UNIVERSIDADE ESTADUAL DE PONTA GROSSA
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

ALAN FERNANDO COELHO GARCIA

SISTEMA DE NARIZ ELETRÔNICO APLICADO NA ANÁLISE DA FLORAÇÃO E RALEIO EM PESSEGUEIROS

PONTA GROSSA - PR

2023

ALAN FERNANDO COELHO GARCIA

SISTEMA DE NARIZ ELETRÔNICO APLICADO NA ANÁLISE DA FLORAÇÃO E RALEIO EM PESSEGUEIROS

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, como requisito parcial para obtenção do título de mestre em Computação Aplicada na área de Computação para Tecnologias em Agricultura.

Orientador: Prof. Dr. Sergio Luiz Stevan Junior
Coorientador: Prof. Dr. Ricardo Antonio Ayub

PONTA GROSSA - PR

2023

Garcia, Alan Fernando Coelho
G216 Sistema de nariz eletrônico aplicado na análise da floração e raleio em pessegueiros / Alan Fernando Coelho Garcia. Ponta Grossa, 2023.
106 f.

Dissertação (Mestrado em Computação Aplicada - Área de Concentração: Computação para Tecnologias em Agricultura), Universidade Estadual de Ponta Grossa.

Orientador: Prof. Dr. Sergio Luiz Stevan Junior.
Coorientador: Prof. Dr. Ricardo Antonio Ayub.

1. Nariz eletrônico. 2. Aprendizado de máquina. 3. Pêssego. 4. Compostos orgânicos voláteis (COVs). 5. Raleio. I. Stevan Junior, Sergio Luiz. II. Ayub, Ricardo Antonio. III. Universidade Estadual de Ponta Grossa. Computação para Tecnologias em Agricultura. IV.T.

CDD: 004

Ficha catalográfica elaborada por Maria Luzia Fernandes Bertholino dos Santos- CRB9/986

TERMO

TERMO DE APROVAÇÃO

Alan Fernando Coelho Garcia

SISTEMA DE NARIZ ELETRÔNICO APLICADO NA ANÁLISE DA FLORAÇÃO E RALEIO EM PESSEGUEIROS

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora:

Prof. Dr. Sérgio Luiz Stevan Júnior(UTFPR-PG) Presidente)

Prof. Dr. José Carlos Ferreira da Rocha (UEPG)

Prof. Dr. Hugo Valadares Siqueira (UTFPR-PG)

Ponta Grossa, 26 de setembro de 2022.

Documento assinado eletronicamente por **Hugo Valadares Siqueira, Usuário Externo**, em 30/09/2022, às 10:31, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Jose Carlos Ferreira da Rocha, Coordenador(a) do Programa de Pós-Graduação em Computação Aplicada - Mestrado**, em 06/03/2023, às 16:04, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Sergio Luiz Stevan Junior, Usuário Externo**, em 19/05/2023, às 11:24, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

A autenticidade do documento pode ser conferida no site https://sei.uepg.br/autenticidade informando o código verificador **1137626** e o código CRC **52D7DEB5**.

Dedico este trabalho à minha mãe, Marilene Coelho. E ao meu irmão, Alisson Felipe Coelho Garcia.

AGRADECIMENTOS

Primeiramente gostaria de agradecer a minha Mãe, Marilene Coelho, pelo amor, cuidado e inspiração que hoje fundamentam o que sou e meus sentimentos mais profundos e vivos de orgulho, gratidão e amor a você mãe. Muito obrigado.

Em seguida meus agradecimentos são para a influência direta que tenho na vida até o momento, meu irmão, Alisson Felipe Coelho Garcia. Agradeço pelos conselhos, pela didática, pela amizade, e também pela inspiração. Além disso, agradeço e reconheço o privilégio que tive e tenho em ser seu irmão, principalmente pelo aprendizado, seja observando, imitando ou aperfeiçoando seus atos (maioria dos casos, :)). Privilégio que se não existisse não seria o que sou hoje. A você meu irmão, muito obrigado.

Ao meu orientador, Professor Dr. Sergio Luiz Stevan Junior pela orientação, pelo ensinamento e pela paciência durante os anos do mestrado. Também agradeço ao nosso co-orientador, o Professor Dr. Antonio Ricardo Ayub por toda contribuição e apoio teórico e prático ao desenvolvimento deste trabalho.

Aos membros da banca, o Professor Dr. José Carlos Ferreira da Rocha, o Professor Dr. Hugo Valadares Siqueira, a Professora Dra. Rosane Falate, e a Professora Dra. Christiane Gonçalves pelo aceite e contribuições neste trabalho.

À Universidade Estadual de Ponta Grossa e à coordenação do PPGCA pelo apoio e paciência durante o mestrado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa que foi fundamental no desenvolvimento teórico e prático desta pesquisa.

Por fim, agradeço a todos os educadores que fizeram parte da minha educação. Muito obrigado.

RESUMO

O pessegueiro durante seu período reprodutivo apresenta brotações floríferas que se desenvolvem passando por diferentes estágios, como a abertura das pétalas e o inhaço do ovário, até a formação do fruto. Para melhorar a distribuição dos recursos em plantas frutíferas o raleio é empregado. A técnica reduz a quantidade de brotações ou frutos, sendo uma atividade que traz benefícios, porém é onerosa e em alguns casos impraticável em toda extensão de um pomar. O momento adequado para execução e a intensidade são questões complexas que envolvem conhecimento específico sobre as condições de um pomar. Portanto, informações que possam auxiliar o planejamento e a tomada de decisão sobre raleio são relevantes e podem melhorar o custo benefício da técnica. Nesta pesquisa a informação sobre o estágio das brotações foi considerada a mais importante, sendo usada para definir o estado geral de desenvolvimento do pessegueiro. No entanto, o desenvolvimento das brotações ocorre simultaneamente de acordo com as características da planta e as condições do ambiente. Uma abordagem para o problema é analisar odor do pessegueiro e as possíveis variações devido o desenvolvimento das brotações. Nesse sentido, um sistema de nariz eletrônico pode ser usado, já que é um equipamento que utiliza uma matriz de sensores de gás e reconhecimento padrões para relacionar amostras de odor com o objetivo da análise. Para isso, um prótotipo foi desenvolvido e coletas foram executadas durante o período reprodutivo de um pessegueiro. A hipótese da pesquisa foi de que o nariz eletrônico seria capaz de identificar o estágio geral das brotações. Para verificar a hipótese, uma análise de classificação foi efetuada. O resultado mostrou que todos os modelos apresentaram acurácia balanceada acima 90%, com melhor desempenho para os modelos baseados em *K-Nearest Neighbors* (KNN), e *Random Forest* (RF), do que os baseados em *Extreme Learning Machine* (ELM) e *Support Vector Machine* (SVM). No total de 24 modelos, 11 não obtiveram erro dado as métricas empregadas. Em relação à hipótese da pesquisa, considera-se que o nariz eletrônico foi capaz de distinguir os diferentes estágios propostos, sendo eles de queda das pétalas (QP), formação de fruto inicial (FFI), formação de fruto avançada (FFA), e fruto formado (F). Os resultados desta pesquisa indicam para um potencial uso do nariz eletrônico no auxílio da tomada de decisão de atividades em pomares.

Palavras-chave: Nariz eletrônico, Aprendizado de máquina, Pêssego, Floração, Compostos Orgânicos Voláteis (COVs), Raleio.

ABSTRACT

The peach tree during its reproductive period presents flowering buds that develop through different stages, such as the opening of the petals and the swelling of the ovary, until the formation of the fruit. To improve the distribution of resources in fruiting plants thinning is employed. The technique reduces the number of buds or fruits, being an activity that brings benefits, but it is costly and in some cases impractical in the entire extension of an orchard. The right moment for execution and the intensity are complex issues that involve specific knowledge about the conditions of an orchard. Therefore, information that can help planning and decision-making about thinning is relevant and can improve the cost-effectiveness of the technique. In this study, information about the budding stage was considered the most important, being used to define the general state of development of the peach tree. However, the development of buds occurs simultaneously according to plant characteristics and environmental conditions. One approach to the problem is to analyze peach tree odor and possible variations due to bud development. In this sense, an electronic nose system can be employed, since it is a device that uses an array of gas sensors and pattern recognition to relate odor samples with the purpose of analysis. For this, a prototype was developed and collections were performed during the reproductive period of a peach tree. The research hypothesis was that the electronic nose can identify the general stage of the buds. To verify the hypothesis, a classification analysis was performed. The result showed that all models presented balanced accuracy above 90%, with better performance for models based on K-Nearest Neighbors (KNN), and Random Forest (RF), than those based on Extreme Learning Machine (ELM) and Support Vector Machine (SVM). In the total of 24 models, 11 did not obtain error given the metrics employed. Regarding the research hypothesis, it is considered that the electronic nose was able of distinguishing between the different proposed stages, namely petal fall (QP), initial fruit formation (FFI), advanced fruit formation (FFA), and formed fruit (F). The results of this study indicate a potential use of the electronic nose in aiding decision-making in activities in orchards.

Keywords: Electronic nose, Machine learning, Peach, Flowering, Volatile Organic Compounds (VOCs), Thinning.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

COV	Composto Orgânico Volátil
GC-MS	*Gas Chromatography–Mass Spectrometry* - Espectrometria de Massa por Cromatografia Gasosa
SSC	*Soluble Solid Content* - Sólidos Solúveis Totais
BAW	*Bulk Acoustic Wave* - Onda Acústica Volumétria
SAW	*Surface Acoustic Wave* - Onda Acústica Superficial
CP	*Conductive Polymer* - Polímero Condutivo
PLSR	*Partial Least Squares Regression* - Regressão de Mínimos Quadrados Parciais
DWT	*Discret Wavelet Transform* - Transformada Wavelet Discreta
PCA	*Principal Component Analysis* - Análise de Componentes Principais
ID3	*Iterative Dichotomiser 3* - Dicotomizador Iterativo 3
CART	*Classification And Regression Tree* - Árvore de Classificação e Regressão
LDA	*Linear Discriminant Analysis* - Análise de Discriminantes Lineares
KNN	*K-Nearest Neighbors* - K-Vizinhos Mais Próximos
ELM	*Extreme Learning Machine* - Máquina de Aprendizado Extremo
RNA	Rede Neural Artificial
ReLU	*Rectified Linear Unit* - Ativação Linear Retificada
RF	*Random Forest* - Floresta Aleatória
CART	*Classification And Regression Tree* - Árvore de Classificação e Regressão
ID3	*Iterative Dichotomiser 3* - Dicotomizador Iterativo 3
SVM	*Support Vector Machine* - Máquina de Vectores de Suporte
KKT	Condições de Karush-Kuhn-Tucker
OVR	*One-Versus-Rest* - Abordagem Um-Contra-todos
OVO	*One-Versus-One* - Abordagem Um-Contra-Um
UART	*Universal Asynchronous Receiver Transmitter* - Transmissor Receptor Universal Assíncrono

RAM	*Random Access Memory* - Memória de Acesso Aleatória
LAN	*Local Area Network* - Rede de Área Local
SPI	*Serial Peripheral Interface* - Interface Periférica Serial
I^2C	*Inter-Integrated Circuit* - Circuito Inter-Integrado
DAPF	Dias Após a Plena Floração
CE	Compensação por Equações
CTU	Compensação por Temperatura e Umidade como Entrada
HP	Hiperparâmetros
ANOVA	Análise de Variância

SUMÁRIO

1 INTRODUÇÃO

O pêssego (Prunus persica (L.) Batsch) é uma fruta comercialmente importante na região sul do Brasil. Em 2021, o valor total comercializado ultrapassou os 500 milhões de reais em todo o país. Um valor obtido com o trabalho de mais de 4500 produtores. Naquele ano, os três estados do Sul foram responsáveis por 77% da produção nacional. O estado do Paraná foi o quinto maior produtor nacional com 9.985 toneladas (5%) (IBGE, 2023).

Ao longo do ano o pessegueiro necessita de calor e de frio para o desenvolvimento adequado e produção satisfatória. No período reprodutivo o pessegueiro apresenta brotações floríferas que se desenvolvem passando por diferentes estágios, como a abertura das pétalas e o inhaço do ovário, até a formação do fruto. Durante a primavera e o verão pessegueiro é caracterizado por ter um crescimento vegetativo rápido, desenvolvendo suas estruturas. Durante o outono e o inverno o crescimento é reduzido, sendo o período no qual planta entra em dormência, paralizando seu crescimento vegetativo, com as gemas do pessegueiro apresentando desenvolvimento lento e não perceptível. No final do inverno, com o aumento das temperaturas, o crescimento rápido é retomado, iniciando a floração e posteriormente a frutificação (BARBOSA, 1989).

No contexto de um pomar, o raleio é uma técnica utilizada para melhorar a distribuição dos recursos em plantas frutíferas, reduzindo a quantidade de brotações ou frutos. Os principais objetivos do raleio são melhorar o tamanho e a qualidade das frutas, e evitar a alternância de produção entre safras, podendo ser executado de forma manual, mecânica, ou química em diferentes momentos e de forma complementar (FACHINELLO; NACHTIGAL; KERSTEN, 2009; MAYER; FRANZON; RASEIRA, 2019). O raleio nas flores ocorre, geralmente, na plena floração (50-70% de flores abertas), mas também pode ser aplicado nos frutos, quando a maior parte está no início da fase intermediária de crescimento, ainda verdes e com aproximadamente 2 cm de diâmetro (MAYER; FRANZON; RASEIRA, 2019). As pesquisas indicam que a época de plena floração dos pessegueiros é o momento adequado para a realização do raleio, aumentando a efetividade comparativamente ao raleio em frutas (BARRETO *et al.*, 2019a, 2019b; OLIVEIRA *et al.*, 2017). Porém, na prática, o raleio nas frutas pode ser o mais adequado dependendo de fatores como a possibilidade de ocorrência de geadas ou baixas taxas de floração, polinização e fertilização (MAYER; FRANZON; RASEIRA, 2019).

O raleio é uma atividade que traz benefícios, porém é onerosa e em alguns casos impraticável em toda extensão de um pomar, sendo um dos principais custos de mão de obra durante a produção (MAYER; FRANZON; RASEIRA, 2019). Seu custo e sua efetividade são fatores que afetam o ganho financeiro da aplicação da técnica. O momento adequado para execução e a

intensidade são questões complexas que envolvem conhecimento específico sobre as condições de um pomar. Portanto, informações que possam auxíliar o planejamento e a tomada de decisão sobre o raleio são relevantes e podem melhorar o custo benefício da aplicação da técnica. No entanto, em decorrência das características de cada planta e das condições do ambiente, o desenvolvimento das brotações ocorre em diferentes velocidades ao longo do pomar com algumas brotações mais avançadas do que as outras.

Uma abordagem para o problema do raleio considerando as brotações é analisar odor do pessegueiro e as possíveis variações devido o desenvolvimento destas. Nesse sentido, um sistema de nariz eletrônico pode ser usado. O nariz eletrônico é um sistema olfativo artificial utilizado para analisar a expressão geral dos compostos orgânicos voláteis (COVs) em amostras, isto é, um equipamento que utiliza uma matriz de sensores de gás e reconhecimento padrões para relacionar amostras de odor com o objetivo da análise. O equipamento pode ser utilizado em uma variedade de aplicações na agricultura. Em relação às técnicas convencionais, utilizadas para analisar a qualidade e a composição de produtos e substâncias, o nariz eletrônico é um dispositivo eletrônico, não destrutivo em relação à amostra, de baixo custo, rápido em termos do tempo de análise, de expressão geral do odor, e orientado à aplicação.

Para abordar o problema, um prótotipo foi desenvolvido e coletas foram executadas durante o período reprodutivo de um pessegueiro. Para definir o estágio geral pessegueiro a informação sobre o estágio das brotações foi considerada crucial. A hipótese da pesquisa foi de que o nariz eletrônico seria capaz de identificar o estágio geral das brotações do pessegueiro (estágio em que maior parcela das brotações se encontram). Então, para verificar a hipótese de pesquisa, uma análise de classificação foi efetuada. Nela, a variável de saída determinada foi o estágio geral do pessegueiro assumindo quatro valores qualitativos (QP - Queda das pétalas, FFI - Formação de fruto inicial, FFA - Formação de fruto avançada ou F - Fruto formado), correspondentes ao estágio de maior porcentagem em cada amostragem.

1.1 ORGANIZAÇÃO DA DISSERTAÇÃO

O Capítulo 1 apresenta a introdução à pesquisa descrevendo o problema, a hipótese, as delimitações, e os objetivos da pesquisa. No Capítulo 2, a revisão de literatura é apresentada, contendo na primeira parte as informações sobre os pessegueiros, o raleio e COVs emitidos durante a floração. Na segunda parte, são apresentados os conceitos e as definições sobre o nariz eletrônico, o pré-processamento e processamento dos dados. No Capítulo 3, o material e os métodos utilizados são apresentados. Os resultados das análises executadas e as discussões são expostos no Capítulo 4. As conclusões sobre o trabalho estão no Capítulo 5.

1.2 DELIMITAÇÕES DA PESQUISA

Esta pesquisa se enquadra no contexto em que analisa-se os estágios finais da floração e o estágio inicial de desenvolvimento dos frutos do pessegueiro, compreendendo a floração e a frutificação. O protótipo de nariz eletrônico foi utilizado no local para fazer a coleta dos dados. Esse tipo de aplicação em ambientes complexos, isto é, com pouco ou sem controle, difere-se de ambientes controlados e apresenta desafios devido a interação entre o local e o dispositivo.

A principal questão acerca do ambiente complexo é se os métodos de reconhecimento de padrões são capazes de suprir as múltiplas causas de variações nos sensores devido às interações com o meio, e relacionar as variações adequadas com o que se deseja analisar. Nesta pesquisa o foco esteve nos estágio de desenvolvimento das brotações florais do pessegueiro, que representam o estágio geral da planta durante a fase reprodutiva.

Esta pesquisa é baseada no trabalho de Voss (2019) em que um protótipo foi utilizado em um pomar para analisar o ciclo de crescimento de pessêgos, sendo o trabalho de maior similaridade com a atual pesquisa. Em termos da literatura revisada, a maior parte das pesquisas envolvendo nariz eletrônico e a análise de frutas são realizadas em ambientes controlados visando questões de qualidade alimentícia. Aparentemente, há carência de pesquisa envolvendo nariz eletrônico em ambientes não controlados, como o ambiente agrícola. Ainda, as pesquisa em ambiente controlado geralmente utilizam conjuntamente a análise por GC-MS para identificar os COVs emitidos pelas amostras. Essa identificação não está no escopo dessa pesquisa.

Portanto, este trabalho pode ser descrito como a construção e a utilização de um protótipo de nariz eletrônico para coletar amostras em um pomar, visando identificar os estágios de desenvolvimento das brotações ao longo do experimento. A hipótese da pesquisa sustenta a possibilidade do nariz eletrônico ser capaz de indentificar os diferentes estágios de desenvolvimento, isto é, se o sistema consegue identificar o estágio geral do pessegueiro.

1.3 OBJETIVO GERAL

O objetivo geral deste trabalho é analisar a capacidade de um sistema de nariz eletrônico em identificar os estágios de desenvolvimento das brotações floríferas em pessegueiros, com o propósito de auxílio nas decisões de atividades de manejo como o raleio.

1.4 OBJETIVOS ESPECÍFICOS

Como objetivos específicos deste trabalho têm-se:

- Desenvolver um protótipo de nariz eletrônico para ser utilizado nas coletas em campo;

- Analisar a efetividade do sistema enquanto reconhecimento qualitativo dos estágios ao longo do desenvolvimento das brotações de um pessegueiro;
- Avaliar a efetividade do sistema em auxiliar a definição do momento de raleio.

O intuito de construir o protótipo de nariz eletrônico foi ter um dispositivo que poderia ser utilizado na coleta dos dados em campo e ser avaliado em termos de sua aptidão para as análises, isto é, sua validação dado o conjunto de sensores e a configuração do equipamento. Após a coleta das amostras, uma vez que a base de dados foi formada, o objetivo tornou-se avaliar a base por meio de métodos de reconhecimento de padrões. Dessa forma, para identificar os diferentes estágios de desenvolvimento a análise de classificação foi adequada para o problema, já que utiliza, entre outros, o aprendizado supervisionado e trabalha com dados de múltiplas variáveis. Assim, a avaliação de desempenho do sistema proposto pode ser efetuada. Por conseguinte, a interpretação dos resultados obtidos e as consequências para a atividade de raleio.

2 REVISÃO DE LITERATURA

2.1 A CULTURA DO PÊSSEGO

O pêssego (Prunus persica (L.) Batsch) é uma fruta pertencente ao grupo das drupas ou frutas de caroço, enquanto taxonomia é membro da família Rosaceae, subfamília Amygdaloidae e gênero Prunus L. (BASSI *et al.*, 2016; MAYER; FRANZON; RASEIRA, 2019). A fruta possui estruturalmente o epicarpo, sendo o tecido fino que envolve a fruta, o mesocarpo, que é a polpa, e o endocarpo, o caroço, localizando-se no seu interior a semente, sendo esta dicotiledônea. O pessegueiro é formado pelo sistema radicular, tronco, ramos primários, secundários, folhas, flores e frutos (MAYER; FRANZON; RASEIRA, 2019). Pode viver de 20 a 30 anos, porém em pomares comerciais tem vida de 12 a 15 anos (BASSI *et al.*, 2016). As cultivares de pessegueiro geralmente são autoférteis, não necessitando de polinização cruzada e possuindo flores hermafroditas (BASSI *et al.*, 2016; MAYER; FRANZON; RASEIRA, 2019). A planta durante seu ciclo produtivo anual necessita de frio e de calor para obter uma produção mais adequada, sendo que as necessidades de temperatura, o momento de floração, e a maturação do fruto variam entre os genótipos (BASSI *et al.*, 2016).

Em relação à produção comercial, mundialmente o Brasil localiza-se em 11º com 199.010 toneladas produzidas de pêssego e nectarina em 2021 (FAO, 2023). Nacionalmente, o Paraná foi o quinto maior produtor com 9.985 toneladas no ano de 2021, atrás do Rio Grande do Sul, São Paulo, Santa Catarina e Minas Gerais respectivamente (IBGE, 2023). De acordo com a mesma pesquisa, a cidade de Ponta Grossa foi a oitava maior produtora de pêssego no estado com 340 toneladas.

2.1.1 Dormência

Ao longo do ciclo produtivo anual o desenvolvimento do pessegueiro pode ser definido por duas fases: de crescimento e de formação dos orgãos reprodutivos. Na fase de crescimento, durante as estações quentes, ocorre a formação das estruturas complexas, isto é, o desenvolvimento inicial e final das gemas floríferas e vegetativas, brotação de folhas, ramos e a floração, e o desenvolvimento dos frutos e folhas. Nas estações frias ocorre a fase de desenvolvimento lento dessas estruturas (BARBOSA, 1989).

Durante as estações frias o pessegueiro apresenta um mecânismo fisiológico adaptativo, chamado de dormência, em que para lidar com as condições ambientais adversas, no caso as temperaturas baixas, a planta paraliza seu crescimento vegetativo (BARBOSA, 1989). A dormência pode ser classificada em três tipos: Eco, quando é causada por fatores ambientais extremos; Para, quando o crescimento de determinada estrutura inibe o crescimento de outra;

e Endodormência, quando o crescimento é paralizado por estímulos ambientais, como temperaturas baixas ou redução do período de exposição à luz (fotoperíodo), possivelmente através de reações bioquímicas que ocorrem na planta dado os estímulos ambientais (ANZANELLO; LAMPUGNANI, 2020; BARBIERI, 2018; BARBOSA, 1989). Para pessegueiros em clima temperado a endodormência é a mais presente, ocorrendo no outono e no inverno (BARBOSA, 1989). Para regiões de clima subtropical de inverno ameno, as cultivares podem não entrar em dormência efetiva, dando continuidade ao desenvolvimento lento e podendo ter floração precoce com risco de perdas por geadas (MAYER; FRANZON; RASEIRA, 2019).

2.1.2 Horas de Frio

Relacionadas ao período de dormência estão as horas de frio, sendo definidas como as horas acumuladas em que a temperatura foi inferior ou igual a 7,2 °C. A quantidade acumulada de horas de frio é identificada como fator para a quebra da endodormência das gemas, necessitando de determinada quantidade horas para que a endodormência seja superada e haja um respectivo desenvolvimento adequado, com brotações e florações da gemas com qualidade e em quantidade suficientes. Posteriormente, há a necessidade de horas de calor para o desenvolvimento final das gemas com temperatura entre 15 e 20 °C (CARAMORI *et al.*, 2008; MAYER; FRANZON; RASEIRA, 2019). Conforme relatado, no inverno, durante a necessidade de horas de frio, temperaturas acima de 21 °C podem gerar efeito reverso nas horas de frio, induzindo a planta ao florescimento antecipado e irregular (ABÊ, 2020; BARBIERI, 2018).

Caso as horas de frio não sejam supridas durante a endodormência no outono e no inverno, as plantas apresentam queda na qualidade da próxima etapa do desenvolvimento, como brotação e floração insuficientes ou desuniformes, afetando a distribuição de folhas e ramos na planta, a frutificação adequada e a, respectiva, qualidade dos frutos (ABÊ, 2020; ANZANELLO; LAMPUGNANI, 2020; BARBIERI, 2018; CARAMORI *et al.*, 2008). As cultivares de pêssego apresentam diferentes necessidades podendo variar de 100 a 500 horas de frio (CARAMORI *et al.*, 2008).

Na pesquisa de Caramori *et al.* (2008) foi considerado o estado do Paraná e as necessidades de diferentes cultivares. Já na pesquisa de Anzanello e Lampugnani (2020), efetuada no Rio Grande do Sul, avaliou-se os ramos coletados em campo e, posteriormente, desenvolvidos em ambiente controlado, a fim de avaliar a dormência das cultivares. De acordo com Caramori *et al.* (2008) no zoneamento agroclimático de 2008 para pêssegos e nectarinas no estado do Paraná, a cidade de Ponta grossa acumula de 150 a 250 horas de frio durante o outono e o inverno. Ainda, de acordo com as estimativas da modelagem considerando as geadas (menor ou igual a 1 °C) no estado, a data da última geada de primavera na cidade de Ponta Grossa

compreende o período de 10 a 20 de agosto com 90% de probabilidade, porém com o risco de geada ocorrer inferior a 20%.

Na pesquisa de Anzanello e Lampugnani (2020) foram consideradas como parâmetros de análise a porcentagem de gemas brotadas, a precocidade (dias até a brotação da primeira gema), e a uniformidade (dias entre a primeira e a última gema brotada). O resultado indicou que para os dois cultivares analisados, ao atingir-se as horas de frio necessárias para cada um, aumenta-se a porcentagem de brotação, e diminui-se os a quantidade de dias da precocidade e da uniformidade. Portanto, as horas de frio são um fator determinante no desenvolvimento da floração. Isto é, determinam o desenvolvimento adequado da brotação e floração, em termos de quantidade, uniformidade e precocidade, influenciando sua duração e suas etapas.

2.1.3 Floração

Três etapas ocorrem para que haja floração em pessegueiros: a indução floral ou diferenciação, a iniciação floral, e o desenvolvimento da flor. Na indução floral, que é iniciada no verão após intenso crescimento vegetativo, mudanças metabólicas geram modificações no meristema de algumas gemas, transformando estas de vegetativas para floríferas. Na iniciação floral, o meristema sofre modificações morfológicas, aumentando de tamanho, formando o receptáculo da flor, e as estruturas internas composta das sépalas, pétalas, estames e o pistilo. O desenvolvimento das estruturas da flor possui três períodos principais: o primeiro com crescimento elevado durante a iniciação floral (final do verão), o segundo com redução do crescimento desenvolvendo-se lentamente os orgãos reprodutivos (outono e inverno), e o terceiro com o reinício do crescimento do pessegueiro, iniciando a divisão celular para a produção dos gametas masculino e feminino nas estruturas já formadas das gemas (final do inverno). No desenvolvimento floral, a última etapa, o crescimento dos orgãos florais é retomado após a endodormência, ocorre a maturação dos gametas, e a abertura da flor (BARBOSA, 1989). Os estágios fenológicos da flor de pêssego são apresentados no Quadro 1.

Em cada nó da estrutura dos ramos do pessegueiro, geralmente, são formados duas gemas laterais frutíferas e uma central vegetativa (BASSI *et al.*, 2016). A floração, normalmente, inicia-se antes da brotação das folhas, uma vez que a necessidade de frio das gemas floríferas é menor do que das gemas vegetativas (ABÊ, 2020). Após a abertura das flores inicia-se a florada, posteriormente a plena floração, e por último a queda das pétalas após a fecundação. A plena floração é o momento em que determinada quantidade mínima de flores estão abertas, podendo variar a definição de 50% a 70% das flores abertas. A duração da florada pode ir de 7 até próximo de 30 dias (FACHINELLO; NACHTIGAL; KERSTEN, 2009; MAYER; FRANZON; RASEIRA, 2019). O principal fator que influencia o momento da florada é a temperatura (ABÊ,

Quadro 1: Estágios fenológicos da flor de pêssego.

Classificação [a]	**Estágio**
A	Gema de inverno: dormente
B	Gema inchada: aparecimento das sépalas
C	Botão verde: aparecimento das pétalas
D	Botão rosa: aumento considerável de tamanho
E	Abertura parcial da flor
F	Flor aberta
G	Deiscência das pétalas
H	Deiscência das sépalas
I	Crescimento da fruta

Fonte: (BARBOSA, 1989; MOUNZER *et al.*, 2008).
Nota: *a* - classificação de Baggiolini e Cheller.

2020; BARBIERI, 2018). Assim como, as horas de frio que influenciam diretamente a duração da florada (ANZANELLO; LAMPUGNANI, 2020; MAYER; FRANZON; RASEIRA, 2019).

2.2 A PRÁTICA DE RALEIO

O raleio é uma operação de manejo em que retira-se parte da produção natural frutífera da planta para melhorar a qualidade dos frutos restantes, podendo ser aplicada também às flores e às gemas floríferas. Dentre os principais objetivos descritos por Fachinello, Nachtigal e Kersten (2009) e Mayer, Franzon e Raseira (2019) estão, aumentar o tamanho e melhorar a qualidade das frutas, evitar a alternância de produção, melhorar a distribuição de folhas por fruta, reduzir a quantidade de frutas defeituosas, e evitar a quebra de ramos por excesso de peso. Para o raleio em frutas três critérios são levados em consideração, o espaçamento entre frutas em um ramo, a superfície foliar, e a seção transversal do tronco. Já para o raleio aplicado nas flores, geralmente, busca-se manter 50% das flores na planta, e em caso de baixa floração 80% (MAYER; FRANZON; RASEIRA, 2019).

A execução pode ser realizada de forma manual, mecânica, ou química. O raleio manual é executado com equipamentos como tesouras, sendo preciso, porém oneroso e demorado. No raleio mecânico, geralmente, utiliza-se equipamentos que possam facilitar e aumentar a velocidade da operação como derriçadeiras ou aparadores com fios rotativos. O raleio químico é executado através da aplicação de substâncias que causam a queda de flores ou frutas, tendo menor custo e menor tempo de execução que os demais, porém podendo causar danos as outras estruturas da planta e apresentando resultados inconsistentes em pesquisas. Nem o raleio químico nem o mecânico eliminam totalmente a necessidade da realização do raleio manual, que pode ainda ser necessário como completo de ajuste para ambos (FACHINELLO; NACHTIGAL; KERSTEN, 2009; MAYER; FRANZON; RASEIRA, 2019).

2.2.1 Momento e Efetividade do Raleio

A época mais adequada para a prática do raleio depende da efetividade e dos riscos associados, existindo a relação de que quanto antes for realizado melhores os resultados obtidos, considerando as gemas florais, as flores, e frutos. Porém, quanto antes for executado, dado as três opções anteriores, aumenta-se os riscos com possíveis perdas nas próximas etapas de desenvolvimento, além da possibilidade de ser inviável econonomicamente dado as dimensões dos pomares (FACHINELLO; NACHTIGAL; KERSTEN, 2009).

Segundo Fachinello, Nachtigal e Kersten (2009) quando o raleio é executado dentro do período de divisão celular da fruta, isto é, no primeiro estágio de desenvolvimento, ocorre formação de maior quantidade de células, o que resulta no maior tamanho, se comparado com o raleio executado durante a fase de crescimento celular após a fase de divisão. Ainda, o raleio executado nas flores pode possibilitar maior distribuição de reservas para a fase de divisão celular, o que reduz a competição por carboidratos durante o desenvolvimento das frutas (BARRETO *et al.*, 2019a; OLIVEIRA *et al.*, 2017; BARRETO *et al.*, 2019b).

Filho, Minami e Kluge (2000) analisaram a intensidade do raleio em frutos com 35 dias após plena floração, testando quatro tratamentos variando de 0% até 65% de intensidade de raleio. Os melhores resultados em termos de tamanho, peso e classificação comercial, foram obtidos pelos tratamentos com 56% e 65% de intensidade, em relação aos tratamentos com 0% e 48%, sendo que o maior valor estimado de receita bruta foi obtido pelo tratamento de 65%.

Nas pesquisas de Barreto *et al.* (2019a) e Barreto *et al.* (2019b) foram comparados o raleio manual e o mecânico com diferentes ferramentas, em época de plena floração e nos frutos com 40 dias após plena floração. A intensidade do raleio nas flores variou de 33% a cerca de 62% (BARRETO *et al.*, 2019a), e de 46% a cerca de 53% (BARRETO *et al.*, 2019b). Não houve diferença significativa entre o raleio em plena floração e o aplicado em frutas, em termos de produtividade por planta e massa média das frutas. Porém, o raleio mecânico em plena floração utilizando o equipamento derriçadeira resultou em maior quantidade de frutas com os maiores diâmetros, em relação aos outros dois tratamentos (BARRETO *et al.*, 2019a). Na pesquisa de Barreto *et al.* (2019b) o melhor resultado também foi do tratamento que utilizou derriçadeira em plena floração, em comparativo com outros três tratamentos. Nesse caso, com cerca de 53% das flores raleadas, obteve-se os melhores resultados de produção por planta, massa média por fruta, e tamanho da fruta em relaçao aos demais tratamentos.

Oliveira *et al.* (2017) analisaram tratamentos com raleio nas gemas floríferas em botão rosa, em plena floração, e nos frutos em dois estágios de desenvolvimento. Os resultados não apresentaram diferença significativa para os tratamentos em termos de produção por planta,

isto é, a época não afetou a produção. Porém, em relação à quantidade de frutos de maiores diâmetros e ao peso da fruta, o raleio executado em plena floração apresentou melhor resultado. Além disso, outro aspecto que pode ser considerado, sobre o momento e a efetividade do raleio, é a execução de forma complementar, em que diferentes épocas ou modos de execução são combinados, podendo melhorar os resultados da execução em uma única época.

