UNIVERSIDADE FEDERAL DE ALAGOAS
CENTRO DE TECNOLOGIA
CURSO DE ENGENHARIA QUÍMICA

JARDIEL GUSTAVO DOS SANTOS SILVA

DESENVOLVIMENTO DE FERRAMENTA COMPUTACIONAL PARA RECONHECIMENTO DE PADRÕES NA INDÚSTRIA SUCROENERGÉTICA

Maceió
2022

JARDIEL GUSTAVO DOS SANTOS SILVA

DESENVOLVIMENTO DE FERRAMENTA COMPUTACIONAL PARA RECONHECIMENTO DE PADRÕES NA INDÚSTRIA SUCROENERGÉTICA

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do título de Bacharel em Engenharia Química, sob orientação do Professor Dr. Wagner Roberto de Oliveira Pimentel.

Maceió

2022

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Bibliotecária: Cláudio César Temóteo Galvino – CRB-4/1459

S586d Silva, Jardiel Gustavo dos Santos.
Desenvolvimento de ferramenta computacional para reconhecimento de padrões na indústria sucroenergética / Jardiel Gustavo dos Santos Silva. – Maceió, 2022.
47 f. : il.

Orientador: Wagner Roberto de Oliveira Pimentel.
Trabalho de conclusão de curso (Graduação em Engenharia Química) – Universidade Federal de Alagoas. Centro de Tecnologia. Maceió, 2022.

Bibliografia: f. 44-47.

1. Python. 2. Cana-de-açúcar. 3. Aprendizagem de máquina. I. Título.

CDU: 66.0

AGRADECIMENTOS

Em primeiro lugar, sou grato a aquele que foi meu conselheiro, meu melhor amigo e minha bússola até hoje: o meu Deus. Sem Ele eu não conseguiria fazer nada, nada poderia sonhar ou conquistar. Ele quem me formou, quem conhece tudo que há em mim, desde todos os meus atos até os mais profundos pensamentos. Minha vida é um milagre e tudo que tenho e sou hoje, devo a Ele.

Sou grato às três guerreiras que assumiram a responsabilidade de me conduzir por caminhos de bondade e amor: Minha mãe Jeane, minha avó Juralene e minha Bisavó Antônia. Com muita dificuldade, essas mulheres me proporcionaram a melhor educação que eu poderia receber e nunca mediram esforços em abraçar meus sonhos.

Agradeço também ao amor da minha vida: minha esposa Myllena. Agradeço por sonhar comigo, por entender todo o tempo ausente devido aos estudos, por escolher se arriscar nesse mundo de incertezas ao meu lado.

Sou grato por todos da minha família: meu pai, meus avós, meus tios, minhas irmãs, meus sogros, meus padrinhos, meus cunhados e primo.

Agradeço aos meus líderes durante o período de estágio. O profissional que existe em mim cresceu de uma forma gigantesca ao encontrar com vocês.

Agradeço a todos os colegas de turma e aos professores que proporcionaram grandes momentos de aprendizado durante a minha formação. Por fim, sou grato ao meu orientador, professor Drº Wagner Roberto de Oliveira Pimentel por me escolher como seu orientando, acreditar em mim e no meu potencial! Nunca conseguirei recompensá-lo da forma como mereces! Obrigado por todo o aprendizado e paciência!

RESUMO

A cor do açúcar é um dos principais parâmetros de qualidade avaliados pelos mercados interno e externo. Os precursores de cor podem ser gerados durante as etapas de produção ou encontrados na cana-de-açúcar na forma de aminoácidos, pigmentos e compostos fenólicos. A implementação de metodologias para detecção e quantificação desses corantes nem sempre é possível dentro da rotina de análises dos laboratórios industriais. Como alternativa ao monitoramento analítico, nota-se o crescente uso de ferramentas de aprendizagem de máquina para monitoramento e previsão da qualidade de alguns produtos no contexto industrial e com isso, o presente trabalho buscou o desenvolvimento de uma ferramenta computacional de reconhecimento de padrões utilizando algoritmos de aprendizagem de máquina para identificar relações entre a qualidade do açúcar produzido com as variedades de cana-de-açúcar utilizadas no processo de produção. O conjunto de dados para construção da ferramenta foi composto por sete variáveis referentes a variedades de cana-de-açúcar, cinco variáveis referentes às propriedades físico-químicas do caldo misto resultante e uma variável referente a cor do respectivo açúcar produzido. As amostras de açúcar foram previamente classificadas como "Dentro da Especificação" e "Fora da Especificação" de acordo com as especificações de cor adotadas pela Comissão Internacional para Métodos Uniformes de Análise de Açúcar (ICUMSA) para o açúcar do tipo VHP. Os dados passaram por um pré-processamento conhecido como autoescalamento e foram submetidos a uma análise exploratória por meio da análise de componentes principais (PCA). Durante a análise exploratória foi possível visualizar a estrutura dos dados, reconhecer e eliminar amostras anômalas e verificar relações entre as variáveis. Quatro modelos de classificação foram construídos em ambiente Python, utilizando a biblioteca *scikit-learn* e comparados; dois deles baseados no algoritmo do k-ésimo vizinho mais próximo (KNN e PCA-KNN) com k variando entre 3, 4 e 5 e os outros dois baseados no algoritmo de redes neurais artificiais (MLP e PCA-MLP). Os modelos que apresentaram menor capacidade de predição foram os KNN (k = 3 e 4) com 96,4% das amostras classificadas corretamente. Os demais modelos apresentaram 100% de eficiência na classificação de amostras do conjunto de teste. Uma interface gráfica foi desenvolvida de modo a possibilitar o uso do modelo computacional por pessoas que não dispõem de conhecimentos relacionados a linguagens de programação. O algoritmo utilizado na interface foi o PCA-KNN por apresentar 100% de eficiência e requerer menor tempo computacional para sua construção. Percebeu-se que a ferramenta desenvolvida se tornou acessível para qualquer pessoa que tenha acesso a um computador e apresentou boa capacidade para prever se um açúcar sairá dentro ou fora da especificação com base na matéria-prima processada, tornando possível sua utilização no planejamento estratégico de produção por meio da avaliação prévia dos lotes de cana-de-açúcar enviados para a indústria, evitando assim, a fabricação de produtos fora da especificação.

Palavras-Chave: Python, Cana-de-açúcar, aprendizagem de máquina.

ABSTRACT

Sugar color is one of the main quality parameters for domestic and foreign markets. Color precursors can be generated during production steps or found in sugarcane in the form of amino acids, pigments and phenolic compounds. The implementation of methodologies for detection and quantification of these dyes is not always possible within the routine of industrial laboratory tests. As an alternative to monitoring the use of such detailed development tools, recognizing or increasing machine learning and monitoring the quality of industrial use or a use of work analysis tools or a use of used learning work tools machine to identify quality relationships between the sugar produced and the sugarcane varieties used in the production process. The tools were composed of seven references to sugarcane varieties, five variables identified to the resultant-chemical properties of the mixed dataset and a variable referring to the color of the respective sugar produced. The sugar specifications were originally classified as "In Specification" and "Out of Specification Sugar" according to the color specifications adopted by the International Commission for Uniform Methods of Sugar Analysis (ICUMSA) for the VHP type. The component analysis data known by an autoscaling preprocessing was selected by a middle of the principal exploration (PCA). During an analysis it was possible to visualize a data structure, recognize and eliminate anomalous samples and verify relationships between variables. Four classification models were built in a Python environment, using the scikit-learn library and compared; two of them close networks in the making k-is k-is nearimmost closest (K and k-is close neighbors)3, 45 and the N LPN two N and nm k-is of closest neighbors (K-is LP-K-M-M-N) and in the k-é k-é nextimmost closer (K-ép) The models that present the lowest prediction capacity were the KNN (k = 3 and 4) with 96.4% of the dates correctly classified. The other models showed 100% efficiency in the sample classification of the test set. A graphical interface was developed computationally by people or using models unrelated to knowledge related to programming languages. The one used in the interface was the PCA-KNN for presenting 100% efficiency and requiring less computational time for its construction. The developed tool became accessible to anyone who has access to a useful tool presented and leaves for a specification based on the raw material of the process, making it possible to use it in the strategic planning Sale through the prior evaluation of the lots of sugarcane for production, thus manufacturing off-spec products.

Keywords: Python, Sugarcane, Machine Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AR	Açúcares Redutores
ART	Açúcares Redutores Totais
HMF	Hidroximetilfurfural
ICUMSA	*International Commission for Uniform Methods of Sugar Analysis* (Comissão Internacional para Métodos Uniformes de Análise de Açúcar)
INMETRO	Instituto Nacional de Metrologia, Qualidade e Tecnologia
KNN	*k-nearest neighbor* (K-ésimo vizinho mais próximo)
MLP	*Multi Layer Perceptron* (Perceptron Multicamadas)
NIR	*Near Infrared Spectroscopy* (espectroscopia de infravermelho próximo)
ºBrix	Concentração de sólidos solúveis
P2O5	Concentração de fosfato na forma de P_2O_5
PC	*Principal Component* (Componente Principal)
PCA	*Principal Component Analysis* (Análise de Componentes Principais)
pH	Potencial Hidrogeniônico
PLS	*Partial Least Squares* (mínimos quadrados parciais)
POL	Concentração de Sacarose aparente
R1	Variedade de Cana-de-açúcar RB95 1541
R5	Variedade de Cana-de-açúcar RB86 7515
R9	Variedade de Cana-de-açúcar RB 92579
RF	*Random Forest* (Florestas Aleatórias)
RNA	Rede Neural artificial

S0	Variedade de Cana-de-açúcar SP81 3250
S1	Variedade de Cana-de-açúcar SP79 1011
S9	Variedade de Cana-de-açúcar SP71 6949
SVD	*Singular Value Decomposition* (decomposição por valores singulares)
SVM	*Support Vector Machine* (Máquina de vetores de suporte)
SVR	*Support Vector Regressor*
SX	Somatório das outras variedades de Cana-de-açúcar
VHP	*Very High Polarization* (Polarização muito alta)

LISTA DE SÍMBOLOS

Análise de componentes principais

$\%Var_a$	Porcentagem de variância explicada
E	Matriz de Erros
P^T	Matriz dos pesos transposta
s	Valor singular
T	Matriz dos escores
X	Matriz de dados originais
λ_a	Autovalor da matriz de correlação

K-ésimo vizinho mais próximo

d_{AB}	Distância euclidiana entre as amostras A e B
x_A	Vetor com elementos da amostra A
x_B	Vetor com elementos da amostra B

Rede Neural Artificial

y	Resposta do neurônio artificial
x_{ij}	Dado de entrada no neurônio *i* da camada *j*
w_{ij}	Peso sináptico no neurônio *i* da camada *j*
ϕ	Função de ativação
θ	Limiar de ativação

Autoescalamento

$x_{ij(as)}$	Valor do dado autoescalado para a variável *j* na amostra *i*
x_{ij}	Valor da variável *j* na amostra *i*
$\overline{x_j}$	Média dos valores das amostras em uma coluna *j*
s_j	Desvio padrão dos valores da variável *j*

SUMÁRIO

1 INTRODUÇÃO

Do processo de industrialização da cana-de-açúcar é possível obter vários produtos, dentre eles o açúcar, em diferentes tipos e podendo atender a diversas especificações (ZACURA FILHO; PICCIRILLI, 2012). Dentre outras características, a cor ICUMSA é vista por muitos como o principal parâmetro de qualidade do açúcar. É sabido o impacto que um açúcar de maior cor traz para produtos acabados que tem na sua cor o diferencial de mercado (refrigerantes, biscoitos, entre outros) e devido a isto, é desejado que o açúcar utilizado em tais processos não altere essa característica (OLIVEIRA; ESQUIAVETO; SILVA JÚNIOR, 2007).

A cor final do açúcar pode estar associada a fatores que estão ligados à qualidade da matéria-prima utilizada **(**SMITH *et al.*, 1990**)** e à forma de condução do processo produtivo (SMITH *et al.*, 1981). Devido ao nível de complexidade de monitoramento dos fatores que influenciam a cor, a implementação de metodologias para identificação e quantificação desses corantes nos laboratórios industriais não é simples, dada a rotina de análises que existe nestes ambientes (SANTOS; BORÉM; CALDAS, 2018). Em razão disto, na maioria das vezes, um açúcar fora da especificação só é identificado após algumas horas de produção, gerando prejuízos à indústria em consequência da redução do valor agregado ao produto.

