

**UNIVERSIDADE ESTADUAL PAULISTA JÚLIO DE MESQUITA FILHO
INSTITUTO DE QUÍMICA DE ARARAQUARA**

Ivan Sayed Rossi

**Classificação de ácido oleico e ácido palmítico no processo de microfiltração
tangencial por meio de redes neurais artificiais e árvores de decisão**

**ARARAQUARA
2023**

**UNIVERSIDADE ESTADUAL PAULISTA JÚLIO DE MESQUITA FILHO
INSTITUTO DE QUÍMICA DE ARARAQUARA**

Ivan Sayed Rossi

**Classificação de ácido oleico e ácido palmítico no processo de microfiltração
tangencial por meio de redes neurais artificiais e árvores de decisão**

**Trabalho de conclusão de curso
apresentado ao Instituto de Química, da
Universidade Estadual Paulista "Júlio de
Mesquita Filho" como parte das
exigências para a obtenção do título de
Bacharel em Engenharia Química**

**Orientadora: Prof. Dra. Érica Regina
Filletti Nascimento**

**ARARAQUARA
2023**

IVAN SAYED ROSSI

Classificação de ácido oleico e ácido palmítico no processo de microfiltração tangencial por meio de redes neurais artificiais e árvores de decisão

Trabalho de conclusão de curso apresentado ao Instituto de Química, da Universidade Estadual Paulista "Júlio de Mesquita Filho" como parte das exigências para a obtenção do título de Bacharel em Engenharia Química

Araraquara, 19 de janeiro de 2023

BANCA EXAMINADORA

Prof. Dra. Érica Regina Filletti Nascimento
Instituto de Química – UNESP, Araraquara

Prof. Dr. Elias de Souza Monteiro Filho
Instituto de Química – UNESP, Araraquara



Prof. Dra. Fabiola Manhas Verbi Pereira
Instituto de Química – UNESP, Araraquara

AGRADECIMENTOS

Ao corpo docente, direção e administração da Universidade Estadual Paulista "Júlio de Mesquita Filho" campus Araraquara, pelas oportunidades, ensinamentos e apoio que recebi durante minhas vivências em ensino, pesquisa e extensão.

À CAPES, pelo apoio financeiro e institucional dado para a realização do meu intercâmbio para a École Nationale Supérieure des Ingénieurs en Arts Chimiques et Technologiques em Toulouse - França.

À minha família por nunca terem medido esforços em me apoiar a estudar e atingir meus objetivos.

À minha namorada Larissa por todo apoio, carinho e amizade durante toda a graduação.

Ao meu grande amigo Tiago Bellè por todas as experiências, conhecimentos e momentos compartilhados.

À Profa. Erica Filletti por toda orientação, ensino e paciência durante a elaboração deste trabalho.

LISTA DE FIGURAS

Figura 1 - Representação das diferentes configurações de filtração: a) filtração perpendicular; b) filtração tangencial.....	13
Figura 2 - Neurônio de McCulloch e Pitts.....	15
Figura 3 - Modelo de neurônio não-linear	16
Figura 4 - Função degrau.....	18
Figura 5 - Função limiar por partes	18
Figura 6 - Função sigmóide.....	19
Figura 7 - Redes neurais do tipo feed-foward com uma única camada e múltiplas camadas respectivamente	20
Figura 8 - Rede neural recorrente	20
Figura 9 - Representação de uma árvore de decisão	24
Figura 10 - Performance pelo número de épocas	29
Figura 11 - Matriz de confusão para os conjuntos de treinamento, validação e teste ...	30
Figura 12 - Curva característica de operação do receptor (ROC) para ambos os ácidos	31
Figura 13 - Coeficiente de Gini.....	32
Figura 14 - Erro por número de árvores na simulação	34

LISTA DE TABELAS

Tabela 1 - Parâmetros utilizados no modelo de Florestas Aleatórias	31
Tabela 2 - Matriz de confusão obtida pelo modelo para o período de treinamento.....	33
Tabela 3 - Matriz de confusão obtida pelo modelo para o período de teste	33

SUMÁRIO

1.	INTRODUÇÃO	10
1.1.	OBJETIVO	11
2.	METODOLOGIA	12
2.1.	MEMBRANAS, MICROFILTRAÇÃO E AMOSTRAS DE ÁCIDOS GRAXOS..	12
2.2.	REDES NEURAIS ARTIFICIAIS	14
2.2.1.	MODELOS MATEMÁTICOS DE NEURÔNIOS	15
2.2.2.	FUNÇÕES DE ATIVAÇÃO	17
2.2.3.	ARQUITETURA DAS REDES NEURAIS.....	19
2.2.4.	TIPOS DE APRENDIZAGEM.....	21
2.2.5.	O ALGORITMO DO GRADIENTE CONJUGADO ESCALONADO (SCG).....	21
2.3.	FLORESTAS ALEATÓRIAS	23
3.	RESULTADOS E DISCUSSÕES.....	28
3.1.	RESULTADOS DA RNA	28
3.2.	RESULTADOS DA FLORESTA ALEATÓRIA.....	31
4.	CONCLUSÃO	35
5.	REFERÊNCIAS	36

RESUMO

O processo de microfiltração tangencial com membranas cerâmicas tem sido muito utilizado na clarificação do suco de açaí pois tem as vantagens de operar sem conservante químico, separar seletivamente os materiais e possuir baixo custo operacional (usa apenas a energia elétrica para o funcionamento da bomba hidráulica, não é necessário usar trocadores de calor, evitando problemas de poluição térmica e sobrecarga do sistema). Ainda, esta tecnologia consegue estabilizar as características funcionais e microbiológicas dos produtos à base de açaí, pois utiliza baixas temperaturas reduzindo efeitos prejudiciais de alguns tratamentos térmicos e conservando, assim, as características originais da fruta. Sua polpa possui uma composição complexa com elevada porcentagem de lipídios, proteínas, sais minerais, ácidos graxos e fibra alimentar, que contribuem para a formação de incrustações nos poros das membranas. Assim, visando a melhoria do processo, é útil identificar esses compostos e analisar seus impactos individuais sobre os poros da membrana. Através da base de dados disponibilizada por Camimoto (2012), foram desenvolvidos modelos neurais e florestas aleatórias para classificar os ácidos graxos (ácido oleico e palmítico) no processo de microfiltração tangencial. O projeto busca identificar o ácido oleico e ácido palmítico em misturas bifásicas de óleos vegetais e água no processo de microfiltração por meio de Redes Neurais Artificiais (RNAS) e Florestas Aleatórias (*Random Forest*). Tanto a RNA quanto as Florestas Aleatórias conseguem fornecer resultados altamente precisos com muita rapidez e pouco esforço computacional.

Palavras-chave: Redes Neurais Artificiais. Florestas Aleatórias. Açaí.

ABSTRACT

The tangential microfiltration process with ceramic membranes has been widely used in the clarification of the açai juice because it has the advantages of operating without chemical preservatives, selectively separating the materials and has low operating costs (it uses only electric energy to operate the hydraulic pump, there is no need to use heat exchangers, avoiding problems of thermal pollution and system overload). Furthermore, this technology is able to stabilize the functional and microbiological characteristics of the açai products, because it uses low temperatures, reducing the harmful effects of some heat treatments and preserving the original characteristics of the fruit. Its pulp has a complex composition with a high percentage of lipids, proteins, mineral salts, fatty acids and dietary fiber, which contribute to the formation of incrustations in the pores of the membranes. Thus, aiming at process improvement, it is useful to identify these compounds and analyze their individual impacts on the membrane pores. Using the database provided by Camimoto (2012), neural models and random forests were developed to classify fatty acids (oleic and palmitic acid) in the tangential microfiltration process. The project seeks to identify oleic acid and palmitic acid in two-phase mixtures of vegetable oils and water in the microfiltration process by means of Artificial Neural Networks (ANN) and Random Forests (Random Forest). Both ANN and Random Forests are able to provide highly accurate results very quickly and with little computational effort.

Keywords: Artificial Neural Networks. Random Forest. Açai.

1. INTRODUÇÃO

O açaí, fruto do açaizeiro ou também palmeira *Euterpe oleracea* Mart. é uma planta bem popular na região Norte do Brasil, abundante no estado do Pará e presente na culinária na região. Nos últimos anos, o açaí ganhou imenso mercado em todo Brasil e no mundo. Devido ao seu poder antioxidante e nutritivo, ele é capaz de suprir 65% das necessidades teóricas de lipídeos de um adulto e 25% da necessidade diária de proteína (ROGEZ, 2000).