2.2.2 Custo do Raleio

O raleio tende a ser um dos principais custos de mão de obra durante a produção. Tal fator, juntamente com o tempo de execução e sua efetividade podem reduzir o ganho financeiro da aplicação da técnica. Estima-se que o custo do raleio manual possa chegar a um terço do custo total de mão de obra para a produção, e com tempo de execução de até 150 h/ha, podendo levar cerca de 30 minutos por planta (MAYER; FRANZON; RASEIRA, 2019). Porém, o raleio mecânico pode reduzir o custo da operação e até aumentar a efetividade em relação ao manual.

2.3 COVs DURANTE A FLORAÇÃO E A FRUTIFICAÇÃO

As cultivares de pessegueiro, geralmente, possuem flores rosáceas e hermafroditas. A estrutura reprodutiva da flor é formada pelo pistilo (orgão feminino: composto pelo estigma, estilete e ovário), e pelos estames (orgãos masculinos: compostos pela antera e o filete). Nos estames, cada antera possui em torno de 1000 a 2000 grãos de pólen. As flores podem ser autofecundadas, e serem polinizadas por insetos e pelo vento (BARBOSA, 1989).

No restante da estrutura da flor o conjunto das pétalas formam a corola, e o conjunto das sépalas formam o cálice. Ambos, corola e cálice, constituem o perianto. A flor aberta apresenta cinco pétalas, cerca de 20 a 30 estames, e o pistilo ao centro. A coloração das pétalas pode ser rosa, vermelha, ou branca, e o comprimento de até 45 mm de diâmetro (BASSI *et al.*, 2016). A Figura 1 mostra uma flor de pessegueiro.

No tecido vegetal das flores há compostos químicos endógenos gerados pelos processos naturais, esses compostos são emitidos pelas estruturas quando estão em concentrações elevadas, além do limite da capacidade celular metabólica (HAO *et al.*, 2014). O conjunto de compostos emitidos pelas flores caracteriza o odor floral, e pode ser sentido pelos animais através do sistema olfativo, como por mamíferos e insetos. Questões como a taxa de emissão ao longo do tempo, a relação com fatores ambientais curtos ou longos, os mecânismos metabólicos de produção, além da especificidade relativa à cultivar, são questões para potenciais investigações sobre as flores das prunóideas.

De acordo com Baraldi *et al.* (1999) os grupos de compostos orgânicos voláteis (COVs)

Figura 1: Flor de pessegueiro da cultivar 'Red Haven'.

Fonte: (EL-SAYED *et al.*, 2018).

que foram identificados da emissão de um pessegueiro em plena floração foram os alcanos, alcenos, arenos, monoterpenos, compostos sulfúricos, aldeídos, e cetonas. Os COVs encontrados em maior concentração considerando todo o pessegueiro e em plena floração, isto é, com as flores desenvolvidas abertas e folhas não totalmente desenvolvidas, e mais de 4 horas de emissão de COVs pelo pessegueiro, foram Benzaldeído, 2-Etil Hexanol, e Nonanal. No total 34 compostos foram encontrados.

Ainda, considerando somente as flores, estando estas desenvolvidas completamente, a cultivar *Redheaven* e 24 horas de emissão, os COVs em maior concentração foram 3,5 - Dimetoxitolueno, com cerca de 95,5% da quantidade relativa na amostra, Benzaldeído com 2,8%, e β-Farneceno com 0,9%. No total 7 compostos foram encontrados (EL-SAYED *et al.*, 2018). E, para as flores da cultivar Duplex desenvolvidas completamente, e com 1 hora de emissão de COVs, Benzaldeído representou cerca de 40% da quantidade relativa na amostra (HAO *et al.*, 2014).

Em relação aos compostos endógenos, isto é, os compostos retidos nos tecidos vegetais, Benzaldeído foi o composto mais abundante apresentando concentração acima de 94% para flores do cultivar Duplex (HAO *et al.*, 2014). Assim como para as folhas do cultivar Monroe já desenvolvidas, com mais de 100 dias após plena floração, apresentando concentração acima de 95% da amostra (HORVAT; CHAPMAN, 1990).

O Benzaldeído parece ser um compostos comum emitido pelas flores e folhas de diferentes cultivares de pêssego durante a floração. Isto também é observado em outras espécies do gênero Prunus (EL-SAYED *et al.*, 2018; HAO *et al.*, 2014; RADULOVIĆ *et al.*, 2009). Como comentado por El-Sayed *et al.* (2018), as diferenças encontradas nos COVs emitidos pelas flores

podem ser relativas ao método utilizado para coleta das amostras ou ao cultivar analisado. Além disso, fatores ambientais podem ser investigados como possíveis diferenciadores na emissão de COVs pelas plantas, assim como o tempo de emissão antes da amostragem, o que também pode ser um fator relevante.

Para a frutificação, a pesquisa de Li *et al.* (2021) com frutos e folhas da cultivar *White peach* mostrou que no primeiro estágio considerado (frutos com 2 cm de diâmetro) os COVs de maior concentração foram Ácido acético, Metanol, Acetona e Acetaldeído. E que as folhas emitiram maior concentração de COVs que os frutos durante o desenvolvimento do pessegueiro até a colheita. Além disso, no trabalho de Brandi *et al.* (2011) foi analisado os estágios de desenvolvimento dos frutos das cultivares *Redhaven* e *Redheaven Bianca*. Os resultados para o primeiro estágio considerado (pêssegos verdes em torno de 4 cm de diâmetro e aproximadamente 35 dias após plena floração) mostraram que o composto em maior concentração foi o Benzaldeído para ambos genótipos. Porém, os estágios iniciais (1 e 2) apresentaram um conteúdo baixo de COVs em relação aos demais estágios.

Ademais, as pesquisas indicam que os COVs emitidos durante o desenvolvimento do pessegueiro dependem das características da planta como o genótipo, o que torna difícil estabelecer uma relação direta e precisa utilizando os trabalhos disponíveis na literatura. No Quadro 2 os principais COVs emitidos durante a floração e o estágio inicial de frutificação são apresentados.

Quadro 2: COVs durante a floração e o estágio inicial de frutificação.

COV	**Etapa**
Benzaldeído	Floração
2-Etil Hexanol	Floração
Nonanal	Floração
3,5-Dimetoxitolueno	Floração
Ácido acético	Frutificação
Metanol	Frutificação
Acetona	Frutificação
Acetaldeído	Frutificação
Benzaldeído	Frutificação

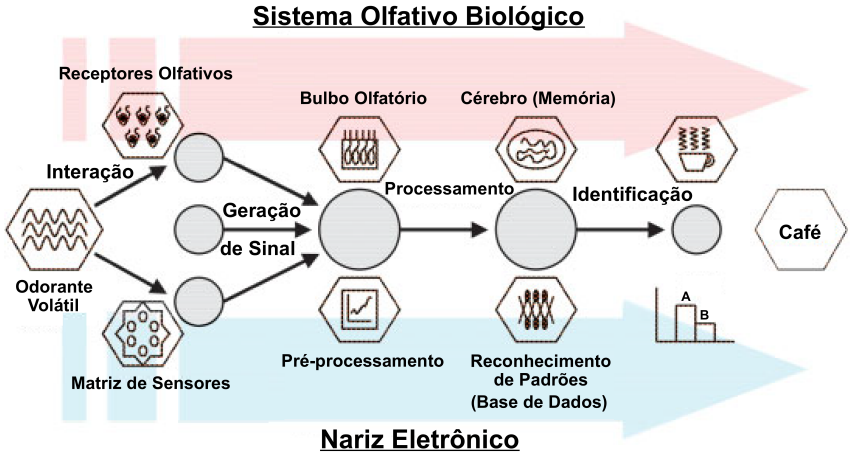
Fonte: (BARALDI *et al.*, 1999; BRANDI *et al.*, 2011; EL-SAYED *et al.*, 2018; HAO *et al.*, 2014; LI *et al.*, 2021).

2.4 NARIZ ELETRÔNICO

O nariz eletrônico é um sistema olfativo artificial que simula o sistema olfativo dos mamíferos, possuindo como *hardware* de entrada a matriz de sensores e o circuito de condicionamento de sinal. O componente de *software*, por sua vez, realiza o pré-processamento e o processamento dos dados através de métodos de análise estatística, e produz, então, como saída

uma predição que relaciona estatisticamente o aspecto geral sensível do odor amostrado com as definições previamente estabelecidas por outras amostras durante o processamento. Na Figura 2 é apresentada uma analogia entre os sistemas olfativos biológico e artificial.

Figura 2: Diagrama da analogia entre o sistema olfativo biológico e artificial.



Fonte: (PEARCE *et al.*, 2003).

Os prováveis primeiros desenvolvimentos sobre dispositivos para análise de odores são da década de 1960, estabelecendo as bases do sensoriamento, para que na década de 1980 os trabalhos de Persaud e Dodd (1982), Ikegami e Kaneyasu (1985), e Kaneyasu *et al.* (1987) desenvolvessem o sistema de nariz eletrônico como é conhecido atualmente, sendo composto por uma matriz de sensores, processamento dos dados, e a predição da amostra como saída, funcionando por meio do reconhecimento de padrões (VOSS, 2019).

As possíveis aplicações do nariz eletrônico vão desde análises de qualidade de produtos alimentícios e o monitoramento ambiental, até aplicações em medicina e em sistemas de segurança (KARAKAYA; ULUCAN; TURKAN, 2020). Em relação à agricultura o nariz eletrônico vem sendo utilizado para aplicações como análises do ambiente agrícola, por exemplo, em pesquisas sobre o estresse hídrico em plantações (FABBRI *et al.*, 2020), doença em colmeias (SZCZUREK *et al.*, 2020), monitoramento do solo (DORJI; POBKRUT; KERDCHAROEN, 2017), e a emissão de poluentes por máquinas agrícolas (VALENTE *et al.*, 2018). Assim como, em análises de plantas e vegetais, por exemplo, em verificação de contaminantes no milho (LEGGIERI *et al.*, 2021) e no arroz (GU *et al.*, 2020), e na classificação de plantas medicionais (WANG *et al.*, 2020). Além de aplicações em análises de frutas, como, na avaliação de qualidade de variedades de maçãs (ZHU *et al.*, 2020), de cultivares de banana (DOU *et al.*, 2020), e de azeitonas após a colheita (GILA *et al.*, 2020). Em análises de produtos agrícolas, por exemplo, na análise de bebidas alcoólicas (SILVELLO; ALCARDE, 2020), na classificação de açucares (WEERAWATANAKORN *et al.*, 2021), e na verificação de adulteração em carnes (HAN *et al.*, 2020).

Convencionalmente os métodos utilizados para analisar a qualidade e a composição

de produtos e substâncias são baseados em análises físico-químicas, químicas, e sensoriais por humanos. A espectrometria de massa por cromatografia gasosa (GC-MS) é um método utilizado em análises físico-químicas, que identifica as moléculas que compõem uma amostra e suas concentrações. Os COVs geralmente são identificados por este método. Diferentes análises químicas podem ser utilizadas para a definição dos índices de qualidade das amostras. Em relação aos produtos alimentícios os índices normalmente analisados são a acidez, o conteúdo total de sólidos solúveis (SSC), a rancidez, e a firmeza. A análise sensorial por humanos é utilizada para avaliar uma amostra, verificando-se fatores de qualidade como o sabor, a textura, o odor, ou a coloração.

Em relação a estes métodos, o nariz eletrônico é um dispositivo eletrônico (o que pode reduzir seu tamanho sem perder qualidade de análise, e com possibilidade de ser embarcado), não destrutivo (não há necessidade de destruição da amostra), de baixo custo (seus componentes podem ser de baixo custo sem comprometer a análise), rápido (uma vez adaptado, isto é, calibrado e treinado, a análise por ser feita em tempo real), de expressão geral do odor (geralmente há a identificação do aspecto geral, e não dos elementos que integram o odor), e orientado à aplicação (pode ser desenvolvido para análises específicas, tendo sua eficiência definida para a aplicação). Essas características garantem ao nariz eletrônico algumas vantagens em relação aos métodos convencionais de análise de qualidade ou característica, pois, geralmente, o GC-MS é caro e lento, a análise química é destrutiva, e a análise sensorial por humanos pode sofrer influência da subjetividade e outros fatores. Embora, normalmente, nas pesquisa os métodos convencionais sejam utilizados como validação dos resultados e de forma complementar com o nariz eletrônico.

2.4.1 Aplicações

São vastas as aplicações de nariz eletrônico na agricultura, que incluem desde análises de qualidade de produtos agrícolas até a complementação à análises de caracterização destes produtos, além de aplicações no ambiente agrícola em si. Em relação aos elementos analisados pode-se diferenciar os grupos em produtos agrícolas, plantas e vegetais, e ambiente agrícola. Na literatura há revisões que abordam tais usos. Nos últimos cinco anos observa-se que as pesquisas utilizaram o sistema na tentativa de facilitar as análises geralmente empregadas na agricultura, com o avanço das tecnologias que compõem os sistemas de nariz eletrônico a análise simplificada torna-se mais plausível e confiável. Embora questões como variações dos sensores devido às condições ambientais e variações ao longo do tempo impliquem na redução da estabilidade, e, consequentemente, na possível inviabilidade de aplicações em ambientes não controlados ou de repetidas análises a longo prazo, as etapas de pré-processamento e

processamento tendem a reduzir os impactos desses fatores no resultado final da aplicação, possibilitando novas aplicações na medida que novos métodos e técnicas são desenvolvidas.

As pesquisas recentes em relação aos produtos agrícolas incluem análises de classificação em açucares (WEERAWATANAKORN *et al.*, 2021), na indentificação de qualidade em vinhos (GAMBOA *et al.*, 2019), classificação de azeite de oliva (OATES *et al.*, 2018), na predição de aditivos alimentares em sucos de fruta (QIU; WANG, 2017), e na análise da quantidade de bactérias na carne (TIMSORN *et al.*, 2016).

Para plantas, frutas, e vegetais as análises abrangem aplicações como a identificação de micotoxinas no milho (LEGGIERI *et al.*, 2021), avaliações de qualidade de azeitonas após a colheita (GILA *et al.*, 2020), de contaminação por fungo em grãos de arroz (GU; WANG; WANG, 2019), e de contaminação por fungo em morangos (LIU *et al.*, 2019), monitoramento do aroma de folhas de menta durante a secagem (KIANI; MINAEI; GHASEMI-VARNAMKHASTI, 2018), a predição da qualidade de tomates pós colheita (FENG *et al.*, 2018), classificação de origem geográfica de acelga chinesa (LEE *et al.*, 2017), na identificação de frutas (UÇAR; ÖZALP, 2017), detecção de qualidade de noz pecan durante o armazenamento (JIANG; WANG, 2016), e na predição de índices de qualidade de bananas (SANAEIFAR *et al.*, 2016).

Em relação ao ambiente agrícola exemplos de aplicações compreendem a classificação do nível de infestação por ácaro em colônias de abelhas (SZCZUREK *et al.*, 2020), em análises do extresse hídrico em plantações de milho e tomate (FABBRI *et al.*, 2020), e do conteúdo de ácidos húmicos e fúlvicos no solo (LAVANYA *et al.*, 2017), na detecção de poluentes emitidos por máquinas agrícolas (VALENTE *et al.*, 2018), e o monitoramento da fertilidade do solo (DORJI; POBKRUT; KERDCHAROEN, 2017).

Pesquisas envolvendo o nariz eletrônico na análise de frutas são frequentes e podem ser encontradas desde a década de 1990, como (HINES; LLOBET; GARDNER, 1999) que analisaram os estágios de maturação de maçãs. Além disso também há revisões na literatura sobre esse tema. As frutas emitem COVs através de suas estruturas e diferentes fatores podem influenciar mudanças no padrão de emissão, como o estágio de maturação, a mudança de qualidade, diferenças entre cultivares, diferenças de origem geográfica, diferenças no tratamento pós colheita, entre outros. As mudanças que ocorrem nas frutas podem ser externas, por exemplo, o aroma, a cor, sabor, ou o tamanho, ou internas como a acidez, o teor de sólidos solúveis, ou a concentração de vitamina C. Os principais grupos de COVs presentes nas frutas são ésteres, álcoois, aldeídos, cetonas, lactonas, terpenóides, apocarotenóides, ácidos graxos e os fenólicos (BAIETTO; WILSON, 2015).

As análises nas frutas ou na fruticultura podem ser definidas como de maturação

ou de desenvolvimento do fruto. O objetivo é avaliar o estágio de desenvolvimento do fruto anteriormente ou posteriormente à colheita. Além das análises de caracterização, em que tenta-se identificar as características dos frutos utilizando o nariz eletrônico de forma completar à análises químicas e físico-químicas. Dentre os parâmetros que definem as características estão os índices de qualidade, os COVs, e o perfil de aroma. E as análises de qualidade, nas quais o objetivo é diferenciar os frutos baseando-se em algum aspecto de qualidade, como a identificação de contaminação ou o nível de deterioração.

Algumas das pesquisas dos últimos cinco anos envolvendo análises de maturação aplicam-se no monitoramento dos estágio de bananas (CHEN *et al.*, 2018), de goiabas (KANADE; SHALIGRAM, 2018), mangas (DEBABHUTI *et al.*, 2019), pêssegos (VOSS; STEVAN JUNIOR; AYUB, 2019), e de kiwis (DU *et al.*, 2019). Para as análises de caracterização algumas das pesquisas recentes visam a classificação do aroma de melões, morangos, limões e cerejas (ADAK; YUMUSAK, 2016), diferenças no perfil de COVs entre pêssegos vermelho e branco (XIN *et al.*, 2018), discriminação de laranjas pela origem geográfica (CENTONZE *et al.*, 2019), a discriminação entre cultivares de maçã (ZHU *et al.*, 2020), e a discriminação de bananas pela época de colheita (DOU *et al.*, 2020). Em relação às análises de qualidade são exemplos de aplicações a discriminação pelo tipo e pelo tempo de armazenamento de lichias (XU *et al.*, 2016), a discriminação pelo grau de contaminação por bactérias em maçãs (EZHILAN *et al.*, 2018), a detecção de deterioração pelo tempo de armazenamento em pitaias, peras, kiwis, e maçãs (DING *et al.*, 2018), detecção e classificação de contaminação por moscas em frutas cítricas (WEN *et al.*, 2019), e a discriminação de danos mecânicos em pêssegos (YANG *et al.*, 2020).

As aplicações envolvendo pessêgos e nariz eletrônico podem ser desde análises do estágio de desenvolvimento anterior a colheita ou pós-colheita, até análises dos produtos derivados, por exemplo, sucos e conservas. A respeito das análises de desenvolvimento do fruto, as pesquisas se concentram na maturação, principalmente, no pós-colheita em que analisa-se a maturação ideal e a deterioração do fruto. A classificação por estágios de desenvolvimento ou estado de maturação apresenta-se como foco principal das pesquisas, como em (BENEDETTI *et al.*, 2008; BREZMES *et al.*, 2000, 2005), sendo geralmente utilizado os dias após a colheita como variável de predição dos estágio de maturação.

Outras pesquisas abragem aplicações como a discriminação entre cultivares (DI NATALE *et al.*, 2002; SU *et al.*, 2013), a indentificação dos índices de qualidade (BREZMES *et al.*, 2000; SU *et al.*, 2013), a predição dos dias até a deterioração (HUANG *et al.*, 2017), análise de contaminação por diferentes fungos (LIU *et al.*, 2018), a predição do tempo e do tipo de armazenamento (WEI *et al.*, 2018), a análise de danos nos frutos (YANG *et al.*, 2020). Voss, Stevan Junior e Ayub (2019) analisaram os estágios de desenvolvimento do fruto até a colheita e pós-colheita,

isto é, os estágio durante o desenvolvimento do fruto no pessegueiro e após a colheita, visando a identificação do momento ideal de colheita dado os estágios de desenvolvimento do fruto no pessegueiro. O resultado indicou que é possível discriminar os estágios e, consequentemente, indentificar o momento ideal de colheita.

As análises de floração com nariz eletrônico geralmente são executadas em ambiente controlado, de forma que a amostra é extraída do ambiente de cultivo e analisada em laboratório. A pesquisa executada por Valeria *et al.* (2009) e recentemente por Yan *et al.* (2021a) são exemplos da análise da floração de pessegueiros utilizando nariz eletrônico.

Valeria *et al.* (2009) analisaram o odor das flores das cultivares Barcelo, Dixiland, Summerprince e Forastero. Com o objetivo de identificar diferenças entre a antese e a pós-antese. O nariz eletrônico MOSES II (GSG Mess- und Analysengeräte GmbH®, Alemanha) foi utilizado, sendo composto de 8 sensores piezoelétricos e 8 sensores MOS. Os resultados da pesquisa mostraram que somente as flores da cultivar Forastero apresentaram diferenças significativa em relação as demais analisadas, e a diferença nos estágios florais de antese e pós-antese foi identificada somente nos sensores MOS. Os autores relatam que a diferença detectada pode estar relacionada ao fato do cultivar Forastero ser macho-estéril. Indicando que o comportamento de emissão de odores ao longo do desenvolvimento das flores pode variar dependendo da cultivar.

Yan *et al.* (2021a) utilizaram o nariz eletrônico para predizer a necessidade de horas de frio dos botões florais de pessegueiros. A diferença nos COVs durante os estágios de desenvolvimento de botões florais de nove cultivares foi avaliada através das respostas dos sensores. O nariz eletrônico PEN3.5 (Airsense Analytics GmbH®, Alemanha) formado de 10 sensores MOS foi utilizado. As análises foram executadas em ambiente controlado e os resultados mostraram que as concentrações dos compostos variam de acordo com a necessidade de horas de frio dos botões florais, e os grupos de COVs em maior concentração foram compostos de metano e compostos orgânicos de enxofre. Os estágios definidos de horas de frio foram discriminados utilizando o LDA, e os cinco estágios foram diferenciados majoritariamente.

Nas demais pesquisas encontradas o nariz eletrônico foi utilizado para analisar outras espécies, como as flores de quatro variedades de maçãs (FAN *et al.*, 2018), e o estágio de desenvolvimento de flores da árvore Cananga odorata (QIN *et al.*, 2014). Nas pesquisas apresentadas não foram realizadas amostragens no local (em campo), geralmente as flores foram colhidas e as amostras realizadas em laboratório, o que difere-se desta pesquisa. No Quadro 3 são apresentados os trabalhos correlatos encontrados.

Quadro 3: Resumo das pesquisas correlatas em que analisou-se pessegueiros com nariz eletrônico.

Aplicação	**Nariz eletrônico**	**Local**	**Resultado**	**Referência**
análise de desenvolvimento dos frutos	Protótipo	em campo	LDA SVM e LDA RF com 98.08% de acurácia na classificação de 4 estágios de maturação	a
análise da floração	MOSES II	ambiente controlado	cv. Forastero $\neq$ demais cultivares, e o único com antese $\neq$ da pós antese	b
análise da floração	PEN3.5	ambiente controlado	LDA discriminou 5 estágios de acúmulo de horas de frio com variância total de 80,49%; PLSR com 100% de qualidadade na validação de horas de necessidade de frio	c

Fonte: a: (VOSS; STEVAN JUNIOR; AYUB, 2019), b: (VALERIA *et al.*, 2009), c: (YAN *et al.*, 2021a).

2.4.2 Sensores

Diversos tipos de sensores podem ser utilizados como elementos de entrada, sendo mais utilizados os baseados em material óxido metálico semicondutor (MOS), em que o elemento sensível tem a resistência alterada quando exposto a determinadas concentrações de compostos voláteis. Os compostos reagem com as moléculas de oxigênio do ar na superfície sensível alterando a disponibilidade de elétrons no material (KARAKAYA; ULUCAN; TURKAN, 2020; TAN; XU, 2020). Os sensores MOS são os mais utilizados para o nariz eletrônico por terem propriedades de alta sensibilidade e seletividade, serem adequados a uma variedade de gases, além de estarem comumente disponíveis no mercado (KARAKAYA; ULUCAN; TURKAN, 2020; TAN; XU, 2020). As principais séries de sensores MOS encontradas nas pesquisas são a TGS (Figaro Engineering, Japão) e a MQ (Hanwei Electronics, China). Além de sensores MOS, sensores piezoelétricos (BAW, SAW) e sensores de polímeros condutivos (CP) também são utilizados.

Ainda, em pesquisas recentes estão sendo analisadas questões como a utilização de nano sensores, tendo como potencialidades a sensibilidade e a seletividade (XU *et al.*, 2018), a combinação do nariz eletrônico e outros sistemas, como a lingua eletrônica e sistemas de visão computacional, para a fusão de sensores (DI ROSA *et al.*, 2017), e a utilização do nariz eletrônico em conjunto com outros métodos de análise como a espectroscopia (BEGHI *et al.*, 2017).

2.4.2.1 Aquecimento

O aquecimento dos sensores é fundamental para o funcionamento adequado durante as medições, e tem como objetivo garantir a temperatura necessária de operação entre o material sensível e os gases que atingem a superfície do sensor (GAJDOSIK, 2014). O aquecimento é efetuado por meio de um resistor próprio no sensor, que é energizado para atingir a temperatura de operação garantindo a sensibilidade adequada. A temperatura indicada para alguns sensores MOS pode variar de 500 K a 700 K (227 a 427 °C), o que corresponde a valores de tensão de aquecimento de 2 V a 5 V, respectivamente (GAJDOSIK, 2014; OATES *et al.*, 2018). Porém, a temperatura de aquecimento de maior sensibilidade é dependente do gás e da concentração deste (GAJDOSIK, 2014).

Diferentes modos de aquecimento podem ser utilizados para a matriz de sensores de um nariz eletrônico. O modo mais simples é manter o aquecimento constante durante as medições, dessa forma a complexidade do circuito do equipamento é reduzida não necessitando de um controle do aquecimento no momento da coleta de dados. Nesse modo, geralmente, a curva de resposta tende a reduzir de forma gradual ao longo do período de aquecimento e se manter estável durante a coleta quando a matriz está no mesmo ambiente, onde a suposição é de que não há variação na concentração dos gases. Dois aspectos que podem afetar a análise com esse modo de aquecimento são uma baixa diferenciação no sinal de resposta à diferentes amostras, e a variação na tensão de alimentação do resistor que pode ser refletida na resposta do sensor (OATES *et al.*, 2018).

Outros modos tem por objetivo controlar o aquecimento por meio de chaveamento para aumentar as respostas dos sensores quando expostos às amostras. Como o modo indicado nas folhas de dados de alguns sensores em que utiliza-se modulação de pulso (PWM) na alimentação do resistor no momento da coleta (HE, 2022a, 2022b; OATES *et al.*, 2018). Essa abordagem exige um circuito mais complexo para o nariz eletrônico. Outro modo possível utiliza forma de onda senoidal na alimentação do resistor para transformar as respostas dos sensores em curvas senoidais e extrair mais informações do sinal (OATES *et al.*, 2018). Tal estratégia necessita de um circuito específico e uma análise de sinal mais complexa.

Portanto, por mais que os modos de controle de aquecimento durante a coleta apresentem alguma melhoria na resposta dos sensores, eles mostram-se não apropriados para uma análise simplificada e contínua em uma ambiente aberto e não controlado como o ambiente de um pomar. Isso se deve, principalmente, pela forma como o modo de aquecimento pode interferir na análise, sendo mais apropriado para os objetivos da pesquisa a abordagem mais simples e de menor interferência possível.

2.4.2.2 Calibração

A calibração dos sensores envolve qualquer procedimento de medida das características de resposta, e a correspondente correção ou não de alguma característica em relação aos valores e condições conhecidas. Os principais aspectos que geralmente são medidos para sensores de gás são a sensibilidade (resposta do sensor ao mesmo gás em diferentes concentrações), a seletividade (resposta do sensor à diferentes gases numa mesma concentração), estabilidade (resposta do sensor ao mesmo gás numa concentração ao longo do tempo), e repetibilidade (resposta do sensor ao mesmo gás numa concentração repetidas vezes).

Uma forma de calibração empregada é em relação à sensibilidade, em que estima-se a concentração utilizando equações obtidas das curvas da folha de dados e o resultado é comparado com o valor conhecido de concentração do gás analisado, então a diferença entre os valores conhecido e estimado pode ser avaliada (DORCEA; HNATIUC; LAZAR, 2018; PIRES, 2018; YAN *et al.*, 2021b). De modo similar, é feita a caracterização utilizando a sensibilidade para descrever a curva que relaciona as respostas do sensor e as concentrações conhecidas do gás (FABBRI *et al.*, 2020; LIN *et al.*, 2016; WU *et al.*, 2020). Ambas as formas de avaliação são restritivas, pois implicam em uma aplicação específica com objetivo de inferir a concentração do gás, isto é, necessitam de um ambiente controlado, um gás isolado e a concentração conhecida, e seus resultados não podem ser extendidos para outros gases, sendo válidos somente para o gás específico.

Outro aspecto para considerar é o de sensibilidade cruzada em que diferentes gases podem gerar repostas similares num sensor, levando à respostas individuais ambíguas (DI NATALE *et al.*, 2006). Além disso, sensores de gás são afetados pela temperatura e umidade influenciando sua sensibilidade, havendo necessidade de compensação desse fator de variação (DI CARLO; FALASCONI, 2012). No entanto, as características mensuráveis dos sensores assim como as variações devido a temperatura e umidade são relativas aos gases e não ao sensor, isto é, um sensor tem comportamento diferente de sensibilidade, estabilidade, repetibilidade e variações devido à temperatura e umidade para cada gás (DI CARLO; FALASCONI, 2012; DORCEA; HNATIUC; LAZAR, 2018; FABBRI *et al.*, 2020; YAN *et al.*, 2021b). O que torna difícil a calibração dos sensores em relação a esses fatores quando tem-se um experimento num ambiente não controlado ou uma aplicação não específica, ou seja, em que não se controla a amostra e o ambiente ou há uma mistura de gases ao invés de um gás isolado.

Nas aplicações, geralmente, o nariz eletrônico é utilizado com o objetivo de analisar o aspecto geral da amostra, isto é, a mistura de compostos de forma qualitativa, não sendo relevante e praticável à análise do nariz eletrônico identificar compostos isolados e concentrações

numa amostra, tarefa que é atribuída à análise por GC-MS. Isso contribuí para um nível de dificuldade na relação entre calibração dos sensores e aplicação, pois a calibração específica de um sensor pode não influenciar na aplicação do nariz eletrônico, ou seja, não auxiliar a aplicação. Uma forma mais adequada de avaliação dos sensores pode ser feita considerando as características do dispositivo que são importantes para a análise dado a aplicação. Isso pode ser compreendido como uma calibração geral ou validação do nariz eletrônico para a análise, isto é, a validação da configuração da matriz de sensores na configuração do equipamento dado a análise de aspecto geral.

2.4.2.3 Compensação

Sensores de gás são passíveis de fatores que afetam sua estabilidade como temperatura e umidade ou deterioração, que produzem variações nas respostas sob a mesma condição de amostra (DI CARLO; FALASCONI, 2012; KASHWAN; BHUYAN, 2005). As estratégias de compensação visam corrigir tais efeitos nos sensores, e como consequência melhorar os resultados da análise de dados.

De acordo com Di Carlo e Falasconi (2012) os métodos de correção podem ser classificados em qautro tipos. O primeiro, é o de pré-processamento de sinal, que inclui as estratégias de manipulação de linha base (transformações no sinais individuais dos sensores utilizando o valor inicial da resposta como base), e de filtragem (remoção da parte do sinal que contém variações por meio de transformada wavelet discreta (DWT), filtro de média móvel, ou filtros de Fourier).

O segundo tipo, é o chamado de calibração periódica, compreendendo as estratégias de correção multiplicativa (correção de cada sensor em relação à calibração para um gás de referência utilizando medidas sequenciais intra e entre execuções, seguida de recalibrações), de correção de componente (análise multivariada em que elimina-se a direção de variação comum do espaço definido pelos sensores relativo a um gás de referência utilizando PCA ou PLSR), e de deflação de componente (análise multivariada em que a variância explicada pelas variáveis de compensação é removida dos dados utilizando CCA ou PLSR).

O terceiro, são os métodos de sintonização aplicados na base de dados, como a correção de componente independente (análise multivariada em que os componentes que possuem maior correlação com as variáveis de compensação são descartados utilizando ICA sem a necessidade de um gás de referência), e a correção ortogonal de sinal (remoção da variância que não é correlacionada com a variável objetivo nas variáveis de entrada). O quarto tipo, são os métodos adaptativos, que abrangem modelos de reconhecimento de padrões para corrigir as variações.

Outra forma empregada para compensar a temperatura e umidade em sensores de gás, semelhante à correção multiplicativa, é utilizar as curvas descritas nas folhas de dados de cada sensor, aproximar uma equação para as curvas, e corrigir os valores de resposta dado os valores de temperatura e umidade (PIRES, 2018; VOSS, 2019). Essa abordagem pode melhorar os resultados em relação aos dados não compensados. Porém, não é adequada para mistura de gases e ambientes não controlados, pois tem como premissa um único gás, ou que o comportamento do sensor é o mesmo para todos os gases ou, ainda, que um determinado gás é o de maior abundância numa mistura. Uma vez que as curvas de temperatura e umidade descritas em uma folha de dados são apenas para um gás específico, não para todos e nem para uma mistura.

A estratégia que parece mais adequada para mistura de gases e ambientes não controlados é a de compensação na etapa de processamento. Nesse caso, as variáveis de temperatura e umidade são consideradas como variáveis de entrada em conjunto com os sensores de gás (HUERTA *et al.*, 2016; ZHANG *et al.*, 2011; TIAN *et al.*, 2016; ZHANG *et al.*, 2021). Isso faz com que o comportamento dinâmico de uma amostra complexa, dado a temperatura e a umidade, seja tratado por métodos de reconhecimento de padrões, ao invés de estratégias diretas nas etapas de análise de sinal e pré-processamento. Essa estratégia pode aumentar o tempo validade do sistema, isto é, reduzir a necessidade de retreinamento do sistema para uma aplicação (DI CARLO; FALASCONI, 2012).

