UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

BARRY MALICK BARQUE

PREDIÇÃO DE MICROBIOMA SAUDÁVEL PRESENTE NA CULTURA DE ARROZ UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA SUPERVISIONADO

MEDIANEIRA

2024

BARRY MALICK BARQUE

PREDIÇÃO DE MICROBIOMA SAUDÁVEL PRESENTE NA CULTURA DE ARROZ UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA SUPERVISIONADO

Prediction of healthy microbiome present in rice culture using supervised machine learning techniques.

Dissertação apresentado como requisito para obtenção do título de Mestre em Tecnologias Computacionais Aplicadas à Produção Agrícola e Agroindústria do Programa de Pós-Graduação em Tecnologias Computacionais Aplicadas à Produção Agrícola e Agroindústria da Universidade Tecnológica Federal do Paraná.

Orientador: Profa. Dra. Deborah Catharine de Assis Leite.

Coorientador: Prof. Dr. Pedro Luiz de Paula Filho.

MEDIANEIRA

2024

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Medianeira

BARRY MALICK BARQUE

PREDIÇÃO DE MICROBIOMA SAUDÁVEL PRESENTE NA CULTURA DE ARROZ UTILIZANDO TÉCNICAS APRENDIZADO DE MÁQUINA SUPERVISIONADO

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Tecnologias Computacionais Para O Agronegócio da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Tecnologias Computacionais Aplicadas À Produção Agrícola E Agroindústria.

Data de aprovação: 29 de Maio de 2024

Deborah Catharine De Assis Leite, - Universidade Tecnológica Federal do Paraná

Dr. Arnaldo Candido Junior, Doutorado - Universidade Estadual Paulista - Unesp

Dra. Denise Da Piedade Silva, Doutorado - Oregon State University

Dr. Oldair Donizeti Leite, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Pedro Luiz De Paula Filho, Doutorado - Universidade Tecnológica Federal do Paraná

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 29/05/2024.

Dedico este trabalho a minha família e aos
meus amigos, pelos momentos de ausência.

AGRADECIMENTOS

Este trabalho não poderia ser terminado sem a ajuda de diversas pessoas e/ou instituições às quais presto minha homenagem. Certamente esses parágrafos não irão atender a todas as pessoas que fizeram parte dessa importante fase de minha vida. Portanto, desde já peço desculpas àquelas que não estão presentes entre estas palavras, mas elas podem estar certas que fazem parte do meu pensamento e de minha gratidão.

A minha família, pelo carinho, incentivo e total apoio em todos os momentos da minha vida.

Ao meu orientador, que me mostrou os caminhos a serem seguidos e pela confiança depositada.

A todos os professores e colegas do departamento, que ajudaram de forma direta e indireta na conclusão deste trabalho.

Enfim, a todos os que de alguma forma contribuíram para a realização deste trabalho.

Primeira Lei: Um robô não pode ferir um ser humano ou, por omissão, permitir que um ser humano sofra algum mal. Segunda Lei: Um robô deve obedecer as ordens que lhe sejam dadas por seres humanos, exceto nos casos em que tais ordens contrariem a Primeira Lei. Terceira Lei: Um robô deve proteger sua própria existência desde que tal proteção não entre em conflito com a Primeira e Segunda Leis (ASIMOV, Isaac, 1950).