Grande parte da estrutura do açaí é formada de lipídeos, no caso, ácidos graxos monoinsaturados e poli-insaturados, sua composição chega a ser comparada com o óleo de oliva, porém o açaí supera em composição de ácidos graxos. Vale citar que os ácidos graxos de maior parcela são, respectivamente: Oleico, Palmitoleico e Palmítico (NASCIMENTO et al., 2008).

A alta demanda de açaí tanto para uso alimentício quanto cosmético, impulsionou a produção industrial bem como a científica. O processo mais utilizado é a microfiltração tangencial com membranas cerâmicas para a clarificação do suco de açaí, que é vantajoso devido ao não uso de conservante químico, separação seletiva e baixo custo operacional, sendo possível devido à ausência de trocadores de calor que é agente de poluição térmica e sobrecarga do sistema. Por fim, vale ressaltar que o uso da microfiltração tangencial permite estabilizar os traços funcionais e microbiológicos dos produtos a base de açaí em decorrência de que em todo o processo utiliza-se baixas temperaturas e ausência de variação dessa, nesse sentido, as características e propriedades da fruta são conservadas (CHERYAN,1998).

Descritas as vantagens do processo de microfiltração tangencial, também há uma desvantagem em relação a este tipo de clarificação que, no caso, é o de decaimento do fluxo de permeado em função do tempo de operação. Tal fato está relacionado a concentração de sólidos presentes na matéria prima as quais dificultam o processo em decorrência de incrustações na membrana.

1.1.OBJETIVO

O objetivo deste trabalho é desenvolver ferramentas computacionais alternativas para identificar ácido oleico e ácido palmítico em misturas bifásicas de óleos vegetais e água no processo de microfiltração tangencial com membranas cerâmicas. O estudo é feito através de Redes Neurais Artificiais com o algoritmo de gradiente conjugado escalonado e Florestas Aleatórias.

A dissertação foi estruturada da seguinte maneira:

1. O capítulo 2 aborda os principais conceitos teóricos de membranas e microfiltração, redes neurais artificiais com os modelos matemáticos de neurônios artificiais, funções de ativação, arquiteturas das RNAs e tipos de aprendizagem, além de florestas aleatórias e suas principais características.
2. No capítulo 3 são apresentados os resultados e discussões da classificação dos ácidos através dos métodos propostos.
3. Por fim, no capítulo 4 encontram-se as considerações finais

2. METODOLOGIA

Para a simulação do modelo de Floresta Aleatória foi utilizado o Software livre Studio R e as seguintes bibliotecas: *readxl*, *caret* e *randomForest*. A partir das quais foi possível a leitura dos arquivos contendo os dados de pressão, fluxo e tempo do ácido palmítico e oléico e, também, execução dos algoritmos. Para implementação da Rede Neural Artificial, foi utilizado o *software* Matlab com a ferramenta *nnstart – pattern recognition app* e o algoritmo do gradiente conjugado escalonado (*trainscg*).

2.1. MEMBRANAS, MICROFILTRAÇÃO E AMOSTRAS DE ÁCIDOS GRAXOS

O processo que envolve a separação por membranas utiliza-se como barreira seletiva as membranas sintéticas através das quais é possível separar, concentrar e purificar fisicamente substâncias (GALVÃO; GOMES, 2015). O sistema separativo propõe-se a separar o líquido dos sólidos que ficam retidos na superfície da membrana. Tal processo exige determinado tipo de força externa, podendo ser a pressão, a sucção ou o potencial elétrico. Ademais, nenhuma reação química ocorre durante o processo. (SCHNEIDER; TSUTYIA, 2001; HABERT et al., 2006).

As membranas são caracterizadas e conhecidas pela sua eficiência na seletividade das partículas que se propõem a separar, nesse sentido, elas são fundamentais na eficiência e qualidade dos processos separativos. A seletividade pode ser caracterizada pela capacidade de seleção das partículas de acordo com a relação entre tamanho dos poros da membrana e a da partícula, diferença de permeabilidade de um componente em outro e pela diferença nas densidades de carga elétrica das partículas. (BHAVE; RAMESH, 1991)

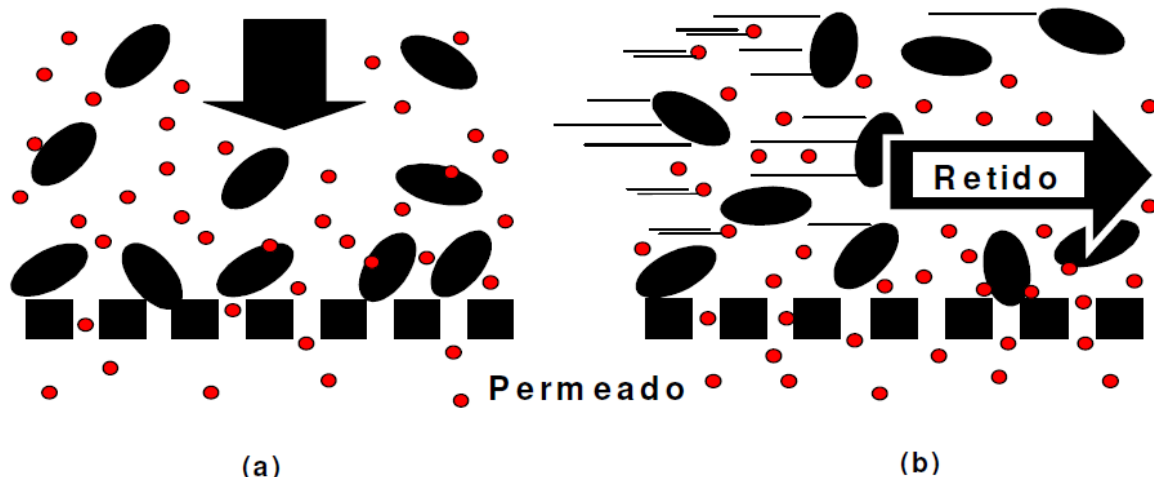
A microfiltração é um processo de separação que utiliza de um gradiente de separação limitado por uma membrana que é utilizado principalmente na separação de misturas do tipo sólido-sólido e líquido-líquido (QUEIROS, 2004, TREVIZOLI, 2010). A separação consiste no efeito peneira molecular e macromolecular, assim a há a retenção

de partículas maiores que a porosidade da membrana, ademais, ressalta-se que as partículas menores podem ser retidas pela formação de uma camada polarizada (PELEGRIIN, 2004). Para líquidos com alta concentração de sólidos tal como polpas de frutas é recomendada a microfiltração tangencial e apresenta três tipos distintos de fluxo, de alimentação, permeado e concentrado. Vale ressaltar que opera normalmente em regime permanente sendo que a alimentação é bombeada de forma ininterrupta por uma membrana. A alimentação em velocidade reduzida evita a formação da camada polarizada e permite maior controle da incrustação do módulo.

O processo de filtração por membranas pode ocorrer de duas formas: perpendicular e tangencial. O primeiro (Figura 1 - a) representa o fluido que se movimenta na direção perpendicular à superfície da membrana de modo que à medida que a filtração acontece, passa a haver um acúmulo de matéria na superfície o que diminui drasticamente a eficiência. Além disso, a filtração perpendicular é indicada para líquidos com baixa concentração de sólidos.

Por outro lado, a filtração tangencial (Figura 1 - b) é vantajosa em relação a primeira justamente por não ocorrer a deposição de partículas na superfície da membrana, nesse sentido, o fluido se movimenta paralelamente a superfície da membrana o que permite que a eficiência seja mantida e não haja acúmulo de matéria. (SANTOS, 1999). Tal processo é recomendado para líquidos com alta concentração de sólidos como polpas de frutas.

Figura 1 - Representação das diferentes configurações de filtração: a) filtração perpendicular; b) filtração tangencial



Fonte: ARMOA, M.H.; JAFELICCI JR, 2020

As amostras empregadas foram de ácidos oleicos e palmíticos obtidos no processo de microfiltração tangencial do açaí, e encontram-se publicadas no trabalho de Camimoto (2012), como parte do projeto de mestrado da discente Karime Bárbara Santo Caminoto, sob a orientação do prof. Dr. Sérgio Rodrigues Fontes da USP. Neste estudo experimental foi investigada a influência dos dois principais ácidos graxos presentes no açaí, ácido oleico e ácido palmítico, em misturas com água e no processo de microfiltração tangencial com membranas cerâmicas de alumina. Ao todo, obteve-se 216 amostras de ácidos graxos sendo 108 amostras de ácido oleico e 108 amostras de ácido palmítico, com variações de tempo, pressão e fluxo.