2.4.2.4 Validação do equipamento

Diferente da avaliação específica dos sensores, a validação tenta explorar as características de resposta da matriz na configuração do equipamento. Dessa forma os fatores que afetam os sensores de forma geral pode ser avaliados e interpretados enquanto influência nos resultados das análises que são objetivos da pesquisa. Este tipo de validação que visa uma verificação geral de efeitos sobre a estabilidade do nariz eletrônico é importante, pois pode indicar de que forma e o quanto as variações afetam o sistema e, consequentemente, as análises e a possível aplicação. Por exemplo, a instabilidade devido às variações nos sensores pode levar a uma necessidade constante de retreinamento do sistema numa aplicação (DI CARLO; FALASCONI, 2012; KASHWAN; BHUYAN, 2005). Então, avaliar os efeitos de variações do equipamento em testes auxiliares pode aumentar a confiabilidade no sistema e na análise executada, além de melhorar a compreensão sobre a viabilidade de aplicações.

Para uma avaliação simplificada as características de estabilidade e repetibilidade podem ser verificadas para a matriz de sensores do nariz eletrônico. A resposta da matriz está relacionada à configuração do equipamento, isto é, o conjunto de sensores (matriz) e a circulação de ar no equipamento, determinando a configuração. A estabilidade e a repetibilidade podem

ser avaliadas como critérios de validação para a configuração do nariz eletrônico independente de medidas diretas de concentração de um ou mais gases, necessárias para a sensibilidade e seletividade.

Em relação aos objetivos da pesquisa, três características do equipamento que contribuem para a validação podem ser avaliadas. A estabilidade e repetibilidade entre execuções de curto prazo, estabilidade e repetibilidade entre execuções de longo prazo, e avaliação de mesmo ambiente. As avaliações de curto e longo prazo têm por objetivo as variações de podem ocorrer na matriz de sensores entre as execuções ao longo do tempo, isto é, os efeitos de uma amostra com diferenças conhecidas sob o regime de execução de curto e longo prazo. E a avaliação de um único ambiente busca interpretar as variações do equipamento exposto ao mesmo ambiente durante cada execução, ou seja, os efeitos de um ambiente que tem como premissa a estabilidade mas que pode sofrer influência de fatores complexos.

Embora essas avaliações ainda sejam afetadas por algum nível de especificidade devido à amostra, gás utilizado, ou ambiente, elas tendem a ser mais representativas sobre o comportamento do disposito para a análise a ser feita do que avaliações específicas de cada sensor, que tendem a ser difíceis e de pouca capacidade de generalização para ambientes complexos.

2.4.3 Pré-processamento

O pré-processamento tem por objetivo preparar os dados para o processamento, tendo como consequência o aumento da eficiência e diminuição da complexidade na etapa seguinte. O pré-processamento pode ser divido entre duas etapas complementares, a primeira sendo a análise de sinal, em que o conjunto de dados é transformado em um conjunto de variáveis representativas ou espaço de características. A segunda sendo a redução de dimensionalidade, na qual o espaço de características é reduzido para um conjunto mais efetivo de acordo com algum critério de relevância.

A análise multivariada é aplicada para encontrar as propriedades internas de múltiplas variáveis, e pode ser utilizada ao longo do pré-processamento para avaliar a aptidão de um sistema operando em determinada aplicação, as técnicas de agrupamento e de representação são as principais utilizadas para esse objetivo (DI NATALE *et al.*, 2006).

De acordo com Han, Pei e Kamber (2012) os métodos utilizados para o pré-processamento podem ser organizados em quatro categorias, sendo elas, limpeza, integração, redução, e transformação dos dados.

A limpeza dos dados tenta reduzir os erros agregados aos dados, tais como os produzidos por valores faltantes, ruído ou valores anormais. Dentre as estratégias para correção de

valores faltantes estão a inserção conforme a média, a mediana ou o valor mais provável. Para a redução de ruído podem ser adotadas estratégias como Binning, em que a média, mediana ou máximo e mínimo de conjuntos de dados são utilizados para representação dos dados, além de análises de regressão e de detecção de valores anormais através de análise de agrupamentos (HAN; PEI; KAMBER, 2012).

A integração compreende a mistura de dados vindos de diferentes bases, tendo como objetivo lidar com redudâncias e inconsistências. A detecção de redudâncias pode ser feita através da análise de correlação, como o teste χ^2 para dados qualitativos, e o coeficiente de correlação de Pearson ou a covariância para dados quantitativos (HAN; PEI; KAMBER, 2012).

A redução visa diminuir a quantidade de dados de forma a manter a representação para aumentar a eficiência do processamento sem prejudicar o resultado deste. Algumas das estratégias são a redução de dimensionalidade, de numerosidade, e a compressão de dados. A redução de dimensionalidade diminui a quantidade de variáveis ou características, compreendendo a extração, na qual um novo conjunto é gerado a partir do original, e a seleção, em que variáveis de menor relevância são removidas. Métodos como a transformada wavelet discreta (DWT) e a análise de componentes principais (PCA) podem ser utilizados para extração de características, e árvores de decisão como ID3 e CART podem ser utilizadas para seleção de características (HAN; PEI; KAMBER, 2012).

A transformação dos dados é o procedimento para adequar os dados ao processamento visando o formato mais apropriado. As estratégias incluem a suavização de ruído, construção de característica, agregação, normalização, e discretização (HAN; PEI; KAMBER, 2012).

2.4.3.1 Análise de sinal

A análise de sinal é a primeira etapa do pré-processamento. O sinal produzido pela matriz de sensores é analisado e transformado no espaço de características. Essa etapa compreende, entre outros, a redução de ruído, o ajuste, a extração de dados do sinal, e a normalização dos dados (HAN; PEI; KAMBER, 2012; DI NATALE *et al.*, 2006).

Na redução de ruído busca-se minimizar as informações incorretas contidas em um sinal vindas de erros aleatórios ou da variância dos dados (HAN; PEI; KAMBER, 2012). A estratégia pode ser executada através de técnicas de suavização de dados como o *Binning*, em que suaviza-se um conjunto de dados conforme seus valores, isto é, os valores do conjunto são substituídos por valores representativos, podendo ser substituídos pela média do conjunto, a mediana ou pela relação entre máximo e mínimo (HAN; PEI; KAMBER, 2012). Outras técnicas que podem ser aplicadas são o filtro de média móvel, no qual os dados são avaliados conforme a

diferença em relação à media e o desvio padrão (SENTHILKUMAR; VENKATAKRISHNAN; BALAJI, 2020; VOSS; STEVAN JUNIOR; AYUB, 2019), e a utilização de transformada *wavelet*, que extraí do sinal coeficientes que podem ser discriminados entre coeficientes de ruído e de sinal (JIA *et al.*, 2016; WIJAYA *et al.*, 2017).

No ajuste de sinal o propósito é corrigir o sinal de acordo parâmetros estabelecidos, como valores obtidos durante a calibração dos sensores ou valores de compensação dado condições ambientais que possam influenciar a resposta dos sensores, como a temperatura e umidade relativa (DORCEA; HNATIUC; LAZAR, 2018; SENTHILKUMAR; VENKATAKRISHNAN; BALAJI, 2020; VOSS, 2019).

A extração de dados do sinal é o procedimento no qual a representação do sinal é formada. Nela, geralmente, diminui-se a quantidade de dados mantendo-se a representação da informação (DI NATALE *et al.*, 2006). A compressão do sinal é utilizada durante a extração podendo ser uma subamostragem, uma extração de parâmetros, ou através do método de identificação de sistemas no qual utiliza-se um modelo matemático para descrever o sinal (GILA *et al.*, 2020). Um conjunto de funções e técnicas podem ser aplicadas no sinal de acordo com objetivo da análise. Alguns exemplos são a média (VOSS; STEVAN JUNIOR; AYUB, 2019; WEI *et al.*, 2018), o valor máximo (GILA *et al.*, 2020; SANAEIFAR *et al.*, 2014), a integral do sinal (HUANG *et al.*, 2017; YANG *et al.*, 2020), a parte ascendente do sinal (JIA *et al.*, 2016), valor instantâneo (QIU; WANG, 2017), valor estável (HUANG *et al.*, 2017), valores extraídos de parâmetros da derivada do sinal (GILA *et al.*, 2020), e técnicas como a transformada *wavelet* (LI *et al.*, 2017), e a transformada discreta de Fourier (OATES *et al.*, 2018).

Na normalização busca-se adequar o formato dos dados padronizando estes em escala curta ou comum. Essa padronização pode evitar que a informação quantitativa esconda a informação qualitativa nos dados, tendo como consequência a possibilidade de aumentar a velocidade durante o treinamento em redes neurais, e melhorar a capacidade de discriminação em métodos baseados na distância como classificação por vizinho mais próximo e agrupamentos (HAN; PEI; KAMBER, 2012; DI NATALE *et al.*, 2006). Em termos dos dados do nariz eletrônico a normalização também é útil para lidar com a seletividade cruzada, em que diferentes amostras geram respostas similares, e compensar variações entre as amostras devido a fatores como mudança de concentração do elemento analisado e variações próprias dos sensores (DI NATALE *et al.*, 2006; SANAEIFAR *et al.*, 2014). Porém, sua efetividade é limitada quando executada em aplicações com a mistura de gases, uma vez que cada composto possuí característica próprias tende-se a ter mudanças específicas, gerando variações de concentração e no padrão analisado quando há variação de temperatura (DI NATALE *et al.*, 2006). Alguns métodos de normalização incluem a normalização por mínimo e máximo, normalização de média zero, e a normalização

de escala decimal. A Equação 1 descreve a normalização por mínimo e máximo (HAN; PEI; KAMBER, 2012). Na equação v_{inorm} é o valor normalizado, v_i é o original, v_{min} o mínimo e v_{max} é o valor máximo do conjunto que v_i pertence.

$$v_{inorm} = \frac{v_i - v_{min}}{v_{max} - v_{min}} \tag{1}$$

2.4.3.2 Extração de características

Na extração de características o objetivo é produzir um conjunto de dados mais adequado para o processamento, gerando um espaço de dados mais apto a ser discriminado e reduzindo a dimensionalidade, isto é, a quantidade de variáveis. A extração de características executa transformações nos espaço de características, produzindo outro espaço composto por variáveis mais relevantes. Os métodos de extração podem ser divididos entre supervionados e não supervisionados, podendo ser categorizados em lineares e não lineares (GHOJOGH *et al.*, 2019; KHALID; KHALIL; NASREEN, 2014). Nos métodos supervisionados utiliza-se as informações das classes para gerar as tansformações sobre os dados, ou seja, as características extraídas expressam as relações entre as classes. Para os métodos não supervisionados utiliza-se as informações contidas nos dados indenpendente da classe, isto é, as características extraídas expressam as relações entre os dados (GHOJOGH *et al.*, 2019).

2.4.3.3 Análise de componentes principais (PCA)

A análise de componentes principais (PCA) é uma técnica utilizada para extrair as informações mais relevantes de um conjunto por meio da variação entre os dados. De modo, que a técnica pode ser utilizada para redução de dimensionalidade, compressão de dados, extração de características, e visualização de dados (BISHOP, 2006). O objetivo do PCA é encontrar direções ortogonais que representam as maiores variações nos dados, utilizando o menor erro ou a maior variância para encontrar cada direção (BISHOP, 2006). Os componentes principais são os vetores ortogonais que representam as direções que mantêm as relações de covariância mínima entre si e variância máxima dos dados projetados (KHALID; KHALIL; NASREEN, 2014). A técnica descreve uma rotação de coordenadas do espaço original para o espaço definido pelos componentes principais (DI NATALE *et al.*, 2006).

Cada componente principal é uma combinação linear das variáveis ou dimensões de um determinado espaço de características, gerando então um mapeamento de um espaço de *n*

dimensões para uma única dimensão, e sobre essa direção ou vetor as amostras são projetadas. Para definir essa direção ou vetor é produzido aquele com o melhor ajuste de menor distância de projeção dado a direção (erro) e de maior comprimento de projeção dado a origem (variância) em relação a todas as amostras de um espaço *n* dimensional, sendo o ponto central das amostras deslocado para a origem. Isto é, em relação às amostras num espaço, o PCA tenta encontrar a direção para projeção que minimiza o erro e maximiza a variância entre as amostras projetadas.

Para uma direção ou vetor unitário **u** a projeção de uma amostra $\mathbf{X}_i = [x_1, ..., x_D]$ nessa direção representa o valor $\mathbf{u}^T\mathbf{X}_i$, a variância das amostras projetadas é definida pela Equação 2, e a matriz de covariância **S** pela Equação 3 (BISHOP, 2006). Sendo D a dimensão do espaço original (quantidade de variáveis). Nas equações, N é a quantidade de amostras e $\bar{\mathbf{X}}$ é a média do conjunto de amostras.

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{u}^T\mathbf{X}_i - \mathbf{u}^T\bar{\mathbf{X}})^2 = \mathbf{u}^T\mathbf{S}\mathbf{u} \tag{2}$$

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T \tag{3}$$

Então, o problema de otimização se resume em encontrar a direção **u** que maximiza a variância projetada $\mathbf{u}^T\mathbf{S}\mathbf{u}$ (BISHOP, 2006). A solução tem a forma de um problema de autovalor descrito pela Equação 4 (BISHOP, 2006; GHOJOGH *et al.*, 2019). Nessa equação, **u** é um autovetor da matriz **S**, e λ é o autovalor de **u**.

$$\mathbf{u}\mathbf{S} = \lambda\mathbf{u} \tag{4}$$

O autovetor de **S** que corresponde ao componente principal é o que tem o maior autovalor λ. Cada próximo componente principal é definido por outro autovetor de **S** ortogonal aos componentes anteriores e com maior autovalor possível. Os componentes principais são todos os M autovetores de **S** correspondentes aos M maiores autovalores (BISHOP, 2006). É possível que $M \leq D$, porém para permitir a visualização dos dados no espaço definido pelos componentes principais geralmente $M = 2$ ou 3 (HAN; PEI; KAMBER, 2012; DI NATALE *et al.*, 2006).

Para a aplicação adequada da técnica é importante que os dados de entrada estejam normalizados, isto é, dentro da mesma amplitude de valores. Isso evita que variáveis com valores de maior amplitude tenham mais influência no resultado final (HAN; PEI; KAMBER, 2012). De acordo com Di Natale *et al.* (2006) a utilização do PCA é útil para representação dos dados do nariz eletrônico, uma vez que os sensores da matriz apresentam alta correlação entre si e o PCA consiste em encontrar uma base ortogonal na qual a correlação entre os sensores é minimizada. Porém, dois aspectos que podem limitar a efetividade da técnica em descrever dados do nariz eletrônico são a consideração do PCA de que o conjunto de dados (cada variável) tenha distribuição normal, e as várias fontes de ruído correlacionado que podem afetar as medidas (DI NATALE *et al.*, 2006). Em relação a primeira, a totalidade das amostras obtidas pode não seguir uma distribuição normal de valores. Sobre a segunda, fatores como a variação de temperatura induzem variações correlacionadas nos sensores, que podem ser descritas pelos primeiros componentes no lugar das propriedades relevantes dos dados (DI NATALE *et al.*, 2006).

2.4.3.4 Análise de discriminates lineares (LDA)

A análise de discriminates lineares (LDA) utiliza a informação sobre as classes para definir a direção de projeção dos dados. Em um espaço *n* dimensional a direção escolhida é aquela que maximiza a razão da variância entre classes sobre a variância intra classes dos dados projetados (GHOJOGH *et al.*, 2019). Isto é, busca-se a representação dos dados na direção de maior variação entre as classes e menor variação dentro de cada classe. Cada direção encontrada é uma combinação linear das variáveis de entrada, e é chamada de discriminante linear. As amostras no espaço *n* dimensional são projetadas sobre cada discriminante, e a quantidade de discriminadores ou dimensão da saída é no máximo uma a menos que a quantidade de classes (GHOJOGH *et al.*, 2019). A direção ou vetor do discriminante linear é produzido por meio do melhor ajuste em relação às amostras projetadas, maximizando a distância entre as médias das classes e minimizando a distância de espalhamento ou dispersão dentro de cada classe.

A técnica é descrita pela maximização do critério de Fisher, que relaciona a variância das amostras projetadas de diferentes classes com a variância projetada das amostras dentro de cada classe. O critério de Fisher é apresentado na Equação 5 (GHOJOGH *et al.*, 2019). As matrizes de dispersão entre classes $\mathbf{S}_B$ e intra classes $\mathbf{S}_W$ são descritas pelas Equações 6 e 7, respectivamente (GHOJOGH *et al.*, 2019).

$$J(\mathbf{u}) = \frac{\mathbf{u}^T \mathbf{S}_B \mathbf{u}}{\mathbf{u}^T \mathbf{S}_W \mathbf{u}} \tag{5}$$

$$\mathbf{S}_B = \sum_{c=1}^{C} N_c(\bar{\mathbf{X}}_c - \bar{\mathbf{X}})(\bar{\mathbf{X}}_c - \bar{\mathbf{X}})^T \tag{6}$$

$$\mathbf{S}_W = \sum_{c=1}^{C} \sum_{i=1}^{N_c} (\mathbf{X}_i - \bar{\mathbf{X}}_c)(\mathbf{X}_i - \bar{\mathbf{X}}_c)^T \tag{7}$$

Nas equações, $\mathbf{u}$ é a direção de projeção ou vetor unitário do discriminante, C é a quantidade de classes, N_c é o número de amostras na classe c, $\bar{\mathbf{X}}_c$ é a média da classe c, $\bar{\mathbf{X}}$ é a média de todas as amostras de treino, e $\mathbf{X}_i$ é uma amostra.

A maximização de $J(\mathbf{u})$ resulta em um problema de autovalor generalizado representado na Equação 8 (GHOJOGH *et al.*, 2019). Onde, λ é o autovalor do autovetor $\mathbf{u}$. Então, os discriminantes lineares são os M autovetores da relação $\mathbf{S}_W^{-1}\mathbf{S}_B$ com os maiores autovalores. É possível que $M \leq C - 1$ (GHOJOGH *et al.*, 2019). Caso a matriz resultante $\mathbf{S}_R = \mathbf{S}_W^{-1}\mathbf{S}_B$ seja não simétrica, isto é $\mathbf{S}_R \neq \mathbf{S}_R^T$, seus autovetores podem não ser ortogonais (STACKEXCHANGE, 2017, 2018). Além disso, o LDA assume que os dados de cada classe seguem uma distribuição normal, o que pode não representar a relação adequada de valores para uma classe (GUTIERREZ-OSUNA, 2022; SCIKIT-LEARN, 2022c). Ainda, a técnica pode não expressar a discriminação entre as classes se a informação discriminatória estiver na variância dos dados e não nas médias, isto é, médias similares e variâncias diferentes (GUTIERREZ-OSUNA, 2022; MATAS; KOSTLIVÁ, 2014).

$$\mathbf{S}_B\mathbf{u} = \lambda\mathbf{S}_W\mathbf{u} \tag{8}$$

A técnica pode discriminar de forma mais adequada do que o PCA o conjunto de amostras, e gerar melhores características extraídas do conjunto original devido a discriminação baseada na relação entre as classes. Para visualização dos dados em um problemas de múltiplas classes ($C \geq 3$) a quantidade de discriminantes pode ser igual a 2 ou 3.

2.4.4 Processamento dos Dados

Nesta etapa o sistema é treinado por meio de técnicas de reconhecimento de padrões, e avaliado através do desempenho em predizer as amostras. As predições podem ser de variáveis qualitativas e discretas (classificação), ou quantitativas e contínuas (regressão).

Uma das estratégias para o reconhecimento de padrões é a utilização do aprendizado supervisionado, em que durante o treinamento utiliza-se as informações dos rótulos dos dados para o modelo aprender, isto é, os padrões que são importantes para o objetivo da análise são considerados enquanto os padrões não importantes são minimizados. Diferentes técnicas de aprendizado de máquina podem ser utilizadas para o reconhecimento de padrões. O modelo de aprendizado de máquina é o resultado do treinamento efetuado atráves da aplicação de uma técnica de aprendizado (algoritmo) sobre um espaço de características. Esse modelo mantém os padrões aprendidos durante o treino através dos parâmetros da técnica utilizada e está apto a fazer predições sobre a variável de saída dado as variáveis de entrada.

Para avaliar o desempenho do modelo executa-se testes e as métricas são computadas. Algumas das métricas que podem ser consideradas em análises de classificação são a acurácia, precisão, ou o *recall*. Esses procedimentos estão inclusos na validação do modelo na qual determina-se a forma como o este é definido e avaliado. Duas estratégias que podem ser utilizadas para validação são os métodos de holdout, em que há uma divisão prévia na base de dados entre treino e teste, e métodos de validação cruzada, nos quais são feitas divisões na base de modo que cada parcela seja utilizada para teste (HAN; PEI; KAMBER, 2012).

2.4.4.1 K-vizinhos mais próximos (KNN)

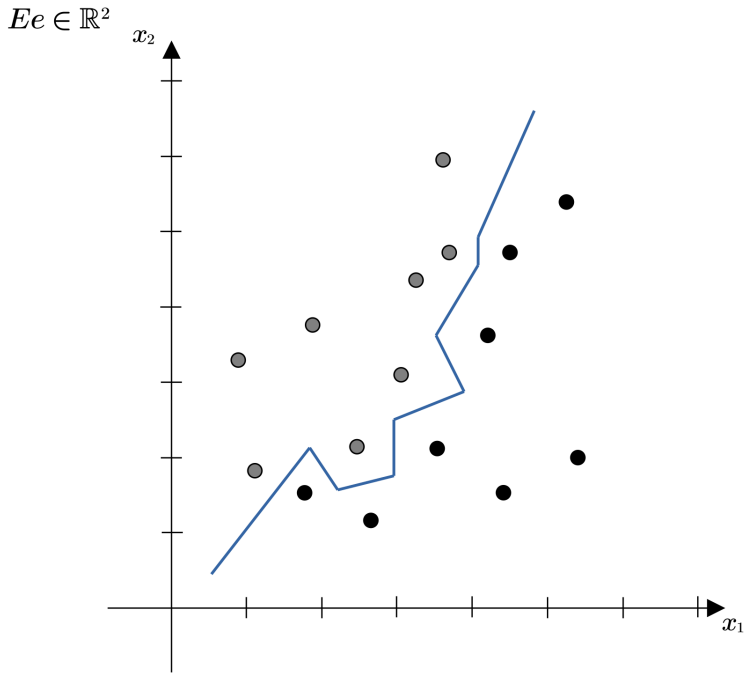
O K-vizinhos mais próximos (KNN) é um algoritmo de aprendizado considerado um *lazzy learner* ou um método baseado em instâncias. Nesse método as amostras de treino são minimamente processadas, sendo armazenadas para então a generalização ou construção do modelo ser efetuada a partir das amostras de teste, quando ocorre o maior esforço computacional devido a buscas no espaço das amostras de treino dado as amostras de teste. O KNN é um método não linear, isto é, pode definir limites de decisão com formato de hiperpolígonos (HAN; PEI; KAMBER, 2012).

Para a classificação as K amostras mais próximas são escolhidas e a classe mais presente é atribuída à amostra. Diferentes métricas podem ser definidas para medir a distância no espaço amostral como distância Euclidiana ou a distância de Manhattan. A Equação 9 descreve a distância Euclidiana entre duas amostras representadas como vetores $\mathbf{X}_1 = [x_{11}, ..., x_{1n}]^T$ e $\mathbf{X}_2 = [x_{21}, ..., x_{2n}]^T$ com n variáveis de entrada (HAN; PEI; KAMBER, 2012). A quantidade vizinhos ou o valor de K pode ser definido experimentalmente começando com K= 1 e avaliando o erro de teste da classificação (HAN; PEI; KAMBER, 2012).

$$dist(\mathbf{X}_1, \mathbf{X}_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \tag{9}$$

Em um espaço bidimensional o algoritmo pode ser interpretado como uma busca por K amostras de treino a partir de uma esfera com a amostra de teste no centro. Dessa forma, a classe atribuída à amostra é a que tem o maior valor da relação K_C/K, em que K_C é a quantidade de amostras de treino no conjunto de K amostras que pertence a classe C. Na Figura 3 é apresentado um exemplo com o limite de decisão criado pelo KNN com K= 1 em um espaço bidimensional. Em relação à K, quanto maior for seu valor mais suavizados tendem a ser os limites de decisão e, consequentemente, as regiões das classes (BISHOP, 2006). De modo que a acurácia de classificação também se altera uma vez que alteram-se as regiões. Fato que pode ser avaliado em termos de custo benefício entre esforço computacional e acurácia.

Figura 3: Representação do limite de decisão (linha azul) formado a partir das amostras de treino num espaço bidimensional com K= 1. Para esse caso os limites tendem a ser formados por retas no centro e ortogonais (hiperplanos formados por mediatrizes) aos pares de amostras de treino de diferentes classes.



Fonte: Adaptado de (BISHOP, 2006).

Questões como a normalização dos dados para evitar desproporcionalidade das amplitudes de valores entre as variáveis no cálculo de distância e a complexidade computacional das buscas executadas devido o tamanho da base de dados devem ser levadas em consideração para a utilização do algoritmo. Além da quantidade de K vizinhos, alguns dos parâmetros que podem ser definidos para o KNN são a métrica usada para calcular a distância, a forma de encontrar e

o modo como ponderar os K vizinhos mais próximos.

2.4.4.2 Máquina de aprendizado extremo (ELM)

A máquina de aprendizado extremo (ELM) é uma RNA em que a correção de seus pesos é simplificada, sendo seus valores definidos analiticamente. A ELM pode ser utilizada para a classificação e para regressão, com a capacidade de obter menores erros de treino e menores pesos, além de ter maior velocidade de operação, o que pode resultar em um melhor desempenho de generalização do que outras estratégias de RNA.

A estrutura de uma rede neural segue o modelo de neurônios e sinapses biológicas simulando o funcionamento das células cerebrais dos animais. Nesse modelo um neurônio recebe e computa informações e, posteriormente, as transmite para outros neurônios através das sinapses. Cada sinapse conecta dois neurônios e possui característica própria atribuída pela conexão. A composição da rede e seu efeito são estabelecidos quando neurônios são conectados, podendo receber informações de sinapses de entrada e produzir informações para transmitir pelas sinapses de saída. O ajuste dos pesos das sinapses é o que garante o efeito de aprendizado à rede, sendo um processo cumulativo conforme a estrutura é exposta a entradas.

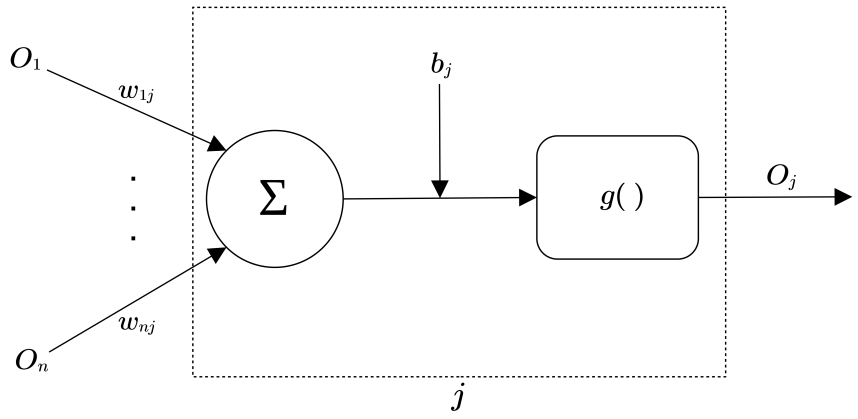
Uma das vantagens das RNAs é que elas podem se comportar como aproximadores universais, pois a saída de uma rede neural pode ser modelada como uma combinação não linear das entradas e, dado a sua configuração, ser capaz de aproximar qualquer função (BISHOP, 2006). Além disso, elas são tolerantes a dados com ruído (HAN; PEI; KAMBER, 2012). A principal questão em relação às RNAs é como encontrar os valores adequados para seus parâmetros conforme os dados de treino (BISHOP, 2006).

Um neurônio é uma unidade da rede responsável por computar a informação que chega até sua entrada. Esse cálculo é efetuado por uma função de ativação, que recebe como entrada a adição entre um valor de viés próprio do neurônio e a soma ponderada das saídas da camada anterior (HAN; PEI; KAMBER, 2012). O modelo de um neurônio é representado na Figura 4, a composição C_j das entradas e sua saída O_j são descritas nas Equações 10 e 11 (HAN; PEI; KAMBER, 2012).

$$C_j = \sum_{i=1}^{n} w_{ij} O_i + b_j \tag{10}$$

$$O_j = g(C_j) \tag{11}$$

Figura 4: Representação de um neurônio artificial. Onde O_n é a saída do neurônio n da camada anterior, w_{nj} é o peso atribuído à conexão entre os neurônios n e j, $\sum$ é o somatório das entradas ($O_n w_{nj}$) do neurônio j, b_j é o viés associado, $g(\ )$ é a função de ativação, e O_j é a saída do neurônio.



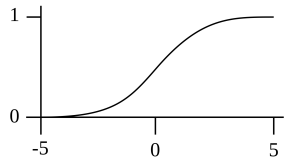
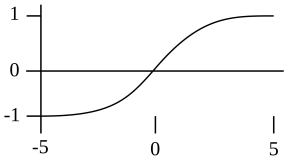
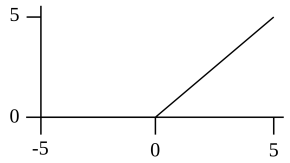
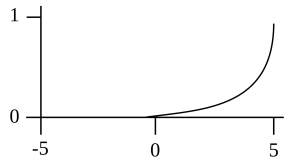
Fonte: Adaptado de (HAN; PEI; KAMBER, 2012).

A função de ativação é utilizada nos neurônios para computar a resposta do neurônio à suas entradas. Essa função, geralmente, é não linear podendo ser do tipo logística, tangente hiperbólica, ReLU, entre outras. A escolha da função de ativação é determinada pela natureza dos dados e a distribuição da variável objetivo, sendo que para os neurônios de saída ela é definida de acordo com o tipo de problema, podendo ser *softmax* ou logística para classificação e identidade para regressão (BISHOP, 2006). As características das quatro funções de ativação comumente utilizadas são apresentadas na Figura 5.

As sinapses são as conexões entre os neurônios, e cada uma expressa a relação entre dois neurônios de camadas distintas. Essa relação é definida pelo valor do peso atribuído à sinapse, que é ajustado após os exemplos serem computados na estrututa da rede. A estrutura básica de uma rede neural possui três camadas, sendo elas, de entrada, intermediária ou oculta, e camada de saída. Essa estrutura é chamada de *multilayer feed-forward* e as informações dos neurônios avançam no sentido da camada de entrada até a camada de saída por meio das sinapses (HAN; PEI; KAMBER, 2012). A topologia de uma rede neural *multilayer feed-forward* é apresentada na Figura 6.

Em redes neurais o método básico de aprendizado é por meio do algoritmo de *backpropagation*, em que para cada entrada inserida na rede os pesos são atualizados, com o objetivo de minimizar o erro quadrático médio entre o valor produzido pela rede e o valor verdadeiro da variável de saída. O erro associado a cada neurônio é calculado e propagado pela rede no sentido da camada de saída até a camada de entrada, isto é, no sentido contrário ao qual a informação inicialmente flui. Esse erro de *backpropagation* é utilizado para calcular os ajustes

Figura 5: Características das funções de ativação.

Logística	$g(a) = \frac{1}{1 + e^{-a}}$	$(g \in (0,1))$	
Tangente Hiperbólica	$g(a) = \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}}$	$(g \in (-1,1))$	
ReLU	$g(a) = max(0, a)$	$(g \in (0, a))$	
Softmax	$g(A_i) = \frac{e^{A\,i}}{\sum_{j=1}^{n} e^{A\,j}}$; $A = A_1, \ldots, A_n$; $i = 1, \ldots, n$	$(g \in (0, 1))$	

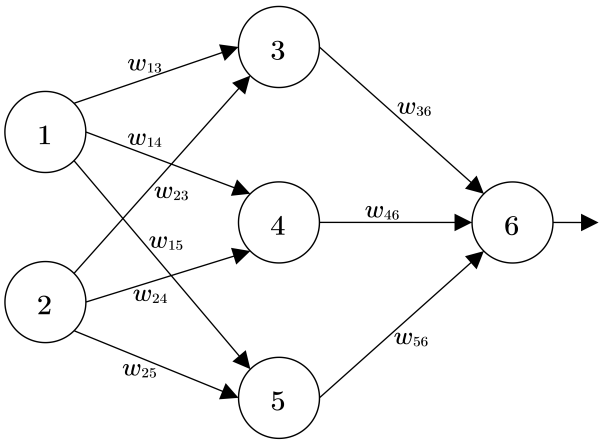
Fonte: Adaptado de (BISHOP, 2006; BROWNLEE, 2021).

que, então, são aplicados para corrigir os valores de peso nas sinapses e de viés nos neurônios. Para encontrar o conjunto de pesos que se ajuste aos dados e minimize a função de perda o método de otimização de gradiente descendente, geralmente, é o utilizado (BISHOP, 2006; HAN; PEI; KAMBER, 2012).

Outro método de aprendizado, e que tende a ser mais vantajoso que o anterior, é o proposto por Huang, Zhu e Siew (2004) chamado de máquina de aprendizado extremo (ELM), em que somente cálculos matriciais são utilizados para encontrar os parâmetros da rede. No ELM, para a camada intermediária, os pesos de entrada são definidos aleatoriamente e os pesos de saída definidos analiticamente. Em relação ao *backpropagation*, o ELM pode encontrar os pesos mais rapidamente, sem a necessidade de épocas para minimizar a função de perda, isto é, evitando a dependência de passar múltiplas vezes todos os exemplos disponíveis para treino pela rede. Além disso, tende a atingir menores erro de treino e norma dos pesos o que geralmente melhora o desempenho. Em termos práticos a rede ELM pode ser treinada mais rapidamente e obter melhor desempenho de generalização (HUANG; ZHU; SIEW, 2004).

A ideia do método é de que ao definir valores arbitrários para o viés dos neurônios e o peso das sinapses de entrada, ambos da camada intermediária, a determinação dos pesos nas

Figura 6: Estrutura topológica de uma RNA *multilayer feed-forward*. Na rede os neurônios 1 e 2 formam a camada de entrada, eles representam as variáveis de entrada do problema e o valor de cada variável numa amostra é a saída de cada um deles, por isso podem ser interpretados como neurônios passivos. A segunda camada, composta pelos neurônios 3, 4, e 5, é a intermediária, na qual geralmente são utilizadas funções não lineares. E o neurônio 6 forma a camada de saída, que produz a resposta da rede à entrada, essa resposta representa a variável objetivo do problema.