RESUMO

Os seres vivos, incluindo as plantas, hospedam uma grande variedade de micro-organismos que podem ter efeitos benéficos ou prejudiciais para o seu desenvolvimento. A identificação e classificação dos conjuntos de micro-organismos que favorecem ou prejudicam as culturas é fundamental para o desenvolvimento da agricultura. Este estudo explorou a influência dos micro-organismos nas plantações de arroz, cruciais para o desenvolvimento das culturas. Empregando redes neurais e algoritmos de inteligência artificial, classificamos a saúde de 110 amostras de arroz, divididas entre 63 doentes e 47 saudáveis. Foi desenvolvida e utilizada a inovadora Micronet e comparou-se com redes já conhecidas como: a rede convolucional MDeep, Árvore de Decisão, Random Forest, MLP e SVM como classificadores. Os resultados revelaram uma notável capacidade de identificar conjuntos de micro-organismos associados à saúde e doença, aprimorada pela utilização dos *SHAP values*. A curva ROC destacou o desempenho superior da Micronet, com uma AUC de 94%. O MDeep seguiu com 91%, enquanto Random Forest e SVM atingiram 88% de AUC. A Árvore de Decisão demonstrou um desempenho sólido, registrando 83% de AUC. Esses achados sugerem que a Micronet, especialmente com o suporte dos *SHAP values*, é um classificador robusto para discernir a saúde das plantações de arroz, apresentando resultados promissores no contexto agrícola. Essa abordagem diversificada com múltiplos classificadores destacou vários microrganismos associado a saúde da planta.

Palavras-chave: redes neurais; microbioma do arroz; inteligencia artificia; *dickeya zeae*; métricas.

ABSTRACT

Living organisms, including plants, host a wide variety of microorganisms that can have beneficial or detrimental effects on their development. The identification and classification of sets of microorganisms that favor or harm crops are crucial for agricultural development. This study delved into the influence of microorganisms on rice plantations, essential for crop development. Employing neural networks and artificial intelligence algorithms, we classified the health of 110 rice samples, comprising 63 diseased and 47 healthy ones. We utilized the innovative Micronet, the convolutional network MDeep, Decision Tree, Random Forest, MLP, and SVM as classifiers. The results revealed a remarkable ability to identify sets of microorganisms associated with health and disease, enhanced by the use of SHAP values. The ROC curve highlighted the superior performance of Micronet, with an impressive AUC of 94%. MDeep followed closely with 91%, while Random Forest and SVM achieved 88% AUC. The Decision Tree demonstrated solid performance, recording an 83% AUC. These findings suggest that Micronet, especially with the support of SHAP values, is a robust classifier for discerning the health of rice plantations, presenting promising results in the agricultural context. This diversified approach with multiple classifiers highlighted several microorganisms associated with plant health.

Keywords: neural networks; rice microbiome; artificial intelligence; *dickeya zeae*; metrics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Siglas

ANNs	Redes Neurais Artificias
CNNs	Redes Neurais Convolucionais
NA	Neurônio Artificial
OTUs	Unidades Taxonômicas Operacionais
SVMs	Máquinas de Vetores de Suportes

SUMÁRIO

1 INTRODUÇÃO

A agricultura é uma atividade econômica fundamental que tem por objetivo atender às necessidades alimentares da população em muitos países. A composição do microbioma presente nas raízes (rizosfera), folhas (filosfera) e sistema de transporte interno (endosfera) das culturas é um dos fatores importantes na saúde e no desenvolvimento das plantas cultivadas em diversos biomas (MONTESINOS, 2003). As plantas são organismos multicelulares que interagem com o ambiente de maneira complexa e dinâmica. Além disso, elas têm a capacidade de hospedar uma variedade de microrganismos, como bactérias, fungos e vírus, que compõem o seu microbioma (CHANDRAN; MEENA; SWAPNIL, 2021). Essa interação entre as plantas e seu microbioma pode ter efeitos significativos na saúde, crescimento e produção de plantas (STEWART *et al.*, 2021; CHANDRAN; MEENA; SWAPNIL, 2021).

O microbioma das plantas pode desempenhar diversos papéis, como a proteção contra patógenos, o aumento da absorção de nutrientes, a melhora na tolerância ao estresse abiótico e a promoção do crescimento. Além disso, a composição e diversidade do microbioma das plantas pode ser influenciada por fatores como a genética da planta, o solo em que a planta cresce, a disponibilidade de nutrientes e a presença de patógenos. Em outras palavras, o patobioma é um conjunto de micro-organismos que torna o hospedeiro mais vulnerável a doenças e pragas, enquanto um core microbioma é um grupo de micro-organismos que favorece a saúde e a resistência de seus hospedeiros a certos tipos de doenças e pragas (STEWART *et al.*, 2021; SHADE; STOPNISEK, 2019).

