9 e 10, respectivamente:

$$r = \frac{\left[\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})\right]}{\sqrt{\left[\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}} \quad (9)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (10)$$

Nas Equações 9 e 10, y_i e $\hat{y}_i$ representam, respectivamente, o valor observado e o valor predito. Já $\bar{\hat{y}}$ refere-se ao valor médio predito e n, ao número de amostras pertencente a cada subconjunto de validação (FERREIRA; PETERNELLI, 2021).

Na classificação de amostras dentro do contexto de melhoramento de plantas, podem existir duas classes possíveis, ou seja, selecionada ou não selecionada. Selecionar uma amostra que não deveria ser selecionada e/ou não selecionar uma amostra que deveria ser selecionada caracterizam-se como erros de classificação (MOREIRA; BARBOSA; PETERNELLI, 2021).

A medição dos erros de classificação e a avaliação dos métodos utilizados na mesma podem ser realizadas através da matriz de confusão (MOREIRA; BARBOSA; PETERNELLI, 2021). A Tabela 1 apresenta o esquema de uma matriz de confusão:

Tabela 1- Exemplo de uma matriz de confusão

	Predito		
Real	NS	S	**TOTAL**
NS	VN	FP	VN+FP
S	FN	VP	FN+VP
TOTAL	VN+FN	FP+VP	N=VN+FP+FN+VP

Em que: NS = Não selecionada; S = Selecionada; VN = Verdadeiro Negativo; FP = Falso Positivo; FN = Falso negativo; VP = Verdadeiro Positivo. O Real e o predito indicam a quantidade de elementos pertencentes a cada classe e a quantidade de elementos selecionados a partir dos modelos utilizados, respectivamente.
Fonte: Moreira, Barbosa e Peternelli (2021).

A avaliação dos processos de classificação, construídos após a predição, pode ocorrer através da taxa de erro aparente (TEA) e da taxa de verdadeiros positivos (TVP). A primeira informa o número de amostras que foram classificadas incorretamente pelo método, ou seja, o complemento da acurácia do modelo. Já a segunda, refere-se à porcentagem de amostras que o método selecionou corretamente (MOREIRA; BARBOSA; PETERNELLI, 2021). Na comparação entre os processos de classificação, obtidos após a predição a partir de cada método, aqueles que apresentam menores valores de TEA e maiores valores de TVP possuem melhor capacidade de classificação (MUÑOZ et al., 2014).

A TEA e a TVP são definidas pelas Equações 11 e 12, respectivamente, em que: FN é o Falso Negativo; FP é o Falso Positivo; VP é o Verdadeiro Positivo; e N é o total de amostras analisadas. Todas essas variáveis encontram-se representadas na matriz de confusão (Tabela 1).

$$\boldsymbol{TEA = \frac{(FN + FP)}{N}} \quad (\mathbf{11})$$

$$\mathbf{TVP = \frac{VP}{FN + VP}} \quad (\mathbf{12})$$

2.10.2.Teste de Shapiro-Wilk, Teste t de Student e Teste de Wilcoxon para dados pareados

O teste t de Student (teste t), paramétrico, e o teste de Wilcoxon (WILCOXON, 1945), não paramétrico, podem ser utilizados quando se deseja averiguar se as médias de duas populações apresentam diferença estatística significativa para uma determinada característica de interesse. As amostras obtidas de cada população podem ser dependentes (pareadas) ou independentes (STEEL; TORRIE; DICKEY, 1996).

Quando as amostras a serem analisadas são dependentes (pareadas), significa que existe algo que as relacione ou, ainda, que elas foram obtidas de um mesmo conjunto de dados. Neste caso, o objetivo de aplicação dos testes é investigar se houve alteração na média de uma população quando a mesma é avaliada sob duas condições diferentes (KIM, 2015). Os testes t e o teste de Wilcoxon também podem ser utilizados, caso deseje-se avaliar se existe diferença significativa entre dois métodos estatísticos aplicados a amostras de um mesmo conjunto de dados (YUAN et al., 2006).

Antes da aplicação dos testes, verifica-se se o conjunto amostral atende à pressuposição de normalidade, o que pode ser feito aplicando-se o teste de Shapiro-Wilk (SHAPIRO; WILK, 1968). Caso a mesma seja atendida, utiliza-se o teste t e, em caso contrário, emprega-se o teste de Wilcoxon. Além da pressuposição de normalidade, alguns autores verificam ainda se os dados possuem variâncias homogêneas, por meio de outros testes. Já outros autores relatam que, quando as amostras possuem o mesmo tamanho, o teste t é robusto e, portanto, a pressuposição de variâncias homogêneas não precisa ser atendida (POSTEN; YEH; OWEN, 1982; SCHEFFÉ, 1970).

Encontram-se ainda, na literatura, trabalhos que sugerem que o teste t é eficiente em lidar com dados que violam a pressuposição de normalidade (BARTLETT, 1935; NEAVE; GRANGER, 1968; RASCH; GUIARD, 2004). Porém, outros estudos consideram o teste de Wilcoxon ainda mais poderoso na análise de dados não normais (HODGES; LEHMANN, 1956; NEAVE; GRANGER, 1968; RANDLES; WOLFE, 1979).

O teste de Shapiro-Wilk, além dos testes t e de Wilcoxon, para duas médias (dados pareados), são brevemente abordados a seguir.

2.10.2.1.Teste de Shapiro-Wilk (W)

O teste de normalidade, proposto por Shapiro e Wilk (1965), é fundamentado na estatística de ordem da distribuição normal e de seus respectivos valores médios. Para obter a estatística do teste, deve-se inicialmente obter uma amostra ordenada, ou seja, obter $Y_1 \leq Y_2 \leq \cdots \leq Y_n$, em que Y_i é a i-ésima estatística de ordem da amostra aleatória $X_1, X_2, \ldots, X_n$ (MOOD; GRAYBILL; BOES, 1974).

Para a realização desse teste, inicialmente deve-se calcular a soma de quadrados dos desvios da amostra ordenada, conforme apresentado na Equação 13:

$$s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{13}$$

Em seguida, verifica-se o número de amostras (n). Caso n seja par, ou seja, n=2k (sendo k ϵ $\mathbb{R}$), utiliza-se a Equação 14, em que os valores de a_i são obtidos na tabela de coeficientes de Shapiro-Wilk. Caso n seja ímpar, ou seja, n = 2k + 1, o cálculo é realizado através da Equação 15:

$$b = \sum_{i=1}^{k} a_{n-i+1}(y_{n-i+1} - y_i) \tag{14}$$

$$b = a_n(y_n - y_1) + \cdots + a_{k+2}(y_{k+2} - y_k) \tag{15}$$

De acordo com Shapiro e Wilk (1965), a estatística do teste de Shapiro-Wilk (W) para normalidade é dada pela Equação 16:

$$W = \frac{b^2}{s^2} \tag{16}$$

2.10.2.2.Teste t de Student para dados pareados

O teste t pode ser utilizado para avaliar se existe diferença significativa entre duas médias populacionais (STEEL; TORRIE; DICKEY, 1996). Para realizar tal teste, a partir das duas amostras disponíveis inicialmente, encontra-se a amostra baseada nos desvios, conforme mostrado na Tabela 2 abaixo:

Tabela 2 - Amostras dos desvios no teste t pareado

Elemento amostral i	**1**	**2**	**....**	**n**
Amostra 1	X_{11}	X_{12}	....	X_{1n}
Amostra 2	X_{21}	X_{22}	....	X_{2n}
$d_i = X_{1i} - X_{2i}$	d_1	d_2	....	d_n

Fonte: Adaptado de Triola (2008).

Os cálculos da variância dos desvios (s_d^2) e da média dos desvios ($\bar{X}_d$) devem ser realizados a partir das equações 17 e 18 (TRIOLA, 2008), respectivamente:

$$s_d^2 = \sum_{i=1}^{n} \frac{(d_i - \bar{d})^2}{n-1} \tag{17}$$

$$\overline{X}_d = \frac{\sum_{i=1}^{n} d_i}{n} \tag{18}$$

A estatística do teste t (t) pode ser obtida a partir da Equação 19 (TRIOLA, 2008):

$$t = \frac{\overline{X}_d - 0}{\sqrt{\frac{s_d^2}{n}}} \tag{19}$$

Esta estatística t segue uma distribuição t de Student com n-1 graus de liberdade (KANJI, 2006).

2.10.2.3.Teste de Wilcoxon

Quando os dados não satisfazem à pressuposição de normalidade, o teste de Wilcoxon substitui o teste t para dados pareados. Este teste foi desenvolvido por Wilcoxon (1945) e baseia-se nos postos das diferenças entre os pares de elementos das amostras. É usado para testar a hipótese nula de que as duas amostras proveem de populações com a mesma distribuição (KANJI, 2006).

Para realizar o teste de Wilcoxon, algumas pressuposições devem ser atendidas (TRIOLA, 2008): os dados consistem em pares combinados mutuamente independentes; a população das diferenças tem uma distribuição aproximadamente simétrica.

Caso o tamanho amostral (n) seja menor ou igual a 30 elementos ($n \leq 30$), a estatística do teste Wilcoxon (z) corresponde à soma dos postos (T) (TRIOLA, 2008). Caso contrário ($n > 30$), z será calculada conforme a Equação 20:

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \qquad (20)$$

3. MATERIAIS E MÉTODOS

3.1.Dados utilizados

Oito conjuntos de dados foram utilizados nesse trabalho. Os dados utilizados foram organizados em uma matriz **X_inicial**, em que cada linha e cada coluna corresponderam, respectivamente, a uma amostra (espectro) e aos números de ondas ou comprimentos de onda (variáveis), e em um vetor **y** que contém valores de alguma propriedade específica das amostras. Destaca-se que o número de linhas de **y** é igual ao número de amostras pertencentes a matriz **X_inicial**. A seguir (Tabela 3), estão detalhados os conjuntos de dados NIR utilizados neste trabalho:

Tabela 3 – Descrição dos conjuntos de dados utilizados (continua)

Dados	Origem/ Descrição	Dimensão (X_inicial)	Faixa
Cana-de-açúcar – Teor de Lignina (Conjunto 1)	Dados obtidos no banco de germoplasma do Programa Genético de Melhoramento de Cana-de-Açúcar da Universidade Federal de Viçosa (PMGCA-UFV), Viçosa, Minas Gerais (MG), Brasil (ASSIS et al., 2017). Os espectros NIR foram obtidos diretamente da folha verde, sem procedimento de preparação da amostra, com o objetivo de predizer o teor de lignina na cana-de-açúcar.	256 ×1038	4000 a 10000 cm^{-1}
Cana-de-açúcar – Teor de Fibra (Conjunto 2)	Dados obtidos de um experimento realizado no PMGCA-UFV, Viçosa, Minas Gerais (MG), Brasil (PETERNELLI et al., 2020). Com o objetivo de predizer o teor de fibra (FIBRA), foram obtidos espectros no terço médio do colmo da cana-de açúcar.	168 ×3113	4000 a 10000 cm^{-1}

Tabela 3 – Descrição dos conjuntos de dados utilizados (continua)

Dados	Origem/ Descrição	Dimensão (X_inicial)	Faixa
Gasolina – Número de Octanos (Conjunto 3)	Conjunto de dados formado por amostras de gasolina, a fim de predizer o número de octanos especificados na mesma (KALIVAS, 1997).	60 × 401	900 a 1700 nm
Leite - Teor de Caseína (Conjunto 4)	Conjunto obtido na biblioteca de dados da *Eigenvector Research Inc*. (https://eigenvector.com/resources/data-sets/). É constituído por amostras de leite, obtidas por um instrumento NIR (NAES et al., 2002). Dentre as diversas variáveis respostas disponíveis para este conjunto de dados, optou-se pela utilização do teor de caseína neste estudo.	231 × 117	1100 a 2500 nm
Cana-de-açúcar – Teor de Brix (Conjunto 5)	Dados obtidos de um experimento realizado no PMGCA-UFV, Viçosa, Minas Gerais (MG), Brasil. As amostras foram obtidas no terço médio do colmo da cana-de-açúcar e analisadas através do NIR portátil, a fim de predizer o teor de brix (BRIX).	128 × 121	900 a 1700 nm
Mangas – Teor de Vitamina C (Conjunto 6)	Dados referentes a amostras de mangas, analisadas por meio do NIR de bancada, com o objetivo de predizer o teor de vitamina C. Foram coletados na Faculdade de Agricultura, *Georg-August University of Goettingen*, Alemanha e podem ser encontrados no *Mendeley* (HAYATI; MUNAWAR; FACHRUDDIN, 2020).	58 × 1557	1100 a 2500 nm

Tabela 3 – Descrição dos conjuntos de dados utilizados (conclusão)

Dados	Origem/ Descrição	Dimensão (X_inicial)	Faixa
Diesel – Densidade (Conjunto 7) e Viscosidade (Conjunto 8)	Dados obtidos da biblioteca de dados da *Eigenvector Research Inc*. (https://eigenvector.com/resources/data-sets/), constituído de espectros NIR, com o objetivo de predizer diversas propriedades do diesel (FOLCH-FORTUNY; ARTEAGA; FERRER, 2015; HUTZLER; BESSEE, 1997). Neste trabalho, foram analisadas as propriedades de densidade e viscosidade. Observou-se alguns valores faltantes no vetor da variável resposta referente a tais parâmetros, o que resultou na eliminação das amostras correspondentes a tais valores. Assim, a matriz **X_inicial** utilizada para a propriedade densidade foi diferente daquela empregada na propriedade viscosidade.	395 × 401	750 a 1550 nm

Fonte: A autora.