2.2. REDES NEURAIS ARTIFICIAIS

As redes neurais artificiais podem ser definidas como múltiplas unidades, no caso neurônios, de processamento orientadas de forma unidirecional que calculam determinadas funções as quais em grande parte são não-lineares. A conexão entre os neurônios e os canais de comunicação estão associados a pesos os quais apresentam orientações de acordo com o "conhecimento" de cada um dele, nesse sentido, ocorre uma ponderação entre a informação recebida por cada neurônio na rede. (FILLETTI, 2007)

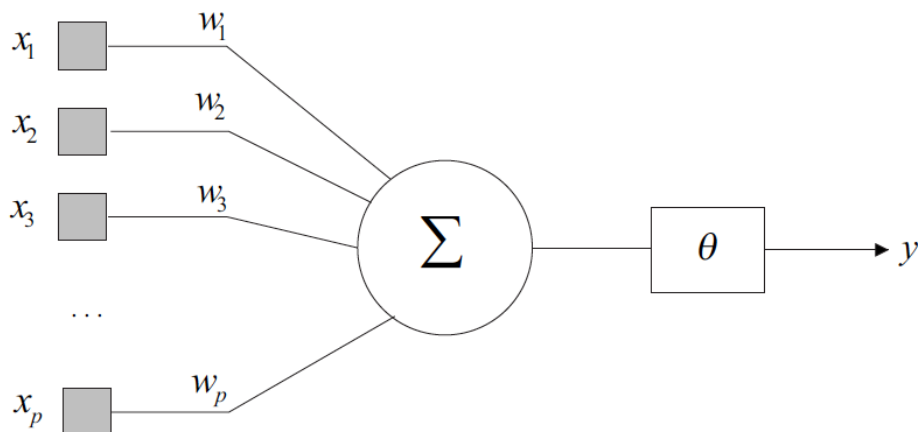
A grande inovação das redes neurais é sua capacidade de aprendizado no ambiente em que é inserida, nesse sentido, os neurônios que as compõem são capazes de melhorar seu desempenho à medida que são treinados. O processo é feito através do treinamento onde os pesos aplicados a cada unidade são ponderados permitindo que o sistema possa ser generalizado. O processo de generalização habilita uma configuração de rede que, uma vez treinada em determinado ambiente, pode atuar de forma similar, porém com dados totalmente novos, logo, atingindo uma solução para uma determinada classe de problemas.

O algoritmo de aprendizado é orientado para a solução de um problema de aprendizado através de um conjunto de regras bem definidas onde, a partir das características que são extraídas do modelo, torna-se possível gerar respostas para um determinado problema.

2.2.1. MODELOS MATEMÁTICOS DE NEURÔNIOS

A representação matemática de um neurônio foi proposta por McCulloch e Pitts em 1943 (Figura 2) a qual pode ser descrita pelas seguintes operações. Primeiro, os sinais são identificados na entrada, em seguida, cada sinal é multiplicado por um fator que representa sua influência no modelo. Posteriormente, soma-se ponderadamente os sinais e obtém-se um nível de atividade. Por fim, se o nível de atividade atingir certo limite, o neurônio produz determinada resposta de saída (ICMC USP, 2019).

Figura 2 - Neurônio de McCulloch e Pitts



Fonte: Filletti, 2006

Na imagem acima, os sinais de entrada são representados por $x_1, x_2, x_3 \dots x_p$ e os pesos sinápticos que recebem esses sinais por $w_1, w_2, w_3 \dots w_p$, por fim, é obtida a resposta y . A ativação do neurônio ocorre a partir da resposta dada pela função de ativação, isto é, a ativação pode ocorrer ou não, dependendo da resposta resultante.

Em linguagem matemática, a resposta de um neurônio é obtida quando a condição abaixo é satisfeita:

$$\sum_{j=1}^p x_j \cdot w_j \geq \theta \quad \text{Equação 1}$$

p = números de entradas na rede neural artificial

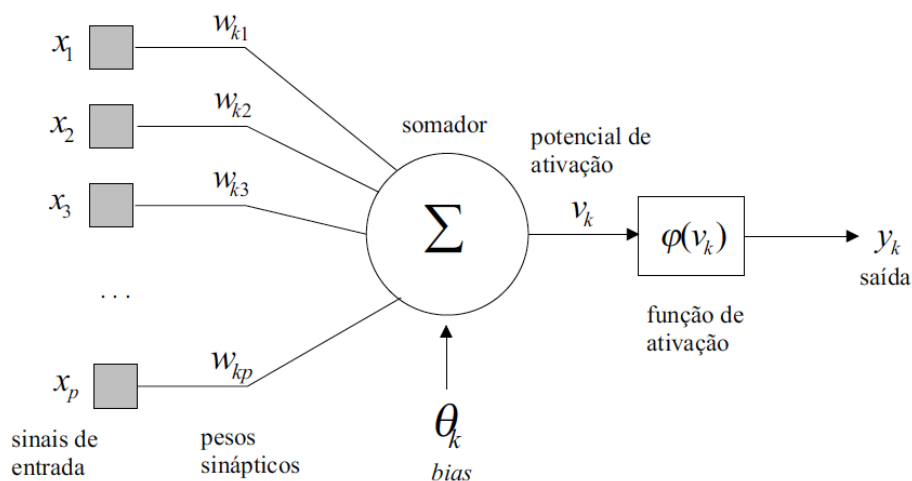
x_j = sinal de entrada

w_j = pesos sinápticos

θ = limiar do neurônio

O modelo mais utilizado atualmente é o não-linear (Figura 3) visto que os pesos sinápticos são ajustáveis, diferentemente do modelo proposto inicialmente onde os pesos são fixos.

Figura 3 - Modelo de neurônio não-linear



Fonte: Filletti, 2006

Seu funcionamento é análogo ao primeiro, porém inclui um *bias* representado por θ_k que se propõem a aumentar ou diminuir a entrada de ativação. Nesse sentido, matematicamente pode-se representar as equações da seguinte maneira:

$$u_k = \sum_{j=1}^p x_j \cdot w_{kj} \quad \text{Equação 2}$$

$$v_k = u_k + \theta_k \quad \text{Equação 3}$$

$$y_k = \varphi(v_k) \quad \text{Equação 4}$$

p = números de entradas na rede neural

x_j = sinal de entrada

w_j = pesos sinápticos

θ_k = bias

$\varphi(v_k)$ = função de ativação

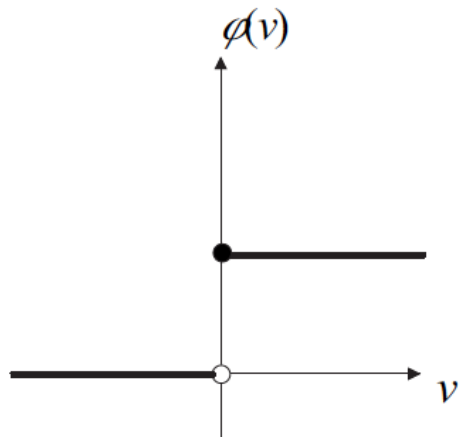
y_k = sinal de saída do neurônio

Como citado anteriormente, o modelo apresenta algumas modificações que permitem maior maleabilidade do neurônio a fim de obter-se determinado comportamento do modelo, isto é, os pesos sinápticos e bias são parâmetros escalares ajustáveis do neurônio. Sendo assim, o modelo matemático permite que a rede neural criada possa ser colocada em treinamento a fim de ajustar os pesos e *bias* de modo que se obtenha o comportamento desejado.

2.2.2. FUNÇÕES DE ATIVAÇÃO

A função de ativação define o sinal de saída do neurônio a partir da combinação linear das entradas e dos pesos das sinapses. Há três tipos de função sendo elas ilustradas abaixo.