Graças ao método de sequenciamento genético, é possível identificar uma quantidade significativa de microrganismos presentes nas plantas, incluindo aqueles que não seriam detectáveis por meio de técnicas tradicionais de cultivo, proporcionando assim benefícios para a agricultura. A implementação de técnicas avançadas como aprendizado de máquina, redes neurais e aprendizado profundo está impulsionando significativamente o progresso em diversas áreas, incluindo a agricultura. Essas abordagens capacitam os sistemas a extrair padrões complexos a partir de conjuntos massivos de dados, possibilitando, por exemplo, a previsão precisa de doenças nas plantas com base em informações macroecológicas (YUAN *et al.*, 2020).

O aprendizado de máquina proporciona flexibilidade para lidar com uma variedade de tarefas, enquanto as redes neurais, inspiradas na estrutura do neurônio biológico, destacam-se em reconhecimento de padrões. Por sua vez, o aprendizado profundo, com suas redes neurais profundas, revela-se particularmente eficaz na análise de dados intricados, como aqueles relacionados a condições climáticas, solo e características das plantas, potencializando avanços significativos na otimização da agricultura e na prevenção de doenças (WILHELM; ES; BUCKLEY, 2022).

Atualmente, graças às facilidades de sequenciamento genético, os estudos nesta área buscam identificar quais microrganismos estão mais associados a doenças e aqueles que favorecem a saúde das culturas.

1.1 Objetivo geral e específicos

Este trabalho tem como objetivo geral utilizar técnicas de aprendizado de máquinas como a árvore de decisão, o *Random Forest*, a SVM, a MLP, a *MDeep* e uma rede neural desenvolvida da Micronet para identificar os micro-organismos que constituem um microbioma saudável e doente de uma determinada cultura.

Este trabalho é composto pelo seguintes objetivos específicos:

- Desenvolver uma rede neural para classificar as amostras e identificar o microbioma saudável e doente;
- Avaliar a rede neural desenvolvida;
- Comparar a rede desenvolvida com outras redes existentes;

2 REVISÃO DA LITERATURA

Neste seção serão definidos conceitos sobre o *core* microbiano das plantas, técnicas de inteligência artificial, redes neurais artificiais, técnicas de classificação utilizando aprendizado de máquina, bem como trabalhos correlatos na área.

2.1 Microbioma associado às plantas

O microbioma das plantas é um conjunto de micro-organismos que residem em uma planta, como: bacteriais, vírus, fungos. Esses micro-organismos podem estar presentes na região das raízes (ambientes rizosfera), na região das folhas (ambientes filosfera) ou ainda na região interna, nos tecidos (ambientes endosferas) (MONTESINOS, 2003). Eles são considerados elementos chaves para manutenção das saúde do hospedeiro, Entre elas pode-se destacar os micro-organismos que ajudam na manutenção da saúde das plantas chamado de *core* microbiano, e os micro-organismos causadoras de doenças chamado de patobioma (MANNAA; SEO, 2021; SHADE; STOPNISEK, 2019).

Estudos recentes mostram que o *core* microbiano das plantas pode desempenhar um papel importante na nutrição e proteção dos vegetais contra doenças e pragas. Por exemplo, algumas bactérias do gênero *Rhizobium* são capazes de fixar nitrogênio atmosférico e fornecer nutrientes essenciais para as plantas (MONTESINOS, 2003). Além disso, outros microrganismos podem estimular o crescimento das raízes e aumentar a absorção de nutrientes pelos vegetais (SHADE; STOPNISEK, 2019).

Por outro lado, certas bactérias e fungos patogênicos podem causar doenças nas plantas e afetar negativamente o seu crescimento. A compreensão da composição e função do *core* microbiano dos vegetais é, portanto, essencial para a promoção do crescimento saudável das plantas e a proteção contra doenças (STEWART *et al.*, 2021).