Como alternativa ao monitoramento analítico dos fatores que influenciam na cor final do açúcar, tem-se o uso de ferramentas computacionais para a previsão de características do produto final com base nos dados coletados no processo (BHARDWAJ *et al.*; 2022). Nota-se que a aquisição de dados no setor sucroenergético, especialmente de natureza química e físico-química, atingiu um ponto bastante sofisticado com o interfaceamento de instrumentos de medida aos computadores produzindo uma enorme quantidade de informação, muitas vezes complexas e variadas, necessitando, pois, de técnicas de tratamentos de dados mais complexas do ponto de vista matemático e estatístico (SENA *et. al.;* 1999).

Dentre as ferramentas computacionais disponíveis, os algoritmos de aprendizagem de máquina têm se tornado cada vez mais populares. Aprendizagem de máquina ou *Machine learning* é o ramo da inteligência artificial que estuda algoritmos capazes de aprender de forma autônoma, diretamente dos dados (GERÓN, 2019). Uma das formas de aplicações dos algoritmos de aprendizagem de máquina é no reconhecimento de padrões onde os modelos são construídos a partir de uma vasta gama de informações sobre uma série de objetos, visando encontrar agrupamento de amostras (objetos) que são similares entre si e assim, detectar tendências nos dados (SOUZA; POPPI 2012).

O processo de construção de modelos de reconhecimento de padrões pode ser supervisionado e não supervisionado. No que diz respeito ao processo de construção não supervisionado, nenhum conhecimento prévio a respeito da classificação das amostras é utilizado e com isso, as amostras são agrupadas naturalmente apenas com base na informação contida nos dados (FERREIRA, 2015).

No desenvolvimento de ferramentas de reconhecimento de padrões supervisionado, cada amostra analisada provém de uma classe pré-estabelecida e essa informação é utilizada durante a análise dos dados e desenvolvimento dos modelos computacionais. Quando, durante seu desenvolvimento as variáveis de saída são categóricas, tem-se um modelo de classificação. Um exemplo clássico de aplicação desses modelos é na detecção de falhas de processo ou na previsão de níveis de qualidade de novos lotes de produção (MASSIMO, 2021). Os modelos de classificação podem ser desenvolvidos utilizando diversos algoritmos, dentre eles o k-ésimo vizinho mais próximo (KNN), árvore de decisão e florestas aleatórias (RF), máquinas de vetores de suporte (SVM) e redes neurais artificiais (RNA) (GERÓN, 2019).

Apesar da diversidade de softwares comerciais existentes no mercado para o desenvolvimento desses modelos, a utilização de ferramentas que não possuem custos associados à sua aquisição apresenta uma vantagem inicial uma vez que torna possível o seu emprego por qualquer indivíduo que tenha acesso a um computador. Devido a isto, tem-se verificado que a implementação dos algoritmos de aprendizagem de máquina acontece em sua maioria utilizando a linguagem de programação Python, que vem ganhando destaque nos últimos anos por ser gratuita e possuir bibliotecas com código aberto que possibilitam sua aplicação em diferentes tipos de problemas (MCKINNEY, 2013).

Dessa forma, o presente trabalho buscou desenvolver uma ferramenta computacional de reconhecimento de padrões utilizando técnicas de aprendizagem de máquina, em ambiente Python, para identificar relações entre a qualidade do açúcar produzido com as variedades de cana-de-açúcar utilizadas no processo de produção.

2 OBJETIVOS

2.1 Geral

Desenvolver ferramenta computacional de reconhecimento de padrões supervisionado para identificar relações entre a qualidade do açúcar produzido com as variedades de cana-de-açúcar utilizadas no processo de fabricação visando a predição e classificação de amostras fora da especificação.

2.2 Específicos

Utilizar a linguagem de programação Python e suas bibliotecas para construção de modelos de reconhecimento de padrões supervisionado.

Realizar análise exploratória dos dados da cana-de-açúcar que entra no processo de produção visando identificar amostras anômalas.

Comparar os algoritmos de aprendizagem de máquina utilizados e identificar qual algoritmo apresentará melhor capacidade de previsão de amostras que estão associadas a produtos fora da especificação.

Desenvolver interface acessível aos profissionais do setor sucroenergético com base no algoritmo de maior eficiência, de modo a facilitar o uso da ferramenta desenvolvida para pessoas que não apresentam conhecimentos sobre linguagens de programação.

3 REVISÃO BIBLIOGRÁFICA

Após todas as etapas de processamento para obtenção do açúcar, a comercialização deste produto pode acontecer de duas formas: para o consumo nos mercados interno e externo. O Açúcar consumido no mercado interno é aquele comercializado por meio de redes de supermercado (para uso doméstico), ou até mesmo comercializado diretamente para empresas que o utilizam como matéria-prima para produção de outros alimentos (bebidas, biscoitos, sorvetes, entre outros produtos). O açúcar comercializado no mercado externo geralmente é do tipo VHP (*Very High Polarization*), podendo ser exportado e utilizado em processos de refino em outros países (ZACURA FILHO; PICCIRILLI, 2012).

A cor é vista como o principal parâmetro de qualidade do açúcar avaliada tanto pelo mercado interno como pelo mercado externo. Este parâmetro é obtido por meio de um método estabelecido pela Comissão Internacional para Métodos Uniformes de Análise de Açúcar (ICUMSA). A cor é avaliada com base na capacidade da solução de açúcar em reter a passagem de luz, ou seja, quanto mais luz é absorvida pela amostra, maior a cor, medida em unidades ICUMSA (UI). Esta capacidade de absorção pode ser mensurada por espectrofotômetros, equipamentos que permitem a passagem da luz, num comprimento de onda de 420 nm, pela solução açucarada, possibilitando com isto a leitura da absorbância/transmitância da respectiva amostra (ICUMSA, 2005).

Caldas (2012) citou em seu trabalho dados de uma pesquisa realizada em 1999 pelo Instituto Nacional de Metrologia, Qualidade e Tecnologia – INMETRO, onde foi verificado no período de estudo que uma a cada três das marcas de açúcar branco disponíveis para comercialização não atendiam aos padrões de cor do mercado interno.

Pesquisas recentes mostraram mudanças de comportamento no consumo de alimentos, dentre elas, notou-se uma maior aceitação de açúcares mais escuros entre os consumidores domésticos do mercado interno. Como exemplo disso, tem-se o trabalho de Rojas (2007) que avaliou a aceitação do açúcar mascavo no mercado de viçosa – MG com base na cor. Observou-se durante as avaliações frente a um grupo de consumidores que não houve diferença significativa nas notas dadas pelos consumidores na cor dos três tipos de açúcar mascavo avaliados.

Apesar do aumento da aceitabilidade de açúcares mais escuros por parte dos usuários domésticos, o mesmo não é observado por parte das indústrias que utilizam o açúcar em seus processos para produzem refrigerantes, biscoitos, fármacos, entre outros, uma vez que um

açúcar de maior cor traz mudanças na cor de produtos acabados que tem esta propriedade como um diferencial de mercado (OLIVEIRA; ESQUIAVETO; SILVA JÚNIOR, 2007).

No contexto da indústria sucroenergética a cor ICUMSA pode sofrer influência de fatores que estão associados à qualidade da matéria-prima utilizada (SMITH *et al.*; 1990) e à forma de condução do processo produtivo (SMITH *et al.*; 1981) e devido a isto, verifica-se que é de extrema importância conhecer e identificar esses fatores dentro do cenário de produção para que se tenha controle sobre as características do produto final.

3.1 Influência da condução do processo na cor do açúcar

No que se refere a forma de processamento, a formação de corantes tem início logo após a colheita da cana-de-açúcar por meio de processos de escurecimento enzimático e continua até o armazenamento do açúcar (LEGENDRE, 1988).

O Quadro 1 descreve os principais mecanismos de formação de corantes em cada etapa do processo produtivo do açúcar VHP. Vemos que na etapa de extração do caldo, o principal mecanismo para a formação de cor se dá por meio do escurecimento enzimático. Esse processo é realizado por enzimas conhecidas como polifenoloxidades que catalisa a oxidação de fenóis contidos na cana a espécies químicas que posteriormente se oxidam na presença do ar, ou mesmo, reagem com grupos aminos para formar pigmentos de melanina de cor escura (GROSS; COOMBS, 1976). Esses compostos constituem cerca de 60% a 75% da cor no caldo clarificado pois não conseguem ser removidos muito bem durante a etapa de clarificação do caldo (PATON, 1992).

Quadro 1 - Mecanismos predominantes de geração de cor por etapa do processo de produção de açúcar

Etapa	Mecanismo
Extração	Escurecimento enzimático
Tratamento do caldo	Degradação alcalina de hexoses
Evaporação	Caramelização e reações de Maillard
Cozimento e cristalização	Reações de Maillard

Fonte: Poel, Schiweck e Schwartz, 1998.

Pode-se perceber, ainda no Quadro 1, que durante a etapa de tratamento de caldo, o principal mecanismo de formação de cor vem por meio da degradação alcalina de hexoses. Uma das sub etapas importantes durante a etapa de tratamento do caldo é a caleação, onde tem-se a correção do pH do caldo. Contudo, a elevação do pH pode gerar produtos da degradação

alcalina de hexoses. O produto dessa degradação são polímeros de ácidos carboxílicos denominados HADP. Segundo Riffer (1988) a degradação alcalina de hexoses é uma reação complexa, com diversos produtos intermediários, dificultando assim a caracterização dos constituintes de cor. Esta reação, em pH entre 6 e 8 pode ser catalisada pelo ferro.

Durante a etapa de evaporação, conforme pode ser observado na Quadro 1, temos como principais mecanismos de formação de cor o fenômeno de caramelização e a reação de Maillard. A caramelização refere-se à degradação térmica da sacarose em temperaturas acima de 120 ºC. Os caramelos são compostos de alto peso molecular formados por moléculas de hidroximetilfurfural (HMF) que apresentam intensa coloração (CELSO, 2012).

As reações de Maillard estão presentes tanto na etapa de evaporação quanto na etapa de cozimento. Essas reações geram compostos que segundo Smith (1990), podem provocar um acréscimo de cor em até 70%, a depender do sistema de cozimento utilizado. As melanoidinas, produto das reações de Maillard, são pigmentos escuros que se desenvolvem em ambientes neutro ou alcalino e concentrações de açúcares superiores a 65 ºBrix (LAROQUE, 2007), condições encontradas justamente durante as etapas de evaporação e cozimento.

Ainda falando sobre a influência do processo produtivo na formação de cor, Smith *et. al.* (1981), fazendo amostragens em vários pontos do processo, mostrou que houve uma redução de cor no caldo de 20 a 25% durante a etapa de clarificação e houve um aumento de 3 a 6% durante a etapa de evaporação. Neste trabalho, mostrou-se ainda que houve um aumento significativo de cor durante a etapa de cozimento devido as reações de Maillard.

3.2 Influência da matéria-prima na cor do açúcar

Apesar de ser verificada a existência de mecanismos que possibilitam o desenvolvimento de substâncias corantes durante as etapas de produção do açúcar, Smith (1990) após fazer um acompanhamento da cor dos produtos das etapas intermediárias do processo (extração, tratamento do caldo, evaporação e cozimento) concluiu que o determinante primordial da cor do açúcar é o nível de cor no caldo que entra no processo, parâmetro este governado pela matéria-prima processada.

Os precursores de cor contidos na cana-de-açúcar (pigmentos naturais, fenóis, aminoácidos, ferro, entre outros) não se apresentam em uma composição constante e dependem de fatores como variedade e estágio de maturação da cana, tipo e umidade do solo, uso de fertilizantes e forma de colheita, que por sua vez está associada com a quantidade de matéria estranha (folhas, copas, solo e raízes) (LEGENDRE, 1988). Com o aumento no teor de matéria estranha, aumenta por sua vez a quantidade de compostos fenólicos que entram no processo

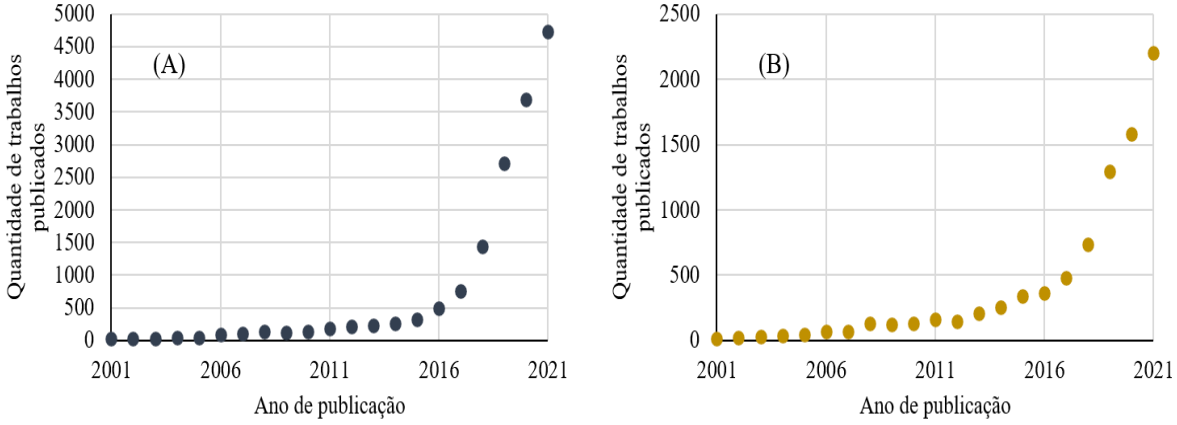
(dentre eles os flavonóides) em até 7,3 vezes quando comparado com a cana limpa (POEL, SCHIWECK E SCHWARTZ, 1998). Conforme já foi mencionado, estes compostos envolvem-se em reações de degradação enzimática, podendo estar associados a até 75% da cor no caldo clarificado.