3.2.Construção dos modelos de predição

Os conjuntos de dados foram utilizados para a construção de modelos de predição, com os métodos PCA-ANN e PLS, sendo avaliados com e sem pré-tratamentos.

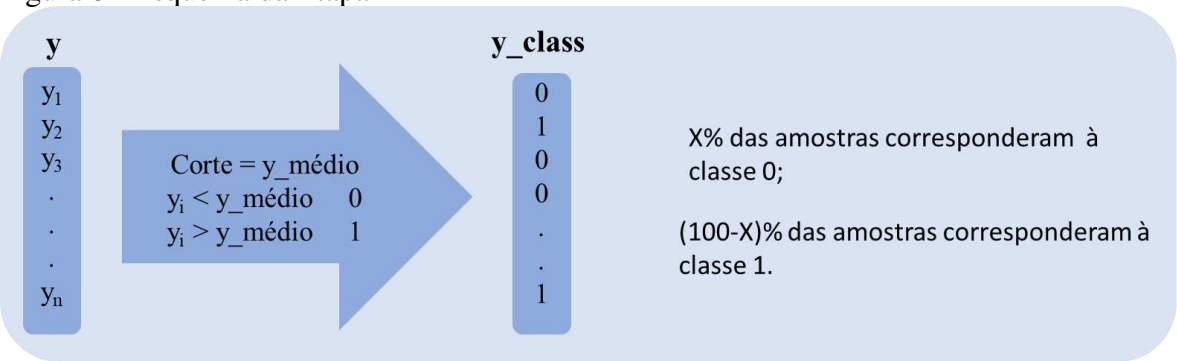
3.2.1.Construção dos modelos de predição com pré-tratamentos de dados

Os modelos de predição, construídos com os métodos PCA-ANN e PLS com pré-tratamentos de dados foram denominados PCA-ANN(CT) e PLS(CT). A seguir, são descritas as etapas realizadas para a obtenção dos modelos propostos:

ETAPA 1: Inicialmente, realizou-se as leituras da matriz **X_inicial** e do vetor **y_inicial** de cada conjunto de dados proposto e, através do método PCA associado à Elipse de Hotelling T^2 (HOTELLING, 1947), foram identificados e eliminados os *outliers* existentes, obtendo-se assim a matriz **X** e o vetor **y** que foram utilizados para a modelagem dos dados.

ETAPA2: Estabeleceu-se o valor de corte para o processo de classificação, por meio da média dos valores pertencentes ao vetor **y**, de forma que aqueles valores acima e abaixo do valor de corte (y_médio) corresponderam, respectivamente, à classe 1 (selecionado) e à classe 0 (não selecionado). O vetor formado por essas classes foi denominado de **y_class**. Diante disso, obteve-se as porcentagens (proporções) X e (100 – X) de amostras pertencentes, nessa ordem, à classe 0 e à classe 1 (houve variação de tais porcentagens de um conjunto para outro). A Figura 8 mostra um esquema desta Etapa 2:

Figura 8 - Esquema da Etapa 2



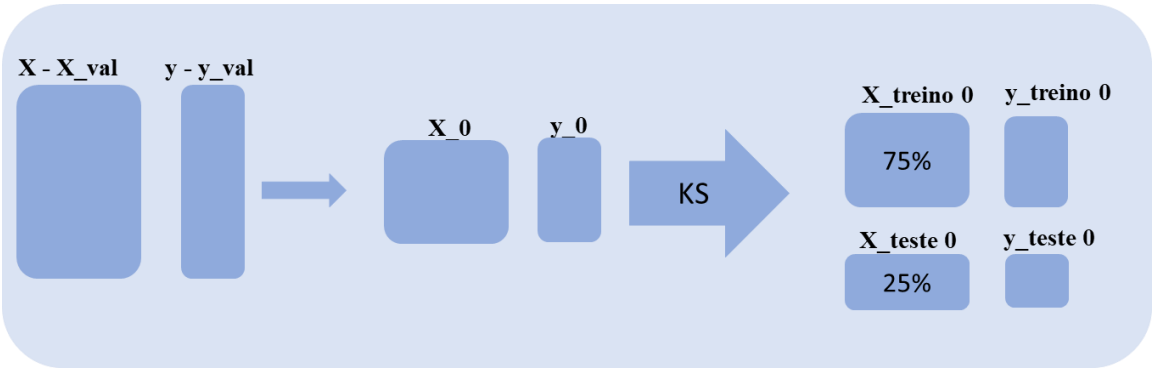
Em que n refere-se à quantidade de amostras pertencentes ao vetor **y** (vetor que contém informações de alguma propriedade de interesse); i (i = 1 a n) refere-se a uma determinada linha do vetor **y**; y_médio refere-se ao valor médio do vetor **y**; **y_class** refere-se ao vetor que contém as classes correspondentes do vetor **y**; X% corresponde à porcentagem de elementos que pertencem a classe 0; e (100-X) % representa a proporção de elementos que pertencem a classe 1.
Fonte: A autora.

ETAPA 3: Utilizou-se o RS para selecionar, aproximadamente, um terço das linhas (amostras) da matriz **X** e do vetor **y,** mantendo-se as porcentagens de amostras pertencentes à classe 0 e à classe 1, conforme definidas na ETAPA 2. Assim, foram formados os subconjuntos de validação (**X_val** e **y_val**) e de validação de classificação (**y_val_class**) contendo, respectivamente, X% de amostras da classe 0 e (100-X) % de amostras da classe 1. Tais subconjuntos foram reservados para validação externa do modelo.

ETAPA 4: Após a retirada das amostras referentes ao subconjunto de validação, as amostras restantes foram utilizadas na obtenção dos subconjuntos de treino e de teste, para a construção

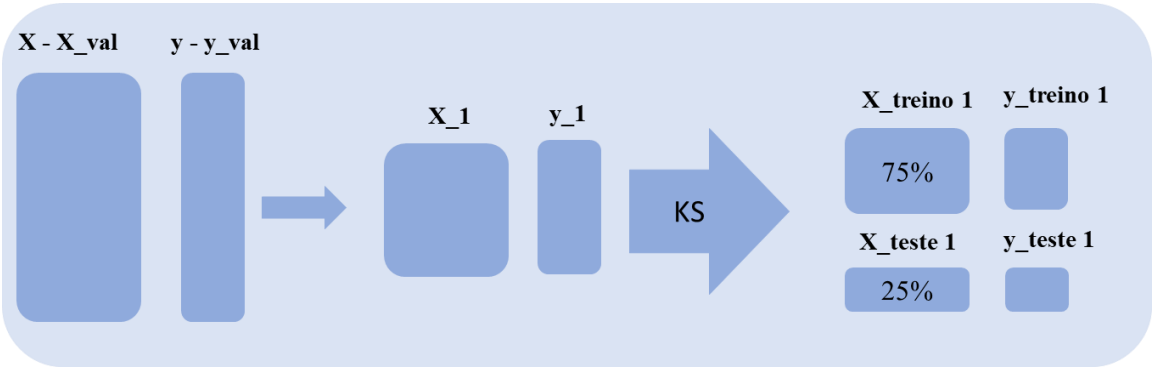
do modelo. Inicialmente, foram retiradas da matriz e do vetor resultantes (**X** – **X_val** e **y** - **y_val**) as linhas (amostras) associadas à classe 0, gerando a matriz **X_0** e o vetor **y_0**. Em seguida, utilizou-se o método KS na matriz **X_0** e no vetor **y_0**, de forma que aproximadamente 75% das amostras pertenceram ao subconjunto de treino (**X_treino 0** e **y_treino 0**) e os outros 25% formaram o subconjunto de teste (X_**teste 0** e **y_teste 0**) (Figura 9). O mesmo procedimento foi executado em relação à classe 1, obtendo-se assim as matrizes **X_treino 1** e **X_teste 1**, além dos vetores **y_treino 1** e **y_teste 1** (Figura 10).

Figura 9 – Esquema da obtenção dos subconjuntos de treino e de teste associados à classe 0



Em que **X** – **X_val** e **y** - **y_val** são, respectivamente, a matriz e o vetor utilizados para a construção do modelo. A matriz **X_0** e o vetor **y_0** referem-se às linhas (amostras) associadas à classe 0. **X_treino 0** e **y_treino 0** referem-se aos dados do subconjunto de treino. **X_teste 0** e **y_teste 0** referem-se aos dados do subconjunto de teste.
Fonte: A autora.

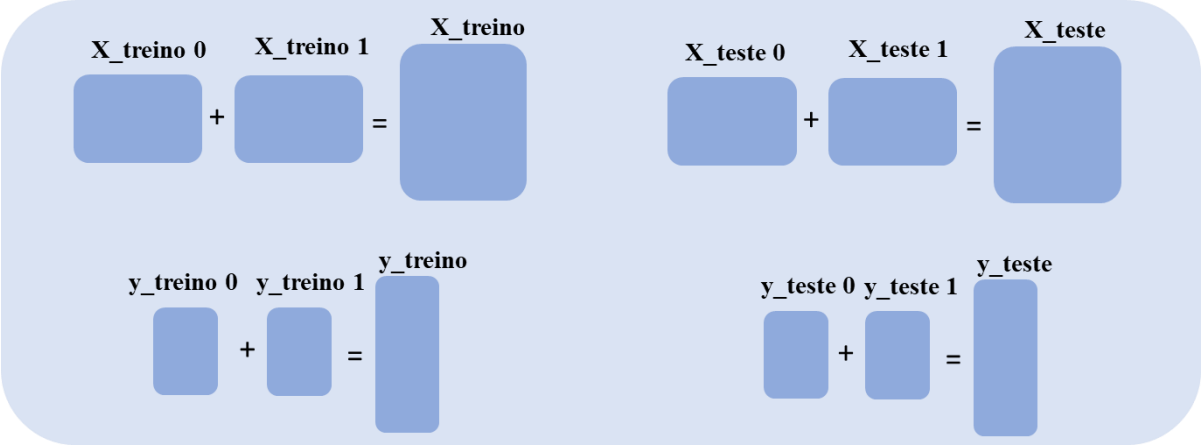
Figura 10 - Esquema da obtenção dos subconjuntos de treino e de teste associados à classe 1



Em que **X** – **X_val** e **y** - **y_val** são, respectivamente, a matriz e o vetor utilizados para a construção do modelo. A matriz **X_1** e o vetor **y_1** referem-se às linhas (amostras) associadas à classe 1. **X_treino 1** e **y_treino 1** referem-se aos dados do subconjunto de treino. **X_teste 1** e **y_teste 1** referem-se aos dados do subconjunto de teste.
Fonte: A autora.