O simbioma é um bioma de relação simbiótica entre duas especies. A simbiose é um tipo de relação mutualística entre dois seres vivos, no qual ambos se beneficiam (SHADE; STOPNISEK, 2019) .

No caso das plantas, o simbioma pode incluir bactérias e fungos que formam simbiose com as raízes dos vegetais, ajudando-as a absorver nutrientes do solo; bactérias que fixam o nitrogênio do ar, ajudando ao vegetal a obter esse nutriente, e insetos que polinizam as flores da planta, ajudando-a a produzir sementes e frutos (MANNAA; SEO, 2021; MONTESINOS, 2003).

O simbioma de um vegetal pode ser influenciado por vários fatores, como o tipo de planta, o ambiente em que ela cresce e as práticas de cultivo. Por exemplo, um vegetal que cresce em um solo rico em nutrientes pode ter um simbioma diferente de uma planta que cresce em um solo pobre em nutrientes (STEWART *et al.*, 2021; MONTESINOS, 2003).

O conhecimento do simbioma de uma planta é importante, porque pode ajudar a entender como a planta se relaciona com seu ambiente e como ela se beneficia da relação simbiótica

com outros organismos principalmente dos micro-organismos relacionados. Isso pode ser usado para desenvolver estratégias para aumentar a produção de alimentos e melhorar a qualidade dos alimentos produzidos (MANNAA; SEO, 2021; STEWART *et al.*, 2021).

O patobioma é um termo utilizado para descrever um conjunto de micro-organismos que podem desfavorecer seu hospedeiro, lhe privando das suas funções vitais em determinadas condições ambientais instáveis. De uma forma geral, uma planta saudável está associada a uma comunidade de micro-organismos estáveis e diversificadas (MANNAA; SEO, 2021).

O patobioma de uma planta pode incluir patógenos que causam doenças comuns, como a ferrugem e a mancha das folhas, e também patógenos raros que só afetam determinadas plantas ou regiões (JAIN *et al.*, 2019). Ele pode ser influenciado por vários fatores, como o clima, o solo e as práticas de cultivo. Por exemplo, em uma região quente e úmida, o patobioma de uma planta pode incluir patógenos que causam doenças fúngicas, enquanto em uma região seca e quente, o patobioma pode incluir patógenos que causam doenças virais (STEWART *et al.*, 2021).

O conhecimento do patobioma de uma planta é importante porque pode ajudar a prevenir e controlar doenças nas plantas. Isso pode aumentar a produção de alimentos e melhorar a qualidade dos alimentos produzidos (MANNAA; SEO, 2021; STEWART *et al.*, 2021).

2.2 Métodos de classificação de sequencias microbianas

2.2.1 Unidades Taxonômicas Operacionais (OTUs)

As Unidades Taxonômicas Operacionais (OTUs) são ferramentas importantes em estudos de biodiversidade, permitindo a análise de diversidade de espécies em ambientes complexos e diversificados (JESKE; GALLERT, 2022). Essas unidades são usadas para agrupar sequências de DNA ou RNA com base em sua similaridade genética, representando uma espécie ou grupo de espécies intimamente relacionadas (SCHLOSS, 2021).

A definição de uma OTU pode variar dependendo do nível de similaridade genética considerado. Em geral, um limite de 97% a 99% de identidade de nucleotídeos é usado para agrupar sequências em uma OTU. Essa abordagem permite que sequências com diferenças genéticas significativas sejam agrupadas em OTUs diferentes, enquanto sequências com alta similaridade genética são agrupadas em uma mesma OTU (QUAIL, 2021).