3.3 Aplicação de métodos de aprendizagem de máquina na indústria

Embora seja de extrema importância conhecer os fatores associados à formação de cor dentro do cenário de produção de açúcar, nota-se uma complexidade quanto a implementação de ferramentas de análises que monitorem todos estes parâmetros simultaneamente. Como alternativa ao monitoramento analítico, tem-se notado um aumento significativo no uso de ferramentas computacionais para predição de parâmetros de qualidade nos últimos anos.

Segundo informações da base de dados *Scopus* (2022), houve um aumento considerável no número de trabalhos publicados que relacionam as palavras "*machine learning*" e "*industry*" nos últimos vinte anos (como pode ser observado na Figura 1.A). Isto pode ser justificado pelo crescimento no desenvolvimento de novas tecnologias voltadas para a indústria de processos e ao fato do aumento do interfaceamento de instrumentos de análise (ou de controle) gerando grandes quantidades de informações muitas vezes complexas e variadas (SENA *et. al.;* 1999). O mesmo comportamento é verificado quando se observa a quantidade de trabalhos publicados nos últimos anos que relacionam as palavras "*machine learning*" e "*quality control*" (Figura 1.B), confirmando assim um aumento significativo do uso de ferramentas computacionais associadas ao controle e monitoramento da qualidade dos produtos (SCOPUS, 2022).

Figura 1: Resultado da pesquisa bibliográfica na base de dados *Scopus*: (A) *"Machine Learning" and "Industry"*; (B) *"Machine Learning" and "quality control"*



Fonte: Autor, 2022.

Como exemplo de aplicação do uso de algoritmos de aprendizagem de máquina para previsão da qualidade de produtos, tem-se o trabalho desenvolvido por Shaw, Suman e Chakraborty (2019). O trabalho consistiu no uso de dados de natureza química e físico-química das amostras de vinhos para desenvolvimento de uma ferramenta capaz de prever a qualidade dessas bebidas. Os algoritmos RF, SVM e a rede neural artificial MLP (*Multilayer Perceptron*) foram utilizados. Observando a acurácia dos modelos, foi verificado que os algoritmos MLP e RF obtiveram os melhores resultados.

Bhardwaj *et al.* (2022) também utilizaram técnicas de aprendizagem de máquina para previsão da qualidade do vinho. Neste caso, foram utilizados os algoritmos KNN, RF, *Adaptative Boosting* (AdaBoost) e SVM. Os algoritmos foram comparados e apenas o modelo baseado no algoritmo KNN alcançou acurácia menor que 90%.

No contexto do agronegócio, o uso desses modelos também tem sido verificado. Como exemplo, tem-se o trabalho de Almeida *et. al.* (2021) que desenvolveu um modelo de classificação com base em ferramentas estatísticas descritivas e algoritmos de aprendizagem de máquina para classificar ambientes de produção em áreas de cultivo de cana-de-açúcar com base em um número reduzido de variáveis de baixo custo. Para isso, foram utilizados dados referentes aos fatores de formação e ao manejo do solo. A ferramenta construída apresentou uma acurácia de 75%, mostrando que a aplicação de técnicas de aprendizagem de máquina verifica-se como uma ferramenta promissora para determinação de ambientes de produção de cana-de-açúcar.

No contexto da indústria sucroenergética, pode ser citado o trabalho de Ramírez-Morales et al. (2016) que utilizou o algoritmo SVR (*support vector regression*), um modelo de SVM destinado a problemas de regressão, para monitorar a qualidade de produtos intermediários no processo de produção de açúcar. As amostras coletadas foram analisadas por espectroscopia de infravermelho próximo (NIR) e os espectros gerados foram usados para construção da ferramenta computacional visando a determinação das variáveis ºBrix (porcentagem mássica de açúcares nos fluidos avaliados) e POL (concentração de sacarose aparente). A ferramenta desenvolvida apresentou melhor capacidade de previsão, quando comparada a outras ferramentas publicadas em outros trabalhos, resultando em uma avaliação mais precisa da qualidade dos produtos intermediários do processo de produção de açúcar.

Outro exemplo de trabalho que envolveu ferramentas de aprendizagem de máquina no contexto sucroenergético foi o trabalho de Orzel, Daszykowski e Walczak (2011). Neste trabalho, desenvolveu-se um modelo de regressão usando técnicas de espectrometria de fluorescência associada ao algoritmo *Partial Least Squares* (PLS) para monitorar a cor e o teor

de cinzas em amostras de açúcar coletadas após o processo de centrifugação da massa cozida. A ferramenta desenvolvida apresentou erros de predição iguais a 3,24% e 4,37% para a cor e teor de cinzas, respectivamente, mostrando assim que o a qualidade do açúcar pode ser monitorada por meio de técnicas de aprendizagem de máquina.

3.4 Fatores que influenciam no desempenho dos modelos de classificação

Os trabalhos citados fortalecem a ideia de que essas ferramentas computacionais podem ser úteis na previsão das características finais dos produtos com base nos dados referentes aos seus respectivos processos de fabricação. No entanto, para que um bom modelo de classificação seja construído, é importante seguir algumas etapas: conhecer a natureza dos dados, selecionar os algoritmos, treinar os modelos e por fim testá-los.

Além dessas etapas, diversos fatores podem influenciar no desempenho das ferramentas construídas. Segundo Gerón (2019), o baixo desempenho do modelo de classificação pode estar associado a quantidade insuficiente de dados para treinamento, dados não representativos, dados com baixa qualidade (presença de ruídos ou amostras anômalas), dentre outros fatores. Dessa forma, percebe-se que é de extrema importância o uso de ferramentas que possibilitem uma análise exploratória para os dados que serão utilizados na construção desses modelos de reconhecimento de padrões, de modo a identificar e conhecer a estrutura dos dados, reduzir a quantidade de ruído e identificar a presença de amostras anômalas.

Dentre as formas disponíveis para realização de análise exploratória dos dados, tem-se as ferramentas estatísticas multivariadas, dentre elas, a análise de componentes principais.

3.4.1 Análise de componentes principais

A Análise de Componentes Principais (*Principal Component Analisys,* PCA) é um dos métodos estatísticos multivariados mais importantes e é a base de diversos métodos de reconhecimento de padrões. Normalmente é utilizada com o objetivo de visualizar a estrutura dos dados, encontrar similaridades entre amostras, detectar amostras anômalas (*outliers*) e reduzir a dimensionalidade do conjunto de dados. Com a redução da dimensionalidade proporcionada pela PCA, as amostras passam a ser pontos localizados em espaços de dimensões reduzidas definidos pelas PCs, por exemplo, bi ou tridimensionais. (SOUZA; POPPI 2012)

Matematicamente, na PCA, a matriz de dados originais "$\boldsymbol{X}$" é decomposta em um produto de duas matrizes, denominadas escores ("$\boldsymbol{T}$") e pesos ("$\boldsymbol{P}$") mais uma matriz de erros ("$\boldsymbol{E}$"), como conforme apresentado na Equação 1:

$$\boldsymbol{X = TP^T + E} \qquad \text{[EQ 1]}$$

Os escores representam as coordenadas das amostras no sistema de eixos formados pelas componentes principais. Cada componente principal é constituído pela combinação linear das variáveis originais e os coeficientes da combinação são denominados pesos. Os pesos representam o quanto cada variável original contribui para uma determinada PC. A primeira componente principal (PC1) é traçada no sentido da maior variação do conjunto de dados; a segunda (PC2) é traçada ortogonalmente à primeira, com o intuito de descrever a maior porcentagem da variação não explicada pela PC1 e assim por diante. A avaliação dos pesos permite entender quais variáveis mais contribuem para os agrupamentos observados no gráfico dos escores (LOVATTI, 2019).

Diversas informações podem ser retiradas das matrizes geradas na PCA, inclusive para analisar as características dos dados e para monitorar mudanças no processo. Muitos trabalhos, já fazem uso da matriz dos escores no próprio treinamento das ferramentas de aprendizagem de máquina. Qin (1997), por exemplo, cita uma forma de treinamento de redes neurais artificiais onde faz-se uma associação entre a matriz dos escores com as variáveis de saída, defendendo ainda que a rede neural combinada com o esquema PCA oferece uma abordagem viável, pois como as componentes principais são ortogonais há uma superação de problemas associado a colinearidade durante o treinamento desses modelos computacionais.

Há várias maneiras para o cálculo das matrizes dos escores e pesos, dentre elas, tem-se o método usado no presente trabalho: decomposição por valores singulares (SVD).

3.4.1.1 Decomposição por valores singulares

A notação adotada para ilustrar os cálculos seguem o mesmo padrão da simbologia utilizada por Ferreira (2015). No método SVD, a matriz de dados, já submetida a algum tipo de pré-tratamento (como por exemplo o autoescalamento, centragem na média ou normalização) é decomposta em três matrizes: U, S e V, de modo que a Equação 2 seja satisfeita:

$$X = USV^T \qquad \text{[EQ 2]}$$

Nota-se ainda que o produto das matrizes US $(i \times j)$ é a matriz dos escores T, enquanto que V é a matriz de pesos P.

As matrizes U e V são quadradas de dimensões respectivas $(i \times i)$ e $(j \times j)$ e são ortonomais. A matriz S é retangular $(i \times j)$ com todos os elementos fora da diagonal principal iguais a zero. Os elementos da diagonal principal são chamados de valores singulares e são ordenados em ordem decrescente, $s_{11} \geq s_{22} \geq s_{33} \geq \cdots \geq s_{kk}$ e o quadrado desses valores

singulares são os autovalores da matriz de correlação $X^T X$ (para o caso onde os dados são pré-processados por autoescalamento ou centragem na média).

O autovalor da matriz de correlação λ_a é numericamente igual à variância dos dados descrita pela a-ésima componente principal. Com isto, a quantidade de informação original contida nessa única componente principal pode ser dada pela porcentagem de variância explicada, $\%Var_a$ de acordo com a Equação 3:

$$\%Var_a = \frac{\lambda_a}{\sum_{a=1}^{k} \lambda_a} \qquad \text{[EQ 3]}$$

A decomposição SVD pode ser aplicada a qualquer matriz real e as matrizes U, S e V são determinadas simultaneamente.

3.5 Algoritmos de aprendizagem de máquina para problemas de classificação

Diversos algoritmos podem ser utilizados para problemas de reconhecimento de padrões supervisionado, dentre eles, os que serão utilizados neste trabalho: k-ésimo vizinho mais próximo e as redes neurais artificiais.