Fonte: O autor.

sinapses de saída possa ser realizada analiticamente por meio de matrizes e a solução de um sistema linear. Isto implica, que os pesos nas sinapses de entrada não precisam ser corrigidos e os pesos nas sinapses de saída podem ser calculados. O treinamento da rede se resume em encontrar a solução de quadrados mínimos com a menor norma para o sistema linear $\mathbf{H}\boldsymbol{\beta} = \mathbf{T}$ (HUANG; ZHU; SIEW, 2004).

Nas Equações 12, 13 e 14, $\mathbf{H}$ é matriz das saídas da camada intermediária num formato Nxn em que cada coluna relaciona um neurônio desta camada com todas as amostras de treino, $\boldsymbol{\beta}$ é a matriz de pesos das sinapses de saída da mesma camada com dimensão nxm, e $\mathbf{T}$ é a matriz dos valores verdadeiros ou ideais para os neurônios da camada de saída na forma Nxm. Ademais, N é a quantidade de amostras para treino, n é o número de neurônios na camada intermediária, m é o número de neurônios na camada de saída, $g(\ )$ é a função de ativação, $\mathbf{x}_i = [x_{i1}, ..., x_{ik}]^T$ é o vetor de valor dos k neurônios de entrada para a amostra i, $\mathbf{t}_i = [t_{i1}, ..., t_{im}]^T$ é o vetor de valor ideal dos m neurônios de saída em decorrência do valor verdadeiro da variável objetivo na amotra i, $\mathbf{W}_j = [W_{j1}, ..., W_{jk}]^T$ é o vetor de pesos de entrada do neurônio intermediário j, $\boldsymbol{\beta}_j = [\beta_{j1}, ..., \beta_{jm}]^T$ é o vetor de pesos de saída de j, b_j é o viés associado, e $\mathbf{W}_j$. $\mathbf{x}_i$ é o produto interno da relação (HUANG; ZHU; SIEW, 2004).

$$\mathbf{H} = \begin{bmatrix} g(\mathbf{W}_1.\mathbf{x}_1 + b_1) & ... & g(\mathbf{W}_n.\mathbf{x}_1 + b_n) \\ : & ... & : \\ g(\mathbf{W}_1.\mathbf{x}_N + b_1) & ... & g(\mathbf{W}_n.\mathbf{x}_N + b_n) \end{bmatrix} \tag{12}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1^T \\ : \\ \boldsymbol{\beta}_n^T \end{bmatrix} \tag{13}$$

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^T \\ : \\ \mathbf{t}_N^T \end{bmatrix} \tag{14}$$

Uma parte da solução é obtida por meio de uma operação inversa na matriz de sinapses de saída $\mathbf{H}$, que produz a matriz inversa generalizada de Moore-Penrose ${}^{+}\mathbf{H}$. A solução de menor norma dentre as soluções de quadrados mínimos para o sistema $\mathbf{H}\boldsymbol{\beta} = \mathbf{T}$ é única, e tem a forma $\boldsymbol{\beta} = {}^{+}\mathbf{HT}$ (HUANG; ZHU; SIEW, 2004). Essa é a solução que consegue atingir o erro de treino mínimo e as menores normas dos pesos (HUANG; ZHU; SIEW, 2004). Então, para encontrar o valor dos pesos de saída $\boldsymbol{\beta}$ da camada intermediária basta calcula-los utilizando a multiplicação matricial ${}^{+}\mathbf{HT}$, e o treinamento é concluído.

O algoritmo ELM pode ser descrito em três passos em relação à camada intermediária (HUANG; ZHU; SIEW, 2004):

1) Definir arbitrariamente os valores de peso das sinapses de entrada e viés dos neurônios.

2) Calcular a matriz de saída $\mathbf{H}$.

3) Calcular a matriz de pesos $\boldsymbol{\beta} = {}^{+}\mathbf{HT}$.

Diferentemente dos algoritmos de aprendizado baseados em gradiente que enfretam questões como mínimo local, taxa de aprendizado inadequada e sobreajuste, o ELM tende a chegar nas soluções de forma direta sem tais dificuldades (HUANG; ZHU; SIEW, 2004).

Apesar da capacidade de encontrar a solução ótima global e de sua velocidade o ELM ainda é passível de sobreajuste. Por exemplo, o desempenho de generalização do algoritmo tende a diminuir quando a camada intermediária é muito pequena ou muito grande (HUANG; ZHU; SIEW, 2006). O sobreajuste pode ocorrer se uma quantidade excessiva de neurônios for definida (LAI *et al.*, 2020). Um dos métodos utilizados para previnir o sobreajuste é a regularização, sendo as estratégias de L1 (*lasso regression*) ou L2 (*ridge regression*) geralmente utilizadas (LAI *et al.*, 2020).

Alguns dos parâmetros que podem ser definidos na rede antes do treino são o número de neurônios na camada intermediária, a função de ativação desta, a densidade das conexões, a quantidade de neurônios na camada de saída e sua respectiva função de ativação.

2.4.4.3 Floresta aleatória (RF)

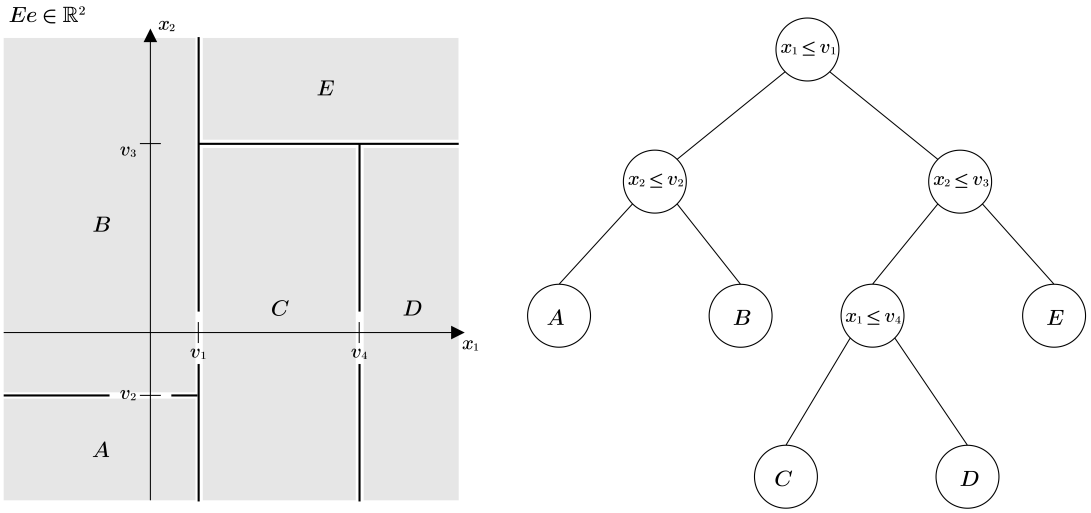
A floresta aleatória (RF) é um método de *ensemble* em que árvores de decisão são construídas a partir de subconjuntos aleatórios. Cada árvore é treinada em um conjunto aleatório de amostras da base de treino. Na construção de uma árvore, a cada nó, um subconjunto das variáveis de entrada é selecionado aleatoriamente para definir qual será atribuída ao nó. Esse procedimento é reproduzido para cada árvore do conjunto, que, por sua vez, obtém a predição final de uma amostra considerando a resposta de cada árvore à entrada. Para a classificação a predição é determinada pelo resultado que ocorrer o maior número de vezes, isto é, o mais indicado pelas árvores. Para regressão a predição é determinada pela média dos resultados.

Em uma árvore de decisão sua estrutura básica segue o modelo topológico de um grafo acíclico com cada nó possuindo somente uma conexão com o antecessor. Na estrutura, o nó inicial é a raiz e os nós finais são as folhas. Cada nó intermediário e a raiz representam uma decisão sobre uma variável do conjunto de entrada utilizado para construir a árvore. Cada decisão produz dois nós, e indica uma separação de regiões no espaço definido pelas variáveis de entrada. Os nós folha representam as regiões específicas relacionadas aos possíveis valores da variável objetivo (BISHOP, 2006). Dois nós folha podem ter relação com o mesmo valor da variável objetivo, porém cada nó é o resultado de um caminho de decisão único através da árvore. A estrutura de árvore de decisão e o espaço definido pelas variáveis entrada são representados na Figura 7.

A estrutura da árvore é definida durante o treinamento. Para isso, em cada nó são selecionadas algumas das variáveis de entrada para formar o critério de separação, no qual o valor do limiar e a variável de separação são definidos. Com desenvolvimento da árvore os valores da variável objetivo dentro de cada região são determinados (BISHOP, 2006). Para construir a árvore de decisão o método CART é um dos que podem ser utilizados, assim como o ID3 e o C4.5. Esses métodos utilizam como estratégia um algoritmo guloso (míope) para construir a árvore de forma recursiva de cima para baixo (raiz até as folhas), em que a cada etapa de decisão o conjunto de treino é subdivido em conjuntos menores (regiões) (BISHOP, 2006; HAN; PEI; KAMBER, 2012). Cada nó, incluindo a raiz, corresponde a uma subdivisão do conjunto de treino (região).

Para o critério de separação, cada combinação de variável de entrada e limiar são

Figura 7: Estrutura de um árvore de decisão e as regiões do espaço Ee definido pelas variáveis de entrada separadas pelos valores de limiar v_i.



Fonte: Adaptado de (BISHOP, 2006).

avaliados de acordo com um medidor de desempenho, que considera o valor da variável objetivo em cada possível subregião. Então, a variável de entrada e o limiar que produzem a melhor separação para a região são os escolhidos (BISHOP, 2006; HAN; PEI; KAMBER, 2012). O critério de separação pode ser resumido como uma busca para encontrar a variável de entrada e o limiar que melhor discriminam uma região dado os valores da variável objetivo na região, seguida da determinação sobre repetir ou não o procedimento em cada subregião (HAN; PEI; KAMBER, 2012). Para o CART o critério de separação produz duas regiões, isto é, dois nós folha são adicionados na árvore a cada decisão (BISHOP, 2006; HAN; PEI; KAMBER, 2012). O crescimento da árvore é interrompido, e o nó definido como folha, se ocorrer um dos critérios de parada, como o valor mínimo ou limite em um medidor de desempenho, a quantidade mínima de amostras em uma região, ou a profundidade máxima da árvore (BISHOP, 2006; HAN; PEI; KAMBER, 2012; STACKEXCHANGE, 2013).

Em problemas de classificação dois medidores de desempenho geralmente utilizados para o critério de separação são o índice Gini e a entropia cruzada. Esses medidores são descritos nas Equações 15 e 16 (BISHOP, 2006; HAN; PEI; KAMBER, 2012).

$$Q(D) = 1 - \sum_{i=1}^{K} p_i^2 \tag{15}$$

$$Q(D) = -\sum_{i=1}^{K} p_i log_2(p_i) \tag{16}$$

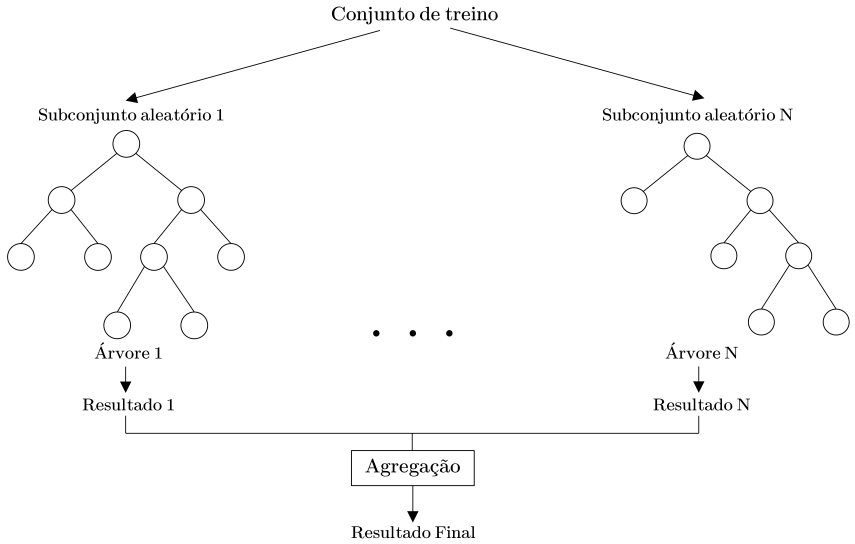
Nas equações, D é um conjunto de amostras de treino dentro da região do espaço definida pelas variáveis de entrada e os limiares de separação, K é a quantidade de classes, e p_i é a proporção de amostras que pertencem a classe i no conjunto D. O índice Gini tem valores no intervalo de $[0, 0{,}5]$, e a entropia no intervalo de $[0, 1]$.

Na prática, uma árvore pode ser configurada para ter uma profundida máxima igual a quantidade de variáveis de entrada, caso o critério de separação impessa que uma variável seja escolhida mais de uma vez (HAN; PEI; KAMBER, 2012). Da mesma forma, é possível que tenha a profundidade de $N - 1$, onde N é o número de amostras de treino (BROWNLEE, 2016; STACKEXCHANGE, 2013). No primeiro caso a árvore pode estar subajustada e no segundo sobreajustada para as amostras de treino de um problema. Uma das estratégias para minimizar o sobreajuste é a poda, em que indicadores de desempenho são utilizados para retirar nós folha ou subárvores, isto é, combinar regiões separadas (BISHOP, 2006; BROWNLEE, 2016; HAN; PEI; KAMBER, 2012).

O RF é um método que utiliza a estratégia de *bagging*, em que cada árvore é construída em um conjunto de mesmo tamanho ou menor, selecionado aleatoriamente e com reposição, dado o conjunto das amostras de treino. Esta estratégia de amostragem é chamada de *bootstrap*, e em média deixa aproximadamente um terço das amostras fora de cada conjunto selecionado para treinamento com mesmo tamanho que o original (HAN; PEI; KAMBER, 2012; QIU; WANG, 2017). Durante o treinamento cada árvore é construída paralelamente, utilizando em cada nó um subconjunto aleatório das variáveis de entrada para o critério de separação. O resultado final é um concenso do resultado de cada árvore, que para problemas de classificação é o resultado mais indicado, e para problemas de regressão é a média dos resultados. O CART geralmente é o algoritmo utilizado para construir as árvores, de forma a atingir o tamanho máximo e sem poda. Uma maneira direta de tratar o sobreajuste é aumentar a quantidade de árvores, uma vez que o erro de generalização da floresta tende a diminuir enquanto o número de árvores for grande o suficiente (HAN; PEI; KAMBER, 2012). Na Figura 8 o método é resumido.

Alguns dos parâmetros que podem ser definidos antes do treino são a quantidade de árvores, a profundidade máxima, o medidor de desempenho, a quantidade de variáveis selecionadas para o critério de separação, e o tamanho do conjunto de amostras utilizado para treinar as árvores.

Figura 8: Operação do método de floresta aleatória. Que pode ser definido como um método de *ensemble* que utiliza a estratégia de *bagging* (*bootstrap* do conjunto de treino e *aggregating* dos resultados).



Fonte: Adaptado de (VOSS; STEVAN JUNIOR; AYUB, 2019).

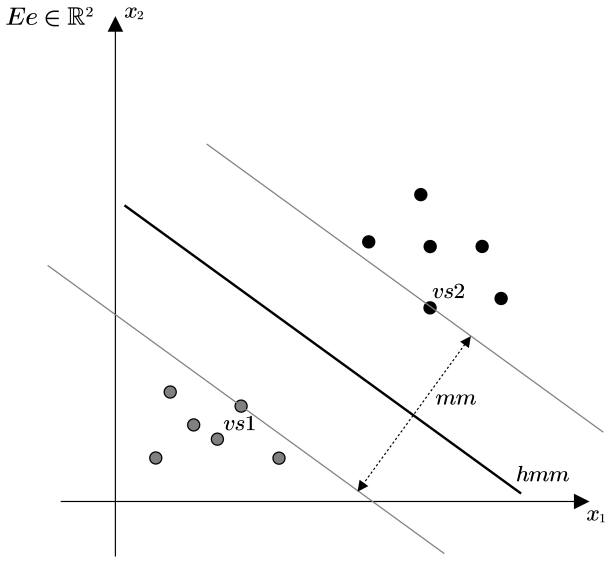
2.4.4.4 Máquina de vetores de suporte (SVM)

Na máquina de vetores de suporte (SVM) hiperplanos de dimensão superior à dimensão do espaço de características são utilizados para descrever relações não lineares no espaço original. As curvas descritas separam as amostras no espaço original de acordo com os valores da variável objetivo. Vetores de suporte definem a relação entre cada hiperplano e as amostras no espaço de dimensão superior, isto é, a direção e o local adequado de separação das amostras dado os valores das variáveis de entrada e da variável objetivo. Uma vez que a transformação do espaço é um mapeamento das variáveis de entrada, uma função é implementada possibilitando representações no espaço de maior dimensão de cálculos no espaço original. Isso evita a transformação dos dados, simplificando os cálculos. Em problemas de classificação o hiperplano define o formato da fronteira das regiões onde estão a maioria das amostras de cada classe. Para problemas de regressão o hiperplano é utilizado para descrever o formato da curva de regressão.

Em relação a uma direção, os vetores de suporte são as amostras que estão mais próximas no espaço definido pelas variáveis de entrada e possuem valor diferente para a variável objetivo. A margem é determinada pela distância entre os hiperplanos que passam por cada vetor de suporte (HAN; PEI; KAMBER, 2012). O hiperplano de separação é perpendicular à direção da distância entre os vetores de suporte e atravessa seu centro. A Figura 9 exemplifica os vetores de suporte e seu hiperplano de separação.

Para separar um espaço em regiões o SVM tenta encontrar o hiperplano de margem máxima, isto é, o hiperplano que maximiza a distância entre dois vetores de suporte. As Equações 17 e 18 descrevem um hiperplano qualquer e a inequação dos vetores de suporte, respec-

Figura 9: Vetores de suporte (vs), hiperplano de margem máxima (hmm), margem máxima (mm), e amostras no espaço Ee definido pelas variáveis de entrada x_1 e x_2.



Fonte: Adaptado de (HAN; PEI; KAMBER, 2012)

tivamente (BISHOP, 2006; HAN; PEI; KAMBER, 2012). Seja $\mathbf{W}$ o vetor de pesos com cada valor associado a uma variável de entrada, a margem máxima de um hiperplano tem o valor $2/||\mathbf{W}||$, e maximizar esse valor é o equivalente a minimizar $2||\mathbf{W}||^2$. Essa busca compreende um problema de otimização que pode ser descrito como a minimização de um função quadrática dado um conjunto de restrições de desigualdade linear, e é representada pela Equação 19 (BISHOP, 2006; HAN; PEI; KAMBER, 2012).

$$y(\mathbf{X}) = \mathbf{W}^T\mathbf{X} + b \tag{17}$$

$$y_{Oi}(\mathbf{W}^T\mathbf{X}_i + b) \geq 1, \quad i = 1, ..., N. \tag{18}$$

$$\underset{\mathbf{W},\, b}{\arg\min}\ ||\mathbf{W}||^2 \tag{19}$$

Nas equações descritas, $y(\mathbf{X})$ é o valor da equação, $\mathbf{W}^T = [w_1, ..., w_n]$ é o vetor de pesos, $\mathbf{X} = [x_1, ..., x_n]^T$ é o vetor das n variáveis de entrada, b é o viés, $y_{Oi} \in \{-1, 1\}$ é o valor da variável objetivo na amostra i, e $\mathbf{X}_i$ é o vetor das variáveis de entrada da mesma amostra, e

N é a quantidade de amostras de treino.

Para encontrar os vetores de suporte e o hiperplano de margem máxima multiplicadores de Lagrange são utilizados e a Equação 19 é reformulada como uma função lagrangiana. As soluções são encontradas usando as condições de Karush-Kuhn-Tucker (KKT), que estabelecem formas de verificar quais pontos no espaço são vetores de suporte do hiperplano de margem máxima (BISHOP, 2006; HAN; PEI; KAMBER, 2012).

Em problemas em que os dados não são linearmente separáveis uma abordagem simples é permitir amostras em locais incorretos em relação ao hiperplano de separação e os valores da variável objetivo. Para isso, uma variável de folga é utilizada para medir o erro em função da distância até o hiperplano, sendo então associada à amostra. Dessa forma, o conjunto de variáveis de folga é considerado na definição das restrições e durante os cálculos para encontrar os vetores de suporte (BISHOP, 2006). Porém, para descrever curvas e separar as amostras de forma mais adequada é necessário utilizar um espaço de maior dimensão que o original. Isso possibilita que as considerações lineares de vetores de suporte e hiperplanos possam ser utilizadas no espaço de dimensão superior, o que acaba descrevendo curvas e vetores de suporte no espaço original. Nessa abordagem, a primeira etapa é a transformação das variáveis de entrada para um espaço de dimensão superior usando um mapeamento não linear. A segunda etapa é encontrar o hiperplano de separação neste espaço. Uma vez encontrado, seus parâmetros podem ser utilizados no espaço original, já que as variáveis do espaço superior são descritas pelas variáveis originais (HAN; PEI; KAMBER, 2012).

A relação entre os dois espaços é realizada por uma função de núcleo, que faz a execução de cálculos com parâmetros que pertencem ao espaço original e a representação de sua saída no espaço de dimensão superior. Essa função é calculada em pares de amostras em um determinado espaço. Para os cálculos no espaço superior as amostras de treino são computadas na forma de um produto escalar entre dois vetores, $\phi(\mathbf{X}_i)^T\phi(\mathbf{X}_j)$, onde $\phi(\ )$ é a função de mapeamento não linear usada para transformar as amostras, e $\mathbf{X}_i$ e $\mathbf{X}_j$ são amostras ou pontos no espaço original (BISHOP, 2006; HAN; PEI; KAMBER, 2012).

A função de núcleo $K(\mathbf{X}_i, \mathbf{X}_j)$ produz o mesmo resultado que o produto escalar dos vetores no espaço superior, porém tem como parâmetros os vetores no espaço original (HAN; PEI; KAMBER, 2012). Ou seja, uma equivalência sem a necessidade de transformar os dados para uma dimensão superior e executar os cálculos nessa dimensão. De fato, os dados não precisam ser transformados apenas representados na dimensão superior pela função de núcleo. Dessa forma, todos os cálculos são executados na dimensão original. Então, para o treinamento o produto escalar é substituído pela função de núcleo, que exige menos cálculos (HAN; PEI; KAMBER,

2012). Essa equivalência entre os espaços é representada na Equação 20 (BISHOP, 2006; HAN; PEI; KAMBER, 2012). Diferentes tipos de funções de núcleo podem ser empregadas, sendo, geralmente, do tipo linear, polinomial, base radial ou sigmoide. Essas funções são definidas nas Equações 21, 22, 23 e 24, respectivamente (HAN; PEI; KAMBER, 2012).

$$K(\mathbf{X}_i, \mathbf{X}_j) = \phi(\mathbf{X}_i)^T \phi(\mathbf{X}_j) \tag{20}$$

$$K(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{X}_i^T \mathbf{X}_j \tag{21}$$

$$K(\mathbf{X}_i, \mathbf{X}_j) = (\gamma \mathbf{X}_i^T \mathbf{X}_j + r)^d \tag{22}$$

$$K(\mathbf{X}_i, \mathbf{X}_j) = e^{-\frac{||\mathbf{X}_i - \mathbf{X}_j||^2}{2\sigma^2}} \tag{23}$$

$$K(\mathbf{X}_i, \mathbf{X}_j) = tanh(\gamma \mathbf{X}_i^T \mathbf{X}_j + r) \tag{24}$$

Para problemas de classificação a Equação 25 é usada para predizer a classe de uma amostra, considerando o SVM após o treinamento. Na equação, $y(\mathbf{X})$ é o valor da equação, S é a quantidade de vetores de suporte, a_i é o multiplicador de Lagrange associado ao vetor de suporte i, y_{Oi} é o valor da variável objetivo para o vetor de suporte i, $K(\mathbf{X}, \mathbf{X}_i)$ é a função de núcleo para o vetor de entrada $\mathbf{X}$ e o vetor de suporte $\mathbf{X}_i$, e b é o viés.

$$y(\mathbf{X}) = \sum_{i=1}^{S} a_i y_{Oi} K(\mathbf{X}, \mathbf{X}_i) + b \tag{25}$$

Em problemas de múltiplas classes, estratégias diferentes podem ser utilizadas. Duas

comumente usadas são a de um-contra-todos (OVR) e de um-contra-um (OVO). Na abordagem OVR, para cada classe uma SVM é construída e tenta-se encontrar o hiperplano de separação entre a classe atual e as demais. Se m é a quantidade de classes, então são treinados m classificadores binários. Para classificar uma amostra um *ensemble* é implementado de modo que a resposta de cada classificador é considerada ou não de acordo com a predição do classificador e o hiperplano de separação. Na abordagem OVO um classificador é treinado para cada possível par de classes. A quantidade de classificadores binários é de $m(m-1)/2$. Para classificar uma amostra cada SVM tem sua resposta considerada e a mais indicada é escolhida (BISHOP, 2006; HAN; PEI; KAMBER, 2012).

O SVM tende a construir modelos precisos devido sua capacidade em descrever formas não lineares, e conseguir encontrar o máximo global da função de otimização durante o treinamento (HAN; PEI; KAMBER, 2012). Apesar disso, a questão do tempo necessário para treino ainda é uma desvantagem do método. Alguns dos parâmetros que podem ser definidos antes do treino são a função de núcleo, os parâmetros da função, e o valor de regularização.

2.4.5 Avaliação de Desempenho

Para avaliar o desempenho dos modelos e selecionar aqueles que apresentam os melhores resultados diferentes métodos podem ser usados. Os métodos de *hold-out* e validação cruzada são comumente empregados. No *hold-out* a base é dividida entre treino e teste, em que a base de treino é utilizada na construção do modelo e a base de teste para avaliar seu desempenho. Na validação cruzada a base é dividida em K partes e cada divisão é utilizada uma vez com base de teste, então são avaliados K modelos e o desempenho médio é medido. Uma forma de tratar classes com quantidades diferentes de amostras, isto é, o desbalaceamento em problemas de classificação, é empregar a estratificação na validação cruzada. Desse modo cada uma das K divisões mantêm a proporção de amostras de cada classe nas K bases de treino e teste geradas, produzindo modelos que podem ser avaliados mais adequadamente de acordo com métricas de desempenho que consideram o desbalanceamento.

A estratégia de *hold-out* é dependende da divisão realizada na base. Um resultado melhor ou pior pode ser obtido de acordo a formação das bases de treino e teste. Dessa forma a divisão feita na base influencia no resultado que pode melhorar ou piorar de acordo com a separação. Por outro lado, a validação cruzada tende a um resultado mais geral sobre a base de dados, uma vez que cada parte é utilizada como teste para um modelo, o que reduz a influência da divisão sobre o resultado final. Por exemplo, dado uma base, a validação cruzada k-*fold* tende apresentar uma estimativa geral com viés da métrica utilizada levemente pessimista em comparação com um modelo treinado na mesma base e avaliado numa base de teste (CAWLEY;

TALBOT, 2010). Porém, quando tenta-se otimizar os hiperparâmetros de um modelo na validação cruzada o resultado tende a um viés otimista devido ao protocolo de avaliação.

2.4.5.1 Otimização de hiperparâmetros

O protocolo de avaliação utilizado na otimização compreende a seleção de modelo e a estimativa de desempenho. Na seleção a otimização é executada, maximizando ou reduzindo um critério de seleção (métrica) sobre uma base de validação, os modelos são treinados com diferentes combinações de hiperparâmetros numa base de treino, e o modelo com melhor resultado é selecionado. Na estimativa de desempenho, o modelo selecionado tem o desempenho estimado sobre a base de validação. Esse protocolo de avaliação é propenso à sobreajuste durante a seleção de modelo e, consequentemente, a um maior viés durante a estimativa de desempenho.

O sobreajuste na seleção de modelo ocorre quando o critério de seleção continua a ser aprimorado, aumentando ou diminuindo dependendo da métrica, mas o desempenho de generalização começa a cair (CAWLEY; TALBOT, 2010). A causa do sobreajuste é que um modelo com determinada configuração de hiperparâmetros pode explorar as peculiaridades da base em que ele é avaliado, ajustando os valores dos hiperparâmetros à base de validação utilizada (BROWNLEE, 2020; CAWLEY; TALBOT, 2010; SCIKIT-LEARN, 2022b). Por exemplo, ao mudar a base de validação o mesmo modelo pode obter um resultado diferente para o critério de seleção. Mais precisamente, para um conjunto de modelos configurados com diferentes combinações de hiperparâmetros o modelo que consegue o maior aprimoramento do critério tende a variar conforme a base de validação, estando tal variação relacionada diretamente com o sobreajuste do critério de seleção, isto é, quanto maior a variação maior tende a ser o impacto do sobreajuste sobre o protocolo de avaliação (CAWLEY; TALBOT, 2010).

Dessa forma, o modelo com os valores de hiperparâmetros que sobreajustaram a base de validação tende a ser o escolhido devido o melhor resultado. Como consequência, os valores encontrados na otimização podem não ser efetivamente os que apresentam o melhor desempenho de generalização para o modelo. Além disso, o sobreajuste do critério de seleção pode provocar tanto o subajuste quanto o sobreajuste de um modelo numa base de treino (CAWLEY; TALBOT, 2010). Seu impacto na seleção de modelo é relativo ao tamanho da base de validação (quanto menor a base maior o grau de sobreajuste) e ao número de hiperparâmetros (quanto maior a quantidade de hiperparâmetros maior o grau de sobreajuste) (CAWLEY; TALBOT, 2010).

Mesmo que não ocorra sobreajuste durante a seleção esse protocolo de avaliação tende a um viés otimista durante a estimativa de desempenho, já que a mesma base de validação

é usada para otimização e posteriormente para estimativa de desempenho (CAWLEY; TALBOT, 2010; SCIKIT-LEARN, 2022b). Então, o sobreajuste na seleção de modelo pode gerar um viés ainda mais otimista (BROWNLEE, 2020; CAWLEY; TALBOT, 2010; SCIKIT-LEARN, 2022b). Portanto, um protocolo de avaliação de desempenho que minimiza o viés é o mais adequado na otimização dos hiperparâmetros. Dessa forma, qualquer erro devido o sobreajuste na seleção de modelo é considerado, o que possibilita uma estimativa de desempenho com o menor viés possível (CAWLEY; TALBOT, 2010). Assim, resultando na escolha adequada dos valores dos hiperparâmetros.

2.4.5.2 Validação cruzada aninhada

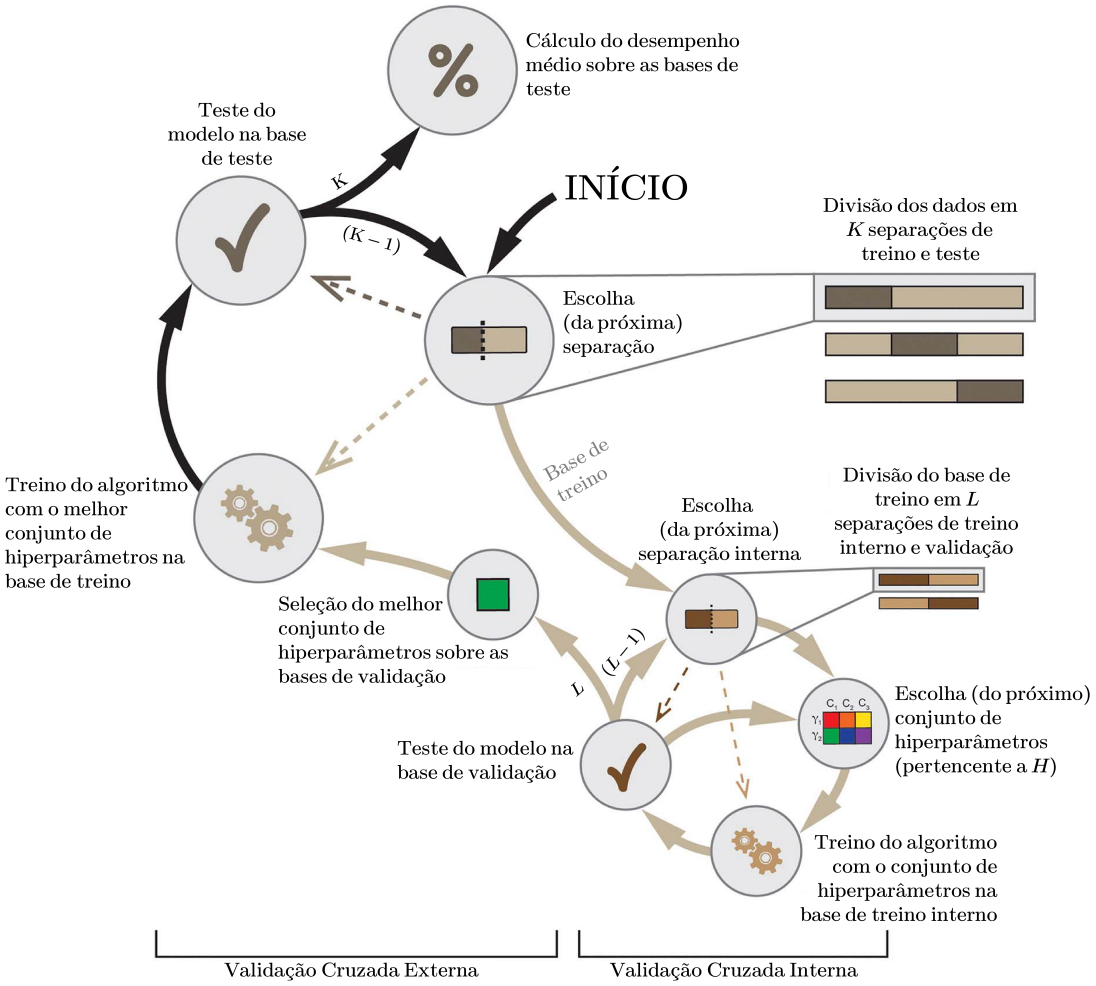
A validação cruzada utilizada para otimização diminuí a chance de sobreajuste durante a seleção devido as diferentes separações, mas ainda apresenta viés de estimativa. Uma alternativa para minimizar o víes é executar a seleção de modelo de forma independente, e avaliar o modelo selecionado numa base de teste com amostras não utilizadas na base de validação (CAWLEY; TALBOT, 2010). A validação cruzada aninhada realiza esse procedimento por meio de uma validação cruzada externa, que separa uma base em bases de treino e teste, e uma série de validações cruzadas internas executadas em cada base de treino. Desse modo, uma seleção de modelo é feita em cada validação cruzada interna e tem seu desempenho avaliado na respectiva base de teste da validação cruzada externa. O resultado final é uma avaliação do desempenho de generalização dos melhores modelos e suas respectivas configurações de hiperparâmetros (BROWNLEE, 2020; SCIKIT-LEARN, 2022b).