As OTUs desempenham um papel crucial em estudos de diversidade microbiana, particularmente em contextos em que a alta diversidade de espécies e a presença de muitas espécies não descritas representam desafios significativos. A utilização de OTUs possibilita a análise da diversidade de espécies em amostras microbianas complexas, permitindo a identificação de padrões de distribuição espacial e temporal das comunidades microbianas (JESKE; GALLERT, 2022; SCHLOSS, 2021). Além disso, as OTUs encontram aplicação em estudos de

ecologia molecular, como análises de metabarcoding, viabilizando a identificação simultânea de múltiplas espécies com base em sequências de DNA ou RNA. Essa abordagem possibilita a investigação das relações entre espécies e seus papéis em ecossistemas específicos.

A utilidade das OTUs se estende à estimativa da riqueza de espécies e à comparação da diversidade de espécies em diferentes amostras ou condições ambientais. Tais análises fornecem visões valiosos sobre a saúde e estabilidade dos ecossistemas, orientando estratégias de conservação e manejo eficazes (JESKE; GALLERT, 2022; QUAIL, 2021). Em síntese, as OTUs constituem ferramentas essenciais em estudos de biodiversidade, oferecendo a capacidade de analisar a diversidade de espécies em amostras complexas e heterogêneas. Seu emprego é fundamental para a compreensão das intricadas interações entre as espécies e seus papéis nos ecossistemas, além de direcionar iniciativas de conservação e manejo (QUAIL, 2021). Vale ressaltar que existe uma abordagem mais recente, denominada Sequência Variante Amplicônica (ASV), que representa uma alternativa à abordagem tradicional baseada em OTUs (JESKE; GALLERT, 2022).

2.2.2 Sequência de Variante de Amplicon (ASV)

A abordagem Sequência de Variante de Amplicon[1] (ASV) tem emergido como uma ferramenta notável na análise de dados provenientes do sequenciamento de amplicons em microbiologia. Sua capacidade de identificar e caracterizar a diversidade microbiana em amostras ambientais ou clínicas é surpreendentes, proporcionando alta resolução taxonômica e precisão analítica (SCHLOSS, 2021).

Diferentemente das técnicas convencionais de análise de dados de sequenciamento, como as Unidades Taxonômicas Operacionais (OTUs), que agrupam sequências com base em similaridades arbitrárias pré-definidas, a técnica ASV atribui uma identificação única a cada sequência individual. Essa abordagem resulta em uma resolução taxonômica mais refinada, permitindo a identificação e quantificação de diferenças sutis entre as sequências (JESKE; GALLERT, 2022).

A aplicação da técnica ASV destaca-se em estudos de comunidades microbianas complexas, como as encontradas em ambientes ou amostras clínicas. Sua eficácia reside na capacidade de identificar espécies raras ou pouco abundantes, geralmente negligenciadas por outras técnicas como a das culturas, e em discernir variações subtis na composição da comunidade entre diferentes amostras (JOOS *et al.*, 2020).

Além disso, a técnica ASV demonstra alta replicabilidade e é passível de ser utilizada na comparação de dados entre diferentes estudos ou conjuntos de dados. Essa característica é crucial para a construção de bancos de dados de sequências microbianas de referência, con-

[1] Um amplicon é um pedaço de DNA ou RNA que é a fonte e/ou produto de eventos de amplificação ou replicação

tribuindo significativamente para a identificação de novas sequências desconhecidas (JESKE; GALLERT, 2022).

Em resumo, a técnica ASV tem se mostrado uma ferramenta poderosa na análise de dados de sequenciamento de amplicons em microbiologia, permitindo uma identificação precisa e quantificação da diversidade microbiana em amostras ambientais ou clínicas. Ela tem o potencial de avançar no entendimento sobre as comunidades microbianas e sua relação com a saúde e o meio ambiente.

2.3 Redes Neurais Artificiais

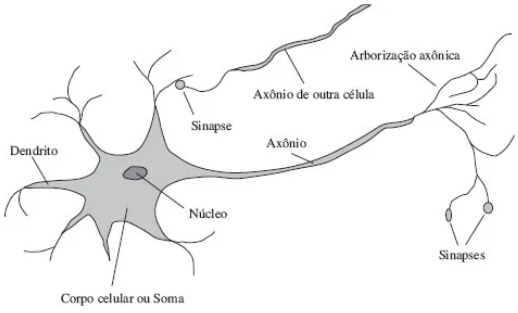
As Redes Neurais Artificias (ANNs) foram criadas com base no funcionamento do cérebro humano, que é composto principalmente de neurônios excitáveis (AMTHOR, 2016).