3.5.1 K-ésimo vizinho mais próximo

A Análise de vizinhos mais próximos (KNN) é um método para classificar casos com base na semelhança com outros casos. Esse classificador atribui um objeto desconhecido à classe à qual a maioria dos *k*- vizinhos mais próximos pertence (BRERETON, 2015). Durante o processo de construção do modelo, cada amostra do conjunto de treinamento é excluída uma única vez e então classificada usando-se para isso as amostras restantes. Nesta etapa são calculadas as distâncias (euclidianas, por exemplo – Equação 4) entre a amostra excluída das demais amostras constituintes do conjunto de treinamento e quanto mais similares forem as amostras, menores serão suas distâncias no espaço (FERREIRA, 2015; LOVATTI, 2019).

$$d_{AB} = [(x_A - x_B)^T (x_A - x_B)]^{1/2} \qquad \text{[EQ 4]}$$

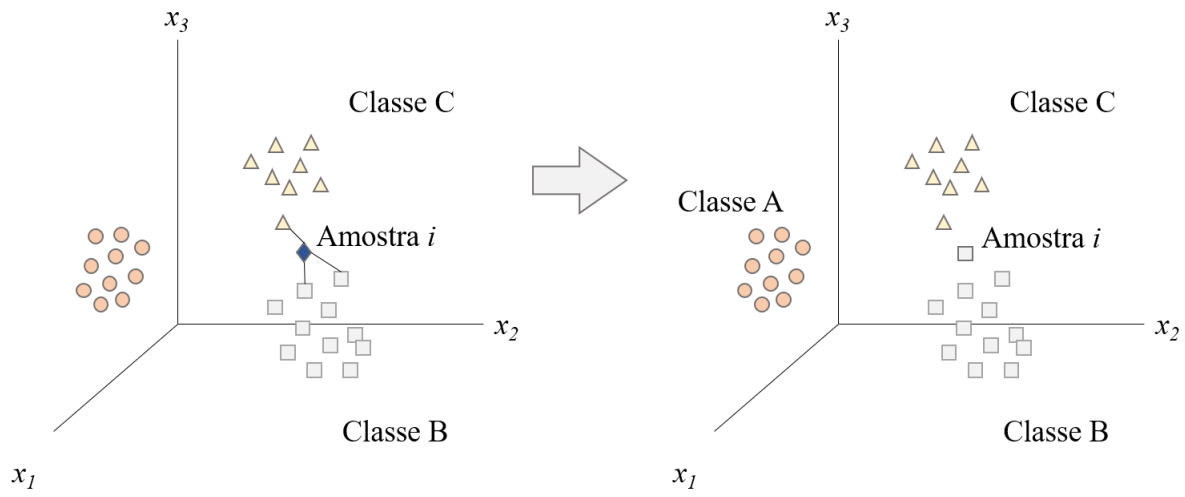
Onde d_{AB} é a distância euclidiana entre as amostras *A* e *B*; x_A é vetor cujo elementos são as respostas da amostra *A*; x_B é vetor cujo elementos são as respostas da amostra *B*; T – subscrito que significa “transposta”.

As distâncias são organizadas em ordem crescente para facilitar a identificação dos k-vizinhos mais próximos. Finalmente, a amostra excluída é então classificada de acordo com a maioria dos “votos” de seus vizinhos mais próximos. Quando a classe designada para amostra excluída com a sua classe verdadeira tem-se êxito na classificação. Esse processo, denominado

validação cruzada, é repetido e aplicado para todas as amostras que compõem o conjunto de treinamento, na construção final do modelo são obtidos os resultados de êxito e erros de classificação (LOVATTI, 2019).

A Figura 2 pode ser usada para exemplificar o mecanismo de aplicação do método KNN na classificação da amostra "*i*". Vê-se nesta imagem que dois dos três vizinhos mais próximos são pertencentes a Classe A e por consequência, a amostra "*i*" também foi classificada como pertencente a classe A. É fácil perceber que um critério importante durante o desenvolvimento do modelo é a escolha do número de vizinhos mais próximo, *k*. O procedimento mais adequado para a escolha do número ótimo de vizinhos do modelo final é construir vários modelos de classificação, variando, pois, o número de vizinhos. Ferreira (2015) afirma que, quando a distribuição dos membros da classe for homogênea e a distância entre as classes for maior do que o espalhamento das amostras dentro da classe, é aconselhável que o valor de *k* esteja entre 3 e 5.

Figura 2 - Princípio de funcionamento do algoritmo KNN



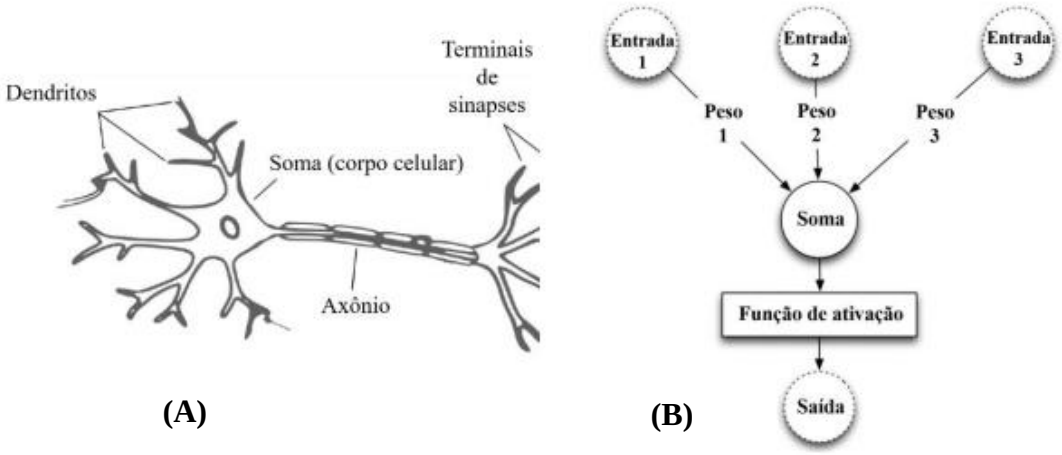
Fonte: Adaptado de Ferreira, 2015.

A maior vantagem desse método está relacionada a sua versatilidade em solucionar problemas de classificação contendo dados de natureza linear como também não linear (LOVATTI, 2019). Ferreira (2015) afirma que ele é um dos poucos algoritmos que funciona bem quando as classes no conjunto de treinamento estão fortemente sub agrupadas ou parcialmente sobrepostas. Apesar disto, este método também apresenta desvantagens como o fato de não possuir habilidade em identificar amostras com comportamento anômalo nos conjuntos de treinamento e teste. Para contornar essa problemática, pode-se fazer uma associação entre a PCA e o algoritmo KNN, permitindo, pois, que a PCA se encarregue pela identificação de amostras anômalas e a classificação seja feita por meio do algoritmo KNN.

3.5.2 Redes neurais artificiais

Redes Neurais Artificiais (RNA's) são modelos computacionais inspirados no sistema nervoso de seres vivos. Conforme ilustra a Figura 3, a tarefa do neurônio biológico e do neurônio artificial é coletar informações nas entradas, processar essas informações e emiti-las (SILVA, SPATTI e FLAUZINO, 2019)**.**

Figura 3 - Redes neurais: (A) Estrutura celular de um neurônio biológico; (B) Representação simplificada de uma rede neural artificial



Fonte: Adaptado de Géron, 2019.

Haykin (2001) afirma que uma rede neural artificial é um processador maciço constituído de unidades simples de processamentos que funcionam para armazenar conhecimentos e torná-los disponíveis para utilização. Para isto, utilizam o seguinte princípio de funcionamento: as informações são captadas do meio exterior (ou de outros neurônios), ponderadas pelos respectivos pesos; subtrai-se do resultado, um valor (limiar de ativação ou bias) e em seguida, o valor gerado destas operações é modificado por uma função de ativação, gerando um sinal de saída do neurônio artificial que será transmitido ao meio exterior ou para outro neurônio. A relação de entrada e saída pode ser representada pela Equação 5:

$$y = \phi\left(\sum_{i=1}^{k} w_i x_i - \theta\right) \qquad \text{[EQ 5]}$$

$i - quantidade\ de\ entradas/neurônios$

$j - quantidade\ de\ camadas\ intermediárias$

Onde x_{ij} e w_{ij} são respectivamente as entradas e os pesos de cada camada intermediária, θ são os limiares de ativação e ϕ uma função de ativação diferenciável.

Devido a extrema capacidade de aprendizado, as RNA's podem ser empregadas em diversos problemas relacionados às engenharias e ciências, como por exemplo: controle de

processos, agrupamento de dados (clusterização) e reconhecimento de padrões (SILVA, SPATTI e FLAUZINO, 2019). As soluções de RNA's são bem aceitas nas indústrias de processo pois são soluções econômicas, de fácil compreensão e utilizam para sua construção, dados oriundos do próprio processo (QIN, 1997).

Um dos primeiros modelos de RNA's proposto foi a rede *Perceptron* (ROSENBLATT, 1958), a forma mais simples de configuração de uma rede neural artificial contendo apenas um neurônio artificial em uma camada neural única. A rede *Perceptron* inicialmente foi proposta para classificar padrões linearmente separáveis (SOUSA, 2014).

Com o passar do tempo, outros modelos de redes neurais artificiais foram criados visando superar limitações existentes na rede *Perceptron*, como por exemplo, a Rede Perceptron Multicamadas (MLP) com o algoritmo de aprendizado *Backpropagation*. As redes MLP são caracterizadas pela presença de pelo menos uma camada intermediária de neurônios, situada entre a camada de entrada e a respectiva camada de saída (BRYSON; HO, 1969).

3.5.2.1 O algoritmo *backpropagation*

O algoritmo de *backpropagation* ou retropropagação usa o método de gradiente descendente generalizado para modificar pesos e limiares de ativação (bias) para que o erro entre a saída desejada e o sinal de saída da rede seja minimizado (QIN, 1997). O aprendizado por retropropagação consiste em duas etapas: propagação e retropropagação. Na etapa de propagação, o padrão de ativação é aplicado nos nós da camada de entrada da rede, e seu efeito é propagado pela rede camada por camada. Na última camada, é produzido um conjunto de saídas que se configura como a verdadeira resposta da rede. Na etapa de retropropagação, todos os pesos sinápticos e bias são ajustados de acordo com as regras de correção de erros pré-determinada. O sinal de erro é retropropagado através da rede, na direção oposta das conexões sinápticas, e com isso, os pesos sinápticos e bias são ajustados estatisticamente para aproximar a resposta real da rede com a resposta desejada. (NIED, 2007).

4 METODOLOGIA

4.1 Obtenção do conjunto de dados

O conjunto de dados utilizado para construção do modelo de reconhecimento de padrões composto por 99 amostras e 13 variáveis foi extraído do trabalho desenvolvido por Silva (2017). As 12 variáveis utilizadas para construção do modelo de classificação referem-se a sete variedades de cana-de-açúcar e cinco propriedades físico-químicas realizadas para as amostras de caldo misto: sacarose aparente (POL), potencial hidrogeniônico (pH), açúcares redutores totais (ART), teor de fosfato (P_2O_5) e cor. A variável classificatória foi a cor do respectivo açúcar produzido. As amostras de açúcar foram classificadas como "Dentro da Especificação" e "Fora da Especificação" de acordo com as especificações adotadas pela ICUMSA para o açúcar do tipo VHP, considerando que não existiram grandes variações no processo durante as etapas de extração e tratamento do caldo misto e também durante a etapa de cozimento do xarope.

4.2 Pré-processamento dos dados

Os dados utilizados para construção dos modelos de classificação passaram pelo pré-processamento conhecido como autoescalamento, onde a média de cada variável foi subtraída de seus respectivos elementos e em seguida cada valor resultante da subtração foi dividido pelos seus respectivos desvios-padrão (Equação 6). A transformação realizada sobre o conjunto de dados permitiu que cada variável apresente média zero e variância igual a um, dando assim a mesma importância para todas as variáveis, independente da sua dimensão (PIMENTEL, 2005).

$$x_{ij(as)} = \frac{x_{ij} - \overline{x_j}}{s_j} \qquad \text{[EQ 6]}$$

Onde $x_{ij(as)}$ é o valor autoescalado para a variável j na amostra i; x_{ij} é o valor da variável j na amostra i; $\overline{x_j}$ é a média dos valores das amostras em uma coluna j; s_j é o desvio padrão dos valores da variável j.

Esta etapa do projeto foi realizada em ambiente Python versão 3.7 (Pacote Anaconda) usando a interface *Jupyter Notebook* (interface 6.0.3) e a ferramenta *StandardScaler* pertencente ao módulo *preprocessing* da biblioteca *scikit-learn*.

4.3 Análise exploratória dos dados

Realizou-se uma análise exploratória com os dados autoescalados por meio da análise de componentes principais (PCA) com o intuito de visualizar a estrutura dos dados e detectar amostras anômalas (*outliers*). Para obtenção das matrizes dos escores e dos pesos utilizou-se o algoritmo de decomposição por valores singulares (SVD). Esta etapa do projeto foi realizada em ambiente Python versão 3.7 (Pacote Anaconda) usando a interface *Jupyter Notebook* (interface 6.0.3) e a função *PCA* pertencente ao módulo *decomposition* da biblioteca *Scikit-learn*.