Para a validação cruzada aninhada baseada na validação cruzada k-*fold* as seguintes etapas são realizadas. Na validação cruzada externa as separações são realizadas produzindo as K bases de treino e teste. Em seguida, na base de treino de cada separação, é executada a validação cruzada interna, dividindo a base em L separações de bases de treino interno e validação. Assim, em cada validação cruzada interna é executada a seleção de modelo. Na seleção, em cada base de treino interno, os modelos são configurados com os hiperparâmetros e a busca pela combinação de valores que otimiza o critério de seleção na respectiva base de validação é feita. Para completar a seleção, a configuração de hiperparâmetros que produzir a melhor média considerando os resultados das L separações internas é escolhida. Então, um modelo configurado com os valores dos hiperparâmetros é construído utilizando a base de treino externa, isto é, todas as amostras que foram empregadas na validação cruzada interna. Por fim, seu desempenho de generalização é avaliado na base de teste da validação externa.

Esse procedimento é reproduzido para cada uma das K separações da validação externa, e o resultado final dos K modelos e suas configurações é avaliado. Durante o procedi-

mento considerando uma busca exaustiva um total de KLH modelos são avaliados na otimização, onde H é a multiplicação entre a quantidade de valores dos hiperparâmetros otimizados. A Figura 10 apresenta o procedimento executado pela validação cruzada aninhada na otimização de hiperparâmetros.

Figura 10: Procedimento da validação cruzada aninhada na otimização de hiperparâmetros. Na figura, K e L são as quantidades de separações da validação cruzada externa e interna, respectivamente. H é a multiplicação entre a quantidade de valores dos hiperparâmetros.



Fonte: Adaptado de (WEAVERDYCK; LIEBERMAN; PARKINSON, 2020).

2.4.5.3 Considerações sobre a avaliação de desempenho

Uma consideração sobre a otimização utilizando a validação cruzada aninhada é que um equilíbrio deve ser feito para tentar evitar o sobreajuste na seleção de modelo e o sobreajuste de treino. Uma vez que aumentar o tamanho da base de validação de cada separação interna reduz a chance de sobreajuste de seleção, porém isso acaba reduzindo o tamanho da base de treino dessa separação tornando os modelos mais propensos ao sobreajuste no treino. Mesmo restringindo a quantidade e os valores dos hiperparâmetros esse fator pode ser um limitante da utilização dessa forma de otimização em análises de problemas com bases de dados pequenas.

Um método de avaliação de desempenho utilizado nas análises de classificação e regressão tende a sofrer a influência da separação da base total por *hold-out* e de qualquer possível sobreajuste na otimização. Porém, mesmo que ambos ocorram não devem ser impactantes na análise se seus efeitos são minimizados para cada estratégia de aprendizado. Por exemplo, cada algoritmo avaliado no mesmo espaço base, mesma divisão *hold-out*, e com as mesmas separações na validação cruzada aninhada. Embora cada algoritmo de aprendizado de máquina tenha seu comportamento dentro da otimização em relação ao sobreajuste de seleção e o grau de seu impacto, a possibilidade de sobreajuste tende a ser baixa para um conjunto pequeno de valores de hiperparâmetros, e, caso ocorra, tende a não influenciar a estimativa desempenho devido a validação cruzada aninhada.

Além disso, para abordar a possibilidade de sobreajuste de treino na otimização os valores dos hiperparâmetros podem ser restritos por um limite superior. Desse modo, apesar da influência da separação por *hold-out* no método de avaliação, a estratificação pela variável de saída e a validação cruzada aninhada tendem a gerar conjuntos representativos de uma base. Logo, espera-se que ambos reduzam a influência da separação por *hold-out*. Por fim, um método de avaliação de desempenho pode compreender inicialmente a otimização dos hiperparâmetros de cada técnica de aprendizado de máquina, seguido do uso dos valores encontrados para configurar os modelos nas demais etapas, em que o objetivo é avaliar o desempenho de generalização de cada modelo final utilizando a base de teste.

2.4.5.4 Métricas de desempenho

As métricas são as diferentes medidas utilizadas para aferir o desempenho de um modelo na predição das amostras. Para serem calculadas utiliza-se os valores verdadeiros da variável objetivo e os valores preditos por um modelo. Em problemas de classificação a acurácia e o indicador F_1 podem ser utilizados. A acurácia descreve a taxa de reconhecimento de um modelo, ou seja, o quão bem um classificador consegue acertar suas predições (HAN; PEI; KAMBER, 2012). Seu cálculo pode ser feito pela divisão entre a quantidade total de acertos e a quantidade total de amostras. O indicador F_β é uma métrica que representa a eficiência de predição por meio da relação entre a efetividade e a eficácia de predição de um modelo para cada classe. No seu cálculo são consideradas duas métricas, a precisão e o *recall*, e a relação é computada pela média harmônica entre elas. Se a precisão e o *recall* são considerados de igual importância seus pesos são equivalentes, então β = 1, e o indicador é chamado de F_1. As Equações 26 e 27 mostram os cálculos da acurácia e do indicador F_1, respectivamente (HAN; PEI; KAMBER, 2012).

$$Acuracia = \frac{TP + TN}{P + N} \tag{26}$$

$$F_1 = \frac{2PrRe}{Pr + Re} \tag{27}$$

Nas equações, cada métrica pode ser calculada pressupondo uma classe i em relação as demais, sendo VP a quantidade de amostras da classe i preditas corretamente, VN a quantidade de amostras das demais classes preditas corretamente, P a quantidade de amostras da classe i, N a quantidade de amostras das demais classes, FP a quantidade de amostras das demais classes preditas erroneamente como pertencentes a classe i, FN a quantidade de amostras da classe i preditas erroneamente como pertencentes as demais classes, $Pr = VP/(VP+FP)$ é a precisão, e $Re = VP/(VP + FN) = VP/P$ é o *recall* (HAN; PEI; KAMBER, 2012).

Para problemas em que há diferentes quantidades de amostras para cada classe, métricas adequadas devem ser empregadas de modo a tonar possível a consideração desse desbalanceamento na estimativa de desempenho de um modelo. A acurácia balanceada e o indicador F_1 ponderado são duas métricas que podem ser utilizadas. A acurácia balanceada evita que a estimativa seja afetada pelo desbalanceamento atribuindo para cada amostra um peso relativo à prevalência inversa de sua classe, ou seja, o inverso da quantidade de amostras. A métrica pode ser calculada pela acurácia levando em conta os pesos de cada amostra ou pela média entre os valores de *recall* de cada classe sem considerar os pesos (SCIKIT-LEARN, 2022a). A acurácia balanceada é apresentada na Equação 28, onde y_i é o valor verdadeiro da variável objetivo na amostra i, $(y_j = y_i)$ é a função indicador que retorna 1 ou 0 de acordo com a igualdade, $w_i = 1/\sum_j(y_j = y_i)$ é o peso associado a cada amostra i, yo_i é o valor predito da variável objetivo, e C é a quantidade de classes (SCIKIT-LEARN, 2022a).

De modo similiar, o indicador F_1 ponderado é a média ponderada dos valores de F_1 de cada classe. A métrica considera a quantidade de amostras de cada classe no cálculo, ponderando cada F_1 com sua respectiva quantidade de amostras. Dessa forma, uma estimativa mais apropriada da média dos valores de F_1 é realizada pela inclusão do desbalanceamento. O indicador F_1 ponderado é representado na Equação 29, onde N é a quantidade de amostras, C a quantidade de classes, N_c a quantidade de amostras da classe c, e F_{1c} é o indicador F_1 da classe c (SCIKIT-LEARN, 2022d).

$$Acuracia\ balanceada = \frac{1}{\sum w_i} \sum_i (yo_i = y_i) w_i = \frac{1}{C} \sum_{c=1}^{C} Re_c \tag{28}$$

$$F_1\ ponderado = \frac{1}{N} \sum_{c=1}^{C} N_c F_{1c} \tag{29}$$

2.4.5.5 Desempenho de generalização

A capacidade de generalização de um modelo é avaliada pelo desempenho obtido em dados não utilizados no treino, e pode ser compreendida como uma tentativa de aproximar o comportamento do modelo em uma situação de aplicação prática. O desempenho de generalização é definido pelas medidas realizadas na base de teste. Uma das avaliações utilizadas para medir o desempenho de generalização é a de sobreajuste, em que calcula-se a diferença entre o desempenho de treino e de teste para determinada métrica (BARBIERO; SQUILLERO; TONDA, 2020; GROSSE, 2018).

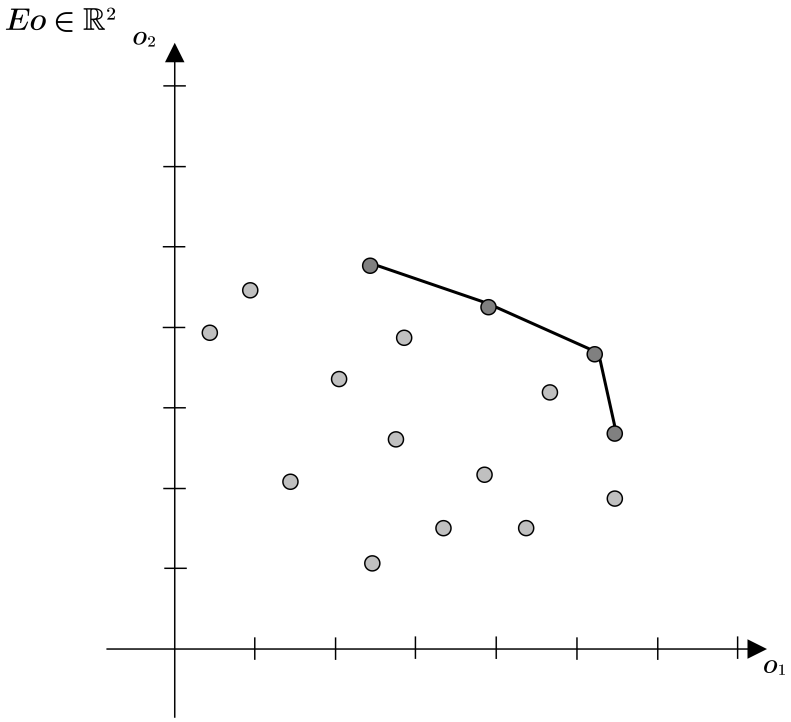
2.4.5.6 Comparação e seleção de modelos

A seleção final de modelo visa comparar os modelos por meio dos indicadores de desempenho e selecionar aqueles com os melhores resultados, isto é, o modelo ou um grupo de modelos mais adequados de acordo com as métricas utilizadas para medir o desempenho de generalização. Uma estratégia que auxilia na seleção é a análise multiobjetivo, uma vez que possibilita um conjunto de métricas necessárias sejam inclusas numa mesma avaliação comparativa. Nessa análise cada métrica avaliada é um objetivo e define uma dimensão do espaço, que por sua vez tem dimensão M correspondente a quantidade de métricas incluídas na avaliação. No espaço dos objetivos um modelo tem sua localização determinada pelos valores em cada uma das métricas.

Uma vez definidos o espaço dos objetivos e o conjunto de modelos é possível estabelecer a fronteira de Pareto, que é formada por todos os modelos que não podem mais ser melhorados em nenhum dos objetivos sem perder desempenho em algum outro (RAIMUNDO, 2018). Um modelo na fronteira é chamado de candidato ótimo de Pareto e representa uma solução eficiente dado os objetivos (RAIMUNDO, 2018). A Figura 11 exemplifica a análise multiobjetivo e a fronteira de Pareto entre modelos.

A fronteira de Pareto é útil para comparar e visualizar a relação de custo-benefício entre os modelos dado as métricas (BARBIERO; SQUILLERO; TONDA, 2020). Dessa forma, os modelos com melhor desempenho geral podem ser selecionados. No entanto, questões como as métricas adequadas para comparar modelos de diferentes espaços base e a interpretação correta dos modelos no espaço objetivo devem ser levadas em consideração.

Figura 11: Exemplo de modelos no espaço objetivo de maximização e a fronteira de Pareto (linha preta) composta pelos modelos Pareto eficientes (cinza escuro).



Fonte: Adaptado de (BARBIERO; SQUILLERO; TONDA, 2020).

3 MATERIAIS E MÉTODOS

Neste capítulo estão apresentados os materiais utilizados no trabalho para o desenvolvimento do protótipo, e os métodos empregados, descrevendo os procedimentos utilizados no pré-processamento e processamento.

3.1 MATERIAIS

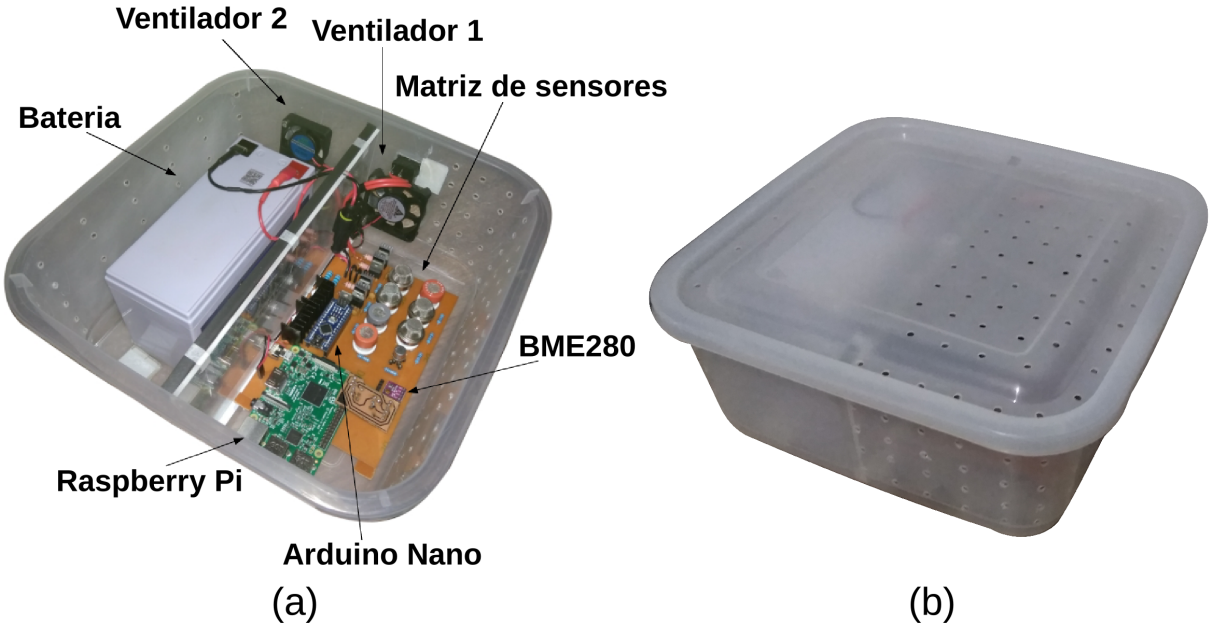
O protótipo de nariz eletrônico foi composto por uma ventoinha que tem como função puxar o ar e os COVs para dentro de uma câmara onde a matriz de sensores está alocada. O dispositivo em si, é composto por 8 sensores químicos comerciais para captura de concentrações de diferentes gases como componentes principais da matriz de sensores, selecionados a partir de uma análise efetuada na revisão de literatura sobre os COVs emitidos pelas flores de pessegueiros e também da disponibilidade comercial dos mesmos. Os sensores utilizados e seus gases sensíveis são apresentados no Quadro 4. Adicionalmente, o dispositivo possui um sensor para monitoramento da temperatura e umidade relativa do ar no interior da camara de aquisição. Para este fim foi utilizado o sensor BME280.

Para realizar a aquisição dos dados dos sensores, foi utilizado um microcontrolador ATmega328, baseado na plataforma de prototipagem rapida Arduino Nano. Já para o armazenamento dos dados e leitura do sensor de temperatura e umidade foi utilizado o microcomputador de placa única Raspberry Pi 3-B. A comunição entre o Raspberry e o Arduino nano é realizada por meio do protocolo serial UART. Duas mini ventoinhas foram utilizados, um para circulação contínua de gases durante a amostragem, e a outra para a ventilação da bateria. A alimentação elétrica do protótipo foi realizada por uma bateria selada de 12 V com capacidade nominal de 5,2 Ah/1 h. Uma caixa de plástico e suportes de acrílico foram utilizados para alocar a circuitaria eletrônica do dispositivo e também a bateria. Na Figura 12 o protótipo desenvolvido é exibido. O Quadro 5 apresenta os principais componentes do protótipo, suas respectivas funções, e principais características.

Quadro 4: Sensores selecionados, concentração de sensibilidade, e elementos sensíveis em escala do mais sensível ao menos sensível.

Sensor	Concentração referenciada	Gases a que são sensíveis
MQ-4	200 - 10000 ppm	CH4, GLP, H2, Fumaça, Etanol, CO
MQ-7	50 - 4000 ppm	H2, CO, GLP, CH4, Etanol
MQ-8	200 - 10000 ppm	H2, Etanol, GLP, CH4, CO
MQ-9	200 - 10000 ppm	CO, GLP, CH4
MQ-135	0,1 - 200 ppm	Acetona, Tolueno, Etanol, CO2, NH3, CO
TGS 813	500 - 10000 ppm	H2, C4H10, Propano, Etanol, CH4, CO
TGS 822	50 - 5000 ppm	Acetona, Etanol, Benzeno, Hexano, C4H10 CO, CH4
TGS 2602	1 - 30 ppm	Tolueno, Etanol, NH3, H2, H2S

Figura 12: Protótipo de nariz eletrônico desenvolvido para o experimento. Componentes básicos (a) e protótipo fechado (b).



Fonte: O autor.

3.2 MÉTODOS

O experimento foi realizado em um pomar da cultivar de pêssego Douradão localizado nas coordenadas 25°06’55.8"de latitude Sul e 50°05’43.4"de longitude Oeste na cidade de Ponta Grossa - PR. A região possui um clima considerado Cfb (subtropical úmido) de acordo com a classificação Koppen-Geiger, apresentando chuva todos os meses e temperatura média anual entre 16 e 19 °C (ALVARES *et al.*, 2013; FABC, 2021). As amostragens foram realizadas no período entre 24/08/2021 e 01/10/2021, abrangendo a floração e o primeiro estágio de desenvolvimento do fruto. O Quadro 6 relaciona os estadios fenológicos da flor e do fruto com os estágios de desenvolvimento da floração e frutificação do pessegueiro considerados nesta pesquisa.

Quadro 5: Principais componentes, suas funções e principais características.

Componente	**Função**	**Principais Características**
Matriz de sensores	sensores dos COVs	5 V, 2 A(máx.); 10 min. de de aquecimento
Arduino Nano	conversor A/D	7-12 V, 500 mA(máx.); conversor A/D 10 bits, 8 entradas analógicas; controlador ATmega328
Raspberry Pi 3 B	armazenamento dos dados, comunicação com bme280	5 V, 2 A(máx.); Broadcom BCM2837, 1 GB RAM, LAN wireless, Bluetooth, 28 pinos digitais
Placa de circuito impresso	conexão entre os componentes	dimensão 150x100 mm
Circuito de condicionamento	condicionamento de sinal, regulação de tensão	resistores 10 kΩ, reguladores de 5 e 8 V
BME280	sensor de temperatura e umidade	5 V; -40 a +85 °C temp., 0 a 100% umidade relativa; comunicação SPI, I^2C.
Ventilador 1	circulação de ar	5 V, 300 mA(máx.); dimensão 50x50x10 mm
Ventilador 2	circulação de ar	12 V, 100 mA(máx.); dimensão 40x40x10 mm
Bateria	alimentação elétrica	12 V; 5,2 Ah/1 h; selada

3.2.1 Procedimento Experimental

As amostragens foram executadas no mesmo local com o protótipo sendo alocado no solo embaixo do pessegueiro selecionado como mostra a Figura 13. No total foram realizadas 22 amostragens durante o período do experimento, iniciando-se em torno das 15h12min, com excessão das amostragens 7, 20 e 22 que foram iniciadas às 15h35min, 15h32min e 10h11min, respectivamente. Cada amostragem teve duração de 30 minutos, sendo 10 minutos para o aquecimento e estabilização dos sensores, e 20 minutos de coleta de dados. A coleta foi composta por dez arquivos, indo de c1 até c10, e cada arquivo foi formado durante 120 segundos e 2 Hz de taxa de amostragem, compondo então dez arquivo de 240 linhas cada por amostragem (dia de coleta). Cada linha foi formada por um conjunto de dez dados separados por vírgula, sendo os oito primeiros os dados dos sensores, em que o valor de cada sensor é a média de cinco leituras a uma taxa de 10 Hz, e os últimos dois dados são os valores de temperatura e umidade relativa vindos do sensor BME280. Após a finalização da amostragem, os arquivos de c1 ao c10 eram transferidos para o *notebook* e armazenados. No total de 22 amostragens, quatro foram incompletas pois ocorreram falhas durante a coleta. As amostragens incompletas foram a 3, 4, 7 e 19, cada uma com total de 1, 8, 2, e 8 arquivos de coleta, respectivamente.

Para acompanhar o desenvolvimento da floração no pessegueiro, três ramos foram

Quadro 6: Relação dos estágios de desenvolvimento da flor, do fruto, e dos considerados nesta pesquisa.

Estágio floral[a]	**Estágio do fruto**[b]	**Estágio considerado**
Flor Aberta (F)	Estágio I	Flor Aberta
Deiscência das pétalas (G)	Estágio I	Queda das Pétalas (queda de uma ou todas as pétalas)
Deiscência das sépalas (H)	Estágio I	Formação de Fruto Inicial (inchaço do ovário)
Deiscência das sépalas (H)	Estágio I	Formação de Fruto Avançada (estrutura do fruto perceptível e cercado pelas sépalas)
Crescimento da fruta (I)	Estágio I e II	Fruto formado (estrutura do fruto é a mais evidente)

Nota: *a* - classificação de Baggiolini e Cheller (BARBOSA, 1989; MOUNZER *et al.*, 2008), *b* - estágio I: até 35 dias após a plena floração, estágio II: posterior a 35 dias após a plena floração (BARBOSA, 1989; VOSS, 2019).

Figura 13: Protótipo durante a amostragem no pomar.

Fonte: O autor.

selecionados e marcados, e a cada amostragem realizada as brotações florais foram analisadas, a fim de estimar o percentual de cada um dos estágio de desenvolvimento para o pessegueiro ao longo do experimento. As marcações são os registros da quantidade e estágio das brotações dos três ramos do pessegueiro, que pode representar de 6% a 10% do total considerando um pessegueiro com 30 a 50 ramos. A quantidade de brotações no início do experimento (24/08/21) somando os três ramos amostrados era de 54 brotações, e no final do experimento (01/10/21) a quantidade era de 19.

Na Tabela 1 é apresentado o conjunto de metadados do experimento. Na Figura 14 são exemplificados os estágios de desenvolvimento considerados para análise. As duas imagens para cada estágio representam os limites considerados durante a análise em campo, seguindo as definições de distinção dos estágios presentes no Quadro 6. Na Figura 15 está representada a

evolução (curvas aproximadas) dos percentuais dos estágios das brotações florais ao longo do experimento.

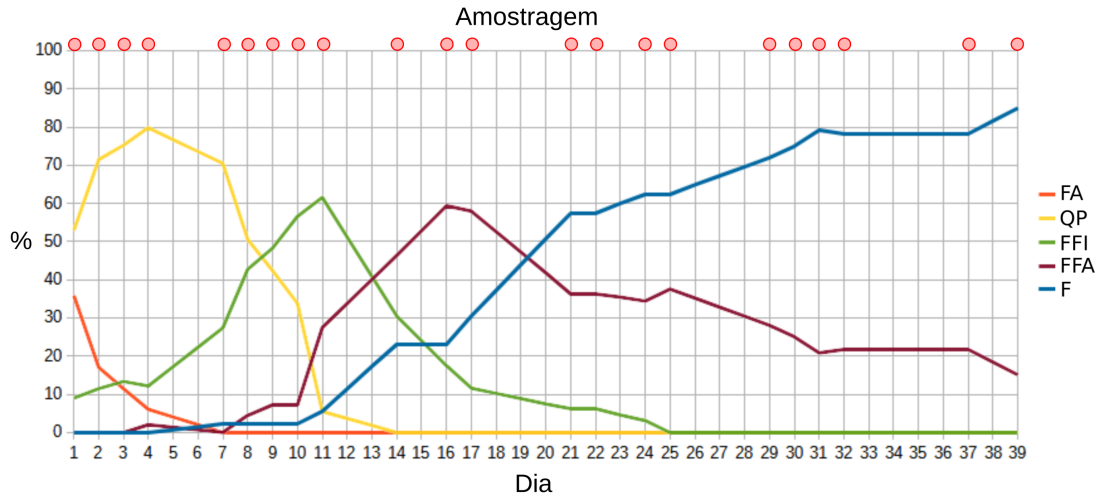
Tabela 1: Metadados do experimento.

Dia	DAPF	Coleta	Data	FA(%)	QP(%)	FFI(%)	FFA(%)	F(%)	Classe
1	8	1	24/08/21	35.90	52.85	9.00	0.00	0.00	QP
2	9	2	25/08/21	17.10	71.45	11.45	0.00	0.00	QP
3	10	3	26/08/21	11.45	75.20	13.35	0.00	0.00	QP
4	11	4	27/08/21	6.05	79.80	12.15	2.00	0.00	QP
5	12		28/08/21						
6	13		29/08/21						
7	14	5	30/08/21	0.00	70.45	27.40	0.00	2.10	QP
8	15	6	31/08/21	0.00	50.55	42.75	4.45	2.25	QP
9	16	7	01/09/21	0.00	42.35	48.25	7.05	2.35	FFI
10	17	8	02/09/21	0.00	33.80	56.55	7.25	2.40	FFI
11	18	9	03/09/21	0.00	5.50	61.55	27.45	5.50	FFI
12	19		04/09/21						
13	20		05/09/21						
14	21	10	06/09/21	0.00	0.00	30.45	46.35	23.20	FFA
15	22		07/09/21						
16	23	11	08/09/21	0.00	0.00	17.45	59.35	23.2	FFA
17	24	12	09/09/21	0.00	0.00	11.60	57.95	30.45	FFA
18	25		10/09/21						
19	26		11/09/21						
20	27		12/09/21						
21	28	13	13/09/21	0.00	0.00	6.05	36.40	57.50	F
22	29	14	14/09/21	0.00	0.00	6.05	36.40	57.50	F
23	30		15/09/21						
24	31	15	16/09/21	0.00	0.00	3.10	34.40	62.45	F
25	32	16	17/09/21	0.00	0.00	0.00	37.55	62.45	F
26	33		18/09/21						
27	34		19/09/21						
28	35		20/09/21						
29	36	17	21/09/21	0.00	0.00	0.00	28.05	71.95	F
30	37	18	22/09/21	0.00	0.00	0.00	25.05	74.95	F
31	38	19	23/09/21	0.00	0.00	0.00	20.85	79.15	F
32	39	20	24/09/21	0.00	0.00	0.00	21.75	78.25	F
33	40		25/09/21						
34	41		26/09/21						
35	42		27/09/21						
36	43		28/09/21						
37	44	21	29/09/21	0.00	0.00	0.00	21.75	78.25	F
38	45		30/09/21						
39	46	22	01/10/21	0.00	0.00	0.00	15.05	84.95	F

Figura 14: Estágios de desenvolvimento das brotações considerados para análise. Compreendendo os estágios de flor aberta (FA), quedas das pétalas (QP), formação de fruto inicial (FFI), formação de fruto avançada (FFA), e fruto formado (F). As imagens apresentam os limites considerados para a análise, sendo que a identificação e distinção de cada estágio segue as informações do Quadro 6. O estágio FFA difere do FFI pois tonar-se perceptível a base do fruto e o deslocamento das sépalas. O estágio F difere do FFA uma vez que há envelhecimento seguido de ruptura ou queda da estrutura das sépalas ao redor do fruto.

Fonte: O autor.

Figura 15: Curvas do desenvolvimento das porcentagens de cada estágio de brotação ao longo dos dias de experimento. As curvas foram construídas ligando por retas os valores de porcentagem obtidos em cada amostragem.



Fonte: O autor.

3.2.2 Procedimento de Pré-processamento

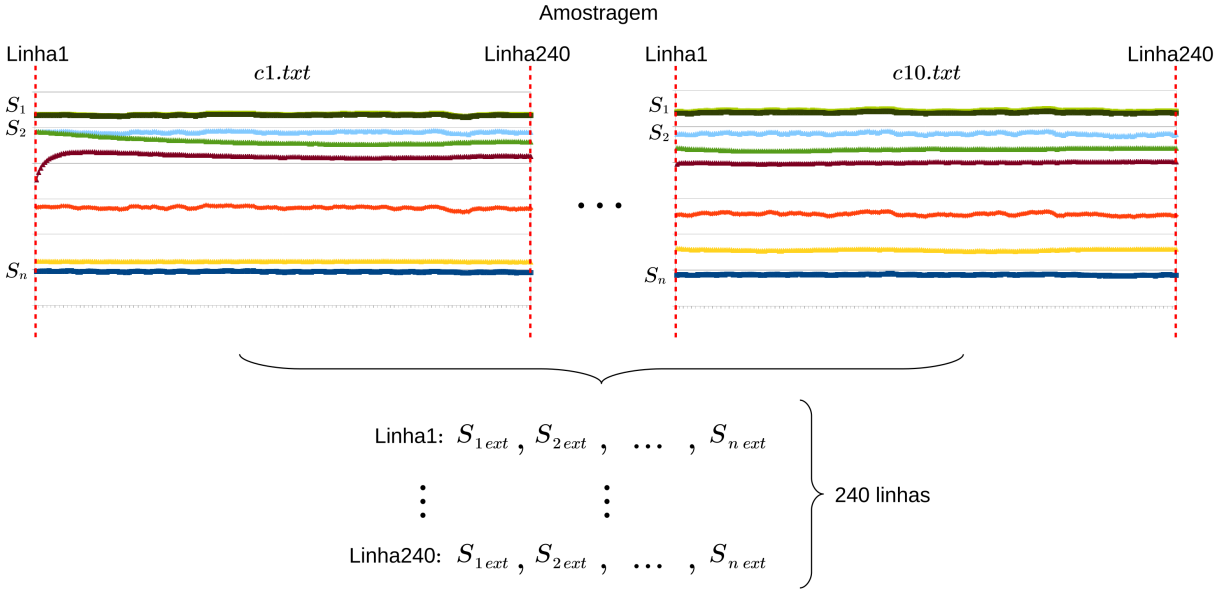
A etapa inicial do procedimento de pré-processamento compreende extrair informações representativas do sinal para formar a base de dados. A forma de extração de dados do sinal adotada é uma composição média dos instantes separados por 2 minutos dentro dos 20 minutos da coleta, chamada de mc10.

Para os dez arquivos de uma amostragem, as linhas equivalentes de cada arquivo são combinadas e a média é calculada. O procedimento produz 240 linhas (amostras) por amostragem (coleta). A Equação 30 exemplifica a extração de dados para o sensor 1 (S_1) dado o conjunto de arquivos de coleta ($c_i.txt$), em que $S_1c_i.txt$ é o valor de S_1 no arquivo $c_i.txt$, e S_{1ext} é o valor extraído. A equação é empregada para cada sensor usando a mesma linha de cada arquivo para compor o somatório. O mc10 aplicado em uma amostragem é representado na

Figura 16. Desse modo, busca-se uma estabilização representativa do período de 20 minutos de amostragem, e não uma variação dentro dessa amostragem. Assim, os valores das respostas dos sensores de gás e de temperatura e umidade também são representados de forma estabilizada pela média da composição.

$$S_{1ext} = \frac{1}{10}\sum_{i=1}^{10} S_1 c_i.txt \tag{30}$$

Figura 16: Método de extração de dados do sinal por meio da composição de 10 valores para cada sensor (mc10).



Fonte: O autor.

O procedimento de extração de dados do sinal utilizado em todas as amostragens (1 até a 22) gerou uma base com 5040 amostras. Cada linha da base tem dez variáveis, na qual as oito primeiras são dos sensores de gás, e as duas últimas do sensor de temperatura e umidade relativa. Os valores dos sensores de gás foram convertidos para seus valores analógico (0 a 5 V). O arquivo de uma base manteve o formato txt com formatação dos valores separados por vírgula.

Na rotulagem dos dados, para a análise de classificação, a variável de saída foi definida pelo maior estágio de brotação do pessegueiro, de forma que para cada amostragem o respectivo estágio de maior porcentagem foi atribuído como valor qualitativo da variável de saída de suas amostras. Como pode ser observado na Tabela 1 a variável de saída pode assumir quatro valores, sendo eles, queda das pétalas (QP), formação de fruto inicial (FFI), formação de fruto avançada (FFA), e fruto (F).

Uma vez definida a variável de saída e sua relação com as amostras, foi formado um espaço composto pelas variáveis de entrada e de saída. Esse arquivo foi chamado de espaço

base original (EBO), já que possui os valores originais das dez variáveis de entrada. Cada transformação aplicada nesse espaço operou sobre as variáveis de entrada e gerou um novo espaço base.

A normalização máximo-mínimo no intervalo de 0 a 1 foi usada para ajustar os valores das variáveis de entrada numa mesma magnitude. A normalização é importante para tratar dados de magnitude desproporcional, geralmente vindos de grandezas diferentes, como os dados de temperatura, de umidade, e os dados dos sensores de gás. Além disso, a normalização é recomendada para obter um resultado adequado do PCA quando tem-se valores desproporcionais na base (HAN; PEI; KAMBER, 2012). A forma comum de realizar a normalização é sua execução individualmente sobre cada variável de entrada. Outro modo que pode ser efetuado, é a execução sobre o conjunto de variáveis de mesma grandeza, isto é, a normalização sobre três conjuntos, sendo os sensores de gás, temperatura e umidade relativa. Assim, são exploradas diferentes representações de valores e variação de cada variável. As duas formas foram aplicadas sobre o espaço base original produzindo o espaço base normalizado (EBN) e o espaço base normalizado 2 (EBN2), respectivamente.