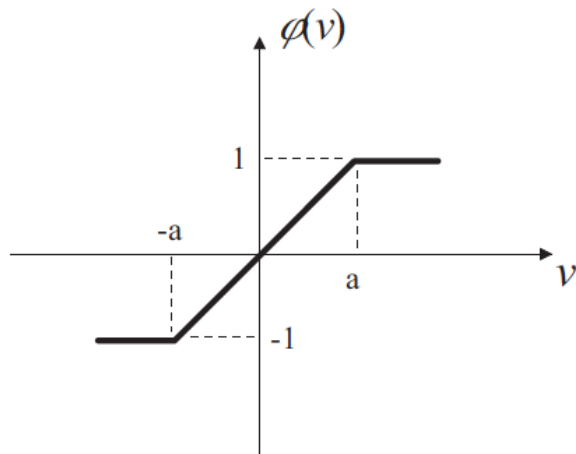
Figura 4 - Função degrau



$$\varphi(v) = \begin{cases} 1 & \text{se } v \geq 0 \\ 0 & \text{se } v < 0 \end{cases}$$

Equação 5

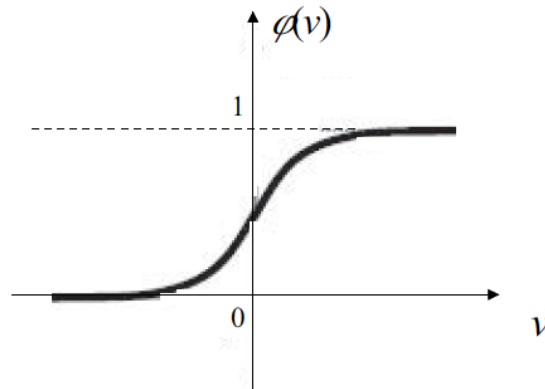
Figura 5 - Função limiar por partes



$$\varphi(v) = \begin{cases} 1 & \text{se } v \geq a \\ \frac{v}{2} & \text{se } -a < v < a \\ -1 & \text{se } v \leq -a \end{cases}$$

Equação 6

Figura 6 - Função sigmóide



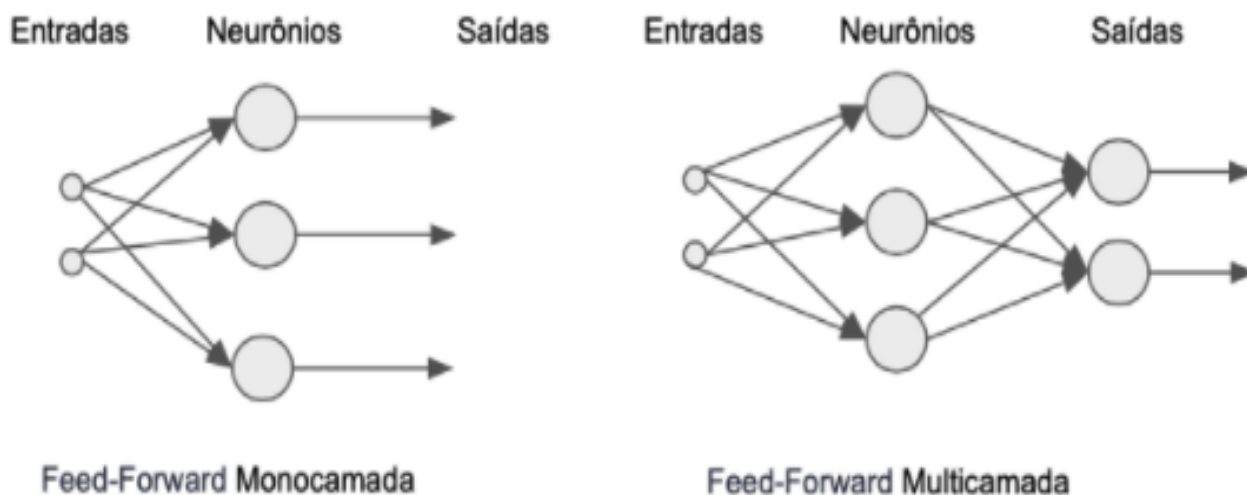
$$\phi(v) = \frac{1 - e^{-av}}{1 + e^{+av}} \quad \text{Equação 7}$$

2.2.3. ARQUITETURA DAS REDES NEURAIS

No escopo das redes neurais, o termo arquitetura refere-se à topologia e ao modelo propriamente dito. Há três tipos de RNAs que são mais difundidas e serão detalhados abaixo.

As redes neurais do tipo *feed-forward* com uma única camada e múltiplas camadas são caracterizadas por camada de entrada, uma ou mais camadas(s) intermediárias e camada de saída. A camada de entrada representa o local onde recebe-se as entradas, a camada intermediária onde é realizado o processamento e, camada de saída, onde o resultado é obtido. Em redes neurais do tipo *feed-forward* monocamada existe apenas uma camada de neurônios ligados diretamente aos nós de entrada e eles mesmos já fornecem os dados de saída. Por outro lado, o do tipo *feed-forward* com múltiplas camadas apresenta mais de uma camada de neurônios na camada oculta que, nesse sentido, possibilita maior interação entre os neurônios (Sazli, 2006).

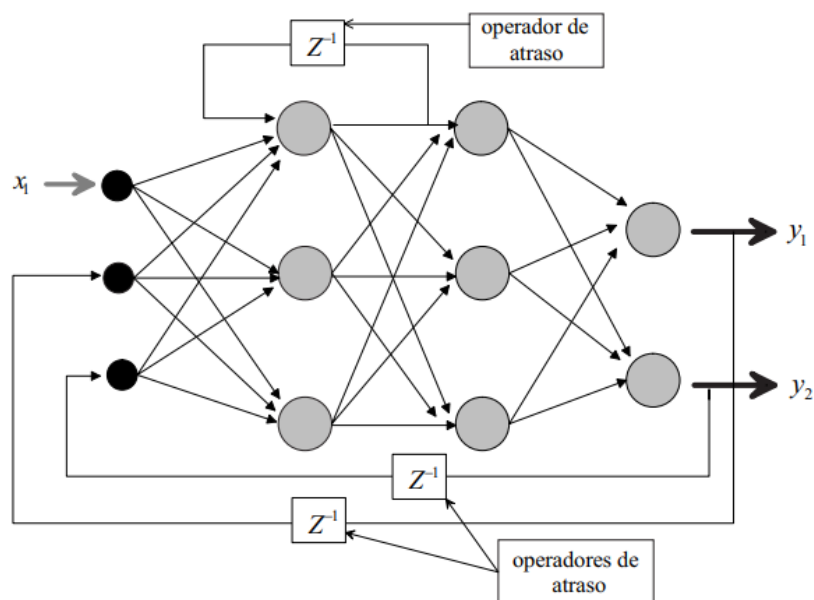
Figura 7 - Redes neurais do tipo *feed-forward* com uma única camada e múltiplas camadas respectivamente



Fonte: Ateliware, 2018

A terceira rede neural é a do tipo recorrente a qual apresenta conexões de realimentação. Essas conexões apresentam como principal característica a criação de um comportamento temporal na rede, como uma memória de curto prazo (FILLETTI, 2007).

Figura 8 - Rede neural recorrente



Fonte: Filletti, 2007

2.2.4. TIPOS DE APRENDIZAGEM

Os tipos de aprendizagem envolvem o conceito de que uma rede neural deve ser treinada para que se obtenha as saídas desejadas. Nesse sentido, uma rede neural quando implementada passa por um período de aprendizado onde haverá o ensino de padrões de entrada e mudança de pesos das conexões, a fim de que possa ser implementada para determinadas aplicações.

Há três tipos de aprendizagem que são utilizadas: aprendizagem supervisionada, aprendizagem não-supervisionada e aprendizagem por reforço.

Na aprendizagem supervisionada, a rede neural recebe um conjunto de dados de entrada e de saída desejados. Em seguida, a partir da resposta de saída gerada e a esperada, calcula-se o erro da saída. Os ajustes nos pesos das conexões ocorrem à medida que se compara os resultados de saída com a saída desejada, assim, é possível encontrar os pesos ideais para as conexões.

A aprendizagem não-supervisionada funciona de modo que não há um controle crítico sobre os dados que estão sendo obtidos. A rede neural trabalha de modo a determinar algumas propriedades dos conjuntos de dados, extraíndo dos conjuntos de dados as informações necessárias para a realização do mapeamento de entrada e saída.

A aprendizagem por reforço utiliza o conceito de recompensa ou de reforço os quais realizam ajustes nas redes a partir dos dados que são obtidos na saída gerada pela rede. Tal processo ocorre a partir da utilização de uma função a qual determina se aquela resposta gerada é positiva ou negativa para rede, assim, realizará ajustes a fim de obter os pesos que convirjam para a resposta esperada (FILLETTI, 2007).

2.2.5. O ALGORITMO DO GRADIENTE CONJUGADO ESCALONADO (SCG)

O método do algoritmo do gradiente conjugado (GC) precede ao do gradiente conjugado escalonado (SCG), nesse sentido, é necessária a introdução do primeiro a fim de apresentar o último. Esse algoritmo funciona como um método intermediário entre o de primeira ordem *backpropagation* e o algoritmo de Newton, sendo seu principal objetivo acelerar a taxa de convergência que é normalmente lenta com o

backpropagation e evitando despesas computacionais como a resolução de matrizes hessianas como ocorre no algoritmo de Newton. O único que tem aplicação para problemas de grande escala, com milhares de parâmetros ajustáveis é o GC quando comparado aos outros algoritmos de segunda ordem. Por fim, vale ressaltar que o SCG é a versão otimizada do GC (ALMEIDA, 2007).