Finalmente, a união entre os subconjuntos de treino e teste, de cada classe, resultou nas matrizes (**X_treino**, **X_teste**, **y_treino** e **y_teste**) que foram utilizadas para a construção do modelo. Um esquema desta etapa pode ser observado na Figura 11:

Figura 11 - Esquema da obtenção dos subconjuntos de treino e de teste, para a construção do modelo



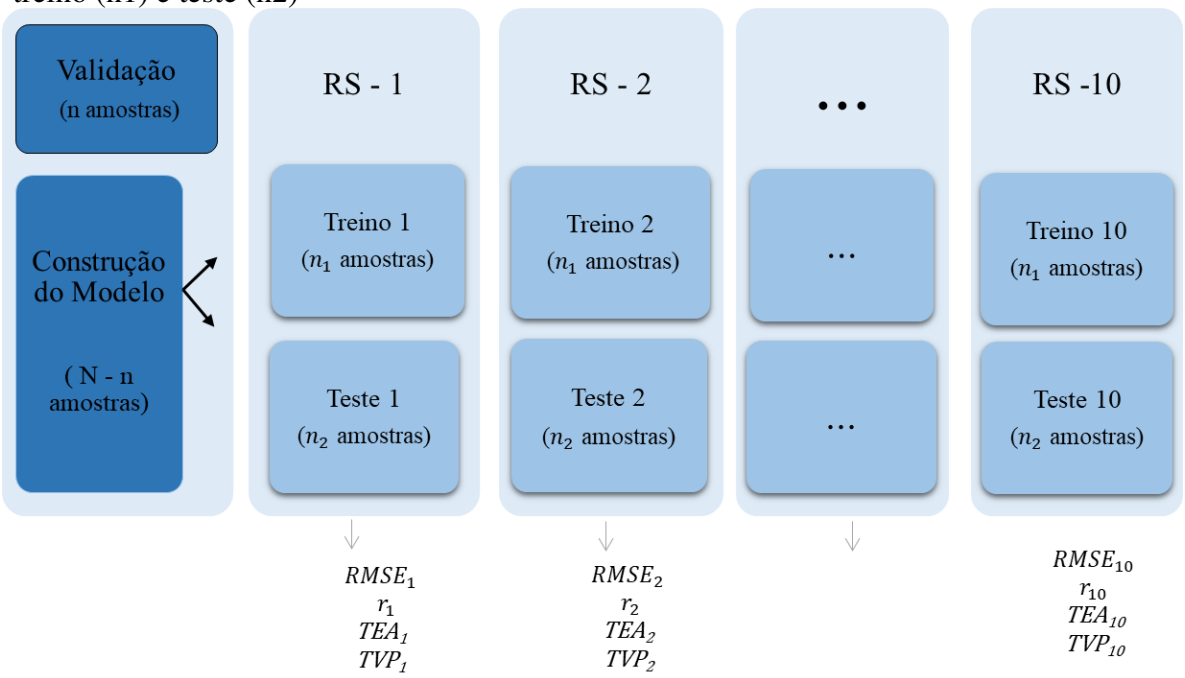
Em que **X_treino 0**, **y_treino 0**, **X_treino 1**, **y_treino 1, X_treino** e **y_treino** referem-se aos dados do subconjunto de treino. **X_teste 0**, **y_teste 0**, **X_teste 1**, **y_teste 1, X_val e y_val** referem-se aos dados do subconjunto de teste.
Fonte: A autora.

ETAPA 5: Aplicou-se os pré-tratamentos nos dados obtidas na Etapa 4. Os pré-tratamentos utilizados neste trabalho foram: (1) centragem na média; (2) primeira derivada; (3) segunda derivada; (4) MSC e (5) alisamento. Os valores de RMSE e do coeficiente de correlação (r), obtidos na validação cruzada, foram usados como parâmetros de verificação do ajuste dos modelos, através dos métodos PLS e PCA-ANN.

ETAPA 6: Uma vez realizado o ajuste dos modelos, utilizou-se o subconjunto de validação (pré-tratado) para verificar a capacidade preditiva dos modelos. Os valores de RMSE, r, TEA, TVP e do tempo computacional (TC) foram obtidos e usados como parâmetros de comparação dos modelos. Vale ressaltar que, enquanto os parâmetros RMSE e r foram utilizados para avaliar o poder preditivo (regressão) do modelo, o TEA e o TVP foram empregados para analisar o poder de classificação após a predição do mesmo.

Todo o processo de construção do modelo, descrito nas etapas de 1 a 6, foi repetido 10 vezes com o propósito de compensar eventuais problemas amostrais e melhorar a significância estatística da comparação entre os métodos PCA-ANN e PLS. Desta forma, foram obtidos 10 valores de RMSE, r, TEA e TVP em cada conjunto de dados, conforme apresentado na Figura 12. As médias de tais valores foram utilizadas na comparação entre os modelos construídos.

Figura 12 - Esquema da divisão e repetição das amostras nos subconjuntos de validação (n), treino (n1) e teste (n2)



Sendo: N = n+n_1+n_2 = total de indivíduos no banco de dados completos; RS = amostragem aleatória simples, sem reposição.
Fonte: Adaptado de Ferreira; Teixeira e Peternelli (2022).

3.2.2.Construção dos modelos de predição sem pré-tratamentos de dados

Todas as etapas descritas no item anterior, excetuando-se a etapa 5 que descreve o uso dos pré-tratamentos de dados, foram utilizadas para a construção dos modelos de predição, a partir da aplicação dos métodos PCA-ANN e PLS sem os pré-tratamentos. Tais métodos foram denominados PCA-ANN(ST) e PLS(ST). Assim como na construção dos modelos com pré-tratamentos, todo o processo foi repetido 10 vezes.

De modo geral, tanto nos modelos construídos com pré-tratamentos de dados, quanto naqueles obtidos sem os mesmos, realizou-se uma análise de *outliers* na matriz **X_inicial,** por meio da Elipse de Hotelling T^2 com 95% de nível de confiança, obtendo-se assim a matriz **X**. Posteriormente, conforme mostrado na Figura 12, os dados foram particionados em três partes: validação, treino e teste. Cada uma destas partes possui tamanho amostral correspondente a n ,n_1 e n_2, respectivamente. Em cada conjunto de dados, retirou-se, de forma aleatória, aproximadamente um terço das amostras ($n \approx 33\%$), para formarem o subconjunto de validação. Das amostras restantes, aproximadamente 75% constituíram o subconjunto de treino e 25% integraram o subconjunto de teste ($n_1 \approx 75\%$ e $n_2 \approx 25\%$).

Para verificar se os valores médios do RMSE, r, TVP e ATR, obtidos por meio da aplicação do método PLS(CT), diferiam estatisticamente daqueles determinados a partir da aplicação do método PCA-ANN(ST), foi utilizado o teste t pareado, para os dados que atenderam à pressuposição de normalidade, e o teste de Wilcoxon, para os dados que não atenderam a tal pressuposição, ambos a um nível de 1% de significância. Testes pareados foram utilizados, visto que em uma mesma repetição, os métodos propostos foram analisados na mesma partição de dados. Para a verificação do pressuposto de normalidade dos dados, aplicou-se o teste de Shapiro-Wilk.

Ressalta-se que não foi aplicado o método *Partial Least Squares Discriminant Analysis* (PLS-DA) para classificação das amostras. Neste trabalho, o processo de classificação foi realizado após a predição, ou seja, os valores preditos pelos modelos foram classificados conforme o valor de corte estabelecido inicialmente (y_médio), de forma que aqueles valores acima do valor de corte (y_médio) corresponderam à classe 1 (selecionado), enquanto que os valores abaixo do valor de corte corresponderam à classe 0 (não selecionado). Em seguida, tais valores foram comparados aos valores de validação de classificação (**y_val_class**) obtidos na ETAPA 3. Finalmente foi construída a matriz de confusão e calculados os valores de TVP e ATR dentro de cada repetição (PETERNELLI et al., 2020).

3.3.Recursos computacionais

Todas as rotinas dos métodos empregados foram implementadas no *software* R (R CORE TEAM, 2021). Os espectros foram obtidos através do *software* Matlab 7.9 (Math Works, Natick, USA). A seguir, são detalhados os recursos computacionais utilizados em cada um dos métodos aplicados neste estudo:

- **Regressão por Quadrados Mínimos Parciais (PLS):** para ajuste do modelo com o método PLS, foi utilizada a função plsr, do pacote pls do *software* R (WEHRENS et al., 2021). Para a escolha do nVL, foi construído um gráfico, no qual o eixo das abscissas e o eixo das ordenadas contêm, respectivamente, os nVL (geralmente de 0 a 20) e os valores de RMSECV. O nVL correspondente ao menor valor de RMSECV foi o escolhido para a construção do modelo.

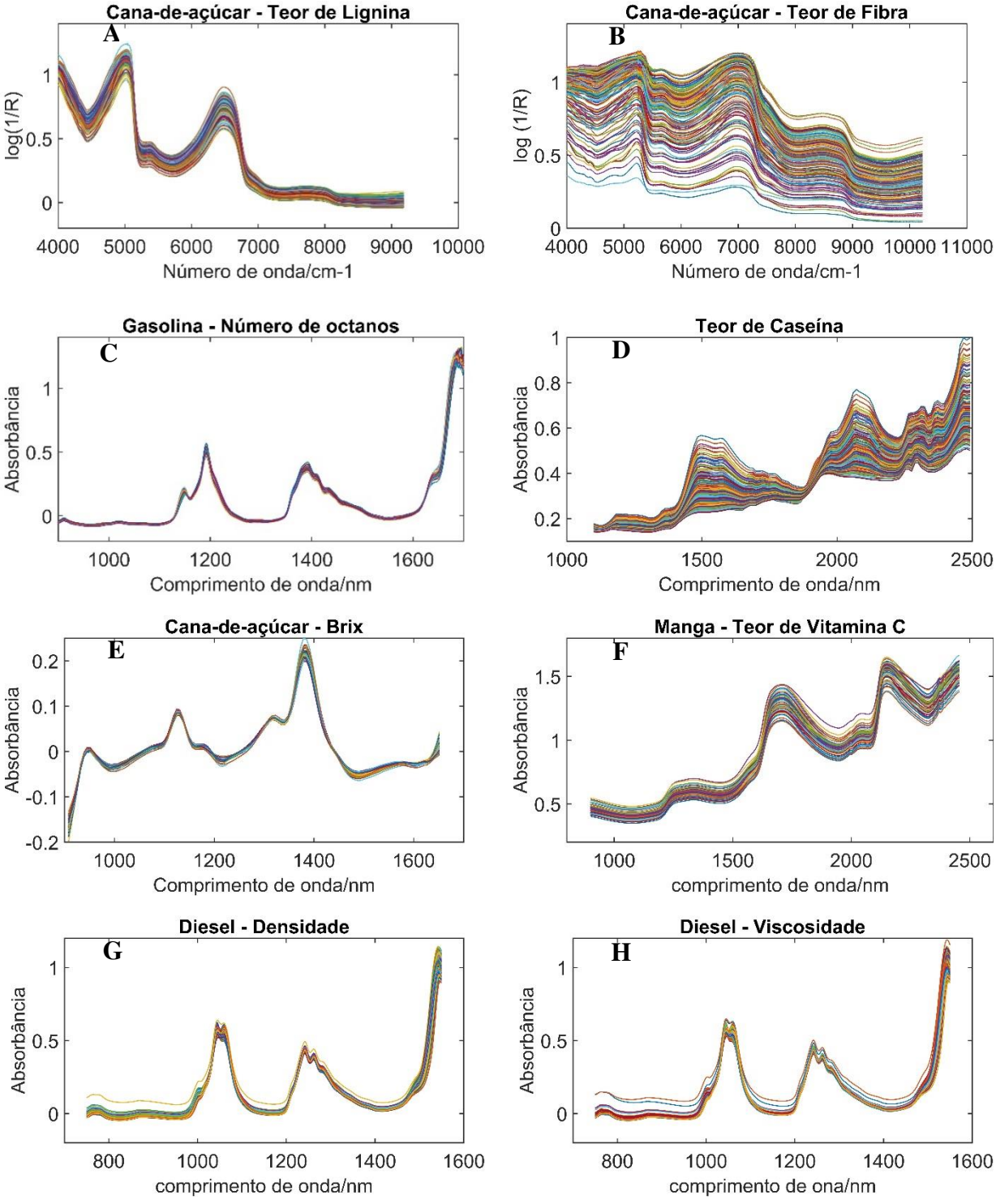
- **Análise de Componentes Principais associada às Redes Neurais (PCA-ANN):** para o ajuste do modelo, aplicando-se o método PCA-ANN, foi utilizada a função

`pcaNNet` do pacote `caret` do *software* R (KUHN, 2016). Inicialmente, a função executa o PCA, calculando a porcentagem de variância acumulada para cada CPs. Em seguida, ela utiliza o argumento `thresh` para determinar o número de CPs necessário para captar tal variância. Estas CPs são, então, usadas como entradas da ANN. A rede utilizada possuía apenas uma camada oculta, com o número de neurônios (0 a 20) variando nesta camada. Um gráfico foi construído para determinar a quantidade de neurônios na camada oculta em função do RMSECV. O ponto de menor valor de RMSECV foi selecionado, e o número de neurônios associado a ele foi escolhido para a construção do modelo (AZADSHAHRAKI et al., 2022).

4. RESULTADOS E DISCUSSÃO

Os espectros brutos NIR das amostras de todos os conjuntos de dados analisados são apresentados na Figura 13.

Figura 13- Curva espectral NIR, referente a cada conjunto de dados analisados



Fonte: A autora.