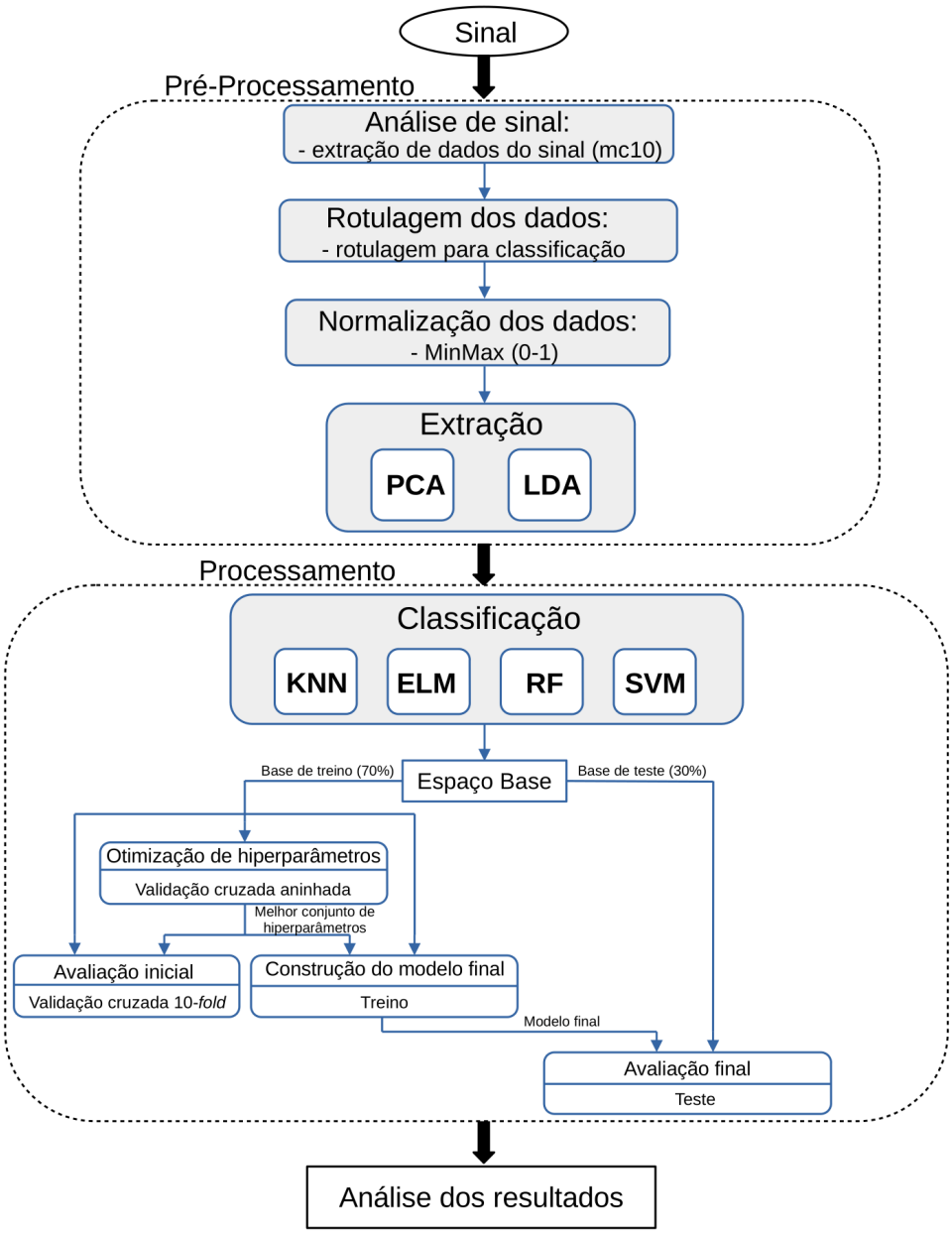
As técnicas de PCA e LDA foram executadas para extrair as características dos espaços bases normalizados. Como o LDA explora as relações entre a dispersão dos dados de entrada considerando as informações das classes, isto é, a variável de saída, seu resultado não é influenciado diretamente pelos valores. Desse modo, quando o LDA é aplicado sobre os espaços bases EBO, EBN e EBN2 o resultado é o mesmo. Então, no total a extração de características gera apenas três espaços base. O espaço resultado do PCA executado sobre o EBN (EBPCA), sobre o EBN2 (EBPCA2), e o espaço base resultado do LDA executado sobre o EBN (EBLDA).

No total 6 espaços base são formados no pré-processamento. Na Figura 17 o fluxograma resume o procedimento de análise de dados empregado, composto pelas etapas de pré-processamento (formação dos espaços base) e processamento (construção e avaliação dos modelos).

3.2.3 Procedimento de Processamento

No processamento a análise de classificação é executada. Nessa etapa os quatro algoritmos de aprendizado de máquina (KNN, ELM, RF, SVM) foram implementados e cada modelo formado pela combinação de algoritmo e espaço base foi avaliado. Então, os modelos construídos tiveram desempenhos medidos e comparados pelas métricas. No Quadro 7 estão representados os algoritmos e os hiperparâmetros otimizados. As métricas utilizadas nas análises estão descritas no Quadro 8.

Figura 17: Procedimento de análise de dados utilizado na pesquisa.



Fonte: O autor.

No Quadro 7, os parâmetros do KNN são *Nn*, que é o número de vizinhos mais próximos, e *We*, que é a forma de ponderar os vizinhos mais próximos. No ELM, *Ne* é o número de neurônios na camada intermidiária, e *g(a)* é a função de ativação dos neurônios. Para o RF, *Tr* é o número de árvores, e *Q(D)* é o critério de separação utilizado para dividir um nó. No SVM, *De* é o grau da função polinomial utilizada como função de núcleo, e *coef0* é o termo independente na função.

O procedimento descrito na etapa de processamento da Figura 17 foi realizado para cada espaço base. Para construir os modelos os quatro algoritmos de aprendizado de máquina foram executados no espaço base. No processamento, 24 modelos finais foram construídos e

seus desempenhos de generalização avaliados sobre as respectivas bases de teste.

Quadro 7: Algoritmos e hiperparâmetros otimizados.

Algoritmo	**Hiperparâmetro 1**	**Valores$_{HP1}$**	**Hiperparâmetro 2**	**Valores$_{HP2}$**
KNN	*Nn*	[1, 10]	*We*	(Uniform, Distance)
ELM	*Ne*	[10, 2000]	*g(a)*	(ReLu, TanH)
RF	*Tr*	[2, 200]	*Q(D)*	(Gini, Entropia)
SVM	*De*	[2, 6]	*coef0*	(0.0, 0.5)

Nota: Os hiperparâmetros 1 (HP1) e 2 (HP2) e suas amplitudes de valores foram definidos para a otimização.

Quadro 8: Métricas de classificação.

Métrica	**Símbolo**	**Descrição**
Acurácia Balanceada	Acc_{bal}	Acurácia considerando o desbalanceamento entre as classes
Sobreajuste Acc$_{bal}$ Regulado	Saj_{Acc}	(1 - |$Acc_{bal(Treino)}$ - $Acc_{bal(Teste)}$|)
F_1 Ponderado	F_{1pon}	Indicador F_1 considerando o desbalanceamento entre as classes
Sobreajuste F_{1pon} Regulado	Saj_{F_1}	(1 - |$F_{1pon(Treino)}$ - $F_{1pon(Teste)}$|)

O princípio da análise multiobjetivo foi empregado para avaliar os modelos usando mais de uma métrica ao mesmo tempo. Dessa forma, é possível melhorar a visualização do comportamento dos modelos. Isso é realizado por meio da representação simples no espaço objetivo, ou seja, o espaço das métricas não é transformado por funções, sendo somente uma correspondência linear em que cada dimensão é uma métrica, e portanto não há transformação nem preferência entre elas. No espaço objetivo, dado a minimização ou a maximização, é possível definir os modelos dominantes sobre os demais, isto é, os mais eficientes de acordo com o critério de Pareto.

Nas análises, para a implementação dos algoritmos foi utilizada a linguagem python no ambiente gratuito Colaboratory do Google e a biblioteca Scikit-Learn. A biblioteca Optuna foi usada para a visualização da fronteira de Pareto.

3.2.4 Procedimento de Validação do Equipamento

Para verificar o comportamento do nariz eletrônico e avaliar sua resposta para um elemento de análise específico dois conjuntos de testes foram executados. O objetivo da análise de validação foi determinar a estabilidade e repetibilidade dos sensores na configuração do equipamento. O primeiro foi realizado paralelamente as coletas do experimento principal (pessegueiro no pomar). Nesse teste 18 execuções foram realizadas em um mesmo local considerado mais estável que o ambiente do experimento principal. A premissa foi de que as amostras vindas deste teste seriam afetadas pela condições e características que produzem variações nos sensores. Isso permite considerar que essas amostras sejam referência de como as variações

afetam as medidas com o equipamento. Assim, o teste possibilita comparar as amostras do experimento com as amostras de referência e visualizar distinções e equivalências. As execuções foram realizadas do dia 27/08/2021 (execução 1) ao dia 01/10/2021 (execução 18), e no total 4320 amostras de referência foram geradas.

O segundo teste foi realizado com amostras de álcool de três composições diferentes (46%, 70%, 99%). As execuções foram realizadas em três intervalos distintos. O intervalo de execução curto, em que no mesmo dia cada composição de álcool foi amostrada numa execução com período de 15 minutos entre as execuções. O intervalo médio, no qual as três execuções foram efetuadas em três dias seguidos, alterando a sequência das composições durante cada dia. Por último, o intervalo longo, compreendendo três dias seguidos de execuções separados por meses. As primeiras 9 execuções foram realizadas do dia 14 ao 16/09/2021, as seguintes (10-18) nas datas de 04 a 06/03/2022, e as últimas (19-27) nos dias 18, 19 e 20/05/2022. Nesse teste 27 amostragens foram realizadas, 9 de cada composição alcoólica, produzindo um total de 6480 amostras (2160 de cada composição).

Para ambos os testes o mesmo método de extração de dados do sinal foi efetuado (mc10), diferentes formatações dos espaços de entrada foram verificadas (EBO, EBN e EBN2), e análises de agrupamento com PCA e LDA foram realizadas. No teste com as amostras de referência o objetivo foi visualizar a diferenças em relação as amostras do experimento. No teste com álcool foi para verificar o quanto o equipamento pode variar entre execuções e sua capacidade de distinguir as três composições de álcool.

3.2.5 Procedimento de Comparação de Estratégias de Compensação

Para verificar a diferença entre as estratégias de compensação de temperatura e umidade uma análise comparativa foi realizada. Os dois modos de compensação avaliados foram de espaços base gerados de dados compensados por equações (CE) e espaços base de dados não compensados e com as variáveis de temperatura e umidade na entrada (CTU). Os espaços base originais (EBO) de CE e CTU possuem oito e dez variáveis de entrada respectivamente. Além desses, os espaços base gerados pela extração de características utilizando PCA (EBPCA) e LDA (EBLDA) também foram avaliados nesta análise. No total para cada estratégia de compensação (CE, CTU) três espaço bases foram avaliados (EBO, EBPCA, EBLDA).

A primeira componente da análise é uma avaliação de discriminação entre CE e CTU para o PCA e o LDA aplicados sobre o espaço base original (EBO) para CE e sobre o espaço base normalizado (EBN) para CTU. A segunda, é uma avaliação de classificação considerando os três espaços bases (EBO, EBPCA, EBLDA) e os quatro algoritmos (KNN, ELM, RF, SVM).

Nesta avaliação para cada estratégia de compensação 12 modelos foram construídos e avaliados, seguindo os mesmos procedimentos de pré-processamento e processamento. De modo complementar, uma análise de variância foi efetuada nos resultados da classificação para avaliar se há diferença significativa (95% confiança) entre as métricas de acurácia balanceada (Acc_{bal}) e indicador F1 ponderado ($F1_{pon}$) de ambas estratégias (CE, CTU).

Para corrigir os valores dos sensores e formar os dados utilizados na estratégia CE as equações foram aproximadas para cada sensor aplicando as informações disponíveis nas folhas de dados. Tal estratégia é descrita no trabalho de Voss (2019) e utiliza uma regressão polinomial para relacionar o valor de resistência do sensor Rs e o valor de resistência corrigida Ra. O polinômio quadrático é descrito pela Equação 31 (VOSS, 2019). Na equação, x e y são a temperatura e a umidade relativa, respectivamente. O conjunto α, β e γ são os coeficientes, e ϵ é a constante de equação. Os valores encontrados nas folhas de dados de cada sensor para os coeficientes da equação são apresentados na Tabela 2.

$$Ra = \frac{Rs}{\alpha_1 x + \alpha_2 x^2 + \beta_1 y + \beta_2 y^2 + \gamma xy + \epsilon} \tag{31}$$

Tabela 2: Coeficientes obtidos pela regressão polinomial para cada sensor.

Sensor	$\boldsymbol{\alpha_1}(10^{-2})$	$\boldsymbol{\alpha_2}(10^{-4})$	$\boldsymbol{\beta_1}(10^{-2})$	$\boldsymbol{\beta_2}(10^{-4})$	$\boldsymbol{\gamma}(10^{-5})$	$\boldsymbol{\epsilon}$	$\mathbf{R}^2$
TGS 813	-1,518	1,97	-1,408	0,9637	-7,693	1,804	0,9533
MQ-4	-0,9197	0,8391	-0,3146	0,00	-0,1841	1,262	0,9935
TGS 822	-5,186	5,392	-1,732	0,7324	7,698	2,533	0,9559
TGS 2602	-3,259	1,077	-0,7488	-0,4812	1,128	1,374	0,9882
MQ-135	-2,557	3,127	-0,2533	0,00	2,257	1,473	0,9894
MQ-8	-0,9859	1,272	-0,103	0,00	0,7672	1,18	0,9898
MQ-7	-1,553	1,412	-0,378	0,00	3,574	1,361	0,9844
MQ-9	-1,555	1,416	-0,3783	0,00	3,712	1,362	0,9836

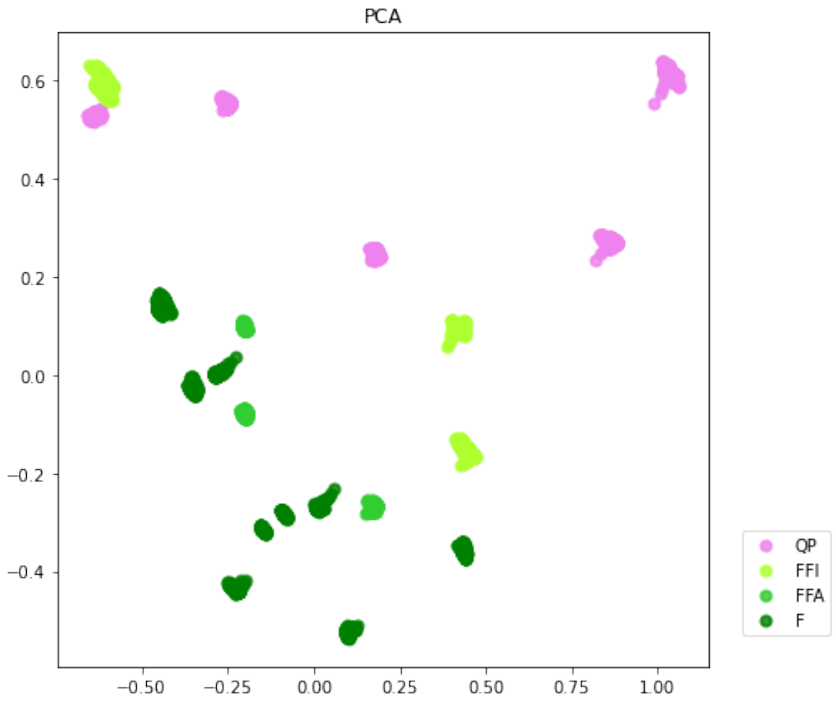
4 RESULTADOS

4.1 ANÁLISE DE COMPARAÇÃO ENTRE COMPENSAÇÕES

Nesta análise as duas estratégias de compensação para temperatura e a umidade relativa foram avaliadas. Os dados compensados pelas equações são referenciados por CE, enquanto os dados sem compensação com as variáveis de temperatura e umidade relativa inclusas na entrada são descritos por CTU.

As Figuras 18 e 19 mostram o resultado do PCA e do LDA aplicados no espaço base original (EBO) para a estratégia CE. Nas Figuras 20 e 21 é apresentado o resultado do PCA e do LDA para o espaço base normalizado (EBN) para a estratégia CTU. A Tabela 3 descreve os valores encontrados na otimização dos hiperparâmetros (HP) dos modelos para ambas estratégias (CE, CTU). Os resultados da classificação são apresentados na Tabela 4. O resultado da análise de variância e sua configuração estão nas Tabelas 5 e 6, respectivamente.

Figura 18: Execução do PCA no espaço base original (EBO) das amostras produzidas pela estratégia CE. Com PC1 de 56,70% e PC2 de 37,94%.

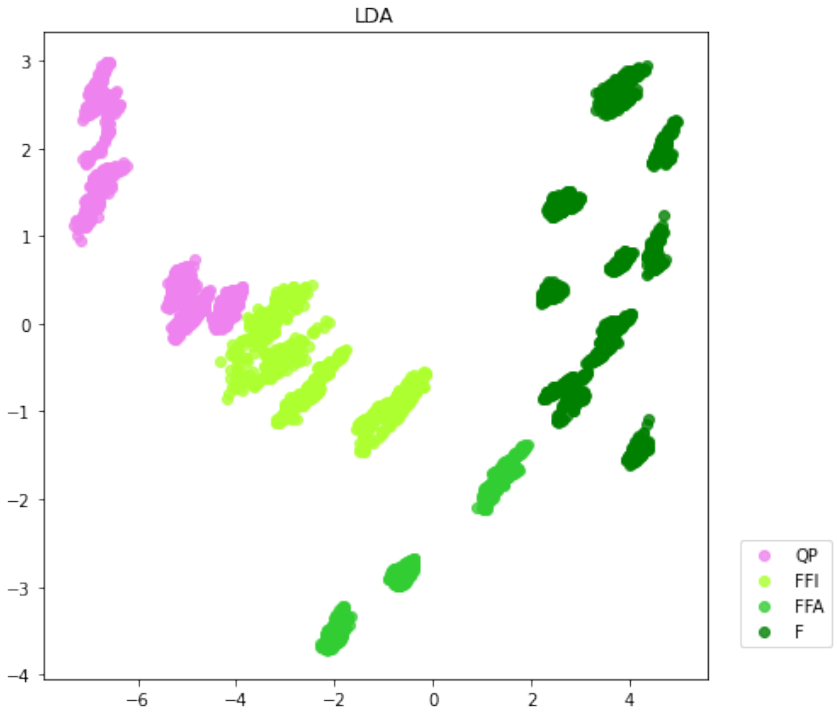


Fonte: O autor.

Na análise de discriminação para o PCA a configuração CE (compensaçãp por equações) apresentou melhor resultado (> 5,76%), e para o LDA a configuração CTU (compensação por temperatura e umidade na entrada) apresentou melhor desempenho (> 2,19%). Na análise de classificação a configuração CE apresentou melhor desempenho em cinco modelos, e a configuração CTU em dois, enquanto o desempenho foi o mesmo em outros cinco modelos.

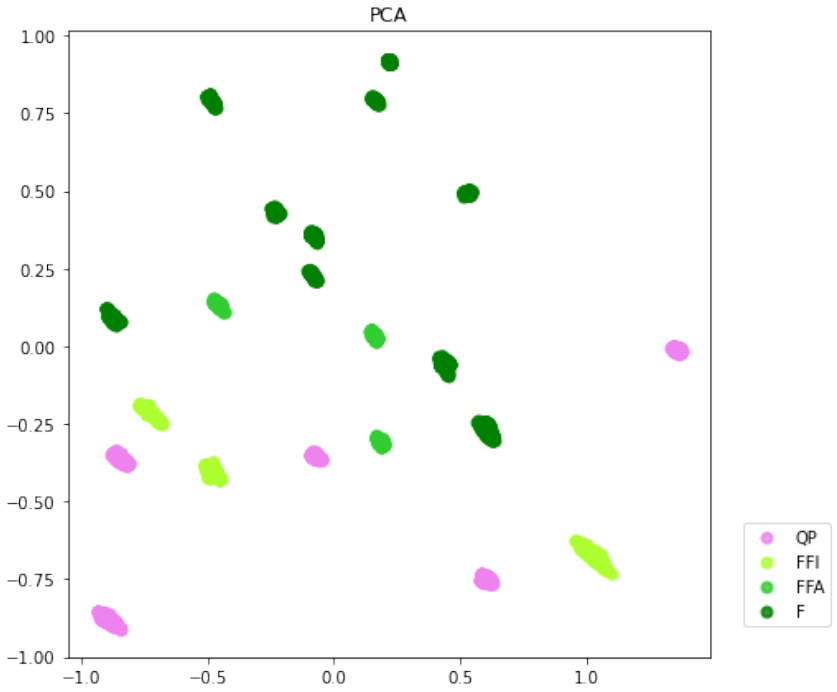
Na análise de variância das médias de desempenho considerando as configurações (CE,

Figura 19: Execução do LDA no espaço base original (EBO) das amostras produzidas pela estratégia CE. Com LD1 de 87,38% e LD2 de 8,79%.



Fonte: O autor.

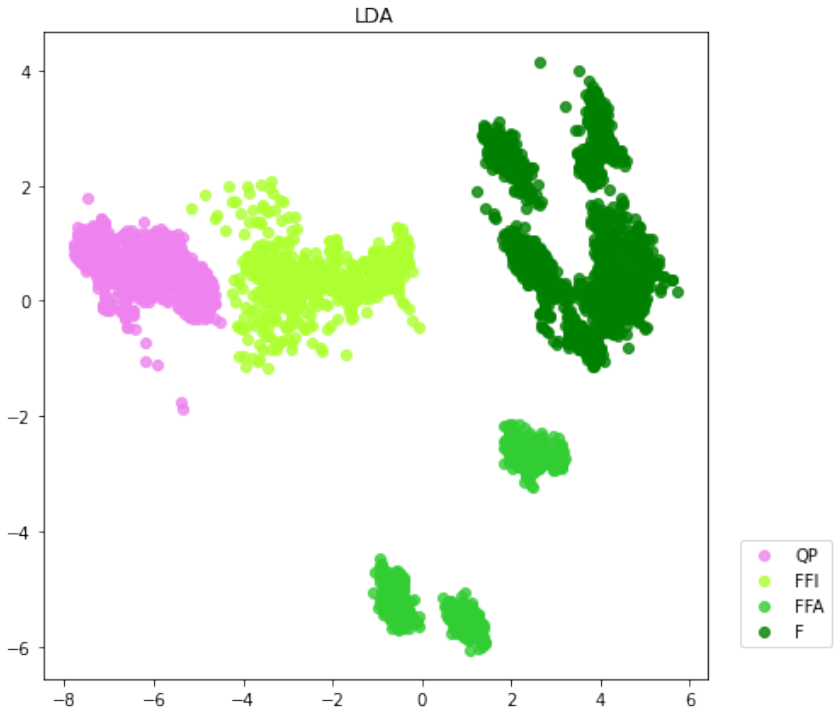
Figura 20: Execução do PCA no espaço base normalizado (EBN) das amostras produzidas pela estratégia CTU. Com PC1 de 53,59% e PC2 de 35,29%.



Fonte: O autor.

CTU) e as métricas (Acc_{bal}, $F1_{pon}$) o resultado demonstrou que não há diferença significativa (95% de confiança) entre as médias das métricas utilizadas dado as configurações CE e CTU. O resultado da análise de variância corrobora com o resultado descrito no trabalho de Huerta *et al.*

Figura 21: Execução do LDA no espaço base normalizado (EBN) das amostras produzidas pela estratégia CTU. Com LD1 de 81,64% e LD2 de 16,72%.



Fonte: O autor.

Tabela 3: Valores encontrados na otimização dos hiperparâmetros (HP) dos modelos para ambas estratégias (CE, CTU).

Modelo	**CE (HP1, HP2)**	**CTU (HP1, HP2)**
KNN EBO	(1, uniform)	(1, uniform)
ELM EBO	(20, tanh)	(30, relu)
RF EBO	(10, gini)	(10, gini)
SVM EBO	(2, 0.0)	(6, 0.5)
KNN EBPCA	(1, uniform)	(1, uniform)
ELM EBPCA	(40, relu)	(100, relu)
RF EBPCA	(10, gini)	(10, gini)
SVM EBPCA	(5, 0.5)	(3, 0.5)
KNN EBLDA	(1, uniform)	(1, uniform)
ELM EBLDA	(20, tanh)	(50, tanh)
RF EBLDA	(20, gini)	(10, gini)
SVM EBLDA	(5, 0.5)	(4, 0.5)

(2016) em que não houve diferença significativa de desempenho entre os dados compensados e os dados não compensados mais temperatura e umidade como entrada.

4.2 ANÁLISE DE VALIDAÇÃO DO EQUIPAMENTO

O objetivo da análise de validação foi determinar a estabilidade e repetibilidade dos sensores na configuração do equipamento. Dois testes foram executados e para ambos as análi-

Tabela 4: Resultado dos modelos na base de teste para a classificação.

Modelo	CE $Acc_{bal}(\%)$	CE $F_{1pon}(\%)$	CTU $Acc_{bal}(\%)$	CTU $F_{1pon}(\%)$
KNN EBO	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
ELM EBO	99,988 ± 0,081	99,993 ± 0,046	99,674 ± 1,272	99,787 ± 0,791
RF EBO	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
SVM EBO	100 ± 0,000	100 ± 0,000	91,435 ± 0,000	94,720 ± 0,000
KNN EBPCA	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
ELM EBPCA	99,545 ± 0,575	99,505 ± 0,622	98,491 ± 0,735	99,128 ± 0,423
RF EBPCA	99,990 ± 0,032	99,993 ± 0,020	100 ± 0,000	100 ± 0,000
SVM EBPCA	99,965 ± 0,000	99,934 ± 0,000	100 ± 0,000	100 ± 0,000
KNN EBLDA	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
ELM EBLDA	99,973 ± 0,083	99,975 ± 0,078	99,819 ± 0,052	99,828 ± 0,049
RF EBLDA	100 ± 0,000	100 ± 0,000	99,874 ± 0,037	99,873 ± 0,041
SVM EBLDA	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000

Tabela 5: Análise de variância do resultado da classificação entre CE e CTU a 5% de significância.

Compensação (métrica)	Média	Resultado
CE (Acc_{bal})	99,955	a
CE ($F1_{pon}$)	99,950	a
CTU (Acc_{bal})	99,108	a
CTU ($F1_{pon}$)	99,445	a

Tabela 6: Configuração utilizada para análise de variância.

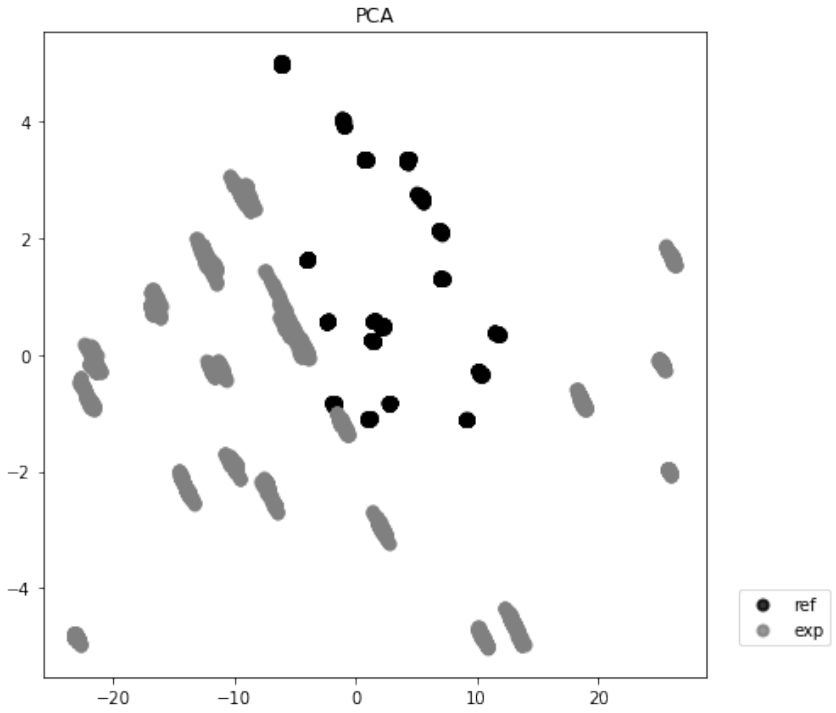
Modelo de ANOVA	Hipóteses	Significância	Tratamentos	Repetições
EBC simples	H_0: $\mu1$=$\mu2$=$\mu3$=$\mu4$ H_1: há diferença	0,05	4	12

Nota: As médias são indicadas por μ.

ses de agrupamento com PCA e LDA foram realizadas. No teste com as amostras de referência para visualizar a diferenças em relação as amostras do experimento. No teste com álcool para verificar o quanto o equipamento pode variar entre execuções e capacidade de distinguir as três composições de álcool. As Figuras 22 e 23 mostram os resultados do PCA nas bases original e normalizada para distinção entre as amostras de experimento e de referência. Nas Figuras 24 e 25 os melhores resultados de PCA e LDA no teste com álcool são apresentados, mostrando as variações entre execuções e a discriminação entre as composições de álcool.

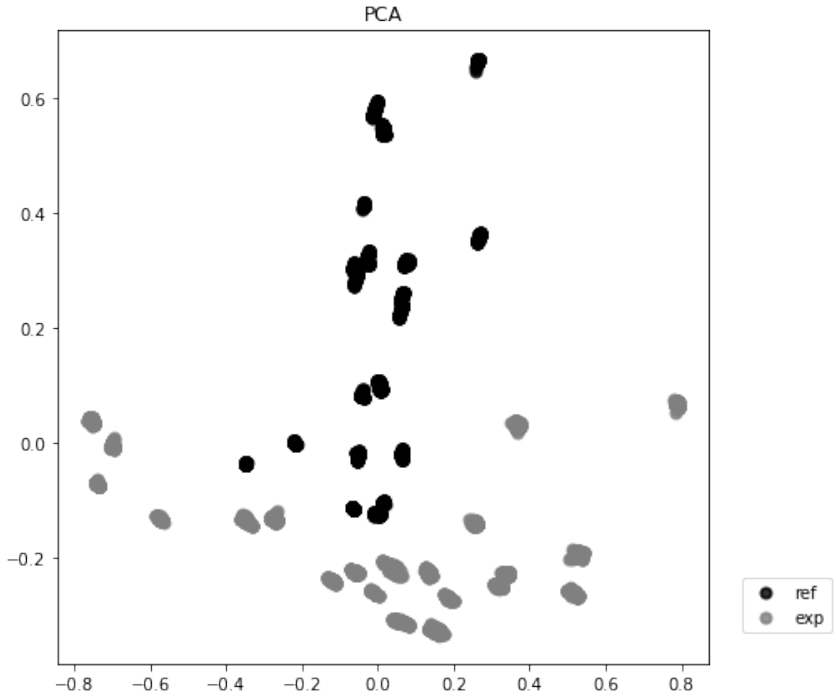
No teste de referência foi possível distinguir as amostras do experimento (cinzas) e de referência (pretas). Isto é indicativo de que os ambientes do pomar e o de referência podem ser diferenciados pelo equipamento, uma vez que foram discriminados de forma perceptível pelo PCA. Além disso, o padrão observado no PCA sugere que as maiores variações tem pouca ou nenhuma correlação, pois estão dispostas em eixos diferentes. Assim, provavelmente, as variações são específicas do pomar e do ambiente de referência. Por exemplo, quando avalia-se a direção de maior variação é possível observar que as amostras de referência têm variação

Figura 22: Execução do PCA no espaço base original (EBO) das amostras de referência (pretas) e amostras do experimento (cinza). Com PC1 de 96,24% e PC2 de 3,25%.



Fonte: O autor.

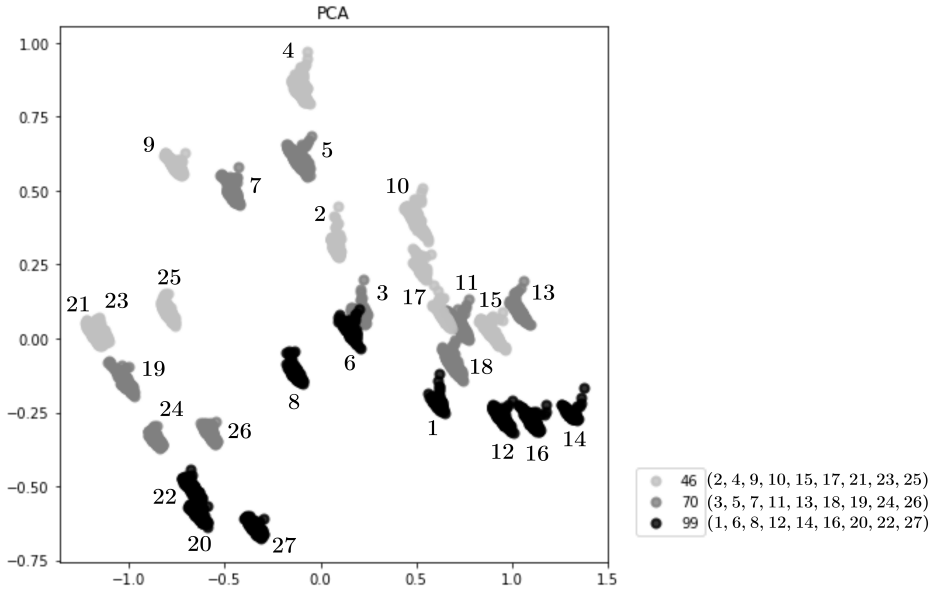
Figura 23: Execução do PCA no espaço base normalizado (EBN2) das amostras de referência (pretas) e amostras do experimento (cinza). Com PC1 de 43,15% e PC2 de 30,71%.



Fonte: O autor.