Em suma, no método de *backpropagation* o ajuste dos pesos é feito no negativo do gradiente, na direção onde a função de custo decresce de forma acelerada, porém apesar do fato da função decrescer mais rapidamente na direção negativa do gradiente, a velocidade de convergência é lenta. Por outro lado, no algoritmo de gradiente conjugado, a minimização da função dos custos ocorre em diversas direções conjugada de modo a produzir uma taxa de convergência maior que o algoritmo de *backpropagation* além de apresentar uma taxa de aprendizagem que é constante e usada de modo a definir qual será a intensidade de mudança de cada peso. No GC, a mudança dos pesos que são aplicados é ajustada em cada interação que resultada em uma sequência de estimativas η_k as quais finalizam quando uma solução satisfatória é encontrada (ALMEIDA, 2007).

A Equação 8, exibe como os algoritmos de Gradiente Conjugado atuam, isto é, partem no sentido da descida do gradiente a partir da primeira interação (ALMEIDA, 2007).

$$P_0 = -g_0 \quad \text{Equação 8}$$

A distância ótima é determinada a partir da busca em linha de modo a se mover no sentido da atual direção de busca. Por fim, após a determinação do parâmetro da taxa de aprendizado (η), os pesos da Equação 9 abaixo são atualizados.

$$w_k + 1 = w_k + \eta_k P_k \quad \text{Equação 9}$$

Em seguida, a fim de que os sentidos das buscas anteriores sejam conjugados é realizada uma nova direção de busca. Para obter-se a nova direção, é realizada uma nova busca de modo que combine o sentido da descida mais íngreme com o sentido precedente da busca como descrito na Equação 10.

$$P_k = -g_0 + \beta_k P_{k-1}$$

Equação 10

A constante β_k é o fator determinante que distingue as diferentes versões dos algoritmos de Gradiente Conjugado. A pesquisa de linha em cada etapa da iteração não é necessária quando comparada a outras funções de treinamento conjugado devido ao fato de que o algoritmo SCG foi projetado para evitar o custo computacional da busca em linha. O SCG atua de modo a unir o conceito do modelo de região de confiança utilizado no algoritmo Levenberg-Marquardt, com a abordagem do GC em razão de utilizar uma aproximação do cálculo da matriz Hessiana que deve ser positiva.

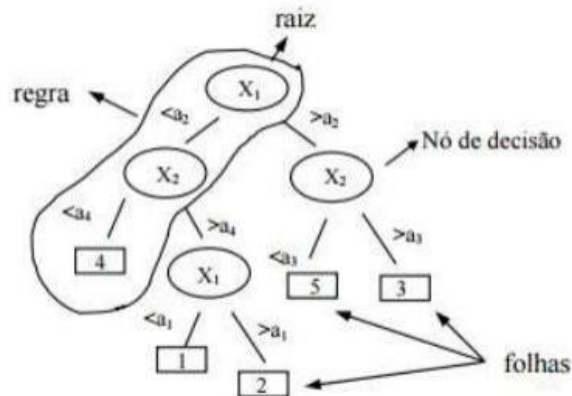
Nesse sentido, esse mecanismo permite que o algoritmo seja mais rápido que qualquer outro de segunda ordem. Apesar da função *trainscg* requerer que mais iterações sejam necessárias para convergir quando comparado a outros algoritmos, o número de cálculos em cada iteração é reduzido pois não é realizada a pesquisa de linha.

Por fim, ressalta-se que há dois parâmetros no algoritmo SGC sendo os parâmetros: σ_k e λ_k . O primeiro é uma ponderação para cálculo de aproximação de segunda ordem e o segundo um parâmetro que regula a falta de definição da matriz hessiana. Logo, a definição desses parâmetros é necessária a fim de que a convergência seja rápida e encontre resultados relevantes (ALMEIDA, 2007).

2.3.FLORESTAS ALEATÓRIAS

As florestas aleatórias são um conjunto de diversas árvores de decisão. A árvore de decisão pode ser definida como modelo estatístico que é formado por treinamento supervisionado para projetar e classificar dados. Os modelos baseiam-se na técnica da criação de subproblemas de menor complexidade a partir de um problema altamente complexo e, recursivamente, a técnica é aplicada para cada subproblemas de menor complexidade (GAMA, 2004). A imagem abaixo representa uma árvore de decisão onde em cada nó de decisão há um teste para determinado atributo, além disso, vale ressaltar que cada conjunto de ramos são distintos, cada folha está relacionada a uma classe e cada percurso está ligado a uma regra de classificação.

Figura 9 - Representação de uma árvore de decisão



Fonte: GAMA, 2004

Como citado anteriormente, o modelo de florestas aleatórias é o conjunto de árvores de decisão onde cada árvore de decisão utiliza um subconjunto de atributos extraídos dos dados originais. Os algoritmos de árvores de decisão são criados a fim de encontrar os atributos que provêm máxima segregação dos dados, de acordo com o atributo que está sendo classificado e a cada nível da árvore. Nesse sentido, a árvore de decisão segue as seguintes etapas. Primeiro, a partir dos dados originais é criado um nó raiz (nó inicial) e, a partir de um teste lógico, esse nó é dividido em outros dois ou mais ramos. Em seguida, os nós que foram gerados a partir da divisão contêm novos testes lógicos os quais promovem nova divisão e ramificação da árvore. Por fim, a divisão dos nós internos continua até que se atinja um nó folha o qual não se ramificará mais. Nesse sentido, a repetição das etapas citadas caracteriza a recursividade da árvore de decisão (BREIMAN et al., 1984).

A construção de árvores é feita pela seleção aleatória dos atributos em que são aplicados em cada um dos nós das árvores. A partir da criação de diversas árvores, é possível determinar qual atributo, isto é, teste de lógica, é determinante para a tomada de decisão. Os subconjuntos recebem um peso, entenda-se peso por relevância, o qual determinará sua importância na classificação, isto é, quanto mais precisa for uma árvore, maior será a contribuição para a decisão de classificação (NETO, 2014).

Além disso, o algoritmo é escalar e comporta um número grande de atributos. O uso de subconjuntos torna o algoritmo mais complexo e robusto, apresentando, assim,

boa assertividade quando exposto a diferentes conjuntos de dados (BELLE, 2008). A técnica é considerada de alta precisão comparada a redes neurais e máquina de vetores suporte (CARUANA et al., 2008). Vale dizer também que as florestas aleatórias são efetivas e evitam o sobre ajuste, além de serem pouco sensíveis a ruídos (BREIMAN, 2001).

O coeficiente de Gini é amplamente utilizado em diversos campos de estudos tal como sociologia, ecologia, estatística e outros. O índice é empregado em diversas análises de dados de modo a determinar a importância de uma variável, capturando a sensibilidade do modelo para cada preditor, isto é, quanto mais sensível, mais importante será a variável para o desempenho do modelo (MORAIS, 2010). O coeficiente de Gini é definido por:

$$Q_m(y_i, c) = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk}) = 1 - \sum_{k=1}^K \hat{p}_{mk}^2 \quad \text{Equação 11}$$

Sendo que $\hat{p}_{mk}$ representa a proporção da categoria k na região R_m . A medida é utilizada como estimativa da probabilidade de cada classe, sendo definida como:

$$\hat{p}_{mk} = \frac{1}{n_m} \sum_{x_i \in R_m^P} I(y_i = k) \quad \text{Equação 12}$$

onde $\hat{p}_{mk}$ assume valores entre 0 e 1

Por fim, vale descrever as vantagens e desvantagens das Florestas Aleatórias. Diferentemente de outros, este é de fácil implementação e eficiente em ressaltar a relevância dada para cada característica em sua entrada. Vale dizer também que não são necessários vários parâmetros para serem ajustados até mesmo quando utilizados valores *default* já que produzem bons resultados. Por outro lado, uma das desvantagens desse método é o sobreajuste (*overfitting*) em casos em que não existam árvores suficientes na floresta e essas tenham sido escolhidas aleatoriamente, porém com

variáveis não ideias, nesse sentido, o classificador sobre ajusta o modelo (DONGES apud SILVA, 2018).