Os espectros NIR das amostras de cana-de-açúcar, na faixa de 4000 a 10000 cm^{-1}, para predizer os teores de Lignina e Brix, são apresentados nas Figuras 13-A e 13-B. Nas Figuras 13-C e 13-E, têm-se os espectros NIR das amostras referentes à gasolina e à cana-de-açúcar, na faixa de 900 a 1700 nm, com o intuito de predizer o número de octanos e o teor de Brix, respectivamente. Nas Figuras 13-D e 13-F, estão os espectros NIR relacionados às amostras de leite e manga, na faixa de 1100 a 2500 nm, a fim de predizer o teor de caseína e o teor de vitamina C, nessa ordem. Finalmente, as Figuras 13-G e 13-H representam os espectros NIR das amostras de diesel, na faixa 750 a 1550 nm, para predizer a densidade e a viscosidade, respectivamente.

A Tabela 4 mostra as dimensões das matrizes **X_inicial** e **X**, a quantidade de *outliers* eliminados e os tamanhos amostrais dos subconjuntos de validação, treino e teste, representados por n ,n_1 e n_2, respectivamente, obtidos em cada conjunto de dados avaliado.

Tabela 4 **-** Informações dos dados utilizados, em que **X_inicial** é a matriz original de dados, **X** é a matriz obtida após a retirada de *outliers*, n, n_1 e n_2 representam a quantidade de elementos pertencentes aos subconjuntos de validação, treino e teste, respectivamente

Dados	**X_inicial**	***Outliers***	**X**	**n**	$\boldsymbol{n_1}$	$\boldsymbol{n_2}$
1. Cana-de-açúcar – Teor de Lignina	256 × 1038	18	238 × 1038	79	120	39
2. Cana-de-açúcar – Teor de Fibra	168 × 3113	7	161 × 3113	54	81	26
3. Gasolina – Número de Octanos	60 × 401	2	58 × 401	19	29	10
4. Leite - Teor de Caseína	231 × 117	8	223 × 117	74	112	37
5. Cana-de-açúcar – Teor de Brix	128 × 121	5	123 × 121	41	61	21
6. Manga – Teor de Vitamina C	58 × 1557	5	53 × 1557	18	27	8
7. Diesel – Densidade	395 × 401	61	334 × 401	111	168	55
8. Diesel – Viscosidade	395 × 401	41	354 × 401	118	177	59

Fonte: A autora.

Observa-se (Tabela 4) que os conjuntos 7 e 8 representaram os maiores números de *outliers* removidos do conjunto inicial de dados (**X_inicial**), ou seja, aproximadamente 15,4 % e 10,4 %, respectivamente, das amostras de tais conjuntos foram eliminadas pela análise de

outliers, através da Elipse de Hotelling T^2, ao nível de confiança de 95%. Dentre todos os conjuntos analisados, os representados pelos números 3 e 6 foram aqueles dos quais foram removidos o menor número de *outliers*, o que significa uma vantagem, em razão do pequeno número de amostras disponíveis inicialmente nesses conjuntos.

Sabe-se que, de maneira geral, a multicolinearidade é um problema presente em dados NIR (FREUND; WILSON; SA, 2006), sendo portanto uma característica dos conjuntos de dados analisados neste estudo. Além disso, observa-se (Tabela 4) que a maioria dos conjuntos de dados analisados, com exceção daqueles representados pelos números 4 e 5, também apresentou problemas de alta dimensionalidade. Contudo, como a multicolinearidade e/ou a alta dimensionalidade são problemas que estiveram presentes em todos os conjuntos de dados estudados, devem ser utilizados métodos específicos para ajuste dos modelos a serem construídos (FERREIRA; PETERNELLI, 2021).

A Tabela 5 apresenta os valores médios de RMSE, r, TEA, TVP e TC, obtidos a partir de cada repetição realizada nos conjuntos de predição, utilizando-se os métodos PCA-ANN(CT), PCA-ANN(ST), PLS(CT) e PLS(ST) em todas as variáveis da matriz de dados.

Tabela 5 **-** Valores médios da raiz quadrada do erro quadrático médio (RMSE), do coeficiente de correlação (r), da taxa de erro aparente (TEA), da taxa de verdadeiros positivos (TVP) e do tempo computacional (TC), avaliados pelos métodos: PCA-ANN, com e sem pré-tratamentos; PLS, com e sem pré-tratamentos, definidos respectivamente por: PCA-ANN(CT), PCA-ANN(ST), PLS(CT) e PLS(ST). Os valores médios das melhores estatísticas obtidas em cada conjunto de dados estão destacados em negrito (continua)

	1. Cana-de-açúcar – Teor de Lignina				**2. Cana-de-açúcar – Teor de Fibra**			
	PCA-ANN (CT)	**PCA-ANN (ST)**	**PLS (CT)**	**PLS (ST)**	**PCA-ANN (CT)**	**PCA-ANN (ST)**	**PLS (CT)**	**PLS (ST)**
RMSE	**1,78**	1,79	1,84	1,98	2,37	**1,64**	1,85	1,87
r	**0,64**	0,57	0,51	0,40	0,30	**0,44**	0,40	0,39
TEA	**0,23**	0,35	0,32	0,37	0,38	**0,30**	0,34	0,31
TVP	**0,79**	0,67	0,70	0,65	0,59	**0,65**	0,63	0,63
TC	441s	26s	333s	22s	1130s	56s	664s	35s
	3. Gasolina - Número de octanos				**4. Leite - Teor de Caseína**			
	PCA-ANN (CT)	**PCA-ANN (ST)**	**PLS (CT)**	**PLS (ST)**	**PCA-ANN (CT)**	**PCA-ANN (ST)**	**PLS (CT)**	**PLS (ST)**
RMSE	0,56	0,33	**0,35**	0,70	0,50	**0,32**	0,34	0,70
r	0,90	**0,97**	**0,97**	0,86	0,93	**0,98**	**0,98**	0,86
TEA	0,07	0,06	**0,05**	0,26	0,07	0,06	**0,05**	0,27
TVP	**0,97**	0,94	0,96	0,82	**0,95**	0,93	**0,95**	0,82
TC	97s	15s	74s	13s	98s	13s	73s	11s

Tabela 5 **-** Valores médios da raiz quadrada do erro quadrático médio (RMSE), do coeficiente de correlação (r), da taxa de erro aparente (TEA), da taxa de verdadeiros positivos (TVP) e do tempo computacional (TC), avaliados pelos métodos: PCA-ANN, com e sem pré-tratamentos; PLS, com e sem pré-tratamentos, definidos respectivamente por: PCA-ANN(CT), PCA-ANN(ST), PLS(CT) e PLS(ST). Os valores médios das melhores estatísticas obtidas em cada conjunto de dados estão destacados em negrito (conclusão)

	5. Cana-de-açúcar – Teor de Brix				**6. Mangas - Teor de Vitamina C**			
	PCA-ANN (CT)	**PCA-ANN (ST)**	**PLS (CT)**	**PLS (ST)**	**PCA-ANN (CT)**	**PCA-ANN (ST)**	**PLS (CT)**	**PLS (ST)**
RMSE	2,00	**1,30**	**1,30**	1,64	1,63	1,22	**1,21**	1,46
r	0,53	0,77	**0,78**	0,60	0,32	0,54	**0,55**	0,53
TEA	0,37	**0,12**	0,23	0,35	0,43	0,49	**0,24**	0,26
TVP	0,71	**0,81**	0,80	0,67	0,51	0,49	**0,70**	**0,70**
TC	71s	12s	62s	10s	224s	33s	165s	22s
	7. Diesel – Densidade				**8. Diesel – Viscosidade**			
	PCA-ANN(CT)	**PCA-ANN(ST)**	**PLS(CT)**	**PLS(ST)**	**PCA-ANN(CT)**	**PCA-ANN(ST)**	**PLS(CT)**	**PLS(ST)**
RMSE	0,006	0,002	**0,001**	0,003	0,28	**0,16**	**0,16**	0,20
r	0,90	0,98	**0,99**	0,97	0,83	**0,94**	**0,94**	0,89
TEA	0,11	0,11	**0,03**	0,06	0,18	0,10	**0,08**	0,16
TVP	0,93	0,92	**0,98**	0,96	0,86	0,90	**0,92**	0,83
TC	332s	37s	230s	25s	279s	37s	225s	21s

Em que: s = segundos.
Fonte: A Autora.

Nota-se em todos os conjuntos de dados, descritos na Tabela 5, que o menor TC correspondeu àquele obtido por meio da aplicação do método PLS(ST). O segundo menor TC foi aquele referente à utilização do método PCA-ANN(ST), seguido dos tempos relacionados ao uso dos métodos PLS(CT) e PCA-ANN(CT), nesta sequência. Os menores TC relacionados aos métodos PLS(ST) e PCA-ANN(ST) justificam-se pelo fato destes últimos não contarem com a etapa de pré-tratamento de dados. Observa-se ainda que os valores médios das melhores estatísticas obtidas em cada conjunto de dados avaliados encontram-se destacados em negrito na Tabela 5.

Comparando-se os valores médios obtidos no conjunto de dados 1, percebeu-se que o PCA-ANN(CT) apresentou os melhores parâmetros estatísticos, ou seja, maiores valores de r e TVP (r = 0,64 e TVP = 0,79), e menores valores de RMSE e TEA (RMSE = 1,78 e TEA = 0,23). A fim de prever propriedades do solo, o estudo desenvolvido por Ramadan et al. (2005) também comparou o uso dos métodos PLS e PCA-ANN, associados a pré-tratamentos de dados, obtendo melhores valores preditivos com o PCA-ANN(CT), assim como ocorreu na análise do conjunto de dados 1.

No conjunto 2, em que os dados aparentemente eram bastante ruidosos (Figura 13-B), o PCA-ANN(ST) apresentou as melhores estatísticas (RMSE = 1,64, r = 0,44, TEA = 0,3 e TVP = 0,65) e um TC aproximadamente 11 vezes menor, quando comparado àquele referente ao PLS(CT). No conjunto de dados 5, em geral o PCA-ANN(ST) também apresentou melhor desempenho, observando-se os valores médios dos parâmetros estatísticos analisados neste trabalho. Já nos conjuntos de dados 3, 4, 6, 7 e 8, o PLS(CT) em geral apresentou as melhores estatísticas em termos de valores médios.

O estudo desenvolvido por Mireei e Sadeghi (2013) utilizou a espectroscopia NIR para classificação de frutos de tâmaras como não infectados e infectados. Um valor de corte de 0,5 foi estabelecido de forma que valores preditos maiores que o mesmo, foram classificados como não infectados, e aqueles menores, como infectados. Foram utilizados diversos processos de classificação, incluindo o PCA-ANN associado a pré-tratamentos que apresentou bons resultados quando comparado aos demais, o que também foi observado em alguns resultados de conjuntos de dados analisados neste trabalho.

Percebeu-se que entre as repetições realizadas em cada conjunto de dados avaliado, os pré-tratamentos, o nVL, o número de componentes principais e o número de neurônios na camada oculta apresentaram poucas variações entre si. Isso é relevante, caso seja necessário obter um único modelo para predição de alguma propriedade de interesse do pesquisador. A Tabela 6 informa a variação do número de componentes principais (nCPs), o número de

neurônios na camada oculta e o nVL, utilizados pelos métodos PCA-ANN(CT), PCA-ANN(ST), PLS(CT) e PLS(ST) em cada conjunto de dados nas repetições realizadas.

Tabela 6 - Informação da variação do número de componentes principais (nCPs), do número de neurônios e do número de variáveis latentes (nVL), utilizadas pelos métodos PCA-ANN(CT), PCA-ANN(ST), PLS(CT) e PLS(ST) em cada conjunto de dados nas repetições realizadas

Dados	**PCA-ANN(CT)**		**PCA-ANN(ST)**		**PLS(CT)**	**PLS(ST)**
	nCPs	**Número de Neurônios**	**nCPs**	**Número de Neurônios**	**nVL**	**nVL**
1. Cana-de-açúcar – Teor de Lignina	94 a 109	17 a 20	3 e 4	5 a 20	4 a 10	6 a 10
2. Cana-de-açúcar – Teor de Fibra	99 a 102	14 a 20	2	1 a 18	2 a 6	5 a 9
3. Gasolina – Número de Octanos	6 a 32	9 a 19	6 a 8	2 a 8	5 a 10	8 a 10
4. Leite - Teor de Caseína	93 a 102	13 a 20	3 e 4	3 a 7	5 a 10	7 a 10
5. Cana-de-açúcar – Teor de Brix	10 a 24	13 a 19	10 e 11	1 a 5	7 a 9	8 a 10
6. Manga – Teor de Vitamina C	3 a 33	10 a 20	3	1 a 16	5 a 10	5 a 10
7. Diesel – Densidade	8 e 35	13 a 20	8	1 a 6	10	10
8. Diesel – Viscosidade	36 a 99	17 a 20	7 e 8	1 a 8	10	10

Fonte: A autora.