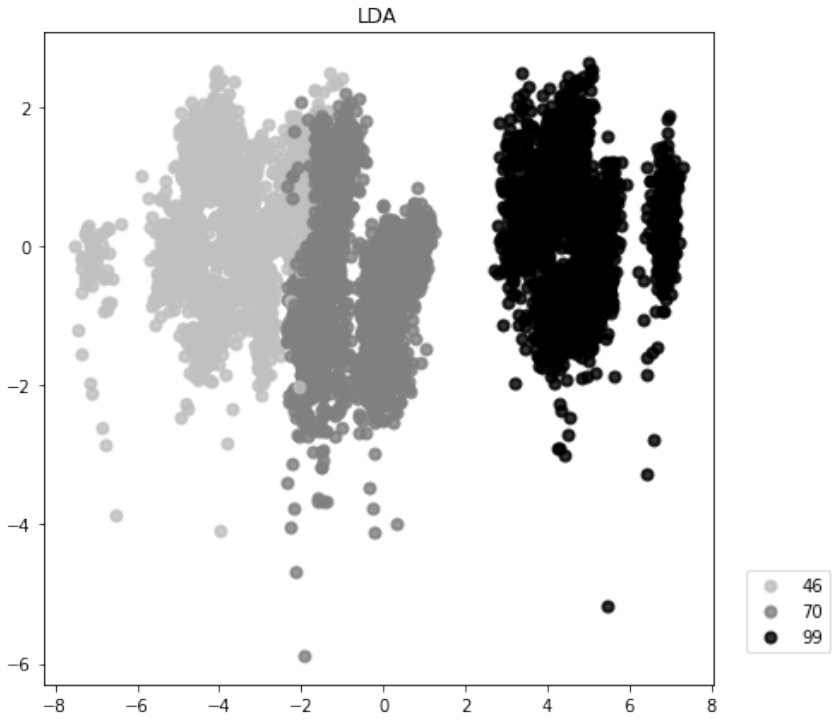
reduzida nessa direção, enquanto as amostras de experimento têm sua variação extensa. No entanto, não é possível interpretar os resultados enquanto somente uma direção de variação, necessitando de duas ou mais. Ainda, mesmo no espaço de entrada normalizado (Figura 23),

Figura 24: Execução do PCA para base normalizada (EBN) das amostras de álcool. Com PC1 de 72,42% e PC2 de 16,99%.



Fonte: O autor.

Figura 25: Execução do LDA para base normalizada (EBN) das amostras de álcool. Com LD1 de 98,92% e LD2 de 1,08%.



Fonte: O autor.

em que a influência das variações dos dados de temperatura e umidade relativa é considerada de forma adequada, é possível que a temperatura e a umidade estejam afetando de forma intensa os resultados. Desse modo, dificultando uma interpretação visual sobre as variações de referência e experimento.

Para o teste com álcool a estabilidade e a repetibilidade podem ser avaliadas de forma

simples pelo resultado do PCA no espaço de entrada normalizado (Figura 24). As execuções de médio prazo (dias) podem ser vistas com uma tendência de agrupamento. Assim como as execuções de longo prazo (meses) que estão formando três grupos. A variação de cada composição demonstra a tendência de estabilidade e repetibilidade do equipamento para as amostras de álcool. Foi possível observar que para as três composições as variações de longo prazo produzem maior deslocamento das amostras, e que as de médio prazo também são impactantes. Ainda, as variações de curto prazo não são perceptíveis dado as diferentes composições, ou seja, a diferença que ocorre em minutos entre as execuções não foram maiores que as diferenças produzidas pelas composições de álcool. Dessa forma, mesmo executadas em sequências distintas nos três dias, o padrão permaneceu.

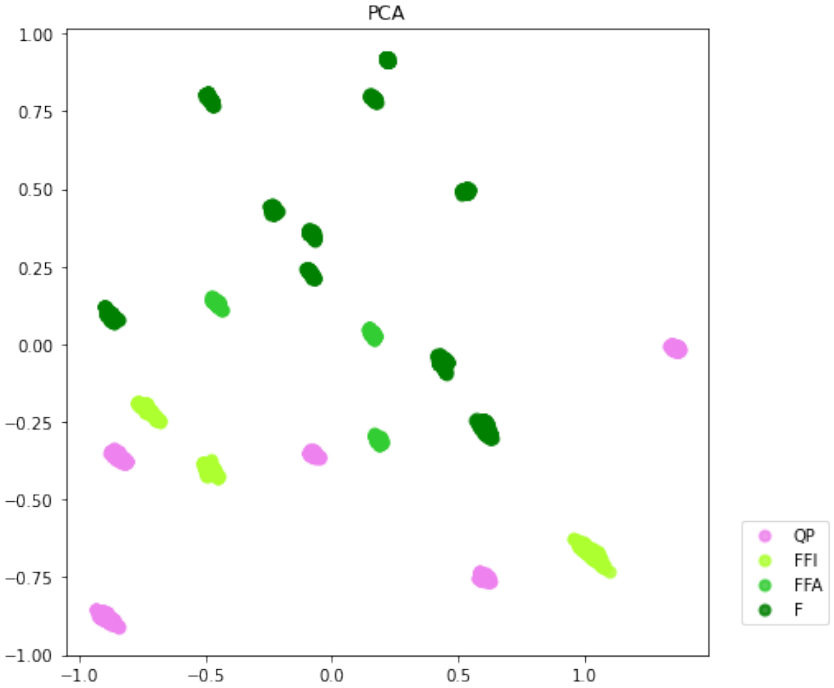
Apesar das variações, foi possível verificar que há um padrão nas composições, com o álcool 99% na região inferior, e nas regiões acima o 70% e o 46%, respectivamente. Além disso, o LDA independente da representação do espaço de entrada apresentou os agrupamentos das três composições alcoólicas de forma nítida, sendo capaz de discriminar as composições apesar das diferenças temporais. Isso indica que questões de estabilidade e repetibilidade de curto (minutos), médio (dias) e longo período (meses) entre execuções afetam os dados, mas podem ser tratadas no pré-processamento de modo efetivo.

4.3 ANÁLISE DE DISCRIMINANTE

As técnicas PCA e LDA foram empregadas para verificar se o nariz eletrônico seria capaz de discriminar os estágio de desenvolvimento considerados durante o experimento por meio da visualização de agrupamentos. Ambas as técnicas utilizam a variância tentando maximiza-la, no PCA em relação aos dados, e no LDA em relação às classes definidas. Duas características de saída foram consideradas para as técnicas. As Figuras 26 e 27 mostram os resultados da execução das extrações para o espaço base normalizado (EBN).

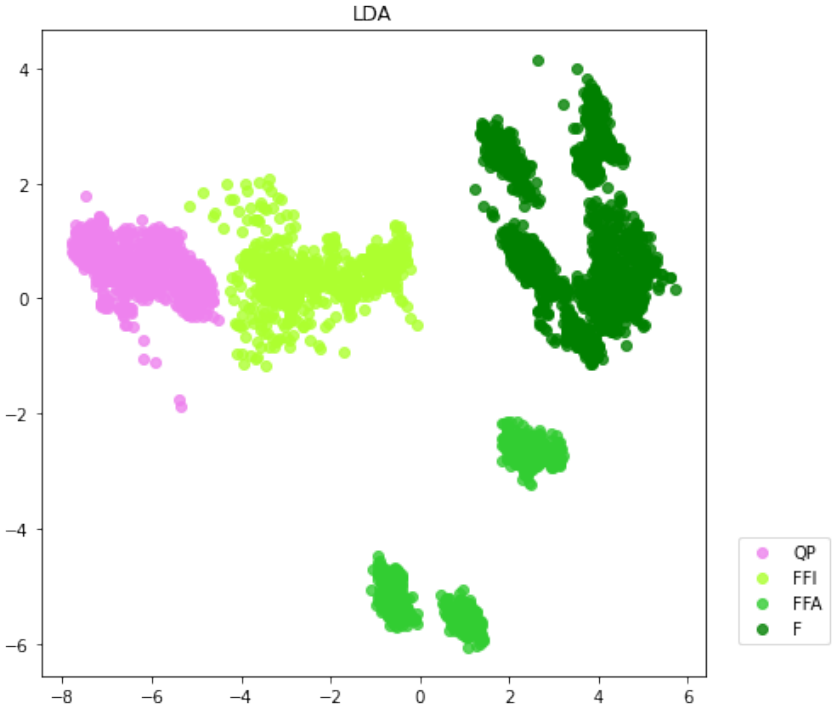
Para o PCA foi possível observar uma tendência nos dados, com PC1 e PC2 de 53,59% e 35,29% de variância explicada, respectivamente. No LDA, a discriminação foi mais evidente com a formação de agrupamentos. A variância explicada é de 81,64% no LD1 e 16,72% no LD2. Para as quatro classes definidas não houve interseção entre amostras, sendo um indicador da aptidão do sistema em discriminar os estágios. Tais resultados sugerem que a próxima etapa, de classificação, pode ser realizada pelo nariz eletrônico de forma adequada dado os agrupamentos obtidos pelo LDA.

Figura 26: Representação da relação entre a base de dados e os 2 componentes principais do PCA extraídos da base. Os estágios são queda das pétalas (QP), formação de fruto inicial (FFI), formação de fruto avançada (FFA), e fruto formado (F). PC1 com 53,59% e PC2 com 35,29% de variância explicada.



Fonte: O autor.

Figura 27: Representação da relação entre a base de dados e os 2 discriminantes lineares do LDA extraídos da base. Os estágio são queda das pétalas (QP), formação de fruto inicial (FFI), formação de fruto avançada (FFA), e fruto formado (F). LDA1 com 81,64% e LD2 com 16,72% de variância explicada.



Fonte: O autor.

4.4 ANÁLISE DE CLASSIFICAÇÃO

O objetivo da análise de classificação foi verificar se seria possível distinguir e identificar corretamente as classes através das características (comportamento) dos dados das variáveis de entrada. A variável de saída era do tipo qualitativa ordinal e podia assumir os valores (QP), (FFI), (FFA), e (F).

Para avaliação de desempenho as métricas adotadas foram a acurácia balanceada (Acc_{bal}), o indicador F_1 ponderado (F_{1pon}), e o cálculo dos respectivos sobreajustes regulados para maximização (Saj_{Acc}), (Saj_{F_1}). As métricas foram cálculadas em relação às amostras de teste. As configurações dos algoritmos estão descritas no Quadro 9. Na Tabela 7 estão os valores encontrados na otimização dos hiperparâmetros para os algoritmos KNN, ELM, RF, e SVM, em cada espaço base. A divisão das amostras em cada espaço base foi de 3528 amostras para treinamento (70%) e 1512 para teste (30%).

Quadro 9: Configuração dos algoritmos e amplitude de busca dos hiperparâmetros otimizados.

Algoritmo	**Configuração**
KNN	*Nn*: [1, 10], *We*: (Uniform, Distance), *algorithm*: brute, *p*: 2, *metric*: minkowski
ELM	*Ne*: [10, 2000], *g(a)*: (ReLu, TanH), *density*: 1, *alpha*: 10^{-7}
RF	*Tr*: [2, 200], *Q(D)*: (Gini, Entropy), *max-depth*: max, *max-features*: sqrt, *min-sample-split*: 2, *min-samples-leaf*: 1, *bootstrap*: true, *max-samples*: max
SVM	*De*: [2, 6], *coef0*: (0.0, 0.5), *kernel*: poly, *C*: 1.0, *gamma*: scale, *shrinking*: true, *tolerance*: 10^{-3}

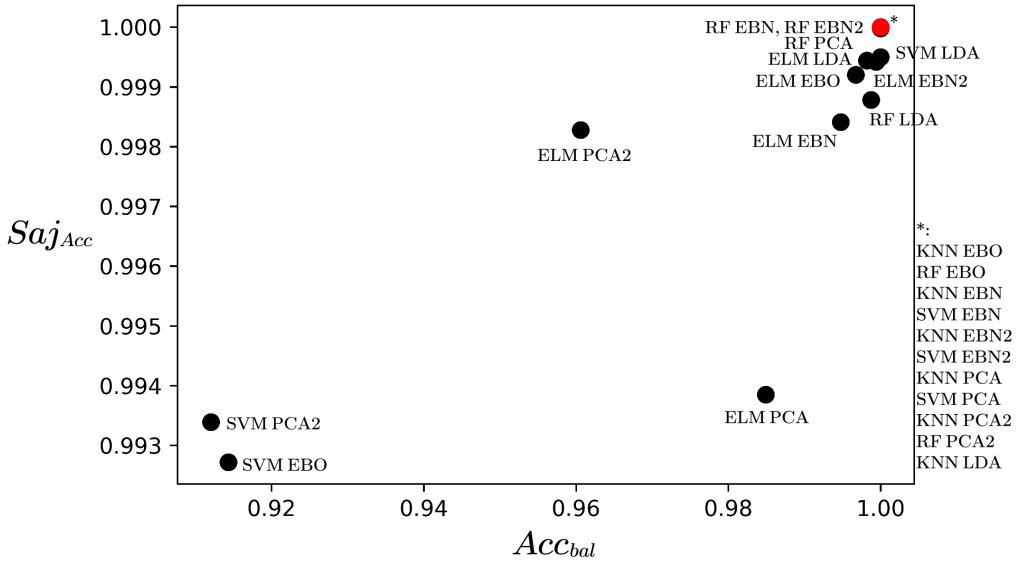
Tabela 7: Valores de hiperparâmetros (HP) encontrados na otimização (Acc_{bal}) para os algoritmos de acordo com o espaço base.

Espaço base	**KNN (HP1,HP2)**	**ELM (HP1,HP2)**	**RF (HP1,HP2)**	**SVM (HP1,HP2)**
EBO	(1, uniform)	(30, relu)	(10, gini)	(6, 0.5)
EBN	(1, uniform)	(20, relu)	(10, gini)	(2, 0.0)
EBN2	(1, uniform)	(30, tanh)	(10, gini)	(2, 0.0)
EBPCA	(1, uniform)	(100, relu)	(10, gini)	(3, 0.5)
EBPCA2	(1, uniform)	(50, relu)	(10, gini)	(4, 0.5)
EBLDA	(1, uniform)	(50, tanh)	(10, gini)	(4, 0.5)

A Figura 28 apresenta os modelos no espaço multiobjetivo definidos pela acurácia balanceada (Acc_{bal}) e o sobreajuste regulado para maximização (Saj_{Acc}). Os modelos foram configurados com os valores encontrados na otimização (Tabela 7). Na Figura 29 estão os modelos no espaço definido pelo indicador F_1 ponderado (F_{1pon}) e o sobreajuste regulado (Saj_{F_1}). No espaço multiobjetivo a fronteira de Pareto pode ser determinada verificando quais modelos são Pareto eficientes. Nesse caso, para as métricas utilizadas o espaço objetivo é de maximização. Nas figuras o agrupamento definido com (*) é o aglomerado de modelos Pareto eficientes de mesmo valor dominantes sobre os demais.

Na Tabela 8 o desempenho dos modelos finais na base de teste é exibido. Os valores apresentados são as médias com seus respectivos desvio padrão. Para os algoritmos que são influenciados por aleatoriedade na construção dos modelos (ELM, RF) foram efetuadas 50 execuções (treinamento e teste) com um conjunto de diferentes números para produzir os valores aleatórios utilizados na construção dos modelos. Para os algoritmos que não dependem de aleatoriedade (KNN, SVM) apenas uma execução de teste foi necessária.

Figura 28: Modelos configurados com hiperparâmetros otimizados (modelos finais) representados no espaço objetivo definido por Acc_{bal} e Saj_{Acc}.



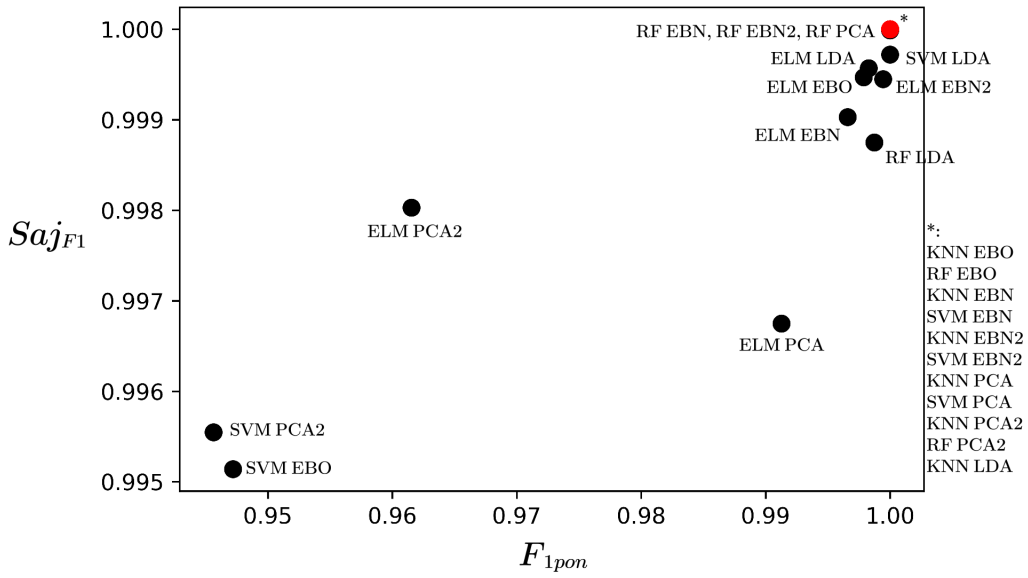
Fonte: O autor.

As avaliações indicaram que os modelos baseados em KNN e RF foram os mais eficientes para o problema, apresentando os melhores resultados em termos de desempenho. Porém, pode ser observado que para os quatro algoritmos avaliados há modelos que se aproximam do resultado com 100% de acurácia. Isso indica que as combinações entre algoritmos e espaços base produzem resultados satisfatórios para a classificação. Em relação aos espaços base, os modelos construídos no espaço base EBN2 obtiveram desempenho geral acima dos demais.

A diferença efetiva entre os espaços base e entre os algoritmos é de difícil avaliação devido a otimização dos hiperparâmetros que tende a levar os modelos ao desempenho máximo possível. Dessa forma, se as métricas têm valores próximos torna-se inviável a comparação de modelos com o intuito de selecionar o melhor, sendo adequado avaliar os resultados individuais e o geral, já que a interpretação destes atende o objetivo da análise.

Ainda, o desempenho mais baixo obtido foi de modelos baseados em SVM nos espaços base EBO e EBPCA, enquanto que os modelos baseados em KNN obtiveram em todos os espaços base um resultado ideal (sem erro). Por fim, em relação à análise de Pareto, todos os

Figura 29: Modelos configurados com hiperparâmetros otimizados (modelos finais) representados no espaço objetivo definido por $F1_{pon}$ e Saj_{F1}.



Fonte: O autor.

Tabela 8: Desempenho dos modelos finais na base de teste.

Modelo	$\boldsymbol{Acc_{bal}(\%)}$	$\boldsymbol{Saj_{Acc}(\%)}$	$\boldsymbol{F_{1pon}(\%)}$	$\boldsymbol{Saj_{F_1}(\%)}$
KNN EBO	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
ELM EBO	99,674 ± 1,272	99,920 ± 0,140	99,787 ± 0,791	99,947 ± 0,087
RF EBO	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
SVM EBO	91,435 ± 0,000	99,272 ± 0,000	94,720 ± 0,000	99,514 ± 0,000
KNN EBN	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
ELM EBN	99,475 ± 1,083	99,841 ± 0,272	99,659 ± 0,684	99,903 ± 0,163
RF EBN	100 ± 0,000	99,999 ± 0,007	100 ± 0,000	99,999 ± 0,004
SVM EBN	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
KNN EBN2	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
ELM EBN2	99,942 ± 0,051	99,942 ± 0,050	99,944 ± 0,048	99,945 ± 0,048
RF EBN2	100 ± 0,000	99,999 ± 0,007	100 ± 0,000	99,999 ± 0,004
SVM EBN2	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
KNN EBPCA	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
ELM EBPCA	98,491 ± 0,735	99,385 ± 0,289	99,128 ± 0,423	99,675 ± 0,169
RF EBPCA	100 ± 0,000	99,998 ± 0,010	100 ± 0,000	99,999 ± 0,006
SVM EBPCA	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
KNN EBPCA2	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
ELM EBPCA2	96,061 ± 1,062	99,828 ± 0,147	96,154 ± 1,120	99,803 ± 0,168
RF EBPCA2	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
SVM EBPCA2	91,204 ± 0,000	99,339 ± 0,000	94,563 ± 0,000	99,555 ± 0,000
KNN EBLDA	100 ± 0,000	100 ± 0,000	100 ± 0,000	100 ± 0,000
ELM EBLDA	99,819 ± 0,052	99,944 ± 0,049	99,828 ± 0,049	99,957 ± 0,030
RF EBLDA	99,874 ± 0,037	99,878 ± 0,043	99,873 ± 0,041	99,875 ± 0,044
SVM EBLDA	100 ± 0,000	99,950 ± 0,000	100 ± 0,000	99,972 ± 0,000

modelos avaliados localizam-se numa região eficiente com acurácia balanceada acima de 0,90 e sobreajuste inferior a 0,01. Em resumo, os modelos acertaram pelo menos 90% das amostras de teste de cada classe durante a avaliação.

4.5 DISCUSSÕES

No que tange a identificação e a discriminação do estágio geral do pessegueiro, os modelos construídos para classificação, assim como as análises de agrupamento, mostraram resultados satisfatórios em relação às métricas avaliadas (Acc_{bal}, F_{1pon}) e à formação de agrupamentos entre as classes consideradas (QP, FFI, FFA, F). Isso indica que a análise de classificação, isto é, o reconhecimento de padrão utilizado para relacionar os dados do nariz eletrônico com a variável de saída definida, foi realizada de forma adequada pelos classificadores (modelos construídos pela combinação dos espaços base (diferentes formas de representar os dados) e algoritmos de aprendizado de máquina (diferentes formas de reconhecer os padrões)).

As avaliações de generalização forneceram informações importantes sobre o comportamento individual dos modelos e enquanto grupo. Desse modo foi possível observar a tendência de agrupamento dos modelos dado as métricas consideradas. A análise utilizada foi a de eficiência de Pareto em um espaço multiobjetivo simplificado. Além desta, outra análise que pode ser realizada quando se quer informações sobre a característica da base de dados e do desempenho dos modelos é de casco convexo produzido pela base de treino (BARBIERO; SQUILLERO; TONDA, 2020).

Ademais, questões mais intrínsecas como a efetividade dos resultados enquanto desempenho de generalização e a reprodutibilidade devido à especificidade do protótipo de nariz eletrônico utilizado podem ser discutidas. A efetividade de generalização pode ser prejudicada devido às variações dos sensores entre as execuções. O que, dependendo a intensidade, pode inviabilizar os dados e consequentemente os modelos construídos, significando que os resultados são válidos somente para o conjunto de medidas realizadas. Nesse caso, para a solução ainda ser válida para o problema haveria necessidade de atualizações constantes para tratar as variações, ou abordar o problema com outras análises mais robustas utilizando o conhecimento produzido em relação aos impactos das variações. A respeito da reprodutibilidade da análise, a principal questão é que ela está relaciona ao equipamento utilizado e a complexidade da amostra analisada. Variações em ambos tendem a reduzir a reprodutibilidade mesmo se um deles for fixado. Essa complexidade da relação entre nariz eletrônico e análise ambiental, isto é, o equipamento num ambiente não controlado, pode ser um limitante neste tipo de pesquisa.

A validação executada com os dois testes (referência e álcool) demonstrou que de fato

as variações dos sensores e do ambiente impactam os resultados, mas que ambos provavelmente estão relacionados. Isso implica que as variações mudam de acordo com ambiente analisado (observado no teste de referência), e que as variações nos sensores entre períodos de execução curtos (minutos), médios (dias) e longos (meses) podem ser corrigidas com os métodos adequados de pré-processamento (observado no teste de álcool). No entanto, a análise de validação não abrange totalmente as questões de efetividade de generalização e reprodutibilidade, mas permite que as características da relação entre sensores e configuração do equipamento, como a estabilidade e repetibilidade nas condições relevantes para o experimento, sejam avaliadas. Desse modo, a validação do equipamento é importante para assegurar a adequação do protótipo, seus resultados e a correta interpretação destes.

Os resultados da validação indicaram que a estabilidade e a repetibilidade do equipamento são dependentes dos sensores e do ambiente, uma vez que provavelmente as variações de ambos estão correlacionadas nos dados. Isto torna a coleta de informações como temperatura e umidade relativa importantes para as análises, do mesmo modo torna-se relevante explorar as diferentes formas de trabalhar com a base de dados por meio das representações possíveis do espaço de entrada. Porém, mesmo se tais resultados não forem satisfatórios, isso não necessariamente significa um impeditivo ou mesmo impossibilidade de investigar as análises posteriores, como a de classificação, pois a técnica de PCA tem suas limitações e não consegue explorar todos os padrões nos dados, apenas a variação (mais precisamente, somente as maiores). O mesmo ocorre com o LDA que pode ter um resultado ruim conforme os dados estão determinados no espaço e as classes definidas. Na prática um resultado ruim em termos de estabilidade e repetibilidade indicaria que durante o processamento o reconhecimento de padrões seria mais complexo, ressaltando a importância de testar diferentes formas de representar os dados de entrada do problema (diferentes espaços base), e levar ao processamento o maior conjunto de informações possíveis (variáveis de entrada).

Além disso, a comparação entre os dados corrigidos por equações (CE) e os dados com temperatura e umidade relativa inclusos na entrada (CTU) não apresentou diferença significativa (5% de significância). Apesar disso, os modelos construídos com os dados corrigidos por equações apresentaram mais vezes valores superiores para as métricas avaliadas. No total de doze modelos, em cinco os que utilizaram dados de CE foram superiores, em dois os que usavam dados de CTU, e nos outros cinco o desempenho foi o mesmo. Esse comportamento sugere que a correção de temperatura e umidade nos sensores pode ser também realizada a nível de processamento, durante a construção dos modelos. Dessa forma, não há necessidade de descrever equações durante o pré-processamento e corrigir os dados antes de processá-los. Como resultado, isso pode evitar a necessidade de fazer considerações sobre os gases

para descrever as equações utilizando as folhas de dados, além de possibilitar o aumento da viabilidade de um aplicação devido o tratamento de temperatura e umidade a nível de modelo.

Os resultados indicam que uma aplicação em campo para identificação dos diferentes estágios do desenvolvimento reprodutivo do pessegueiro é viável a medida que às variações dos sensores entre as execuções sejam adequadamente tratadas. De forma que sejam descartadas ou minimizadas as variações não relevantes. Tal tratamento pode ser feito durante o processamento, porém posteriormente ainda pode haver a necessidade de retreinamento constante do modelo durante a aplicação para se adequar à novas variações. Portanto, um indicativo do momento de raleio pode ser obtido avaliando a resposta do modelo para uma amostra, que deve ser associada ao estágio adequado, assim como as demais informações externas ao modelo, suficientes para compor uma análise efetiva sobre o momento conveniente para execução da técnica. Ainda, o sistema pode ser aplicado em diferentes áreas do pomar e suas respostas sobre o provável estágio fenológico de pessegueiros nesses pontos serem obtidas de modo remoto auxiliando o produtor na tomada de decisão. Dessa maneira, as informações em sequência sobre o estágio de brotação do pessegueiro, isto é o monitoramento contínuo, podem ser úteis para o tomador de decisão, e os dados produzidos podem ser usados para incrementar o modelo na aplicação atual e nas próximas.

Ainda, a interpretação dos resultados dado o problema abordado na pesquisa não é simples. Não sendo possível afirmar que a natureza das variações ao longo do experimento é devido às brotações. A hipótese mais plausível parece ser de que uma parcela da variação venha das brotações e seus diferentes COVs. Isto é, no ambiente do pomar no qual a maior parcela de COVs venha do pessegueiro analisado, em que as variações na emissão de COVs no decorrer do desenvolvimento do pessegueiro estejam reproduzidas nos dados. Espera-se, então, que uma componente dessa variação seja relativa as variações das brotações. Pois, como visto na revisão, em pessegueiros as estruturas que emitem maior quantidade de COVs são as folhas. Porém, diversas possibilidades podem ser discutidas como a estabilidade da emissão de COVs pelas folhas e variação dos COVs das brotações em termos de quantidade e compostos específicos, ou ainda que o padrão de emissão de COVs pelas folhas seja influenciado pelo desenvolvimento das brotações.

A respeito das possíveis contribuições desta pesquisa três componentes podem ser destacadas. A primeira, em relação à estratégia de validação do equipamento que permite uma avaliação geral dos sensores no protótipo sobre como as variações do ambiente e dos sensores podem dificultar ou inviabilizar uma análise. Dessa forma, ao invés de olhar o valor dos sensores individualmente é avaliado o conjunto representativo dos sensores utilizando técnicas de extração de características e agrupamento. A segunda, as contribuições da estratégia de validação,

incorporando a otimização de hiperâmetros, e da avaliação do desempenho de generalização, utilizando representação no espaço objetivo e as métricas adequadas. Essa composição pode facilitar a comparação, a seleção e a observação de tendências entre os modelos. A terceira, a contribuição da pesquisa e seus resultados que incrementam possíveis soluções a problemas abordados com nariz eletrônico em análises ambientais, como a forma de corrigir variações de temperatura e umidade, e as estratégias propostas de pré e processamento dos dados. Porém, ainda há necessidade de maior produção de pesquisas utilizando o nariz eletrônico no ambiente agrícola para fundamentar uma base metodológica robusta e, consequentemente, resultados mais propensos à comparações.

5 CONCLUSÕES

Neste trabalho foi desenvolvido um protótipo de nariz eletrônico e amostras foram coletadas em um pomar (cv. Douradão) ao longo de 39 dias durante o período reprodutivo de um pessegueiro, com o objetivo de analisar a capacidade do sistema em identificar os estágios das brotações floríferas para auxílio à atividade de raleio. Em relação à hipótese da pesquisa, de que seria possível diferenciar os estágios de brotações ao longo do desenvolvimento do pessegueiro por meio COVs produzidos utilizando sensores de gás e reconhecimento de padrões, considera-se que o nariz eletrônico foi capaz de distinguir entre os diferentes estágios propostos (QP, FFI, FFA, F). Apesar da classificação ser realizada de forma adequada, os dados oriundos do nariz eletrônico em um ambiente complexo como um pomar são de difícil interpretação devido às diferentes fontes de variação. Dessa forma, há exigência de um pré-processamento apropriado e um processamento que relacione corretamente as variações adequadas.

O resultado da análise de classificação demonstrou que no pior caso mais de 91% das amostras de teste são classificadas corretamente. Além disso, houveram modelos que atingiram 100% de acurácia balanceada na base de teste e não apresentaram sobreajuste em relação à base de treino. De modo que a otimização de hiperparâmetros foi capaz de levar os algoritmos avaliados (KNN, ELM, RF, SVM) a produzirem modelos com desempenho próximo do ideal (sem erro) para as métricas utilizadas. Em relação aos trabalhos correlatos, em que se analisa as brotações floríferas em laboratório ou o desenvolvimento do pessegueiro em ambiente não controlado, a atual pesquisa apresenta um método computacional que abrange a combinação de algoritmos, espaços base, otimização de hiperparâmetros e a avaliação de desempenho. Além de uma validação de equipamento (análises de amostras de referência e de álcool), uma comparação entre estratégias de correção (compensação por equação e no processamento), e resultado da classificação com menor erro (com parcela considerável dos modelos, 11 de 24, sem erro para as métricas empregadas).

Em suma, nesta pesquisa foi observado que definir o momento de execução do raleio em pessegueiros é uma tarefa naturalmente difícil, sendo uma atribuição do técnico ou do especialista. Mais especificamente, as informações que influenciam a decisão são avaliadas a nível humano. Por exemplo, a necessidade ou não de execução de raleio devido as condições de desenvolvimento dos pessegueiros ao longo do ano. Então, para abordar o problema do momento do raleio de uma forma geral porém precisa, é necessário observar a informação básica utilizada para determinar momento de raleio. Desse modo, analisar os estágios das brotações de pessegueiro representa uma indicação do desenvolvimento deste e permite a tomada de decisão sobre o raleio. Nesse sentido, objetivo da pesquisa se resumiu em analisar as brotações e seu desenvolvimento, e o quanto seria possível reconhecer por meio do aprendizado de máquina

suas características utilizando dados de sensores de gás.

Então, em termos de uma potencial aplicação, levando em consideração as complexidades próprias do ambiente agrícola e do nariz eletrônico, dado o que foi exposto na pesquisa, é possível abordar o problema do raleio com um sistema de suporte à decisão. Dessa forma, utilizar o aprendizado de máquina para construir modelos e compor um sistema de suporte a decisão simples e verificável, levando o maior número de informações possíveis ao tomador de decisão, parece o mais adequado para solucionar o problema. Por fim, conclui-se que foi possível identificar os diferentes estágios de desenvolvimento das brotações ao longo do desenvolvimento reprodutivo do pessegueiro utilizando um protótipo de nariz eletrônico e aprendizado de máquina. O que pode, por meio das informações fornecidas, facilitar o gerenciamento de atividades de raleio em pomares.

Pesquisas futuras podem ser desenvolvidas para explorar outras estratégias de análise com o intuito de compor uma abordagem mais complexa e dinâmica, por exemplo o aprendizado adaptativo para lidar com possíveis variações dos sensores, ou uma análise preditiva mais completa como a análise série temporal para investigar relações mais profundas não abrangidas na classificação, como o desenvolvimento simultâneo dos estágios. Além disso, a pesquisa atual pode ser utilizada como base para trabalhos com testes práticos usando o conjunto específico de equipamento técnica (nariz eletrônico específico e os modelos construídos previamente), e formar bases de dados, assim como análises de aplicação real para o problema. Outra forma de trabalho é desenvolver um modo mais prático de lidar com o raleio por meio da implantação de um sistema de suporte a decisão, e uma equipe multidisciplinar exercendo uma análise ampla com a inclusão de experimentos em laboratório e em campo para avaliar o sistema.

REFERÊNCIAS

ABÊ, V. S. P. **Caracterização vegeto-produtiva de cultivares de pessegueiro em estádio inicial de desenvolvimento**. 2020. Trabalho de Conclusão de Curso (Graduação em Agronomia) - Universidade Federal de Santa Catarina. Curitibanos, 2020.

ADAK, M. F.; YUMUSAK, N. Classification of e-nose aroma data of four fruit types by ABC-based neural network. **Sensors**, MDPI, v. 16, n. 3, p. 304, 2016.

ALVARES, C. A.; STAPE, J. L.; SENTELHAS, P. C.; GONÇALVES, J. L. d. M.; SPAROVEK, G. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, Stuttgart, v. 22, n. 6, p. 711–728, 2013.

ANZANELLO, R.; LAMPUGNANI, C. S. Necessidade de frio no período da dormência em pessegueiros. **Revista Scientia Rural**, v. 1, 2020.

BAIETTO, M.; WILSON, A. D. Electronic-nose applications for fruit identification, ripeness and quality grading. **Sensors**, MDPI, v. 15, n. 1, p. 899–931, 2015.

BARALDI, R.; RAPPARINI, F.; ROSSI, F.; LATELLA, A.; CICCIOLI, P. Volatile organic compound emissions from flowers of the most occuring and economically important species of fruit trees. **Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere**, Elsevier, v. 24, n. 6, p. 729–732, 1999.