A identificação das amostras anômalas se deu por meio da ferramenta *knn* do módulo *models* da biblioteca *PyOD* (biblioteca do Python específica para remoção de *outliers*) que utilizou os valores dos escores gerados pela análise de componentes principais para calcular as distâncias entre os pontos da distribuição e com isso, avaliar o comportamento das amostras. Dessa forma, as amostras com comportamentos discrepantes das demais foi reconhecida pelo algoritmo como *outliers*. Após identificação, o algoritmo forneceu os índices das respectivas amostras, possibilitando a remoção.

4.4 Construção dos modelos de classificação

Os algoritmos de aprendizagem de máquina *k*-ésimo vizinho mais próximo (KNN) e redes neurais artificiais (RNA) foram utilizados para construção dos modelos de reconhecimento de padrões supervisionado. Para construção dos modelos, as amostras foram divididas em um conjunto de treinamento (contendo 70% das amostras) e um conjunto para teste (contendo as amostras restantes). A seleção das amostras para cada conjunto aconteceu de forma randômica para evitar vícios durante o processo de construção dos modelos. A divisão das amostras em treinamento e teste também aconteceu em ambiente Python, por meio da função *train_test_split* pertencente ao módulo *model_selection* da biblioteca *sckit-learn*.

Os algoritmos foram avaliados por meio de duas ferramentas: a matriz de confusão, ilustrada pelo Quadro 2 e com base no valor de eficiência, definido conforme Equação 7:

$$\text{Eficiência (\%)} = \frac{\text{número de amostras classificadas corretamente}}{\text{número total de amostras classificadas}} * 100 \qquad \text{[EQ 7]}$$

Quadro 2 - Representação da matriz de confusão

		Valor predito	
		Classe A	**Classe B**
Valor Real	**Classe A**	Verdadeiro Positivo	Falso Negativo
	Classe B	Falso Positivo	Verdadeiro Negativo

Fonte: Autor, 2022.

4.4.1 Construção dos modelos baseados no algoritmo KNN

Dois modelos de classificação foram construídos utilizando o algoritmo K-ésimo vizinho mais próximo. O primeiro utilizou a matriz de dados autoescalados para treinamento e teste do modelo e o segundo, a matriz dos escores obtida na análise exploratória por meio da análise de componentes principais. O modelo construído com os dados autoescalados foi nomeado simplesmente de modelo KNN, já o modelo construído com a matriz dos escores da PCA foi nomeado modelo PCA-KNN.

Para construção dos modelos KNN e PCA-KNN, as amostras dos conjuntos de treinamento foram associadas à classificação prévia do açúcar produzido (dentro e fora da especificação). Para treinamento dos modelos os valores de $k = 3, 4$ e 5 vizinhos mais próximos foram testados de modo a encontrar o número de vizinhos mais próximo que proporcionasse uma maior eficiência para o modelo. Após o processo de treinamento, as amostras pertencentes ao conjunto de teste passaram pelo processo de classificação, permitindo a avaliação dos modelos construídos por meio da comparação da classificação predita com a classificação real. Os modelos de classificação baseados no algoritmo KNN foram construídos em ambiente Python (versão 3.7) utilizando a função *KNeighborsClassifier* do módulo *neighbors* da biblioteca *scikit-learn*.

4.4.2 Construção dos modelos baseados no algoritmo RNA

Assim como nos modelos baseados no algoritmo KNN, duas ferramentas de reconhecimento de padrões supervisionado foram construídas baseando-se em redes neurais artificiais (RNA). Dois modelos de RNA do tipo *Multi Layer Perceptron* (MLP) foram desenvolvidas. O primeiro (denominado simplesmente de MLP) utilizando a matriz de dados autoescalados para treinamento e teste e o segundo (denominado PCA-MLP), a matriz dos escores obtida na PCA.

Ambas as redes apresentaram apenas uma camada oculta e para treinamento das redes neurais foi utilizado o algoritmo *backpropagation*. A função de ativação escolhida para as camadas de entrada, oculta e de saída foi a *ReLU*, definida pela Equação 8.

$$ReLU = \max(w_i x_i - \theta, 0) \qquad \text{[EQ 8]}$$

Percebe-se pela Equação 8 que a função *ReLU* é contínua, mas infelizmente não é diferenciável em $w_i x_i - \theta = 0$ (a inclinação muda abruptamente, podendo fazer com que o gradiente descendente salte). Apesar disso, a função tem sido bastante empregada pois possibilita bons resultados e cálculos rápidos (GÉRON, 2019).

As redes neurais foram construídas em ambiente Python (versão 3.7) utilizando a função *MLPClassifier* do módulo *neural_network* da biblioteca *scikit-learn*.

4.5 Desenvolvimento da interface gráfica

O algoritmo de melhor desempenho foi utilizado no desenvolvimento de uma interface gráfica. O *software* "*Qt Designer*" (acesso livre) foi utilizado para construção do *layout* da interface, em seguida, utilizou-se a biblioteca "*PyQt5*" do Python para decodificar o arquivo gerado pelo *software* para um código em Python, permitindo, pois, a inclusão do algoritmo referente ao modelo de classificação construído. A interface permite a utilização da ferramenta de classificação por qualquer pessoa que tenha acesso a um computador, sem a necessidade de um conhecimento prévio sobre linguagens de programação.

O *layout* da interface foi constituído de células para preenchimento e dois botões para iteração com usuário. Cada célula, devidamente identificada, possuiu a finalidade de receber o valor das respectivas variáveis das amostras que passariam pelo processo de classificação. Após preenchimento das células, deu-se ao usuário a opção de clicar em dois botões. O botão nomeado "Verificar Classificação" foi responsável por acionar o processo de previsão para amostra inserida. O botão nomeado "Nova Consulta", por sua vez, foi responsável por esvaziar todas as células antes preenchidas, possibilitando a inserção de novos dados para novas previsões.

5 RESULTADOS E DISCUSSÃO

5.1 Aquisição do conjunto de dados

O trabalho desenvolvido por Silva (2017) forneceu o conjunto de dados necessários para o desenvolvimento dos modelos de classificação. Os dados foram extraídos e tratados previamente por meio das planilhas eletrônicas do Excel. A estrutura dos dados originais apresentou-se como uma matriz de dimensão 99×13. Cada linha da matriz referiu-se a um dia de produção de açúcar onde foi possível a coleta das informações de cada variável. As doze primeiras colunas referiram-se as 12 variáveis avaliadas para construção dos modelos. A última coluna referiu-se à classificação do açúcar produzido no respectivo dia.

As nomenclaturas de algumas variáveis foram modificadas de modo a facilitar a manipulação dos dados durante o processo de construção dos modelos computacionais. A Tabela 1 mostra as novas nomenclaturas adotadas para as variáveis avaliadas assim como as respectivas unidades de medida. Ainda na Tabela 1 é possível observar a média e o desvio padrão de cada variável.

Tabela 1 - Descrição das variáveis envolvidas

Descrição Geral	Nome da variável	Nome da variável abreviada	Unidades	Média	Desvio Padrão
Variedades de cana utilizadas no processo	RB 92579	R9	%	43,63	14,23
	SP79 1011	S1	%	11,52	10,51
	SP81 3250	S0	%	8,01	9,81
	RB86 7515	R5	%	4,14	5,41
	RB95 1541	R1	%	9,49	9,84
	SP71 6949	S9	%	1,67	6,28
	SPX	SX	%	21,55	14,67
Propriedades físico-químicas do caldo misto	POL	POL	%	12,76	0,88
	pH	pH	-	5,06	0,29
	ART	ART	kg ART/tc*	14,25	2,42
	P_2O_5	P2O5	g/tc*	44,55	18,10
	COR ICUMSA	COR	UI	14189,36	1798,70

*tc refere-se a tonelada de cana processada

Fonte: Adaptado de Silva, 2017.

5.2 Pré-processamento dos dados

Conforme pode ser observado na Tabela 1, existem diferenças nas escalas das unidades das variáveis utilizadas. Dessa forma, o procedimento de pré-processamento dos dados tornou-se necessário para igualar a magnitude das variáveis antes do procedimento de análise exploratória. Conforme descrito na metodologia deste trabalho, utilizou-se o pré-processamento conhecido como autoescalamento. Ao final desta etapa, cada variável apresentou média zero e desvio-padrão igual a um (conforme pode ser observado no Tabela 2), permitindo, pois que cada coluna da matriz de dados seja centrada na média e escalada pelo desvio-padrão.

Tabela 2 - Resultado do pré-processamento dos dados

Variável	Média original	Desvio-padrão original	Média após pré-processamento	Desvio-padrão após pré-processamento
R9	43,63	14,23	$-1{,}03\times10^{-10}$	1,00
S1	11,52	10,51	$-4{,}10\times10^{-10}$	1,00
S0	8,01	9,81	$5{,}61\times10^{-12}$	1,00
R5	4,14	5,41	$-9{,}01\times10^{-11}$	1,00
R1	9,49	9,84	$-3{,}59\times10^{-11}$	1,00
S9	1,67	6,28	$-1{,}91\times10^{-11}$	1,00
SX	21,55	14,67	$-1{,}88\times10^{-10}$	1,00
POL	12,76	0,88	$1{,}10\times10^{-09}$	1,00
pH	5,06	0,29	$-1{,}71\times10^{-09}$	1,00
ART	14,25	2,42	$-6{,}99\times10^{-09}$	1,00
P2O5	44,55	18,10	$5{,}97\times10^{-10}$	1,00
COR	14189,36	1798,70	$2{,}16\times10^{-10}$	1,00

Fonte: Autor, 2022.

5.3 Análise exploratória dos dados

A análise exploratória dos dados se deu por meio da análise de componentes principais. A Tabela 3 mostra o valor da variância explicada por cada componente principal obtida. Durante a PCA, tornou-se possível a obtenção de duas matrizes: escores e pesos. Com o intuito de visualizar a estrutura dos dados, gerou-se o gráfico de dispersão dos escores para as três primeiras componentes principais, representando 51,43% da variância contida nos dados originais, conforme pode ser observado na Figura 4.A.

Por meio da Figura 4.A, observa-se que as amostras pertencentes a classe “Dentro da especificação” tenderam-se a se agrupar ao lado direito da distribuição, para valores de PC1 maiores que -4,0, enquanto as amostras pertencentes a classe “Fora da especificação” tenderam a se agrupar no lado esquerdo da distribuição, para valores de PC1 menores que -4,0. Apesar

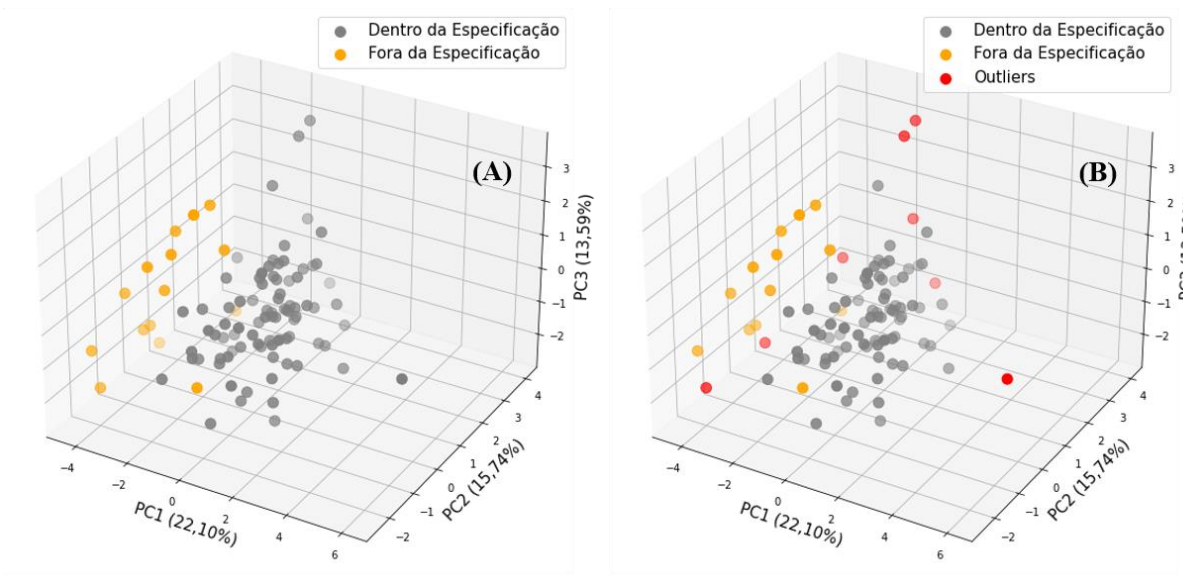
dessa tendência apresentada pelas amostras, não foi observada uma completa separação entre as classes.