A Tabela 7 apresenta o valor-p, obtido por meio do teste de Shapiro-Wilk, do teste t e do teste de Wilcoxon, além da conclusão alcançada.

Tabela 7 - Valor-p dos testes de normalidade e da comparação entre as médias do coeficiente de correlação (r), da raiz quadrada do erro quadrático médio (RMSE), da taxa de erro aparente (TEA) e da taxa de verdadeiros positivos (TVP) obtidos pelos métodos: PLS com pré-tratamento (PLS(CT)) e PCA-ANN sem pré-tratamentos (PCA-ANN(ST)) (continua)

		Valor-p da comparação entre as médias do PCA-ANN(ST) e do PLS(CT)			
		Normalidade (Shapiro-Wilk)	**Teste t**	**Teste de Wilcoxon**	**Conclusão**
1. Cana-de-açúcar (Teor de Lignina)	RMSE	0,660	0,230	-	NS
	r	0,240	0,131	-	NS
	TEA	0,624	0,399	-	NS
	TVP	0,687	0,352	-	NS
2. Cana-de-açúcar (Teor de Fibra)	RMSE	0,950	0,181	-	NS
	r	0,040	0,125	-	NS
	TEA	0,440	0,810	-	NS
	TVP	0,170	0,900	-	NS
3. Gasolina (Número de Octanos)	RMSE	0,242	0,408	-	NS
	r	0,126	0,722	-	NS
	TEA	0,006	-	1	NS
	TVP	0,119	0,302	-	NS
4. Leite (Teor de Caseína)	RMSE	0,126	0,520	-	NS
	r	0,242	0,408	-	NS
	TEA	0,005	-	1	NS
	TVP	0,112	0,302	-	NS

Tabela 7 - Valor-p do teste de normalidade e da comparação entre as médias do coeficiente de correlação (r), da raiz quadrada do erro quadrático médio (RMSE) e da taxa de erro aparente (TEA), da taxa de verdadeiros positivos (TVP) obtidos pelos métodos: PLS com pré-tratamento (PLS(CT)) e PCA-ANN sem pré-tratamentos (PCA-ANN(ST)) (conclusão)

		Valor-p da comparação entre as médias do PCA-ANN(ST) e do PLS(CT)			
		Normalidade (Shapiro-Wilk)	**Teste t**	**Teste de Wilcoxon**	**Conclusão**
5. Cana-de-açúcar (Teor de Brix)	RMSE	0,198	0,904	-	NS
	r	0,001	-	0,232	NS
	TEA	0,001	-	0,075	NS
	TVP	0,255	0,919	-	NS
6. Manga (Teor de Vitamina C)	RMSE	0,102	0,873	-	NS
	r	0,273	-	1	NS
	TEA	0,617	0,001	-	*
	TVP	0,886	<0,001	-	*
7. Diesel (Densidade)	RMSE	0,030	0,080	-	NS
	r	0,000	-	0,090	NS
	TEA	0,000	-	0,006	*
	TVP	0,246	<0,001	-	*
8. Diesel (Viscosidade)	RMSE	0,004	-	0,539	NS
	r	0,020	0,837	-	NS
	TEA	0,070	0,027	-	NS
	TVP	0,855	0,025	-	NS

Em que: NS = Não significativo ao nível de 1%; * = Significativo ao nível de 1%.
Fonte: A Autora.

Na Tabela 7, a partir da análise dos parâmetros RMSE e r, nota-se que em todos os conjuntos avaliados, há indícios de que não existe diferença (valor-p > 0,01), em média, entre os métodos PCA-ANN(ST) e PLS(CT). Ou seja, os métodos são em média estatisticamente iguais, o que implica na possibilidade de utilização de qualquer um deles para o ajuste dos

modelos de predição. Porém, o método PCA-ANN(ST) deve ser aplicado em tais análises, pois apresentou TC inferior àquele relacionado ao uso do PLS(CT).

Assim como neste estudo, no trabalho desenvolvido por Mireei, Mohtasebi e Sadeghi (2014), que objetivou investigar o desempenho de métodos lineares e não lineares, incluindo os métodos PLS e PCA-ANN, este último apresentou resultados superiores àqueles obtidos pelo PLS. Apesar de não mencionarem o uso de testes de hipótese, os autores também afirmaram que tal diferença não foi significativa. Ressalta-se que em tal trabalho, cujo objetivo foi predizer a firmeza de frutos de tâmara, tanto o método PLS(CT) quanto o método PCA-ANN(ST) foram aplicados a dados de espectroscopia NIR.

Já na pesquisa desenvolvida por Pudełko e Chodak (2020), cujo objetivo foi comparar a eficiência de modelos na predição de teores de carbono orgânico e nitrogênio total em solos de minas, através do NIR, utilizou-se os seguintes métodos: PLS, PCA, ANN, PCA-ANN e ANN, cujas entradas corresponderam às variáveis latentes obtidas no PLS (PLS-ANN). Os autores relataram que a ANN aplicada a todas as variáveis do conjunto de dados gerou os piores resultados. Porém, com a aplicação do método PCA-ANN, os modelos foram mais precisos no treino e na validação, devendo o mesmo ser utilizado para a predição das variáveis resposta de interesse do estudo.

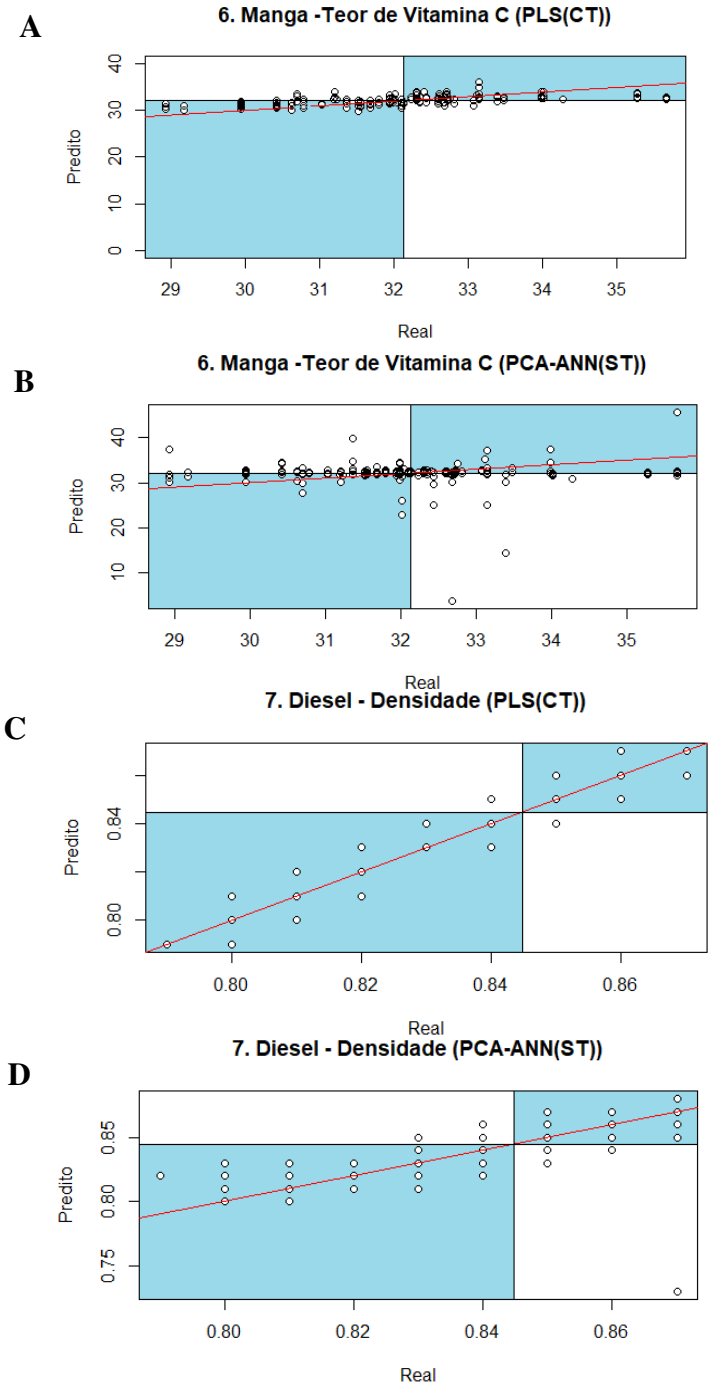
A partir da avaliação dos parâmetros relacionados aos processos de classificação (TEA e TVP), nota-se (Tabela 7) que apenas os conjuntos de dados 6 e 7 apontaram diferença significativa entre os métodos PCA-ANN(ST) e PLS(CT) (valor-p < 0,01). O conjunto de dados 6 apresentou valores médios de TEA = 0,24 e TVP = 0,70, ao utilizar o método PLS(CT), e valores médios de TEA = 0,49 e TVP = 0,49, com a aplicação do método PCA-ANN(ST). Já o conjunto de dados 7 apresentou valores médios de TEA = 0,03 e TVP = 0,98, por meio do uso do método PLS(CT), e valores médios de TEA = 0,11 e TVP = 0,92, empregando-se o método PCA-ANN(ST).

Portanto, o método PLS(CT) mostrou-se superior, em comparação ao método PCA-ANN(ST), na avaliação dos conjuntos de dados 6 e 7 mencionada acima, com maior acurácia de classificação (menores valores de TEA) e maior porcentagem de amostras corretamente classificadas (maiores valores de TVP). Nos demais conjuntos de dados, o PLS(CT) foi estatisticamente igual ao PCA-ANN(ST), devendo este último ser utilizado em razão do seu menor TC.

A fim de avaliar a superioridade do PLS(CT) em comparação ao método PCA-ANN(ST), identificada nos conjuntos 6 e 7, para os parâmetros relacionados aos processos de classificação (TEA e TVP), foram construídos os gráficos apresentados na Figura 14. Os

mesmos mostram como ficaram distribuídos os valores reais em função dos valores preditos pelos métodos, em todas as repetições realizadas.

Figura 14 - Gráficos da distribuição dos valores reais em função dos valores preditos pelos métodos PLS(CT) e PCA-ANN(ST) dos conjuntos de dados 6 (Manga – Teor de Vitamina C) e 7 (Diesel – Densidade), em todas as repetições realizadas

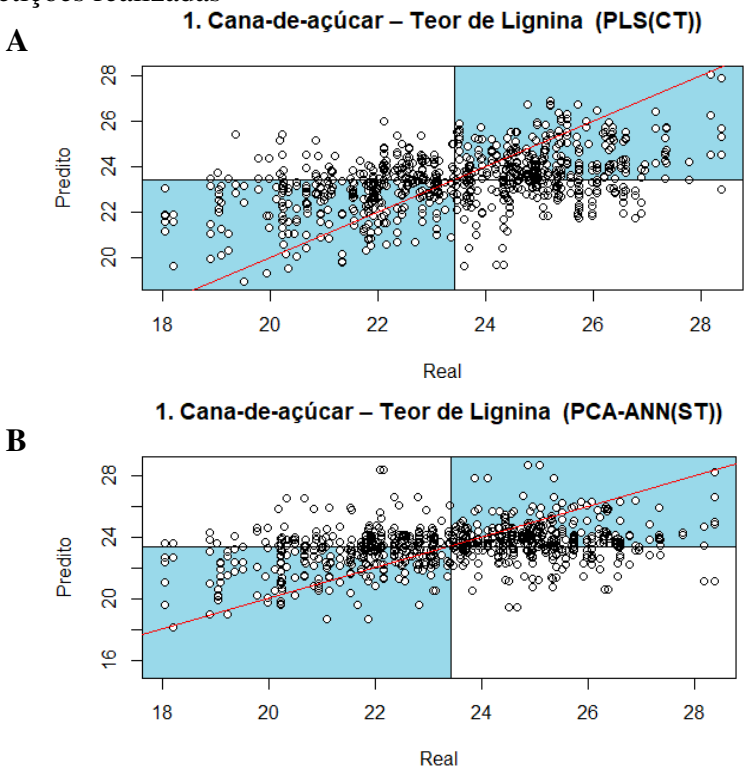


Em azul, tem-se a região das amostras classificadas corretamente. A função da reta destacada em vermelho é dada por x = y, em que x representa o valor real e y, o valor predito.
Fonte: A autora.