Florestas Aleatórias apresentam um classificador semelhante ao *bagging* (combinação de *bootstrap* e agregação) em razão de considerar diferentes amostras *bootstrap* para cada árvore de classificação. O termo *bootstrap* se refere a uma estimativa de distribuição da amostra de interesse de modo que se obtenha um modelo de florestas aleatórias de baixo viés de cada classificador individual à medida que reduz a variância (MORAIS, 2010).

A partir de um conjunto de amostras e B variáveis aleatórias distribuídas com variância (σ^2) e correlação ρ para cada par, expressa-se a variância média descrita na Equação 13.

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \quad \text{Equação 13}$$

As florestas aleatórias promovem mais árvores com parâmetros que as tornam mais independentes entre si de modo a reduzir a correlação entre elas o que promove o aumento da redução da variância. Tal processo passa a levar em consideração o número de características que são utilizadas nas iterações, isto é, cria um novo parâmetro a ser considerado (MORAIS, 2010).

A probabilidade de predição (p) é dada pela Equação 14 abaixo:

$$p = \frac{1}{B} \sum_{j=1}^B p_i \quad \text{Equação 14}$$

Ademais, à medida que o número de árvores aumenta, B tende ao infinito, de modo que o erro converge para a Equação 15 (Breiman, 2001 apud Morais, 2010).

$$P_{X,Y}(M(X,Y) \leq 0) \quad \text{Equação 15}$$

Dado a Equação 16

$$m(x, y) = P_{\theta}(f(X, \theta) = Y) - \max P_{\theta}(f(X, \theta) = k) \quad \text{Equação 16}$$

A medida que expressa quanto a classificação correta excede a média do número de votos de qualquer outra classificação é dada pelo termo $m(X, Y)$. O resultado da equação demonstra que o aumento do número de árvores corrobora com a convergência do erro de generalização para um limite (MORAIS, 2010).

Na aplicação do Random Forest em um conjunto de dados, parte não são utilizados em um primeiro momento. Tal conjunto de dados é definido por *Out-Of-Bag* (OOB). Tendo em vista que os parâmetros das árvores são criados a partir de parte do conjunto de dados, o excedente é utilizado para atestar a precisão do modelo a fim de verificar se o modelo é eficiente em sua aplicação (ICHIPRO, 2018).

O conjunto de dados OOB além de fornecer uma estimativa de erro são usados também para verificar a relevância das variáveis do modelo e proximidade de observações. A importância de cada variável para o modelo é dada pelo índice de Gini que determina a sensibilidade do modelo para cada preditor, ou seja, quanto mais sensível for o modelo para determinada variável, mais determinante essa será para o desempenho das árvores aleatórias (MORAIS, 2010).

As proximidades são definidas por uma matriz de dimensão $n \times n$. A medida de proximidade considera a importância de cada variável bem como a similaridade entre elas. Tal medida é determinada pela seguinte forma: se a proximidade fica em torno de zero, então as duas variáveis não foram classificadas nenhuma vez na mesma região, e 1 quando as variáveis foram classificadas sempre na mesma região (BREIMAN, 2003).

A medida de discrepância é descrita pela Equação 17

:

$$D_{is}^0 = \frac{n_s}{\sum_{ij}^n I(y = s) \times \text{prox}(x_i, x_j)^2} \quad \text{Equação 17}$$

sendo $i = 1, 2, 3, \dots, n$ e n_s o número de observações referentes a categoria s e $\text{prox}(x_i, x_j)^2$.

De acordo com a Equação 18, é definida a discrepância padronizada:

$$D_{is} = \frac{D_{is}^0 - m_s}{MAD * (D_{is}^0)} \quad \text{Equação 18}$$

onde $i = 1, 2, \dots, n$ e m_s é dado pela mediana de D_{is}^0 e $MAD * (D_{is}^0)$ é o desvio absoluto mediano multiplicado pela constante 1,4826 de modo que seja um estimador consistente da variância. Recomenda-se que valores acima de 10 de discrepância sejam considerados como discrepantes, porém para alguns casos o valor de corte pode variar (Breiman, 2003 apud Moraes, 2010).

3. RESULTADOS E DISCUSSÕES

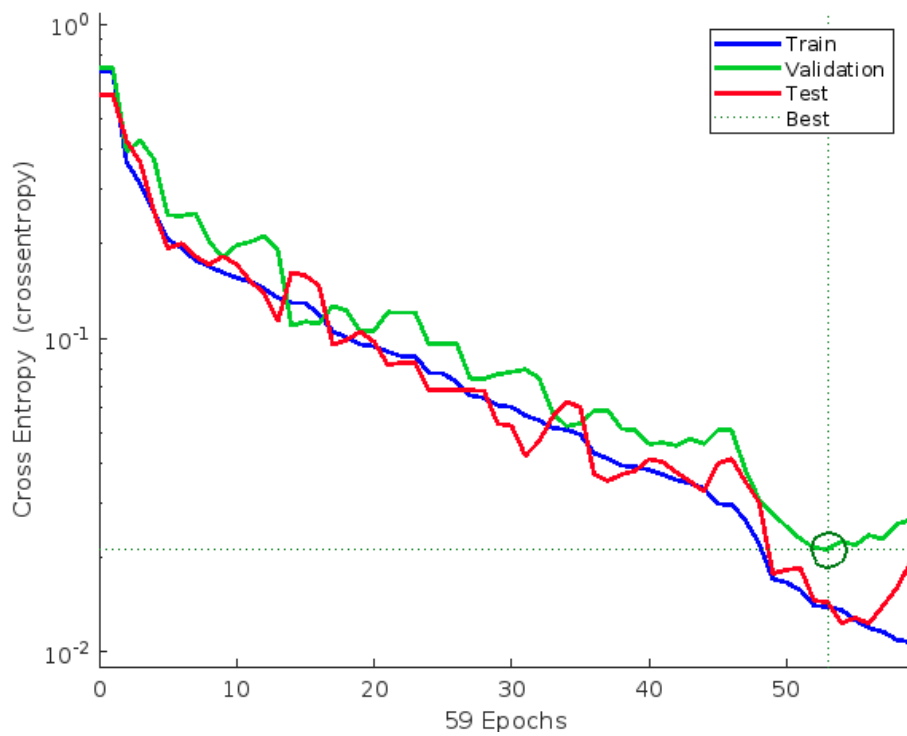
Neste capítulo serão apresentados os resultados obtidos através dos modelos neurais e florestas aleatórias.

3.1. RESULTADOS DA RNA

A base de dados utilizada para realizar o primeiro teste também foi a mesma utilizada no estudo com árvores aleatórias. O estudo de redes neurais contou com 152 amostras para treinamento, 32 para validação e 32 para teste, ou seja, divisão do tipo 70% - 15% - 15% sendo que a RNA que obteve o melhor resultado contou com 10 neurônios na camada intermediária.

A partir dos estudos realizados, obteve-se o gráfico abaixo (Figura 12), onde é possível depreender que, das 59 iterações realizadas, a partir da iteração 53 o erro apresentou aumento, como é observável na figura. Nesse sentido, a melhor performance ocorreu na 53ª iteração com erro de entropia cruzada na validação das amostras.

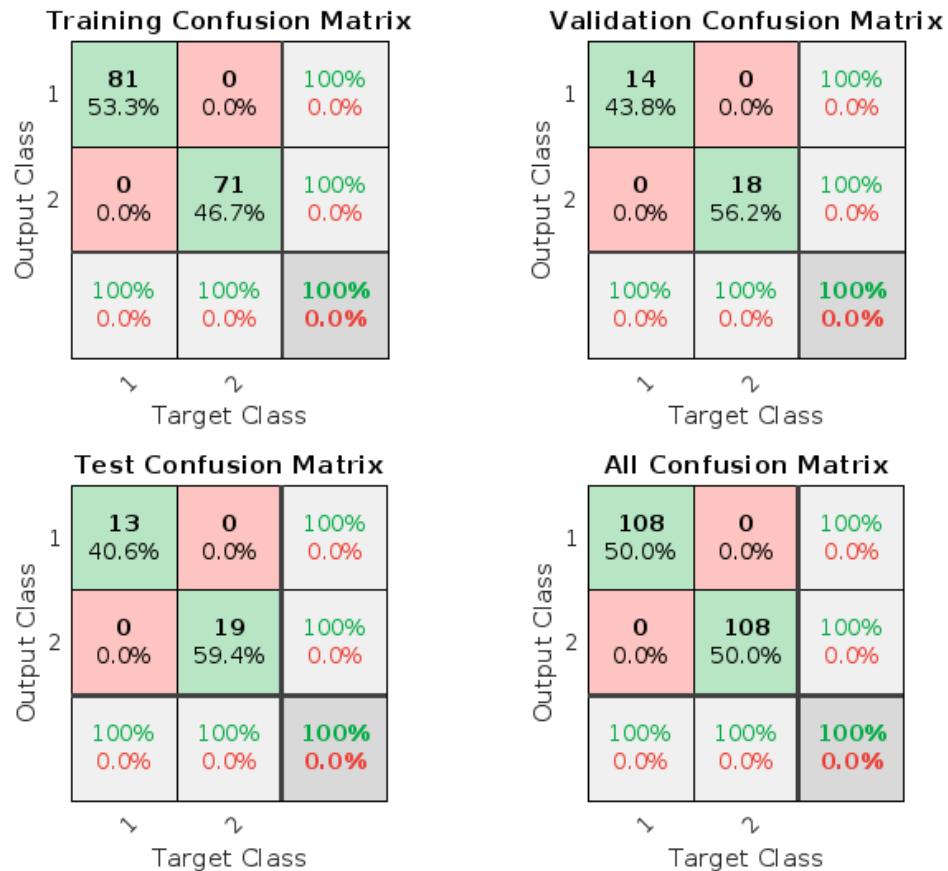
Figura 10 - Performance pelo número de épocas