Tabela 3 - Variância explicada por cada componente principal

Componente Principal	Variância Explicada (%)	Variância Acumulada (%)
PC1	22,10	22,10
PC2	15,74	37,84
PC3	13,59	51,43
PC4	9,93	61,36
PC5	8,51	69,87
PC6	7,19	77,06
PC7	6,96	84,01
PC8	5,49	89,50
PC9	4,71	94,21
PC10	3,32	97,53
PC11	2,47	100,00
PC12	0,00	100,00

Fonte: Autor, 2022.

Figura 4 - Gráfico dos escores: (A) - Dispersão das amostras no espaço gerado pelas componentes principais; (B) Destaque para as amostras anômalas

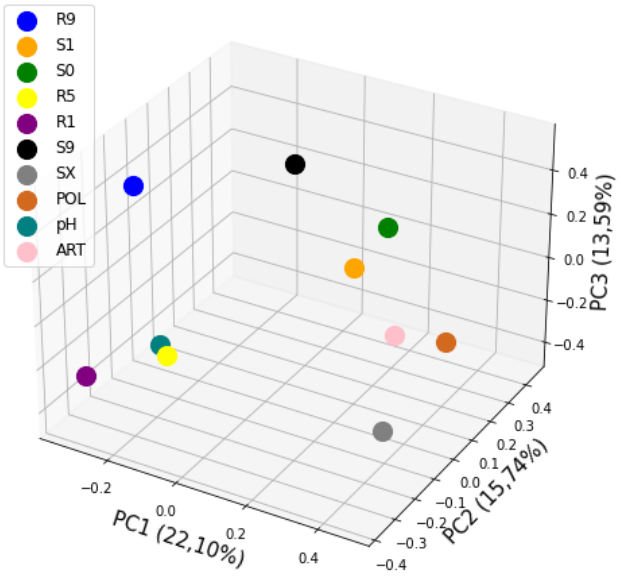


Fonte: Autor, 2022.

Nota-se, ainda na Figura 4.A, que alguns pontos da distribuição apresentaram comportamentos atípicos pelo o fato de se distanciarem de forma considerável do centro das suas respectivas classes. Utilizando a ferramenta *knn* do módulo *models* da biblioteca *PyOD*, 8 amostras foram identificadas como anômalas (destacadas na Figura 4.B como *outliers*) e removidas do conjunto de dados.

A matriz de pesos, também obtida durante a PCA, permitiu a avaliação do quanto cada variável contribuiu dentro do espaço gerado pelas componentes principais, permitindo inclusive a verificação das relações existentes entre as variáveis. Na Figura 5 é possível observar a distribuição das variáveis no espaço gerado pelas três primeiras componentes principais. Percebe-se que apenas as variáveis POL e ART apresentaram um grau de proximidade considerável, indicando assim que essas variáveis estavam relacionadas.

Figura 5 - Gráfico dos pesos



Fonte: Autor, 2022.

Conhecendo o significado químico dessas variáveis no processo, percebe-se que de fato, há uma relação entre a porcentagem de sacarose aparente (POL) com a concentração de açúcares redutores totais (ART), uma vez que o ART é definido como a concentração de açúcares redutores (glicose e frutose) contida no caldo somada com a concentração de sacarose hidrolisada, segundo a Equação 9 (CALDAS; LARRAHONDO; SILVA, 2017).

$$ART = 1{,}0526 * POL + AR \qquad \text{[EQ 9]}$$

Onde ART é a concentração de açúcares redutores totais (kg/tonelada de cana), POL é a concentração de sacarose aparente (kg/tonelada de cana) e AR a concentração dos açúcares redutores glicose e frutose (kg/tonelada de cana). O fator estequiométrico 1,0526 é obtido por meio da reação de hidrólise da molécula de sacarose, sendo transformada em glicose e frutose como pode ser observado nas Equações 10 e 11.

$$\underset{\substack{Sacarose \\ 342\ g/mol}}{C_{12}H_{22}O_{11}} + \underset{\substack{Água \\ 18\ g/mol}}{H_2O} \rightleftharpoons \underset{\substack{Glicose \\ 180\ g/mol}}{C_6H_{12}O_6} + \underset{\substack{Frutose \\ 180\ g/mol}}{C_6H_{12}O_6} \qquad \text{[EQ 10]}$$

$$\frac{AR}{POL} = \frac{360\ g\ de\ AR}{342\ g\ de\ sacarose} = 1{,}0526 \qquad \text{[EQ 11]}$$

No que se refere as outras variáveis avaliadas, nenhuma outra relação significativa foi identificada. Após essas observações, optou-se por excluir a variável POL do processo de construção do modelo de classificação, uma vez que, dada a relação identificada entre as variáveis POL e ART, a inclusão de ambas as variáveis durante o desenvolvimento dos modelos de classificação poderia gerar duplicidade de informação, podendo influenciar negativamente no desempenho desses modelos.

5.4 Construção dos modelos de classificação

5.4.1 Construção dos modelos baseados no algoritmo KNN

Conforme descrito na metodologia, dois modelos de reconhecimento de padrões baseados no algoritmo do k-ésimo vizinho mais próximo foram construídos: KNN e PCA-KNN. Os modelos KNN foram desenvolvidos testando três diferentes valores de k (3, 4 e 5). Como primeira forma de avaliação, tem-se as matrizes de confusão de cada modelo KNN, dispostas na Tabela 4. Observa-se que o modelo com k = 5 apresentou melhor eficiência de classificação quando comparado aos demais. Nos modelos com k = 3 e k = 4, todas as 26 amostras pertencentes a classe “Dentro da especificação” do conjunto de teste foram classificadas corretamente. Apesar disso, uma das duas amostras pertencentes a classe “Fora da especificação” foi classificada incorretamente, gerando um falso positivo.

No caso do modelo KNN com k = 5, todas as amostras foram classificadas corretamente em suas respectivas classes. Com base nessas observações, verificou-se que, de fato, a quantidade k de vizinhos escolhido para treinamento do modelo pode influenciar no seu desempenho. Percebeu-se que nos modelos com k = 3 e k = 4, 27 das 28 amostras foram classificadas corretamente e com isso, as ferramentas construídas alcançaram uma eficiência de 96,4%. O modelo com k = 5, por sua vez, alcançou eficiência de 100% na capacidade preditiva.

Para treinamento do modelo PCA-KNN, utilizou-se a matriz dos escores referentes as três primeiras componentes principais. A escolha das três primeiras PCs derivou-se do processo de análise exploratória, onde foi verificado que o espaço gerado por elas possibilitou a observação de uma tendencia na separação das classes. A Tabela 5 mostra as respectivas matrizes de confusão dos três modelos PCA-KNN construídos (um para cada k adotado). Diferente do que foi observado na construção dos modelos KNN, observou-se que a quantidade de k vizinhos mais próximos não afetou a eficiência dos modelos. Para os valores de $k = 3$, 4

e 5, todas as amostras foram classificadas corretamente em suas respectivas classes, obtendo um valor de eficiência de 100%.

Tabela 4 - Matriz de confusão dos modelos KNN

			Valor predito	
			Dentro da especificação	Fora da especificação
$k = 3$	**Valor Real**	Dentro da especificação	26	0
		Fora da especificação	1	1

			Valor predito	
			Dentro da especificação	Fora da especificação
$k = 4$	**Valor Real**	Dentro da especificação	26	0
		Fora da especificação	1	1

			Valor predito	
			Dentro da especificação	Fora da especificação
$k = 5$	**Valor Real**	Dentro da especificação	26	0
		Fora da especificação	0	2

Fonte: Autor, 2022.

Comparando os modelos construídos, nota-se que os modelos KNN (com $k = 5$) e os PCA-KNN alcançaram melhores resultados. Percebe-se que a estratégia de utilizar a matriz de escores como dados de entrada para construção dos modelos de classificação foi bem sucedida. Observou-se que apesar de representar apenas 51,43% da variância explicada pelos dados originais, o espaço gerado pelas três primeiras PCs facilitou o processo de treinamento do modelo, permitindo assim a construção de uma ferramenta de classificação mais eficiente.

5.4.2 Construção dos modelos baseados no algoritmo RNA

Dois modelos de classificação foram construídos com base no algoritmo de redes neurais artificiais. A Figura 6 mostra uma representação da arquitetura utilizada para construção da rede MLP. A camada de entrada foi composta por onze neurônios, um para cada variável pertencente ao conjunto de dados autoescalados. Optou-se por utilizar apenas uma camada oculta contendo a mesma quantidade de neurônios que a camada de entrada. A camada de saída, com apenas um neurônio, é responsável por enviar um sinal ao meio externo, identificando as amostras como pertencentes as classes “Dentro da especificação” ou “Fora da especificação”.

O resultado do desempenho da rede MLP pode ser observado através da Tabela 6 que mostra a matriz de confusão do respectivo modelo. Como pode ser observado, todas as 28 amostras pertencentes ao conjunto de teste foram classificadas corretamente em suas respectivas classes, gerando assim um modelo com 100% de eficiência na predição.

Tabela 5 - Matriz de confusão dos modelos PCA-KNN

			Valor predito	
			Dentro da especificação	Fora da especificação
k = 3	**Valor Real**	Dentro da especificação	26	0
		Fora da especificação	0	2

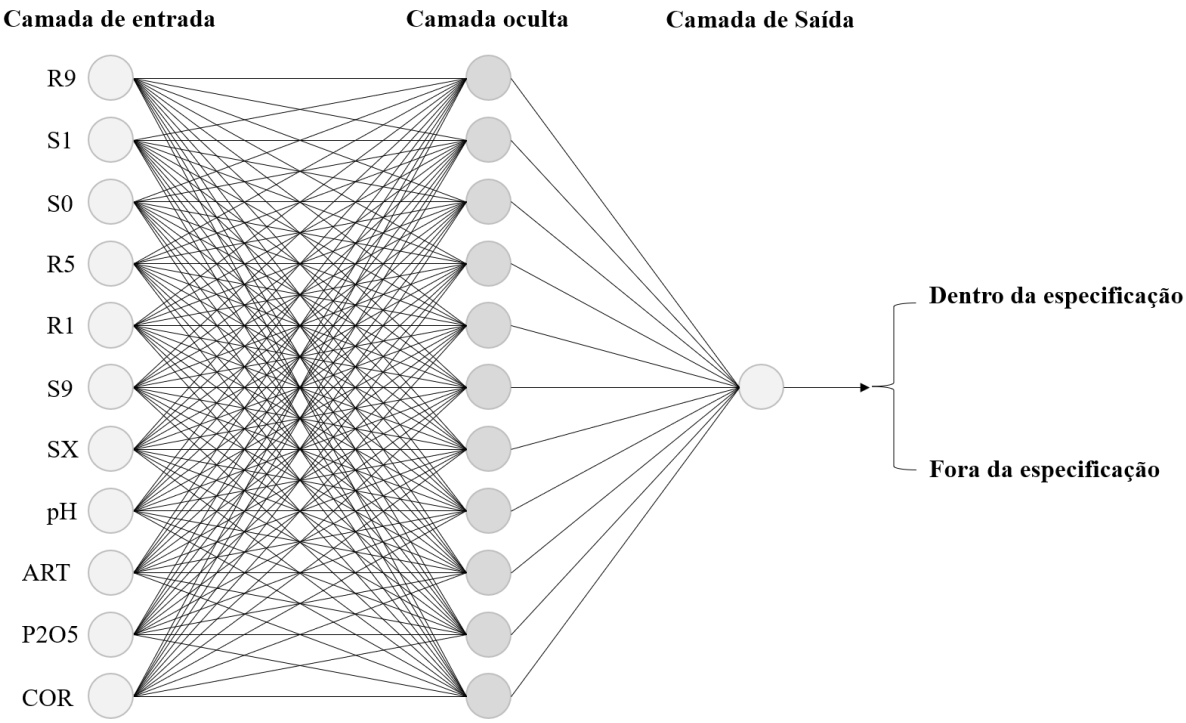
			Valor predito	
			Dentro da especificação	Fora da especificação
k = 4	**Valor Real**	Dentro da especificação	26	0
		Fora da especificação	0	2

			Valor predito	
			Dentro da especificação	Fora da especificação
k = 5	**Valor Real**	Dentro da especificação	26	0
		Fora da especificação	0	2

Fonte: Autor, 2022.