Nota-se nas Figuras 14-A e 14-B que a maioria dos pontos se aproxima da reta vermelha, cuja função é dada por x = y (em que x representa o valor real e y, o valor predito), ou seja, os métodos PCA-ANN(ST) e PLS(CT) conseguiram predizer a maior parte dos valores da propriedade em estudo, com apenas alguns pontos classificados incorretamente (aqueles localizados fora da região pintada de azul nas figuras). Porém, no gráfico correspondente à aplicação do método PCA-ANN(ST) ao conjunto 6 (Figura 14-B), observa-se que os pontos classificados incorretamente estão bem distantes da reta (vermelha), e podem ter influenciado no fato do método PCA-ANN(ST) ter se mostrado inferior ao PLS(CT). O mesmo comportamento foi identificado nos gráficos referentes à aplicação dos métodos ao conjunto 7, representados pelas Figuras 14-C e 14-D. Vale ressaltar que tais gráficos apresentam poucos pontos, pois muitos dos mesmos estão sobrepostos, em razão de alguns valores serem coincidentes.

A Figura 15 exibe os gráficos referentes ao conjunto de dados 1 (Cana-de-açúcar - Teor de Lignina), que correspondem à distribuição dos valores reais em função dos valores preditos pelos métodos PLS(CT) e PCA-ANN(ST).

Figura 15 - Gráficos da distribuição dos valores reais em função dos valores preditos pelos métodos PLS(CT) e PCA-ANN(ST) do conjunto de dados 1 (Cana-de-açúcar - Teor de Lignina) em todas as repetições realizadas

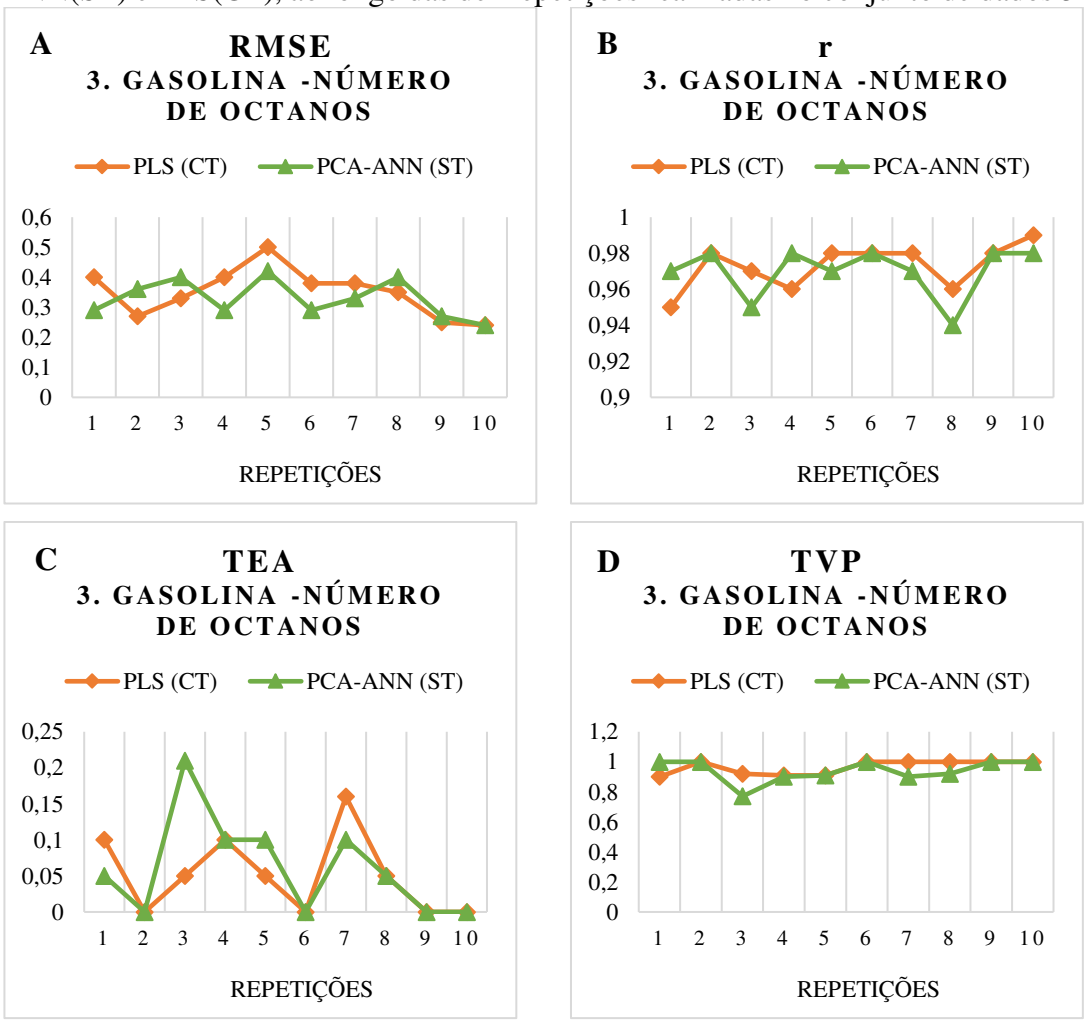


Em azul, tem-se a região das amostras classificadas corretamente. A função da reta destacada em vermelho é dada por x = y, em que x representa o valor real e y, o valor predito.
Fonte: A autora.

Nas Figuras 15-A e 15-B, nota-se um comportamento similar dos pontos, o que justifica os métodos PLS(CT) e PCA-ANN(ST) serem, em média, estatisticamente iguais. Observa-se, em tais figuras que, de maneira geral, os pontos se localizam próximo à reta (vermelha), o que se configura como um comportamento diferente daquele apresentado pelos gráficos referentes à aplicação do método PCA-ANN(ST), exibidos pelas Figuras 14-B e 14-D.

Na Tabela 7, nota-se que os parâmetros estatísticos médios do RMSE, r, TEA e TVP, referentes aos modelos de predição obtidos a partir da aplicação dos métodos PCA-ANN(ST) e PLS(CT), em cada repetição (1 a 10) realizada no conjunto de dados 3 (gasolina – número de octanos), não apresentaram diferença significativa entre si. A Figura 16 tem a finalidade de avaliar a tendência do comportamento de tais parâmetros no decorrer dessas dez repetições.

Figura 16 - Tendência do comportamento dos parâmetros RMSE, r, TEA e TVP, referentes aos modelos de predição obtidos a partir da aplicação dos métodos PCA-ANN(ST) e PLS(CT), ao longo das dez repetições realizadas no conjunto de dados 3



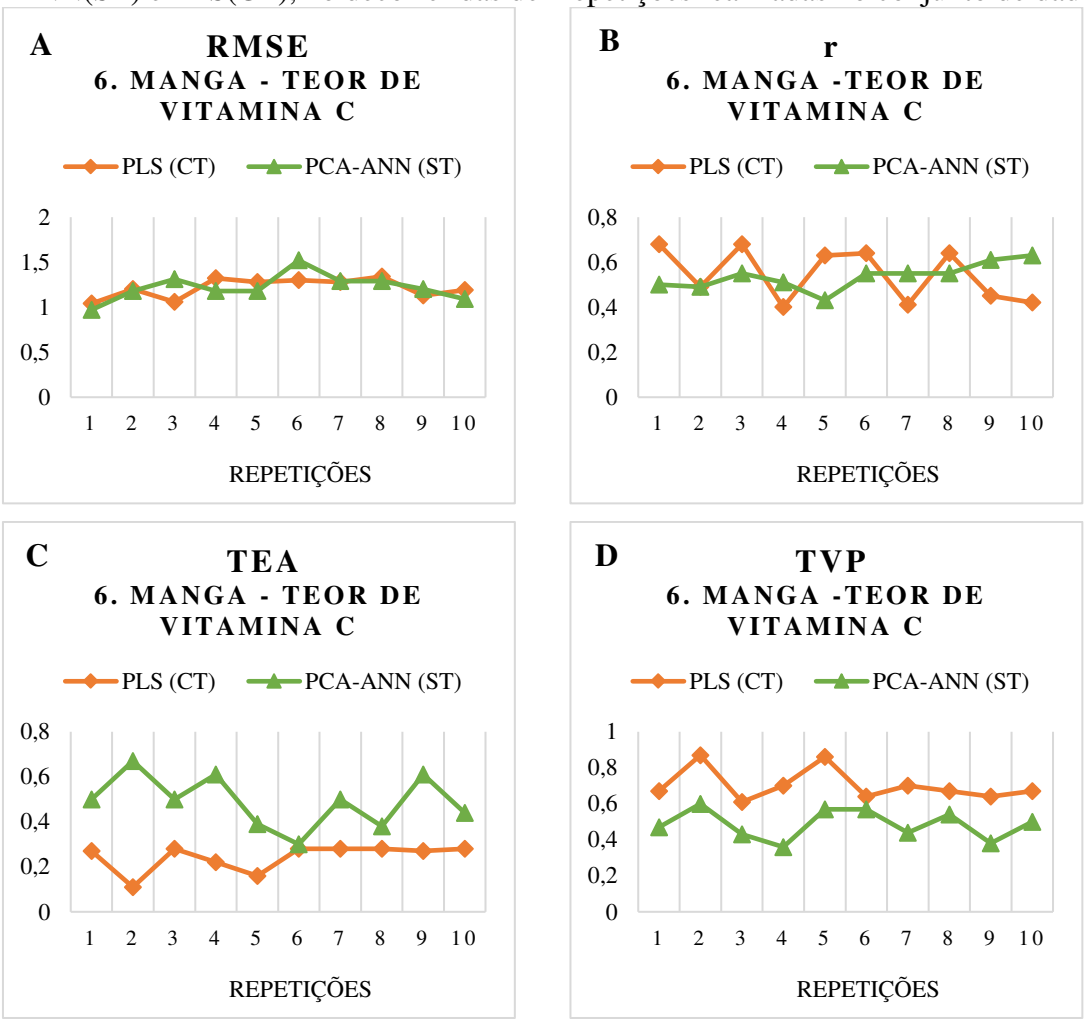
Em que: (A), (B), (C) e (D) representam, respectivamente, os valores da raiz quadrada do erro quadrático médio (RMSE), o coeficiente de correlação (r) entre os valores preditos pelo modelo e os observados, a taxa de erro aparente (TEA), e a taxa de verdadeiros positivos (TVP), nas dez repetições realizadas no conjunto de dados 3 (Gasolina – Número de Octanos). A linha laranja representa os valores obtidos pelo método PLS com pré-tratamento (PLS(CT)), e a linha verde representa os valores obtidos pelo PCA-ANN sem pré-tratamentos (PCA-ANN(ST)), em cada repetição.
Fonte: A autora.

Na Figura 16-A, na repetição 1, o PCA-ANN(ST) e o PLS(CT) apresentaram RMSE = 0,29 e RMSE = 0,40, respectivamente. Já na repetição 2, obteve-se um RMSE = 0,27 para o PLS(CT) e um RMSE = 0,36 para o PCA-ANN(ST). Na repetição 10, ambos os valores de RMSE foram iguais a 0,24. De maneira geral, em todos os parâmetros estatísticos apresentados na Figura 16, observou-se que os valores de RMSE, r, TEA e TVP, obtidos por cada um dos métodos, oscilaram de uma repetição para outra. Ou seja, dentro de algumas repetições, o método PLS(CT) foi superior ao PCA-ANN(ST), enquanto em outras, este último apresentou

melhores estatísticas. Além disso, dentro de algumas repetições não houve diferença expressiva entre os parâmetros avaliados. Tais fatos podem justificar a ausência de uma diferença significativa entre os métodos.

De acordo com a Tabela 7, no conjunto de dados 6 (manga – Teor de Vitamina C), para os parâmetros estatísticos médios relacionados aos processos de classificação (TEA e TVP), o método PLS(CT) foi significativamente superior ao método PCA-ANN(ST). Já em relação àqueles referentes à predição (RMSE e r), a diferença estatística entre os métodos não foi significativa ($p > 0,01$). A Figura 17 mostra o comportamento de tais parâmetros estatísticos, obtido em cada repetição.

Figura 17 - Tendência do comportamento dos parâmetros RMSE, r, TEA e TVP, referentes aos modelos de predição obtidos a partir da aplicação dos métodos PCA-ANN(ST) e PLS(CT), no decorrer das dez repetições realizadas no conjunto de dados 6



Em que: (A), (B), (C) e (D) representam, respectivamente, os valores da raiz quadrada do erro quadrático médio (RMSE), o coeficiente de correlação (r) entre os valores preditos pelo modelo e os observados, a taxa de erro aparente (TEA), e a taxa de verdadeiros positivos (TVP), nas dez repetições realizadas no conjunto de dados 6 (Manga – Teor de Vitamina C). A linha laranja representa os valores obtidos pelo método PLS com pré-tratamento (PLS(CT)), e a linha verde representa os valores obtidos pelo PCA-ANN sem pré-tratamentos (PCA-ANN(ST)), em cada repetição.
Fonte: A autora.