Fonte: O autor, 2022

Através da matriz de confusão abaixo (Figura 13), é possível depreender que, como observado acima, todas as amostras foram classificadas corretamente. A matriz de confusão exibe para todos os conjuntos de treinamento, validação e teste, além de todas as amostras juntas.

Figura 11 - Matriz de confusão para os conjuntos de treinamento, validação e teste

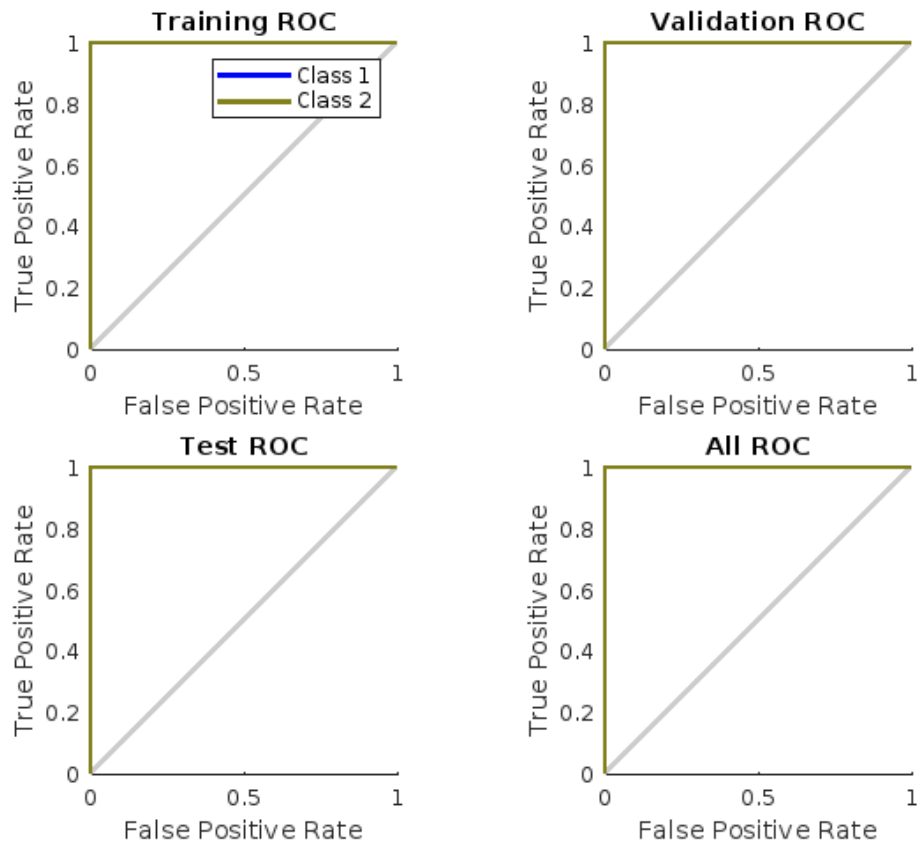


Fonte: O autor, 2022

Observa-se que, no treinamento, as 81 amostras de ácido palmítico foram classificadas corretamente e as outras 71 de ácido oleico também, o mesmo ocorreu para as matrizes de validação e de teste. Nesse sentido, 100% das amostras foram classificadas corretamente e o erro obtido foi nulo.

Além disso, foi obtido o gráfico da curva característica do receptor (ROC) para ambos os ácidos. Na figura abaixo (Figura 14), depreende-se que, a RNA apresentou ótimo desempenho na classificação dos ácidos devido ao fato de que os valores da área abaixo da curva serem iguais a 1 para os três conjuntos: treinamento, validação e teste.

Figura 12 - Curva característica de operação do receptor (ROC) para ambos os ácidos



3.2.RESULTADOS DA FLORESTA ALEATÓRIA

Os parâmetros utilizados no modelo de florestas aleatórias foram: número de árvores que vão compor a floresta aleatória, quantidade de variáveis aleatórias, *bootstrapping* onde a reamostragem é feita com reposição, número mínimo e máximo de observação de nós. Os parâmetros utilizados estão representando na Tabela 1 abaixo.

Tabela 1 - Parâmetros utilizados no modelo de Florestas Aleatórias

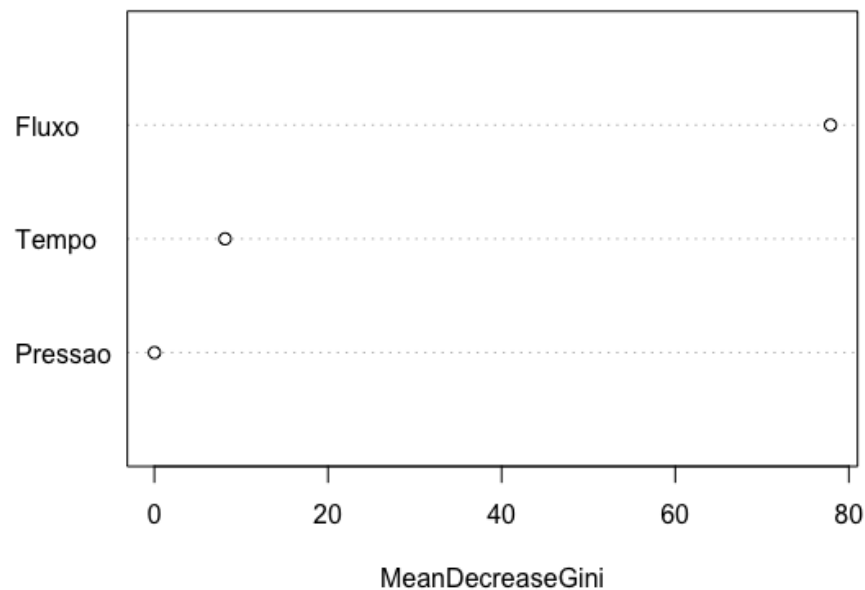
Parâmetros	Valores
Número de árvores	400
Número de variáveis	3

Número mínimo nós	5
Número máximo nós	10
Bootstrapping	Sim

Fonte: O autor, 2022

A partir da implementação do modelo de árvores aleatórias, aplicou-se o coeficiente de Gini a fim de determinar quais são as variáveis mais relevantes no modelo. Quanto maior o valor do Gini, maior a importância da variável no modelo. Sendo assim, é possível observar a partir da imagem abaixo (Figura 10), que o fluxo e o tempo são os mais importantes, por outro lado, a pressão não apresenta importância relevante para o modelo.

Figura 13 - Coeficiente de Gini



Fonte: O autor, 2022

Além disso, a simulação de Florestas aleatórias parametrizada com 400 árvores e 3 variáveis aleatórias gerou a matriz de confusão ilustrada abaixo (Tabela 2). A proporção utilizada foi do tipo 80% - 20% onde 42 foram utilizadas para teste e 174 para treinamento. A partir da qual, depreende-se que, o modelo apresentou ótima capacidade de classificação em decorrência dos poucos falsos positivos e falsos negativos, o erro foi

de aproximadamente 1.7%. Dentre as 174 amostras, 171 foram classificadas corretamente, sendo apenas 1 falso positivo e 2 falsos negativos.

Tabela 2 - Matriz de confusão obtida pelo modelo para o período de treinamento

Respostas	Sim	Não
Sim	85	2
Não	1	86

Fonte: O autor, 2022

A matriz de confusão obtida para o período de teste com os mesmos parâmetros da simulação anterior é representada na Tabela 3 abaixo. Observa-se que as 42 amostras foram classificadas sem erro, sendo 21 corretamente classificadas como "sim" que representa o ácido palmítico e as outras 21 como "não" que representa o ácido oleico.