BARBIERI, C. R. **Germinação de polén, floração e frutificação efetiva de cultivares de pessegueiros no sudoeste do Paraná**. 2018. Dissertação (Mestrado em Agroecossistemas) - Universidade Tecnológica Federal do Paraná. Dois Vizinhos, 2018.

BARBIERO, P.; SQUILLERO, G.; TONDA, A. Modeling generalization in machine learning: a methodological and computational study. ArXiv preprint: 2006.15680, 2020.

BARBOSA, W. **Desenvolvimento vegetativo e reprodutivo do pessegueiro em pomar compacto sob poda drástica anual**. 1989. Dissertação (Mestrado em Agronomia) - Universidade de São Paulo. Piracicaba, 1989.

BARRETO, C. F.; ANTUNES, L. E. C.; FERREIRA, L. V.; NAVROSKI, R.; BENATI, J. A.; PEREIRA, J. F. M. Mechanical flower thinning in peach trees. **Revista Brasileira de Fruticultura**, SciELO Brasil, v. 41, 2019.

BARRETO, C. F.; FERREIRA, L. V.; NOVROSKI, R.; PEREIRA, J. F. M.; ANTUNES, L. E. C. Raleio mecânico como alternativa no cultivo de pessegueiros. **Revista de Ciências Agrárias**, Sociedade de Ciências Agrárias de Portugal, 2019.

BASSI, D.; MIGNANI, I.; SPINARDI, A.; TURA, D. Peach (Prunus persica (L.) Batsch). *In*: SIMMONDS, M. S. J.; PREEDY, V. R. **Nutritional composition of fruit cultivars**. Amsterdam: Elsevier, p. 535–571, 2016.

BEGHI, R.; BURATTI, S.; GIOVENZANA, V.; BENEDETTI, S.; GUIDETTI, R. Electronic nose and visible-near infrared spectroscopy in fruit and vegetable monitoring. **Reviews in Analytical Chemistry**, De Gruyter, v. 36, n. 4, 2017.

BENEDETTI, S.; BURATTI, S.; SPINARDI, A.; MANNINO, S.; MIGNANI, I. Electronic nose as a non-destructive tool to characterise peach cultivars and to monitor their ripening stage during shelf-life. **Postharvest biology and technology**, Elsevier, v. 47, n. 2, p. 181–188, 2008.

BISHOP, C. M. **Pattern recognition and machine learning**. Singapura: Springer, 2006.

BRANDI, F.; BAR, E.; MOURGUES, F.; HORVÁTH, G.; TURCSI, E.; GIULIANO, G.; LIVERANI, A.; TARTARINI, S.; LEWINSOHN, E.; ROSATI, C. Study of Redhaven peach and its white-fleshed mutant suggests a key role of CCD4 carotenoid dioxygenase in carotenoid and norisoprenoid volatile metabolism. **BMC plant biology**, BioMed Central, v. 11, n. 1, p. 1–14, 2011.

BREZMES, J.; FRUCTUOSO, M. L.; LLOBET, E.; VILANOVA, X.; RECASENS, I.; ORTS, J.; SAIZ, G.; CORREIG, X. Evaluation of an electronic nose to assess fruit ripeness. **IEEE Sensors Journal**, IEEE, v. 5, n. 1, p. 97–108, 2005.

BREZMES, J.; LLOBET, E.; VILANOVA, X.; SAIZ, G.; CORREIG, X. Fruit ripeness monitoring using an electronic nose. **Sensors and Actuators B: Chemical**, Elsevier, v. 69, n. 3, p. 223–229, 2000.

BROWNLEE, J. Classification and regression trees for machine learning. **Machine Learning Mastery**. 2016. Disponível em: https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/. Acesso em: 21 jun. 2022.

BROWNLEE, J. Nested cross-validation for machine learning with python. **Machine Learning Mastery**. 2020. Disponível em: https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/. Acesso em: 09 ago. 2022.

BROWNLEE, J. How to choose an activation function for deep learning. **Machine Learning Mastery**. 2021. Disponível em: https://machinelearningmastery.com/choose-an-activation-function-for-deep-learning/. Acesso em: 17 jun. 2022.

CARAMORI, P. H.; CAVIGLIONE, J. H.; WREGE, M. S.; HERTER, F. G.; HAUAGGE, R.; GONÇALVES, S. L.; CITADIN, I.; RICCE, W. da. S. Zoneamento agroclimático para o pessegueiro e a nectarineira no Eestado do Paraná. **Revista Brasileira de Fruticultura**, SciELO Brasil, v. 30, n. 4, p. 1040–1044, 2008.

CAWLEY, G. C.; TALBOT, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. **The Journal of Machine Learning Research**, JMLR.org, v. 11, p. 2079–2107, 2010.

CENTONZE, V.; LIPPOLIS, V.; CERVELLIERI, S.; DAMASCELLI, A.; CASIELLO, G.; PASCALE, M.; LOGRIECO, A. F.; LONGOBARDI, F. Discrimination of geographical origin of oranges (Citrus sinensis L. Osbeck) by mass spectrometry-based electronic nose and characterization of volatile compounds. **Food chemistry**, Elsevier, v. 277, p. 25–30, 2019.

CHEN, L. Y.; WONG, D. M.; FANG, C. Y.; CHIU, C. I.; CHOU, T. I.; WU, C. C.; CHIU, S. W.; TANG, K. T. Development of an electronic-nose system for fruit maturity and quality monitoring. *In*: IEEE International Conference on Applied System Invention (ICASI), IEEE. **Proceedings**, p. 1129–1130, 2018.

DEBABHUTI, N.; SHARMA, P.; ALI, S. B.; TUDU, B.; BANDYOPADHYAY, R.; SARKAR, M. P.; BHATTACHARYYA, N. Discrimination of the maturity stages of indian mango using QCM based electronic nose. *In*: IEEE International Symposium on Olfaction and Electronic Nose (ISOEN), IEEE. **Proceedings**, p. 1–2, 2019.

DI CARLO, S.; FALASCONI, M. Drift correction methods for gas chemical sensors in artificial olfaction systems: techniques and challenges. *In*: WANG, W. **Advances in chemical sensors**. INTECH Open Access Publisher, p. 306–326, 2012.

DI NATALE, C.; MARTINELLI, E.; PENNAZZA, G.; ORSINI, A.; SANTONICO, M. Data analysis for chemical sensor arrays. *In*: BYRNES, J.; OSTHEIMER, G. **Advances in sensing with security applications**. Dordrecht: Springer, p. 147–169, 2006.

DI NATALE, C.; ZUDE-SASSE, M.; MACAGNANO, A.; PAOLESSE, R.; HEROLD, B.; D'AMICO, A. Outer product analysis of electronic nose and visible spectra: application to the measurement of peach fruit characteristics. **Analytica Chimica Acta**, Elsevier, v. 459, n. 1, p. 107–117, 2002.

DING, Q.; ZHAO, D.; LIU, J.; YANG, Z. Detection of fruits in warehouse using electronic nose. *In*: MATEC Web of Conferences, EDP Sciences. **Proceedings**, v. 232, p. 04035, 2018.

DI ROSA, A. R.; LEONE, F.; CHELI, F.; CHIOFALO, V. Fusion of electronic nose, electronic tongue and computer vision for animal source food authentication and quality assessment - A review. **Journal of Food Engineering**, Elsevier, v. 210, p. 62–75, 2017.

DORCEA, D.; HNATIUC, M.; LAZAR, I. Acquisition and calibration interface for gas sensors. *In*: 24th International Symposium for Design and Technology in Electronic Packaging (SIITME), IEEE. **Proceedings**, p. 120–123, 2018.

DORJI, U.; POBKRUT, T.; KERDCHAROEN, T. Electronic nose based wireless sensor network for soil monitoring in precision farming system. *In*: 9th International Conference on Knowledge and Smart Technology (KST), IEEE. **Proceedings**, p. 182–186, 2017.

DOU, T. X.; SHI, J. F.; LI, Y.; BI, F. C.; GAO, H. J.; HU, C. H.; LI, C. Y.; YANG, Q. S.; DENG, G. M.; SHENG, O.; HE, W. D.; YI, G. J.; DONG, T. Influence of harvest season on volatile aroma constituents of two banana cultivars by electronic nose and HS-SPME coupled with GC-MS. **Scientia Horticulturae**, Elsevier, v. 265, p. 109214, 2020.

DU, D.; WANG, J.; WANG, B.; ZHU, L.; HONG, X. Ripeness prediction of postharvest kiwifruit using a MOS e-nose combined with chemometrics. **Sensors**, MDPI, v. 19, n. 2, p. 419, 2019.

EL-SAYED, A. M.; SPORLE, A.; COLHOUN, K.; FURLONG, J.; WHITE, R.; SUCKLING, D. M. Scents in orchards: floral volatiles of four stone fruit crops and their attractiveness to pollinators. **Chemoecology**, Springer, v. 28, n. 2, p. 39–49, 2018.

EZHILAN, M.; NESAKUMAR, N.; BABU, K. J.; SRINANDAN, C.; RAYAPPAN, J. B. B. An electronic nose for royal delicious apple quality assessment a tri-layer approach. **Food Research International**, Elsevier, v. 109, p. 44–51, 2018.

FABBRI, B.; VALT, M.; PARRETTA, C.; GHERARDI, S.; GAIARDO, A.; MALAGÙ, C.; MANTOVANI, F.; STRATI, V.; GUIDI, V. Correlation of gaseous emissions to water stress

in tomato and maize crops: from field to laboratory and back. **Sensors and Actuators B: Chemical**, Elsevier, v. 303, p. 127227, 2020.

FABC. SMA climatologia campos gerais. **Fundação ABC**. 2021. Disponível em: https://sma.fundacaoabc.org/climatologia/cartas_climaticas/campos_gerais. Acesso em: 23 jun. 2021.

FACHINELLO, J. C.; NACHTIGAL, J. C.; KERSTEN, E. **Fruticultura**: fundamentos e práticas. Pelotas: Embrapa, 2009.

FAN, J.; ZHANG, W.; ZHOU, T.; ZHANG, D.; ZHANG, D.; ZHANG, L.; WANG, G.; CAO, F. Discrimination of Malus taxa with different scent intensities using electronic nose and gas chromatography-mass spectrometry. **Sensors**, MDPI, v. 18, n. 10, p. 3429, 2018.

FAO. FAOSTAT: Crops and livestock products. **FAO**. 2023. Disponível em: http://www.fao.org/faostat/en/#data/QC/. Acesso em: 01 fev. 2023.

FENG, L.; ZHANG, M.; BHANDARI, B.; GUO, Z. A novel method using MOS electronic nose and ELM for predicting postharvest quality of cherry tomato fruit treated with high pressure argon. **Computers and Electronics in Agriculture**, Elsevier, v. 154, p. 411–419, 2018.

FILHO, J. A. S.; MINAMI, K.; KLUGE, R. A. Intensidade de raleio de frutos em pessegueiros 'Flordaprince' conduzidos em pomar com alta densidade de plantio. **Pesquisa Agropecuária Brasileira**, SciELO Brasil, v. 35, p. 1109–1113, 2000.

GAJDOSIK, L. The derivation of the electrical conductance/temperature dependency for tin dioxide gas sensor. **Advances in Electrical and Electronic Engineering**, v. 12, n. 5, p. 529–536, 2014.

GAMBOA, J. C. R.; ALBARRACIN, E. S.; SILVA, A. J. da.; LIMA, L. L. de. A.; FERREIRA, T. A. E. Wine quality rapid detection using a compact electronic nose system: application focused on spoilage thresholds by acetic acid. **LWT**, Elsevier, v. 108, p. 377–384, 2019.

GHOJOGH, B.; SAMAD, M. N.; MASHHADI, S. A.; KAPOOR, T.; ALI, W.; KARRAY, F.; CROWLEY, M. Feature selection and feature extraction in pattern analysis: A literature review. ArXiv preprint, 2019.

GILA, D. M. M.; GARCÍA, J. G.; BELLINCONTRO, A.; MENCARELLI, F.; ORTEGA, J. G. Fast tool based on electronic nose to predict olive fruit quality after harvest. **Postharvest Biology and Technology**, Elsevier, v. 160, p. 111058, 2020.

GROSSE, R. CSC 321 Winter 2018 Intro to Neural Networks and Machine Learning - Lecture 9: Generalization. **cs.toronto.edu**. 2018. Disponível em: http://www.cs.toronto.edu/~rgrosse/courses/csc321_2018/. Acesso em: 09 ago. 2022.

GU, S.; CHEN, W.; WANG, Z.; WANG, J.; HUO, Y. Rapid detection of Aspergillus spp. infection levels on milled rice by headspace-gas chromatography ion-mobility spectrometry (HS-GC-IMS) and e-nose. **Food Science and Technology**, Elsevier, v. 132, p. 109758, 2020.

GU, S.; WANG, J.; WANG, Y. Early discrimination and growth tracking of Aspergillus spp. contamination in rice kernels using electronic nose. **Food chemistry**, Elsevier, v. 292, p. 325–335, 2019.

GUTIERREZ-OSUNA, R. L10: Linear discriminant analysis. **DATAJOBS**. 2022. Disponível em: https://datajobs.com/data-science-repo/LDA-Primer-[Gutierrez-Osuna].pdf. Acesso em: 11 jul. 2022.

HAN, F.; HUANG, X.; AHETO, J. H.; ZHANG, D.; FENG, F. Detection of beef adulterated with pork using a low-cost electronic nose based on colorimetric sensors. **Foods**, MDPI, v. 9, n. 2, p. 193, 2020.

HAN, J.; PEI, J.; KAMBER, M. **Data Mining**: concepts and techniques. 3. ed. Waltham: Elsevier, 2012.

HAO, R.; DU, D.; WANG, T.; YANG, W.; WANG, J.; ZHANG, Q. A comparative analysis of characteristic floral scent compounds in prunus mume and related species. **Bioscience, biotechnology, and biochemistry**, Japan Society for Bioscience, Biotechnology, and Agrochemistry, v. 78, n. 10, p. 1640–1647, 2014.

HE. Tecnichal data MQ-7 gas sensor. Data Sheet. Hanwei Electronics. 2022.

HE. Tecnichal data MQ-9 gas sensor. Data Sheet. Hanwei Electronics. 2022.

HINES, E.; LLOBET, E.; GARDNER, J. Neural network based electronic nose for apple ripeness determination. **Electronics Letters**, IET, v. 35, n. 10, p. 821–823, 1999.

HORVAT, R. J.; CHAPMAN, G. W. Comparison of volatile compounds from peach fruit and leaves (cv. Monroe) during maturation. **Journal of Agricultural and Food Chemistry**, ACS Publications, v. 38, n. 7, p. 1442–1444, 1990.

HUANG, G. B.; ZHU, Q. Y.; SIEW, C. K. Extreme learning machine: a new learning scheme of feedforward neural networks. *In*: IEEE international joint conference on neural networks, IEEE. **Proceedings**, v. 2, p. 985–990, 2004.

HUANG, G. B.; ZHU, Q. Y.; SIEW, C. K. Extreme learning machine: theory and applications. **Neurocomputing**, Elsevier, v. 70, n. 1-3, p. 489–501, 2006.

HUANG, L.; MENG, L.; ZHU, N.; WU, D. A primary study on forecasting the days before decay of peach fruit using near-infrared spectroscopy and electronic nose techniques. **Postharvest Biology and Technology**, Elsevier, v. 133, p. 104–112, 2017.

HUERTA, R.; MOSQUEIRO, T.; FONOLLOSA, J.; RULKOV, N. F.; RODRIGUEZ-LUJAN, I. Online decorrelation of humidity and temperature in chemical sensors for continuous monitoring. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 157, p. 169–176, 2016.

IBGE. Produção agrícola - lavoura permanente. **IBGE**. 2023. Disponível em: https://cidades.ibge.gov.br/brasil/pr/ponta-grossa/pesquisa/15/11863. Acesso em: 01 fev. 2023.

IKEGAMI, A.; KANEYASU, M. Olfactory detection using integrated sensors. *In*: 3 rd international conference on solid-state sensors and actuators (Tranducers). **Proceedings**, p. 136–139, 1985.

JIA, X. M.; MENG, Q. H.; JING, Y. Q.; QI, P. F.; ZENG, M.; MA, S. G. A new method combining KECA-LDA with ELM for classification of Chinese liquors using electronic nose. **IEEE Sensors Journal**, IEEE, v. 16, n. 22, p. 8010–8017, 2016.

JIANG, S.; WANG, J. Internal quality detection of Chinese pecans (Carya cathayensis) during storage using electronic nose responses combined with physicochemical methods. **Postharvest Biology and Technology**, Elsevier, v. 118, p. 17–25, 2016.

KANADE, A.; SHALIGRAM, A. Ripening state determination of guava fruit (Psidium guajava) using e-nose with fuzzy logic as pattern recognition tool. **International Journal of Scientific Research Engineering and Technology**, v. 7, n. 4, p. 362–367, 2018.

KANEYASU, M.; IKEGAMI, A.; ARIMA, H.; IWANAGA, S. Smell identification using a thick-film hybrid gas sensor. **IEEE transactions on components, hybrids, and manufacturing technology**, IEEE, v. 10, n. 2, p. 267–273, 1987.

KARAKAYA, D.; ULUCAN, O.; TURKAN, M. Electronic nose and its applications: a survey. **International Journal of Automation and Computing**, Springer, v. 17, n. 2, p. 179–209, 2020.

KASHWAN, K. R.; BHUYAN, M. Robust electronic-nose system with temperature and humidity drift compensation for tea and spice flavour discrimination. *In*: Asian Conference on Sensors and the International Conference on New Techniques in Pharmaceutical and Biomedical Research, IEEE. **Proceedings**, p. 154–158, 2005.

KHALID, S.; KHALIL, T.; NASREEN, S. A survey of feature selection and feature extraction techniques in machine learning. *In*: Science and information conference, IEEE. **Proceedings**, p. 372–378, 2014.

KIANI, S.; MINAEI, S.; GHASEMI-VARNAMKHASTI, M. Real-time aroma monitoring of mint (Mentha spicata l.) leaves during the drying process using electronic nose system. **Measurement**, Elsevier, v. 124, p. 447–452, 2018.

LAI, J.; WANG, X.; LI, R.; SONG, Y.; LEI, L. BD-ELM: A regularized extreme learning machine using biased dropconnect and biased dropout. **Mathematical Problems in Engineering**, Hindawi, v. 2020, 2020.

LAVANYA, S.; DEEPIKA, B.; NARAYANAN, S.; MURTHY, V. K.; UMA, M. V. Indicative extent of humic and fulvic acids in soils determined by electronic nose. **Computers and Electronics in Agriculture**, Elsevier, v. 139, p. 198–203, 2017.

LEE, W. H.; CHOI, S.; OH, I. N.; SHIM, J. Y.; LEE, K. S.; AN, G.; PARK, J. T. Multivariate classification of the geographic origin of Chinese cabbage using an electronic nose-mass spectrometry. **Food science and biotechnology**, Springer, v. 26, n. 3, p. 603–609, 2017.

LEGGIERI, M. C.; MAZZONI, M.; FODIL, S.; MOSCHINI, M.; BERTUZZI, T.; PRANDINI, A.; BATTILANI, P. An electronic nose supported by an artificial neural network for the rapid detection of aflatoxin B1 and fumonisins in maize. **Food Control**, Elsevier, v. 123, p. 107722, 2021.

LI, J.; ZHU, S.; JIANG, S.; WANG, J. Prediction of egg storage time and yolk index based on electronic nose combined with chemometric methods. **Food Science and Technology**, Elsevier, v. 82, p. 369–376, 2017.

LI, S.; YUAN, X.; XU, Y.; LI, Z.; FENG, Z.; YUE, X.; PAOLETTI, E. Biogenic volatile organic compound emissions from leaves and fruits of apple and peach trees during fruit development. **Journal of Environmental Sciences**, Elsevier, v. 108, p. 152–163, 2021.

LIN, T.; SHAH, S. B.; WANG-LI, L.; OVIEDO-RONDÓN, E. O.; POST, J. Development of MOS sensor-based NH3 monitor for use in poultry houses. **Computers and Electronics in Agriculture**, Elsevier, v. 127, p. 708–715, 2016.

LIU, Q.; SUN, K.; ZHAO, N.; YANG, J.; ZHANG, Y.; MA, C.; PAN, L.; TU, K. Information fusion of hyperspectral imaging and electronic nose for evaluation of fungal contamination in strawberries during decay. **Postharvest Biology and Technology**, Elsevier, v. 153, p. 152–160, 2019.

LIU, Q.; ZHAO, N.; ZHOU, D.; SUN, Y.; SUN, K.; PAN, L.; TU, K. Discrimination and growth tracking of fungi contamination in peaches using electronic nose. **Food chemistry**, Elsevier, v. 262, p. 226–234, 2018.

MATAS, J.; KOSTLIVÁ, J. Linear discriminant analysis. **CourseWare Wiki**. 2014. Disponível em: https://cw.fel.cvut.cz/old/media/courses/ae4b33rpz/lectures/lda_2014_06_08.pdf. Acesso em: 11 jul. 2022.

MAYER, N. A.; FRANZON, R. C.; RASEIRA, M. do. C. B. **Pêssego, nectarina e ameixa**: o produtor pergunta, a Embrapa responde. Brasília: Embrapa, 2019.

MOUNZER, O. H.; CONEJERO, W.; NICOLÁS, E.; ABRISQUETA, I.; GARCIA-ORELLANA, Y. V.; TAPIA, L. M.; VERA, J.; ABRISQUETA, J. M.; RUIZ-SÁNCHEZ, M. del. C. Growth pattern and phenological stages of early-maturing peach trees under a Mediterranean climate. **HortScience**, ASHS, v. 43, n. 6, p. 1813–1818, 2008.

OATES, M. J.; FOX, P.; SANCHEZ-RODRIGUEZ, L.; CARBONELL-BARRACHINA, Á. A.; RUIZ-CANALES, A. DFT based classification of olive oil type using a sinusoidally heated, low cost electronic nose. **Computers and Electronics in Agriculture**, Elsevier, v. 155, p. 348–358, 2018.

OLIVEIRA, P. D. de.; MARODIN, G. A. B.; ALMEIDA, G. K. de.; GONZATTO, M. P.; DARDE, D. C. Heading of shoots and hand thinning of flowers and fruits on 'brs kampai' peach trees. **Pesquisa Agropecuária Brasileira**, SciELO Brasil, v. 52, n. 11, p. 1006–1016, 2017.

PEARCE, T. C.; SCHIFFMAN, S. S.; NAGLE, H. T.; GARDNER, J. W. **Handbook of machine olfaction**: electronic nose technology. Weinheim: John Wiley & Sons, 2003.

PERSAUD, K.; DODD, G. Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. **Nature**, Nature Publishing Group, v. 299, n. 5881, p. 352–355, 1982.

PIRES, E. H. de. S. **Projeto de uma unidade de monitoramento e controle ambiental**. 2018. Trabalho de Conclusão de Curso (Graduação em Engenharia Mecatrônica) - Universidade Federal de Uberlândia. Uberlândia, 2018.

QIN, X. W.; HAO, C. Y.; HE, S. Z.; WU, G.; TAN, L. H.; XU, F.; HU, R. S. Volatile organic compound emissions from different stages of Cananga odorata flower development. **Molecules**, MDPI, v. 19, n. 7, p. 8965–8980, 2014.

QIU, S.; WANG, J. The prediction of food additives in the fruit juice based on electronic nose with chemometrics. **Food chemistry**, Elsevier, v. 230, p. 208–214, 2017.

RADULOVIĆ, N. S.; ĐORđEVIĆ, A. S.; ZLATKOVIĆ, B. K.; PALIĆ, R. M. GC-MS analyses of flower ether extracts of Prunus domestica L. and Prunus padus L. (Rosaceae). **Chemical papers**, Springer, v. 63, n. 4, p. 377–384, 2009.

RAIMUNDO, M. M. **Otimização multiobjetivo em aprendizado de máquina**. 2018. Tese (Doutorado em Engenharia Elétrica-Computação). Universidade Estadual de Campinas. Campinas, 2018.

SANAEIFAR, A.; MOHTASEBI, S. S.; GHASEMI-VARNAMKHASTI, M.; AHMADI, H.; LOZANO, J. Development and application of a new low cost electronic nose for the ripeness monitoring of banana using computational techniques (PCA, LDA, SIMCA and SVM). **Czech Journal of Food Sciences**, v. 32, n. 6, p. 538–548, 2014.

SANAEIFAR, A.; MOHTASEBI, S. S.; GHASEMI-VARNAMKHASTI, M.; AHMADI, H. Application of MOS based electronic nose for the prediction of banana quality properties. **Measurement**, Elsevier, v. 82, p. 105–114, 2016.

SCIKIT-LEARN. Metrics and scoring: quantifying the quality of predictions. **Scikit-learn**. 2022. Disponível em: https://scikit-learn.org/stable/modules/model_evaluation.html. Acesso em: 09 ago. 2022.

SCIKIT-LEARN. Nested versus non-nested cross-validation. **Scikit-learn**. 2022. Disponível em: https://scikit-learn.org/stable/auto_examples/model_selection/plot_nested_cross_validation_iris. Acesso em: 09 ago. 2022.

SCIKIT-LEARN. Sklearn.discriminant_analysis.LinearDiscriminantAnalysis. **Scikit-learn**. 2022. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html. Acesso em: 11 jul. 2022.

SCIKIT-LEARN. Sklearn.metrics.f1_score. **Scikit-learn**. 2022. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score. Acesso em: 09 ago. 2022.

SENTHILKUMAR, R.; VENKATAKRISHNAN, P.; BALAJI, N. Intelligent based novel embedded system based IoT enabled air pollution monitoring system. **Microprocessors and Microsystems**, Elsevier, v. 77, p. 103172, 2020.

SILVELLO, G. C.; ALCARDE, A. R. Experimental design and chemometric techniques applied in electronic nose analysis of wood-aged sugar cane spirit (cachaça). **Journal of Agriculture and Food Research**, Elsevier, v. 2, p. 100037, 2020.

STACKEXCHANGE. Cross Validated - Maximal depth of a decision tree. **StackExchange**. 2013. Disponível em: https://stats.stackexchange.com/questions/65893/maximal-depth-of-a-decision-tree? Acesso em: 21 jun. 2022.

STACKEXCHANGE. MATHEMATICS - Are all eigenvectors, of any matrix, always orthogonal?. **StackExchange**. 2017. Disponível em: https://math.stackexchange.com/questions/142645/are-all-eigenvectors-of-any-matrix-always-orthogonal. Acesso em: 11 jul. 2022.

STACKEXCHANGE. Cross Validated - Why are the discriminant axes in linear discriminant analysis (LDA) not orthogonal?. **StackExchange**. 2018. Disponível em:

https://stats.stackexchange.com/questions/354847/why-are-the-discriminant-axes-in-linear-discriminant-analysis-lda-not-orthogon. Acesso em: 11 jul. 2022.

SU, M.; ZHANG, B.; YE, Z.; CHEN, K.; GUO, J.; GU, X.; SHEN, J. Pulp volatiles measured by an electronic nose are related to harvest season, TSS concentration and TSS/TA ratio among 39 peaches and nectarines. **Scientia Horticulturae**, Elsevier, v. 150, p. 146–153, 2013.

SZCZUREK, A.; MACIEJEWSKA, M.; ZAJICZEK, Ż.; BĄK, B.; WILK, J.; WILDE, J.; SIUDA, M. The effectiveness of Varroa destructor infestation classification using an E-nose depending on the time of day. **Sensors**, MDPI, v. 20, n. 9, p. 2532, 2020.

TAN, J.; XU, J. Applications of electronic nose (e-nose) and electronic tongue (e-tongue) in food quality-related properties determination: a review. **Artificial Intelligence in Agriculture**, Elsevier, 2020.

TIAN, F.; ZHANG, J.; YANG, S. X.; ZHAO, Z.; LIANG, Z.; LIU, Y.; WANG, D. Suppression of strong background interference on e-nose sensors in an open country environment. **Sensors**, MDPI, v. 16, n. 2, p. 233, 2016.

TIMSORN, K.; THOOPBOOCHAGORN, T.; LERTWATTANASAKUL, N.; WONGCHOOSUK, C. Evaluation of bacterial population on chicken meats using a briefcase electronic nose. **Biosystems Engineering**, Elsevier, v. 151, p. 116–125, 2016.

UÇAR, A.; ÖZALP, R. Efficient android electronic nose design for recognition and perception of fruit odors using Kernel Extreme Learning Machines. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 166, p. 69–80, 2017.

VALENTE, J.; MUNNIKS, S.; MAN, I. de.; KOOISTRA, L. Validation of a small flying e-nose system for air pollutants control: a plume detection case study from an agricultural machine. *In*: International Conference on Robotics and Biomimetics (ROBIO), IEEE. **Proceedings**, p. 1993–1998, 2018.

VALERIA, M.; SILVIA, R.; ROSA, B.; NOEMÍ, W. de. R. Variation of odour profile detected in the floral stages of Prunus persica (L) Batsch using an electronic nose. *In*: AIP Conference Proceedings, AIP. **Proceedings**, v. 1137, n. 1, p. 465–468, 2009.

VOSS, H. G. J. **Desenvolvimento de um nariz eletrônico aplicado à determinação do estágio de maturação em pessegueiros**. 2019. Dissertação (Mestrado em Computação Aplicada) - Universidade Estadual de Ponta Grossa. Ponta Grossa, 2019.

VOSS, H. G. J.; STEVAN JUNIOR, S. L.; AYUB, R. A. Peach growth cycle monitoring using an electronic nose. **Computers and Electronics in Agriculture**, Elsevier, v. 163, p. 104858, 2019.

WANG, Y.; DIAO, J.; WANG, Z.; ZHAN, X.; ZHANG, B.; LI, N.; LI, G. An optimized deep convolutional neural network for dendrobium classification based on electronic nose. **Sensors and Actuators A: Physical**, Elsevier, v. 307, p. 111874, 2020.

WEAVERDYCK, M. E.; LIEBERMAN, M. D.; PARKINSON, C. Multivoxel pattern analysis in fMRI: a practical introduction for social and affective neuroscientists. **Social Cognitive and Affective Neuroscience**, Oxford University Press, v. 15, p. 487–509, 2020.

WEERAWATANAKORN, M.; ASIKIN, Y.; KAMCHONEMENUKOOL, S.; TAMAKI, H.; TAKARA, K.; WADA, K. Physicochemical, antioxidant, volatile component, and mass spectrometry-based electronic nose analyses differentiated unrefined non-centrifugal cane, palm, and coconut sugars. **Journal of Food Measurement and Characterization**, Springer, v. 15, n. 2, p. 1563–1577, 2021.

WEI, X.; ZHANG, Y.; WU, D.; WEI, Z.; CHEN, K. Rapid and non-destructive detection of decay in peach fruit at the cold environment using a self-developed handheld electronic-nose system. **Food Analytical Methods**, Springer, v. 11, n. 11, p. 2990–3004, 2018.

WEN, T.; ZHENG, L.; DONG, S.; GONG, Z.; SANG, M.; LONG, X.; LUO, M.; PENG, H. Rapid detection and classification of citrus fruits infestation by Bactrocera dorsalis (Hendel) based on electronic nose. **Postharvest Biology and Technology**, Elsevier, v. 147, p. 156–165, 2019.

WIJAYA, D. R.; SARNO, R.; ZULAIKA, E.; SABILA, S. I. Development of mobile electronic nose for beef quality monitoring. **Procedia Computer Science**, Elsevier, v. 124, p. 728–735, 2017.

WU, Z.; WANG, H.; WANG, X.; ZHENG, H.; CHEN, Z.; MENG, C. Development of electronic nose for qualitative and quantitative monitoring of volatile flammable liquids. **Sensors**, MDPI, v. 20, n. 7, p. 1817, 2020.

XIN, R.; LIU, X.; WEI, C.; YANG, C.; LIU, H.; CAO, X.; WU, D.; ZHANG, B.; CHEN, K. E-nose and GC-MS reveal a difference in the volatile profiles of white-and red-fleshed peach fruit. **Sensors**, MDPI, v. 18, n. 3, p. 765, 2018.

XU, K.; FU, C.; GAO, Z.; WEI, F.; YING, Y.; XU, C.; FU, G. Nanomaterial-based gas sensors: A review. **Instrumentation Science & Technology**, Taylor & Francis, v. 46, n. 2, p. 115–145, 2018.

XU, S.; LÜ, E.; LU, H.; ZHOU, Z.; WANG, Y.; YANG, J.; WANG, Y. Quality detection of litchi stored in different environments using an electronic nose. **Sensors**, MDPI, v. 16, n. 6, p. 852, 2016.

YAN, J.; ZHANG, M.; PENG, B.; SU, Z.; XU, Z.; CAI, Z.; YANG, J.; MA, R.; YU, M.; SHEN, Z. Predicting chilling requirement of peach floral buds using electronic nose. **Scientia Horticulturae**, Elsevier, v. 290, p. 110517, 2021.

YAN, M.; WU, Y.; HUA, Z.; LU, N.; SUN, W.; ZHANG, J.; FAN, S. Humidity compensation based on power-law response for MOS sensors to VOCs. **Sensors and Actuators B: Chemical**, Elsevier, v. 334, p. 129601, 2021.

YANG, X.; CHEN, J.; JIA, L.; YU, W.; WANG, D.; WEI, W.; LI, S.; TIAN, S.; WU, D. Rapid and non-destructive detection of compression damage of yellow peach using an electronic nose and chemometrics. **Sensors**, MDPI, v. 20, n. 7, p. 1866, 2020.

ZHANG, J.; XUE, Y.; SUN, Q.; ZHANG, T.; CHEN, Y.; YU, W.; XIONG, Y.; WEI, X.; YU, G.; WAN, H. *et al.* A miniaturized electronic nose with artificial neural network for anti-interference detection of mixed indoor hazardous gases. **Sensors and Actuators B: Chemical**, Elsevier, v. 326, p. 128822, 2021.

ZHANG, L.; TIAN, F.; KADRI, C.; XIAO, B.; LI, H.; PAN, L.; ZHOU, H. On-line sensor calibration transfer among electronic nose instruments for monitoring volatile organic chemicals in indoor air quality. **Sensors and Actuators B: Chemical**, Elsevier, v. 160, n. 1, p. 899–909, 2011.

ZHU, D.; REN, X.; WEI, L.; CAO, X.; GE, Y.; LIU, H.; LI, J. Collaborative analysis on difference of apple fruits flavour using electronic nose and electronic tongue. **Scientia Horticulturae**, Elsevier, v. 260, p. 108879, 2020.