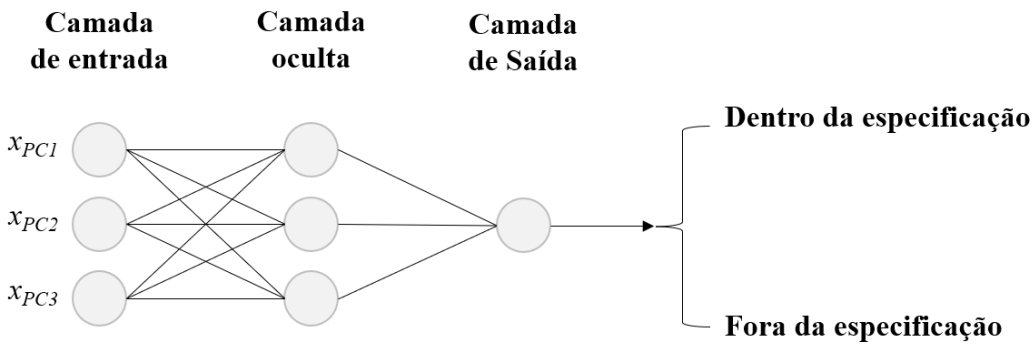
Para construção do modelo PCA-MLP, utilizou-se do mesma estratégia que o modelo PCA-KNN na escolha da quantidade de componentes principais que alimentariam o modelo. A Figura 7 ilustra a arquitetura utilizada para construção da rede PCA-MLP. A camada de entrada da rede neural foi composta por três neurônios, um para cada coordenada dos escores pertencentes ao espaço gerado pelas 3 primeiras PCs. Optou-se pela escolha de apenas uma camada oculta com a mesma quantidade de neurônios que a camada de entrada. Assim como na rede MLP, a camada de saída da rede PCA-MLP foi composta por apenas um neurônio que envia uma resposta ao meio externo, classificando as amostras como “Dentro da especificação” ou “Fora da especificação”.

Figura 6 - Arquitetura da rede MLP



Fonte: Autor, 2022.

Figura 7 - Arquitetura da rede PCA-MLP



Fonte: Autor, 2022.

Na Tabela 7 é possível observar a matriz de confusão da rede PCA-MLP. Como pode ser observado, mesmo com uma arquitetura relativamente simples, o modelo de classificação PCA-MLP possibilitou que todas as 28 amostras pertencentes ao conjunto de teste fossem classificadas corretamente em suas respectivas classes, gerando assim um modelo com 100% de eficiência na predição.

Tabela 6 - Matriz de confusão da rede MLP

		Valor predito	
		Dentro da especificação	Fora da especificação
Valor Real	Dentro da especificação	26	0
	Fora da especificação	0	2

Fonte: Autor, 2022.

Tabela 7 - Matriz de confusão do modelo PCA-MLP

		Valor predito	
		Dentro da especificação	Fora da especificação
Valor Real	Dentro da especificação	26	0
	Fora da especificação	0	2

Fonte: Autor, 2022.

Com isso, percebeu-se que ambos os modelos de classificação construídos a partir do algoritmo de redes neurais artificiais alcançaram 100% de eficiência na capacidade preditiva.

5.5 Desenvolvimento da interface gráfica

Como pode ser observado, os modelos KNN (k = 5), PCA-KNN, RNA e PCA-RNA apresentaram o mesmo valor de eficiência. Dessa forma, foi necessária a utilização de outro indicador, além do valor de eficiência, para escolha do algoritmo de melhor desempenho. Avaliou-se o tempo que cada modelo levou para ser construído. Para quantificação deste tempo, utilizou-se o módulo *time* da biblioteca *time* do Python. A Tabela 8 mostra os respectivos tempos que cada algoritmo utilizou para realizar as etapas de treinamento e teste. Nota-se que o modelo PCA-KNN apresentou menor tempo para construção e devido a isto, esse algoritmo foi selecionado na construção da interface gráfica.

Tabela 8 - Tempo para construção dos modelos de classificação

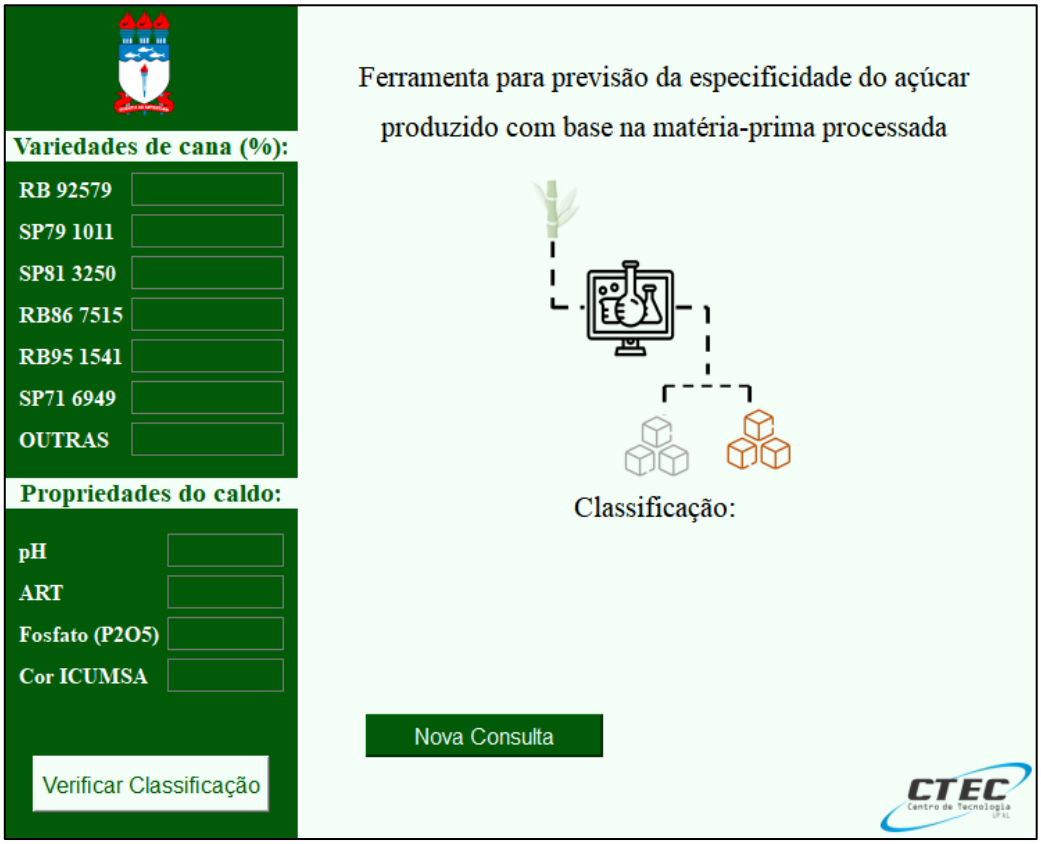
Modelo de classificação	Tempo para construção (s)
KNN (k=5)	0.3784
KNN-PCA	0,0010
RNA	0,5737
PCA-RNA	0,2005

Fonte: Autor, 2022.

Como pode ser observado na Figura 8, o *layout* da interface foi constituído por onze células para preenchimento, cada uma devidamente identificada com as variáveis requiridas para o processo de classificação. O código de programação utilizado na interface para prever a classificação de novas amostras seguiu o algoritmo ilustrado na Figura 9. Para acionar o código, o usuário deve preencher as células vazias com os respectivos valores das variáveis e clicar no botão "Verificar Classificação". Uma vez clicado, o botão "Verificar Classificação" inicia o processo de previsão da amostra inserida, referida na Figura 9 como "nova amostra".

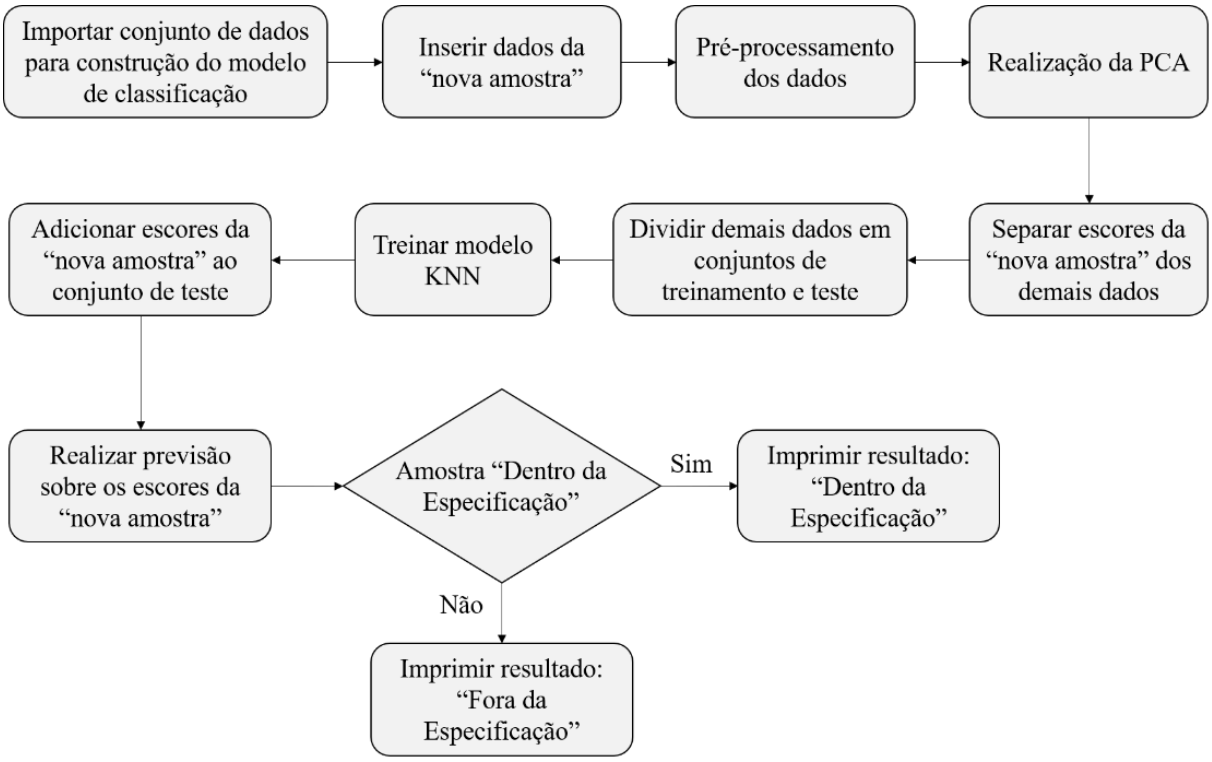
Após obtenção do resultado da classificação da "nova amostra", a interface informa ao usuário se a amostra inserida pertence a classe "Dentro da Especificação" ou "Fora da Especificação" por meio da exibição de uma mensagem, conforme pode ser observado nas Figuras 10 e 11.

Figura 8 - Layout da interface construída



Fonte: Autor, 2022.

Figura 9 – Algoritmo do código utilizado na interface para classificação de novas amostras

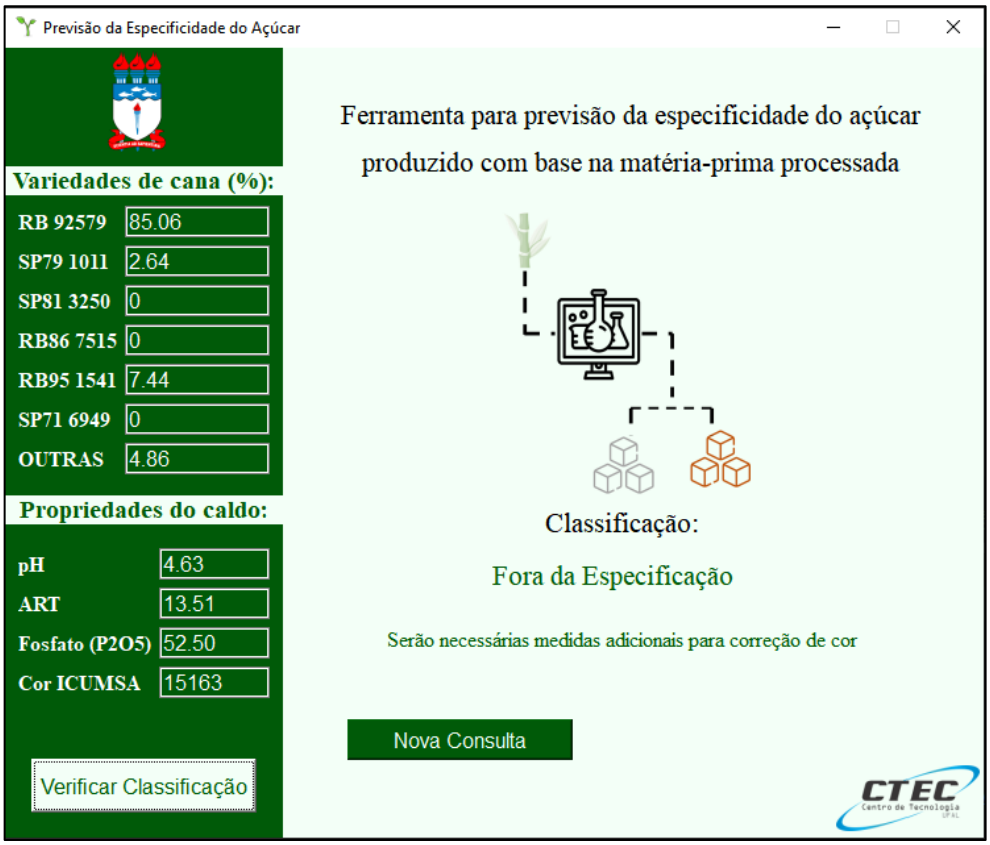


Fonte: Autor, 2022.