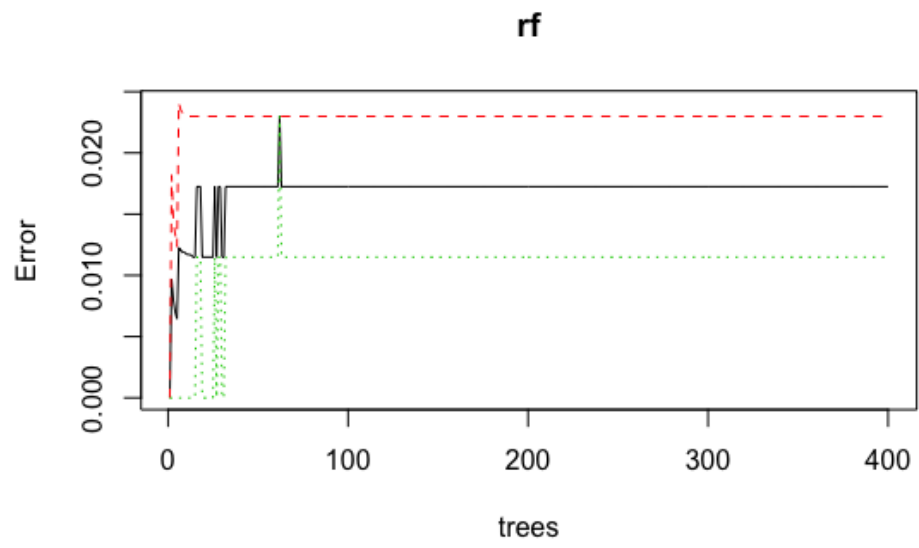
Tabela 3 - Matriz de confusão obtida pelo modelo para o período de teste

Respostas	Sim	Não
Sim	21	0
Não	0	21

Fonte: O autor, 2022

A Figura 11 abaixo representa o erro nas amostras utilizadas no período de teste pelo número de árvores. Como pode ser observado, o erro estabiliza próximo as 80 árvores utilizadas.

Figura 14 - Erro por número de árvores na simulação



4. CONCLUSÃO

Os modelos de aprendizado de máquina são de grande relevância quando aplicados a predição de dados, sendo que foram obtidos resultados satisfatórios para classificação do ácido oleico e ácido palmítico.

O resultado obtido com as redes neurais na classificação dos ácidos palmítico e oleico utilizando fluxo, tempo e pressão foi muito satisfatório tendo em vista que todas as amostras foram classificadas corretamente. Além disso, para os aprendizados realizados com Florestas Aleatórias, também foram obtidos bons resultados com erro de 1,7% no período de treinamento. Vale ressaltar que, apesar do erro, não invalida o modelo já que é um erro muito baixo quando comparado com toda a amostra. Por fim, através dos resultados (Coeficiente de Gini) obtidos com Florestas aleatórias, nota-se que a pressão pode não ter uma atuação relevante na classificação dos ácidos.

Como sugestão de trabalhos futuros, sugere-se implementar tanto as Redes Neurais Artificiais quanto as Florestas Aleatórias como ferramentas preditivas da composição do açaí quanto de outros tipos de óleos vegetais. Sugere-se também fazer uma análise parecida ao Coeficiente de Gini para as Redes Neurais Artificiais, a fim de obter modelos ainda mais robustos na predição de dados.

5. REFERÊNCIAS

SCHNEIDER, René Peter; TSUTIYA, Milton Tomoyuki. Membranas filtrantes para o tratamento de água, esgoto e água de reúso. [S.l: s.n.], 2001.

Redes Neurais Artificiais. ICM USP, São Paulo. Disponível em: <<https://sites.icmc.usp.br/andre/research/neural/>>. Acesso em: 13 de jan. de 2022.

GAMA, J. Notas de aula. USP, 2004.

NETO, C. Di G. Potencial de técnicas de mineração de dados para o mapeamento de áreas cafeeiras. INPE, São José dos Campos, 2014.

CARUANA, R.; KARAMPATZIAKIS, N.; YESSENALINA, A. An empirical evaluation of supervised learning in high dimensions. In: International conference on Machine learning, 25, Helsinki, Proceedings... Helsinki: ACM, p.96-103, 2008.

BREIMAN, L. Random forests. Machine Learning Journal. Hingham, v.45, p.5–32, jan. 2001.

BELLE, V. Detection and recognition of human faces using random forest for a mobile root. 104p. Dissertação (Mestrado em Ciência da Computação) – Rwthachen University, Alemanha. 2008.

PRONI, C.; HANEDA, R. N.; FILLETTI, E. R. Inteligência artificial aplicada ao processo de separação de misturas por membranas cerâmicas. In: ENCONTRO REGIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL, 6., 2019, Bauru. Caderno de trabalhos completos e resumos [...]. Bauru: Unesp, Faculdade de Ciências, 2019. p. 462-463, 2019. Disponível em: <https://www.fc.unesp.br/#!/departamentos/matematica/eventos2341/ermac2019/caderno-de-trabalhos-e-resumos/>. Acesso em: 13 jan. 2022

Proni, Camila & Haneda, Renata & Filletti, Érica. (2020). Desenvolvimento de redes neurais artificiais para análise do fluxo de permeado de uma bebida à base de açaí no processo de microfiltração tangencial. C.Q.D.- Revista Eletrônica Paulista de Matemática. 17. 189-205. 10.21167/cqdv17ermac202023169664cprnherf189205.

Acesso em: 13 jan. 2022

FILLETTI, E. R. Desenvolvimento de modelos neurais para o processamento de sinais acústicos visando a medição de propriedades topológicas em escoamentos multifásicos. Universidade de São Paulo, São Carlos, 2007.

ROGEZ, J. (200). Açaí: Preparo, Composição e Melhoramento da Conservação. Belém: EdUFPA.

NASCIMENTO, R.J.S. et. al. (2008). Composição em ácidos Graxos do óleo da Polpa de açaí extraído com enzimas e com hexano 1, Revista Brasileira de Fruticultura, Jaboticabal, vol. 30m n.2, p. 498-502, jun.

CHERYAN, M.(1998). Ultrafiltration and Microfiltration. Handbook, Technomic Publishing Company, Lancaster, 527 p.

CHERYAN, M. (1998). Ultrafiltration and Microfiltration Handbook. 1 ed. Lancaster: Technomic Publishing Company, 1998. 527 páginas.

SCHNEIDER, R. P.; TSUTYIA, M. T. Membranas Filtrantes para o tratamento de água, esgoto e água de reúso. São Paulo: ABES, 2001. 223p.

GALVÃO, D. F.; GOMES, E.R.S,. Os processos de separação por membranas e sua utilização no tratamento de efluentes industriais da indústria de laticínios: Revisão Bibliográfica, Paraná, 2016.

BHAVE, I.; RAMESH, R. Inorganic membranes: synthesis, characteristics and applications. New York: Chapman & Hall, 129, p. 83-84, 1991.

ALMEIDA, Leandro Maciel. Uma metodologia de busca por redes neurais artificiais quase-ótimas. 2007. 114 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Centro de Informática, Universidade Federal de Pernambuco, Recife, 2007.

CAMINOTO, Karime Bárbara Santo. Estudo do declínio do fluxo transmembrana via microfiltração tangencial de misturas bifásicas de óleos vegetais e água. Universidade de São Paulo, São Carlos, 2012.

Sazli, Murat (2006). A brief review of feed-forward neural networks. Communications, Faculty of Science, University of Ankara.

ARMOA, M. H.; JAFELICCI JUNIOR, M. PRINCÍPIOS E APLICAÇÕES DE PROCESSOS DE SEPARAÇÃO POR MEMBRANAS INORGÂNICAS. Ciência & Tecnologia, [S. l.], v. 2, n. 1, 2011. Disponível em: <https://citec.fatecjab.edu.br/index.php/citec/article/view/67>. Acesso em: 15 jan. 2023.

TREVIZOLI, A. M. S., (2010). Estudo Experimental da Microfiltração Tangencial Aplicado a Clarificação da Vinhaça. Dissertação (Mestrado em Engenharia Mecânica) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 138p.

PELEGRIN, D. C., (2004). Microfiltração Tangencial de Efluente Sanitário após Tratamento Biológico. Dissertação (Mestrado em Engenharia Sanitária e Ambiental), Departamento de Engenharia Sanitária e Ambiental – Universidade de Santa Catarina, Florianópolis, 115p.