Observa-se, nas Figuras 17-A e 17-B, que o comportamento referente aos parâmetros estatísticos de predição foi semelhante àquele observado na Figura 16, em que os métodos não apresentaram diferença significativa. Já nas Figuras 17-C e 17-D, em todas as repetições, os valores médios referentes ao método PLS(CT) foram menores (Figura 17-C) e maiores (Figura 17-D) do que aqueles correspondentes ao método PCA-ANN(ST), o que pode justificar o fato dos testes t e de Wilcoxon terem detectado diferença estatística entre os métodos.

A partir da execução dos métodos propostos, nota-se que tanto o método PLS(CT) quanto o método PCA-ANN(ST) apresentam algumas desvantagens, ou seja, enquanto o primeiro apresenta a necessidade de definição do nVL e da utilização de pré-tratamentos e suas combinações na matriz de dados, o segundo requer a definição do número de neurônios na camada oculta e do número de CPs a ser utilizado. Contudo, apesar de tais métodos demandarem a definição de tais parâmetros, o fato do PCA-ANN(ST) dispensar a utilização dos pré-tratamentos representa uma grande vantagem, visto que esta é uma etapa que além de demandar mais trabalho e atenção do pesquisador, aumenta significativamente o tempo das análises. Desta forma, os resultados deste estudo indicam que o método PCA-ANN(ST) apresenta-se como uma boa alternativa aos métodos comumente utilizados para a predição, na quimiometria, em dados NIR.

5. CONCLUSÕES

Nos modelos de predição, os métodos PCA-ANN sem pré-tratamentos na matriz de espectros (PCA-ANN(ST)) e PLS com pré-tratamentos nessa matriz (PLS(CT)) foram estatisticamente iguais, ou seja, apresentaram a mesma capacidade preditiva em todos os conjuntos de dados avaliados. No processo de classificação utilizado neste estudo, em dois dos oito conjuntos de dados avaliados, o PLS(CT) foi estatisticamente superior ao PCA-ANN(ST), apresentando menor valor médio de TEA e maior valor médio de TVP. Comparado ao PLS(CT), o PCA-ANN(ST) apresentou menor tempo de execução e de recursos computacionais nas análises.

REFERÊNCIAS

ADNAN, A. et al. Rapid prediction of moisture content in intact green coffee beans using near infrared spectroscopy. **Foods**, v. 6, n. 5, p. 1–11, 2017.

AGUSSABTI et al. Data analysis on near infrared spectroscopy as a part of technology adoption for cocoa farmer in Aceh Province, Indonesia. **Data in Brief**, v. 29, p. 105251, 2020.

ARMENTA, S. et al. The Use of Near-Infrared Spectrometry in the Olive Oil Industry. **Critical Reviews in Food Science and Nutrition**, v. 50, n. 6, p. 567–582, 2010.

ASKE, N.; KALLEVIK, H.; SJÖBLOM, J. Determination of saturate, aromatic, resin, and asphaltenic (SARA) components in crude oils by means of infrared and near-infrared spectroscopy. **Energy and Fuels**, v. 15, n. 5, p. 1304–1312, 2001.

ASSIS, C. et al. Prediction of Lignin Content in Different Parts of Sugarcane Using Near-Infrared Spectroscopy (NIR), Ordered Predictors Selection (OPS), and Partial Least Squares (PLS). **Applied Spectroscopy**, v. 71, n. 8, p. 2001–2012, 2017.

AZADSHAHRAKI, F. et al. Diagnosis of Early Blight Disease in Tomato Plant based on Visible / Near- Infrared Spectroscopy and Principal Components Analysis- Artificial Neural Network Prior to Visual Disease Symptoms. **Journal of Agricultural Machinery**, v. 12, n. 1, p. 81–94, 2022.

BALABIN, R. M.; LOMAKINA, E. I. Support vector machine regression (SVR/LS-SVM) - An alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. **Analyst**, v. 136, n. 8, p. 1703–1712, 2011.

BALABIN, R. M.; SAFIEVA, R. Z.; LOMAKINA, E. I. Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction. **Chemometrics and Intelligent Laboratory Systems**, v. 88, n. 2, p. 183–188, 2007.

BARLOW, J. L.; BOSNER, N.; DRMAČ, Z. A new stable bidiagonal reduction algorithm. **Linear Algebra and its Applications**, v. 397, p. 35–84, mar. 2005.

BARTLETT, M. S. The effect of non-normality on the t-distribution. **Proceedings of the Cambridge Philosophical Society**, v. 31, p. 223–231, 1935.

BISHOP, C. M. **Pattern recognition and machine learning**. 1. ed. New York, NY: Springer, 2006.

BLANCO, M.; VILLARROYA, I. NIR spectroscopy: a rapid-response analytical tool. **TrAC Trends in Analytical Chemistry**, v. 21, n. 4, p. 240–250, abr. 2002.

BOGREKCI, I.; LEE, W. S. Spectral measurement of common soil phosphates. **Transactions of the ASAE**, v. 48, n. 6, p. 2371–2378, 2005.

BOKOBZA, L. Near infrared spectroscopy. **Journal of Near Infrared Spectroscopy**, v. 6, n.

1–4, p. 3–17, 1998.

BOLFARINE, H.; BUSSAB, W. O. **Elementos de Amostragem**. São Paulo: Blücher, 2005.

BRAGA, A. P.; CARVALHO, A. P. L. F.; LUDERMIR, T. B. **Redes Neurais Artificiais - Teoria e Aplicações**. 2. ed. Rio de Janeiro: LTC, 2007.

BROWNFIELD, B.; KALIVAS, J. H. Consensus Outlier Detection Using Sum of Ranking Differences of Common and New Outlier Measures Without Tuning Parameter Selections. **Analytical Chemistry**, v. 89, n. 9, p. 5087–5094, 2017.

CHEN, D. et al. A Background and noise elimination method for quantitative calibration of near infrared spectra. **Analytica Chimica Acta**, v. 511, n. 1, p. 37–45, maio 2004.

CHEN, H. et al. A deep learning CNN architecture applied in smart near-infrared analysis of water pollution for agricultural irrigation resources. **Agricultural Water Management**, v. 240, n. April, p. 106303, 2020.

CHEN, Y. et al. Prediction of drug content and hardness of intact tablets using artificial neural network and near-infrared spectroscopy. **Drug Development and Industrial Pharmacy**, v. 27, n. 7, p. 623–631, 2001.

CRUZ, C. D.; NASCIMENTO, M. **Inteligência Computacional Aplicada ao Melhoramento Genético**. Viçosa: Editora UFV, 2018.

DALE, L. M. et al. Hyperspectral imaging applications in agriculture and agro-food product quality and safety control: A review. **Applied Spectroscopy Reviews**, v. 48, n. 2, p. 142–159, 2013.

DASGAONKAR, K. Analysis of multi-layered perceptron , radial basis function and convolutional neural networks in recognizing handwritten digits. **International Journal of Advance Research, Ideas and Innovations in Technology,** v. 4, n. 3, p. 2429–2431, 2018.

DE SOUSA, L. C. et al. Desenvolvimento de modelos de calibração NIRS para minimização das análises de madeiras de eucalyptus spp. **Ciencia Florestal**, v. 21, n. 3, p. 589–597, 2011.

DE SOUZA, A. M.; POPPI, R. J. Teaching experiment of chemometrics for exploratory analysis of edible vegetable oils by mid infrared spectroscopy and principal component analysis: A tutorial, part I. **Química Nova**, v. 35, n. 1, p. 223–229, 2012.

DIAS, M. P. A.; INÁCIO, M. J.; CARVALHO JÚNIOR, Á. B. DE. Aplicação De Redes Neurais Artificiais Para Previsão Da Incidência Solar Na Cidade De Belo Horizonte. **Brazilian Journal of Development**, v. 6, n. 7, p. 52603–52615, 2020.

ENGEL, J. et al. Breaking with trends in pre-processing? **TrAC - Trends in Analytical Chemistry**, v. 50, p. 96–106, 2013.

FERNANDEZ, A. S. T. et al. Autenticação de orégano (origanum vulgare l.) orgânico utilizando espectroscopia NIR e quimiometria. **Química Nova**, v. 43, p. 1500-1504, 2021.

FERREIRA, M. M. C. et al. Chemometrics i: Multivariate calibration, a tutorial. **Química Nova**, v. 22, p. 724-731, 1999.

FERREIRA, M. M. C. Multivariate QSAR. **Journal of the Brazilian Chemical Society**, v. 13, n. 6, p. 742–753, 2002.

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

FERREIRA, R. DE A.; PETERNELLI, L. A. Comparison of variable selection methods in predictive models applied to near-infrared and genomic data. **Genetics and Molecular Research**, v. 20, n. 3, p. 1–16, 2021.

FERREIRA, R. DE A.; TEIXEIRA, G.; PETERNELLI, L. A. Kennard-Stone method outperforms the Random Sampling in the selection of calibration samples in SNPs and NIR data. **Ciência Rural**, v. 52, n. 5, p. 1–11, 2022.

FOLCH-FORTUNY, A.; ARTEAGA, F.; FERRER, A. PCA model building with missing data: New proposals and a comparative study. **Chemometrics and Intelligent Laboratory Systems**, v. 146, p. 77–88, 2015.

FREUND, R. J.; WILSON, W. J.; SA, P. **Regression analysis –** Statistical Modeling of a response variable. San Diego: Elsevier, 2006.

GALVÃO, C. O. et al. **Sistemas inteligentes:** aplicações a recursos hídricos e ciências ambientais. UFRGS/ ABRH, 1999.

GALVÃO, R. K. H. et al. A method for calibration and validation subset partitioning. **Talanta**, v. 67, n. 4, p. 736–740, 2005.

GIANOLA, D. et al. Predicting complex quantitative traits with Bayesian neural networks: A case study with Jersey cows and wheat. **BMC Genetics**, v. 12, p. 4–7, 2011.

GOLUB, G. H.; VAN LOAN, C. F. **Matrix computation**. 2. ed. Johns Hopkins University Press: Baltimore, 1989.

GOURVÉNEC, S. et al. An evaluation of the PoLiSh smoothed regression and the Monte Carlo Cross-Validation for the determination of the complexity of a PLS model. **Chemometrics and Intelligent Laboratory Systems**, v. 68, n. 1–2, p. 41–51, 2003.

GOYAL, S. Artificial Neural Networks in Fruits: A Comprehensive Review. **International Journal of Image, Graphics and Signal Processing**, v. 6, n. 5, p. 53–63, 2014.

GRASSI, S.; ALAMPRESE, C. Advances in NIR spectroscopy applied to process analytical technology in food industries. **Current Opinion in Food Science**, v. 22, p. 17–21, 2018.

GUO, Z. et al. Color compensation and comparison of shortwave near infrared and long wave near infrared spectroscopy for determination of soluble solids content of “Fuji” apple. **Postharvest Biology and Technology**, v. 115, p. 81–90, 2016.

GUYTON, A. **Fisiologia Humana**. 6. ed. Rio de Janeiro: Guanabara Koogan, 1988.

HAYATI, R.; MUNAWAR, A. A.; FACHRUDDIN, F. Enhanced near infrared spectral data to improve prediction accuracy in determining quality parameters of intact mango. **Data in Brief**, v. 30, p. 105571, 2020.

HAYKIN, S. **Redes Neurais:** Princípios e Prática. 2. ed. Bookman, 2001.

HODGES, J. L.; LEHMANN, E. L. The efficiency of some nonparametric competitors of the t test. **Annals of Mathematical Statistic**, v. 27, p. 324–335, 1956.

HOTELLING, H. (1947), "Multivariate quality control, illustra-ted by the air testing of sample bombsights", Techniques of Statistical Analysis, pp.111-184. New York, McGraw Hill.

HUTZLER, S. A.; BESSEE, G. B. Remote Near-Infrared Fuel Monitoring System. **Southwest research inst san antonio txtardec fuels and lubricants research facility**, p. 33, 1997.

JANIK, L. J.; FORRESTER, S. T.; RAWSON, A. The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 97, n. 2, p. 179–188, 2009.

JILANI, T. A. PCA-ANN for Classification of Hepatitis-C Patients. v. 14, n. 7, p. 1–6, 2011. KALIVAS, J. H. Chemometricsand intelligent laboratory systems Two data sets of near infrared spectra. **Chemometrics and Intelligent Laboratory Systems**, v. 37, p. 255–259, 1997.

KANJI, G. K. **100 Statistical Tests**. 3. ed. London: Sage Publications Ltd, 2006.

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

KIM, T. K. T test as a parametric statistic. **Korean journal of anesthesiology**, v. 68, p. 540, 2015.

KODAIRA, M.; SHIBUSAWA, S. Using a mobile real-time soil visible-near infrared sensor for high resolution soil property mapping. **Geoderma**, v. 199, p. 64–79, 2013.