Figura 10 - Resultado da previsão de amostra "Dentro da Especificação" na interface

Fonte: Autor, 2022.

Figura 11 - Resultado da previsão de amostra "Fora da Especificação" na interface



Fonte: Autor, 2022.

6 CONCLUSÃO

O presente trabalho utilizou da linguagem de programação Python e suas respectivas bibliotecas para desenvolver algoritmos de aprendizagem de máquina de modo a relacionar a qualidade do açúcar produzido com base nas características inerentes à matéria-prima processada. Os dados inicialmente foram pré-processados e fez-se uma análise exploratória por meio da análise de componentes principais. Através da análise exploratória foi possível a identificação e eliminação da presença de amostras anômalas que poderiam influenciar negativamente no processo de construção dos modelos de classificação.

Após procedimentos de análise exploratória, quatro modelos de reconhecimento de padrões supervisionado foram construídos e comparados, dois deles utilizando o algoritmo do k-ésimo vizinho mais próximo e os outros dois, baseando-se no algoritmo de redes neurais artificiais. Foi observado que os algoritmos KNN (k = 5), PCA-KNN, MLP e PCA-MLP apresentaram 100% de eficiência de na capacidade preditiva. O modelo PCA-KNN, além de classificar corretamente todas as amostras do conjunto de teste em suas respectivas classes, apresentou menor tempo computacional para sua construção e devido a isto foi selecionado para desenvolvimento da interface gráfica.

A interface construída possibilitou que novas amostras fossem classificadas sem a necessidade prévia de conhecimentos relacionados a linguagem de programação e com isso, a ferramenta de classificação tornou-se acessível por qualquer pessoa que tenha acesso a um computador, uma vez que todas as ferramentas utilizadas para sua construção foram de acesso livre, descartando a necessidade de um investimento financeiro inicial para sua aquisição.

Por fim, percebeu-se que a ferramenta desenvolvida é capaz de prever com boa precisão se um açúcar sairá dentro ou fora da especificação com base na matéria-prima processada, tornando possível sua utilização no planejamento estratégico de produção por meio da avaliação prévia dos lotes de cana-de-açúcar enviados para a indústria. A ferramenta construída auxilia, inclusive, na elaboração de procedimentos que se adaptem a recepção da cana-de-açúcar, de modo a evitar a fabricação de produtos que estejam fora da especificação.

REFERÊNCIAS

ALMEIDA, G. M. *et al*. Machine learning in the prediction of sugarcane production environments. **Computers and Electronics in Agriculture**, v. 190, p. 106452, 22 nov. 2021.

BENTO, L.S. **Colourants through cane sugar production and refining**. **CONFERENCE. Delray Beach, Florida, USA, Sugar Processing Research Institute, Inc**. 2008.

BHARDWAJ, P. *et al*. A machine learning application in wine quality prediction. **Machine Learning with Applications**, v. 8, p. 100261, 28 jan. 2022.

BRERETON, Richard G. Pattern recognition in chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 149, p. 90-96, 26 jun. 2015.

BRISON, A. E.; HO, Y. C. **Applied optimal control**. Vol 1. Blaisdell, New Your. 1969

CALDAS, C. S. **Escurecimento do açúcar branco: influência do processo e tempo de armazenamento**. 2012. Tese (Doutorado) - Universidade Federal de Pernambuco, 2012.

CALDAS, C. S. LARRAHONDO J. E. SILVA, J. R. C. **Cálculos fundamentais para o controle químico das indústrias de açúcar e de álcool.** Central analítica LTDA. Maceió, Alagoas, 2017.

FERREIRA, M. M. C. **Quimiometria**: Conceitos, métodos e aplicações. 1. ed. Campinas, SP: Editora Unicamp, 2015.

FUNAHASHI, Ken-Ichi. On the Approximate Realization of Continuous Mappings by Neural Networks. **Neural Networks**, Japão, v. 2, p. 183-192, 8 mar. 1989.

GERÓN, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. Rio de Janeiro: Alta Books, 2019.

GROSS, D; COOMBS, J. Enzymic colour formation in beet and cane juices. **International Sugar Journal**, p. 69-73 e 106-109, 1976.

HAYKIN, S. **Redes Neurais - Princípios e Práticas**. BOOKMAN, São Paulo, 2ª ed. 2001. 900 p.

HERRERA, A *et al*. The generation of color in the sugar production process: Manufacturing process. **Academia: Accelerating the world's research**., p. 1-9. 2014.

ICUMSA, International Comission for Uniform Methods os Sugar Analysis. **Methods Book**. London, 2005.

LAROQUE, D. *et. al.* Kinetic study on te Maillard reaction. Consideration of sugar reactivity. **Food Chemistry**, v. 111, p. 1032 - 1042, 2007.

LARRAHONDO, J. E. **Composición y características químicas de la caña de azúcar y su impacto en el proceso de elaboración del azúcar**, 1º ed., CENICAÑA, 2012.

LEGENDRE, B. L. Chapter 12 - Varietal Differences in the Chemical Composition of Sugarcane. Sugar Series, v. 9, p. 176-185. 1988.

LOPES, C. H. **Tecnologia de Produção de Açúcar de Cana**. São Carlos: EdUFSCar, 2011.

LOVATTI, B. P. O.. **Métodos de aprendizagem de máquina em química analítica: Floresta Randômica aplicada na avaliação de petróleo.** 1 ed. Vitória –ES, 2019.

MACEDO, M. A. **Potencial do mercado doméstico para o açúcar VHP**. 2015. Trabalho de Conclusão de Curso (MTA) - UNIVERSIDADE FEDERAL DE SÃO CARLOS, 2015.

MASSIMO, B. *et al*. Machine Learning for industrial applications: A comprehensive literature review. **Expert Systems with Applications**. v. 175, p. 114820, 1 ago. 2021.

MCKINNEY, W. **Python for Data Analysis**. 1ed. O'Reilly Media, 2013.

MEAD, G.P; CHEN, J. C. P. **Cane Sugar Handbook**. John Wiley & Sons, Inc: Canada. 13 ed. 1977.

MENEZES, J. F. S. **Balanços de massa na indústria sucroalcooleira**. 2015. Trabalho de conclusão de curso (Graduação em Tecnologia em produção Sucroalcooleira) - Universidade Federal da Paraíba, 2016.

NIED, A. **Treinamento de redes neurais artificiais baseado em sistemas de estrutura variável com taxa de aprendizado adaptativa**. 2007. Belo Horizonte, MG. Tese (Doutorado) – Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais.

OLIVEIRA, D. T; ESQUIAVETO, M. M. M; SILVA JUNIOR, J. F. Impacto dos itens da especificação do açúcar na indústria alimentícia. **Ciênc. Tecnol. Aliment.**, Campinas, SP, v. 27, p. 99-102, 10 ago. 2007.

ORZEL, J.; DASZYKOWSKI , M.; WALCZAK, B. Controlling sugar quality on the basis of fluorescence fingerprints using robust calibration. **Chemometrics and Intelligent Laboratory Systems**, v. 110, p. 89-96, 8 out. 2011.

PATON, N. H. The origin of colour in raw sugar. **Proceedings of the Australian Society of Sugar Cane Technologists**, v. 14, p. 8-17, 1992.

PAYNE, J. H. **Operações unitárias na produção de açúcar de cana**. São Paulo: Nobel/Stab, 1989.

PIMENTEL, W. R O. **Aplicação de redes neurais artificiais e de quimiometria na modelagem do processo de craqueamento catalítico fluído**. 2005. Tese (Doutorado) - Unicamp - SP, 2005.

POEL, P. W; SCHIWECK, H.; SCHWARTZ, T. K. **Sugar technology: beet and cane sugar manufacture**. Verlag Dr Albert Bartens KG, 1998.

QIN, S. Joe. Chapter 8 Neural Networks for Intelligent Sensors and Control — Practical Issues and Some Solutions. **Neural Systems for Control**, [*S. l.*], p. 213-234, 1 jun. 1997.

RAMÍREZ-MORALES, I *et al*. Optimization of NIR calibration models for multiple processes in the sugar industry. **Chemometrics and Intelligent Laboratory Systems**, v. 159, p. 45-47, 5 out. 2016.

RIFFER, R. Chapter 13 - The Nature of Colorants in Sugarcane and Cane Sugar Manufacture. **Sugar Series**, v. 9, p. 186-207, 1 jan. 1988.

ROJAS, E. D. **Avaliação de canais de comercialização e distribuição de açúcar mascavo: estudo de caso em uma associação**. 2008. Magister Scientiae (Mestrado) - Universidade Federal de Viçosa, Viçosa, MG, 2007.

ROSENBLATT, F. **The Perceptron:** A probabilistic model for information storage and organization in the Brai**n. Psychological Review**, 65:386-408.

SANTOS, F.; BORÉM, A.; CALDAS, C. **Cana-de-açúcar**: bioenergia, açúcar e etanol – tecnologia e perspectivas. 3. ed. rev. e aum. Londrina, PR: Mecenas, 2018.

SCOPUS: Elsevier. Disponível em: https://www-scopus.ez9.periodicos.capes.gov.br/. Acesso em: 24 abr. 2022.

SENA, M. M. *et al*. Avaliação do uso de métodos quimiométricos em análise de solos. **Química Nova**, Jaguariúna - SP, v. 23, n. 4, p. 547-556, 13 out. 1999.

SHAW, B.; SUMAN, A.K.; CHAKRABORTY, B. Wine Quality Analysis Using Machine Learning. **Advances in Intelligent Systems and Computing**, v. 937, p. 239–247, 17 jul. 2019.

SILVA, C. S. **Influência de variáveis do campo na qualidade do caldo de cana-de-açúcar para a produção de açúcar e álcool**. 2017. Conclusão de Curso (Graduação) - Universidade Federal de Alagoas, 2017.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. **Redes Neurais Artificiais para engenharia e ciências aplicadas: fundamentos teóricos e aspectos práticos**. 2. ed. rev. e aum. São Paulo: Artliber, 2019.

SMITH, I. A. *et al*. A survey of colour input and formation in process. **Proc S Afr Sug Technol Ass**, v. 64, p. 213-216. 1990.

SMITH, P. *et al.* Colour studies in milling. **Proceedings of the 1981 Conference of the Australian Society of Sugar Cane Technologists, held at Bundaberg, Queensland from 11th May to 15th May, 1981/edited by OW Sturgess**. Brisbane, Qld.: Watson Ferguson and Co., 1981., 1981. (Qualquer coisa, substituir pelo Peter Rein)

SOUSA, Fabiano Bernardo. **Análise de modelo de Hopfield com topologia de rede complexa**. 2014. Dissertação (Mestrado) - USP - São Carlos, 2014.

SOUZA, A. M.; POPPI, R. J. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espcetroscopia no infravermelho médio e análise de componentes principais: um tutorial, parte 01. **Química Nova**, Campinas, SP, ano 2012, v. 35, n. 1, p. 223-229, 22 jul. 2012.

TORRES, R. C. **Estudo do processo de cristalização de açúcar**: modelagem e estratégias de controle. 2007. Dissertação (Mestrado em Engenharia Química) - Universidade Federal de São Carlos, 2007.

WANG, L. Quality analysis, classification, and authentication of liquid foods by near-infrared spectroscopy: A review of recent research developments. **Critical Reviews in Food Science and Nutrition**, v. 57, p. 1524-1538, 21 fev. 2017.

ZACURA FILHO, G.; PICCIRILLI, J. P. **O processo de fabricação do açúcar e do álcool**: desde a lavoura da cana até o produto acabado. 1 ed. Santa Cruz do Rio Pardo - SP: Viena, 2012.