KUHN, M. **Caret: classification and regression training**Astrophysics Source Code Library, 2016. Disponível em: <http://topepo.github.io/caret/train-models-by-tag.html>. Acesso em: 13 de mar. de 2022.

LEE, L. C.; LIONG, C. Y.; JEMAIN, A. A. Iterative random vs. Kennard-Stone sampling for IR spectrum-based classification task using PLS2-DA. **AIP Conference Proceedings**, v. 1940, 2018.

LOPES, S. **Bio**. 11. ed. São Paulo, 2000.

MANLY, B. F. **Multivariate Statistical Methods: A Primer**. 3. ed. Chapman and Hall/CRC, 2004.

MARTENS, H.; NAES, T. **Multivariate Calibration**. John Wiley & Sons: New York, 1989; p 419.

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.

MAURO, L.; CAMPOS, L. DE. Aplicação de redes neurais recorrentes e profundas em tarefas de classificação. **Revista Gestão & Tecnologia,** v. 20, n. 3, 2020.

MCCARTY, G. W. et al. Mid-Infrared and Near-Infrared Diffuse Reflectance Spectroscopy for Soil Carbon Measurement. **Soil Science Society of America Journal**, v. 66, n. 2, p. 640–646, 2002.

MCCULLOCH, W. S. .; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bull Math Bio**, v. 5, p. 115–133, 1943.

MIREEI, S. A.; MOHTASEBI, S. S.; SADEGHI, M. Comparison of linear and non-linear calibration models for non-destructive firmness determining of mazafati date fruit by near infrared spectroscopy. **International Journal of Food Properties**, v. 17, n. 6, p. 1199–1210, 2014.

MIREEI, S. A.; SADEGHI, M. Detecting bunch withering disorder in date fruit by near infrared spectroscopy. **Journal of Food Engineering**, v. 114, n. 3, p. 397–403, 2013.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. **Introduction to the theory of statistics**. New York - US: McGraw-Hill, 1974.

MOREIRA, É. F. A.; BARBOSA, M. H. P.; PETERNELLI, L. A. Can statistical learning models make early selection among sugarcane families easier and still efficient? **Crop Science**, v. 61, n. 1, p. 456–465, 2021.

MORGANO, M. A. et al. Determinação de umidade em café cru usando espectroscopia NIR e regressão multivariada. **Ciencia e Tecnologia de Alimentos**, v. 28, n. 1, p. 12–17, 2008.

MOZAFFARI, M.; SADEGHI, S.; ASEFI, N. Prediction of the quality properties and maturity of apricot by laser light backscattering imaging. **Postharvest Biology and Technology**, v. 186, p. 111842, abr. 2022.

MUÑOZ, E. et al. Prediction of PM10 and SO_2 exceedances to control air pollution in the Bay of Algeciras, Spain. **Stochastic Environmental Research and Risk Assessment**, v. 28, n. 6, p. 1409–1420, 28 ago. 2014.

MUTLU, A. C. et al. Prediction of wheat quality parameters using near-infrared spectroscopy and artificial neural networks. **European Food Research and Technology**, v. 233, n. 2, p. 267–274, 2011.

NAES, T. et al. **A user-friendly guide to multivariate calibration and classification**. Vol. 6. UK: Chichester: NIR, 2002.

NAGY, B. et al. Application of artificial neural networks for Process Analytical Technology-based dissolution testing. **International Journal of Pharmaceutics**, v. 567, n. March, 2019.

NASCIMENTO, M. et al. Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. **Crop Breeding and Applied Biotechnology**, v. 13, n. 2, p. 152–156, 2013.

NEAVE, H. R.; GRANGER, C. W. J. A monte carlo study comparing various two sample tests for differences in means. **Technometrics**, v. 10, p. 509–522, 1968.

NOVAIS, B. V.; VALVERDE, K. P. Virtual analyzer of extractive content in Eucalyptus wood based on hybrid modeling approach for the pulp and paper industry. **Wood Science and Technology**, v. 55, n. 3, p. 777–795, 29 maio 2021.

OLIVEIRA-ESQUERRE, K. P.; MORI, M.; BRUNS, R. E. Simulation of an industrial wastewater treatment plant using artificial neural networks and principal components analysis. **Brazilian Journal of Chemical Engineering**, v. 19, n. 4, p. 365–370, 2002.

OLIVEIRA, A. A. et al. Identificação de Madeiras utilizando a Espectrometria no Infravermelho Próximo e Redes Neurais Artificiais. **TEMA (São Carlos)**, v. 16, n. 2, p. 81, 2015.

PASQUINI, C. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. **Journal of the Brazilian Chemical Society**, v. 14, n. 2, p. 198–219, 2003.

PASQUINI, C. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. **Analytica Chimica Acta**, v. 1026, p. 8–36, 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. Selection of sugarcane clones via multivariate models using near-infrared (NIR) spectroscopy data. **Australian Journal of Crop Science**, v. 14, n. 6, p. 889–896, 2020.

POSTEN, H. O.; YEH, H. C.; OWEN, D. B. Robustness of the two-sample t-test under violations of the homogeneity of variance assumptions. **Communications in Statistics: Theory and Methods**, v. 11, p. 109–126, 1982.

PUDEŁKO, A.; CHODAK, M. Estimation of total nitrogen and organic carbon contents in mine soils with NIR reflectance spectroscopy and various chemometric methods. **Geoderma**, v. 368, n. February, 2020.

R CORE TEAM. **R: A Language and Environment for Statistical Computing.** Vienna, AustriaR Foundation for Statistical Computing, 2021. Disponível em: <https://www.r-project.org/>

RAJER-KANDUČ, K.; ZUPAN, J.; MAJCEN, N. Separation of data on the training and test set for modelling: A case study for modelling of five colour properties of a white pigment. **Chemometrics and Intelligent Laboratory Systems**, v. 65, n. 2, p. 221–229, 2003.

RAMADAN, Z. et al. Application of PLS and Back-Propagation Neural Networks for the estimation of soil properties. **Chemometrics and Intelligent Laboratory Systems**, v. 75, n. 1, p. 23–30, 2005.

RANDLES, R. H.; WOLFE, D. A. **Introduction to the theory of nonparametric statistics.** New York: Wiley, 1979.

RASCH, D.; GUIARD, V. The robustness of parametric statistical methods. **Psychology Science**, v. 46, p. 175–208, 2004.

RICHARD G. BRERETON. **Chemometrics: Data Analysis for the Laboratory and Chemical Plant**. v. 8 ed. England: John Wiley & Sons, Ltda., 2003.

ROQUE, J. V. **Desenvolvimento de modelos de regressão multivariada para determinação de ésteres de forbol em sementes de jatropha curcas L. usando espectroscopia e quimiometria.** 2015. 84 f. Dissertação (Mestrado em Agroquímica) - Universidade Federal de Viçosa, Viçosa, 2015.

ROSSEL, R. A. V.; BEHRENS, T. Using data mining to model and interpret soil diffuse reflectance spectra. **Geoderma**, v. 158, n. 1–2, p. 46–54, 2010.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, p. 533–536, 1986.

SABIN, J. G.; FERRÃO, M. F.; FURTADO, J. C. Análise multivariada aplicada na identificação de fármacos antidepressivos. Parte II: Análise por componentes principais (PCA) e o método de classificação SIMCA. **Revista Brasileira de Ciências Farmacêuticas**, v. 40, n. 3, p. 387–396, 2004.

SACHDEVA, J. et al. Segmentation, feature extraction, and multiclass brain tumor classification. **Journal of Digital Imaging**, v. 26, n. 6, p. 1141–1150, 2013.

SANTOS, A. M. DOS et al. Using Artificial Neural Networks and Logistic Regression in the Prediction of Hepatitis A. **Rev Bras Epidemiol**, v. 8, n. 2, p. 117–126, 2005.

SANTOS, P. et al. Determinação da Autenticidade de Amostras de Azeite Comerciais Apreendidas no Estado do Espírito Santo Usando Um Espectrofotômetro Portátil Na Região Do Nir. **Química Nova**, v. 43, n. 7, p. 891–900, 2020.

SAVITZKY, A.; GOLAY, M. Smoothing and differentiation of data by simplified least squares procedures. **Analytical Chemistry**, v. 36, n. 8, p. 1627–1639, 1964.

SCHEFFÉ, H. Practical Solutions of the-Behrens-Fisher Problem. **Journal of the American Statistical Association**, v. 65, n. 332, p. 1501–1508, 1970.

SHAPIRO, S. S.; WILK, M. B. Approximations for the Null Distribution of the W Statistic. **Technometrics**, v. 10, n. 4, p. 861–866, 1968.

SILVA, G. N. et al. Evaluation of the efficiency of artificial neural networks for genetic value prediction. **Genetics and Molecular Research**, v. 15, n. 1, p. 1–11, 2016.

SIQUEIRA-BATISTA, R. et al. artificial Neural Networks and medical Education as redes Neurais artificiais e o Ensino da medicina Andréia Patrícia Gomes III Alcione de Paiva Oliveira III Ricardo dos Santos Ferreira III Vanderson Esperidião-Antonio III. **Revista Brasileira de Educação Médica**, v. 38, n. 4, p. 557–565, 2014.

STEEL, R.; TORRIE, J.; DICKEY, D. **Principles and Procedures of Statistics:** A Biometrical Approach. New York, NY, USA: McGraw-Hil, 1996.

STWART, A. D. et al. Spectroscopy measurement of used lubricating oil quality. **Applied Spectroscopy**, v. 43, p. 55–60, 1995.

TAFNER, A. T.; XEREZ, M.; FILHO, E. R. **Redes Neurais artificiais : introdução e princípios de neurocomputação**. Blumenau : EKO, 1995.

TEÓFILO, R. F.; MARTINS, J. P. A.; FERREIRA, M. M. C. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. **Journal of Chemometrics**, v. 23, n. 1, p. 32–48, 2009.

TEÓFILO, R. F. **Métodos quimiométricos em estudos eletroquímicos de fenóis sobre filmes de diamante dopado com boro**. 2007. 329 f: Tese (Doutorado em Química) - Universidade Estadual de Campinas, 2007.

TRIOLA, M. F. **Introdução à Estatística**. 10. ed. Rio de Janeiro: LTC, 2008.

VALDERRAMA, P.; BRAGA, J. W. B.; POPPI, R. J. Validation of multivariate calibration models in the determination of sugar cane quality parameters by near infrared spectroscopy. **Journal of the Brazilian Chemical Society**, v. 18, n. 2, p. 259–266, 2007.

VASQUES, G. M.; GRUNWALD, S.; SICKMAN, J. O. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. **Geoderma**, v. 146, n. 1–2, p. 14–25, 2008.

VENDRUSCOLO, D. G. S. et al. Estimativa Da Altura De Eucalipto Por Meio De Regressão Não Linear E Redes Neurais Artificiais. **Revista Brasileira de Biometria**, v. 33, n. 4, p. 556–569, 2015.

VENTURA, R. V. et al. Uso de redes neurais artificais na predição de valores genéticos para peso aos 205 dias em bovinos da raça tabapuã. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 64, n. 2, p. 411–418, 2012.

VERHOEVEN, G. Imaging the invisible using modified digital still cameras for straightforward and low-cost archaeological near-infrared photography. **Journal of Archaeological Science**, v. 35, n. 12, p. 3087–3100, dez. 2008.

WANG, Y. et al. Big data driven outlier detection for soybean straw near infrared spectroscopy. **Journal of Computational Science**, v. 26, p. 178–189, 2018.

WEHRENS, R. et al. Package ‘ pls ’. 2021.

WILCOXON, F. Individual Comparisons by Ranking Methods. **Biometrics Bulletin**, v. 1, n. 6, p. 80, dez. 1945.

WOLD, H. Soft modeling: the basic design and some extensions. **Systems under indirect observation**, v. 2, p. 343, 1982.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: A basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v.58, p. 109-130, 2001.

XU, L. et al. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. **Analytica Chimica Acta**, v. 616, n. 2, p. 138–143, 2008.

YOPLAC, I. et al. Determination of the superficial citral content on microparticles: an application of NIR spectroscopy coupled with chemometric tools. **Heliyon**, v. 5, n. 7, p. e02122, 2019.

YUAN, J. S. et al. Statistical analysis of real-time PCR data. **BMC Bioinformatics**, v. 7, p. 1–12, 2006.

ZHAO, J. et al. Rapid quantification of active pharmaceutical ingredient for sugar-free Yangwei granules in commercial production using FT-NIR spectroscopy based on machine learning techniques. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, v. 245, p. 118878, 2021.

ZONTOV, Y. V. et al. Chemometric aided NIR portable instrument for rapid assessment of medicine quality. **Journal of Pharmaceutical and Biomedical Analysis**, v. 131, p. 87–93, 2016.