2.3.1 Neurônio biológico e Neurônio Artificial

O cérebro humano é estimado possuir cerca de 10^{11} (10 bilhões) de neurônios (GURNEY, 1997; RUSSELL; NORVIG, 2010). Assim como qualquer máquina complexa, o cérebro é composto por vários tipos de neurônios que se comunicam entre si para realizar tarefas complexas (AMTHOR, 2016).

Os neurônios do cérebro podem ser classificados em vários tipos, como neurônios sensoriais que informam ao restante do cérebro sobre o estado do ambiente externo e interno do corpo, neurônios de comunicação que transmitem sinais entre diferentes regiões do cérebro, e neurônios motores que controlam o comportamento dos músculos e alguns órgãos (AMTHOR, 2016). O neurônio é uma célula nervosa composta por dendritos, que recebem estímulos, um axônio que é a saída do neurônio, sinapses inibidoras e excitatórias que o conecta com outros neurônios, e o corpo celular que contém o núcleo da célula, na qual as informações recebidas pelos dendritos são processadas por meio de reações eletroquímicas, como representado na Figura 1 (AMTHOR, 2016).

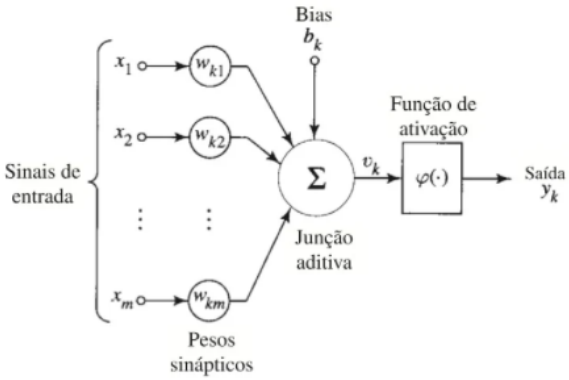
Figura 1 – Representação de um neurônio biológico



Fonte: (RUSSELL; NORVIG, 2010).

Um Neurônio Artificial (NA) tem uma estrutura funcional parecida com o neurônio biológico, como representado na Figura 2 (HAYKIN, 2007).

Figura 2 – Representação de um neurônio artificial não-linear



Fonte: (HAYKIN, 2007).

Um neurônio k é definido por um conjunto de sinais de entrada $[x_1, x_2, \ldots + x_m]$, que são ponderados pelos respectivos pesos sinápticos $[w_{k1}, w_{k2}, \ldots + w_{km}]$. Esses sinais de entrada ponderados são somados e o resultado é armazenado em (u_k), conforme a Equação 1. Esse valor é somado ao *bias* (b_k) do neurônio e o resultado é o potencial de ativação (v_k), conforme a Equação 2. Esse potencial de ativação é aplicado à função de ativação $(\varphi(.))$, que gera a saída do neurônio (y_k), conforme a Equação 3 (FURTADO, 2019).

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \tag{1}$$

$$v_k = u_k + b_k \tag{2}$$

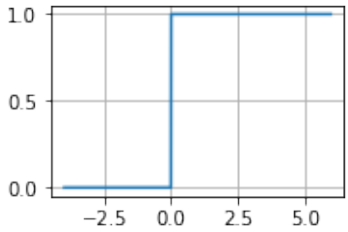
$$y_k = \varphi(v_k) \tag{3}$$

A principal tarefa da função de ativação $\varphi(.)$ é limitar a saída do neurônio para evitar o crescimento infinito (FURTADO, 2019). Existem vários tipos de funções de ativação:

- Função limiar: que é definida pela Equação 4 e é representada graficamente na Figura 3. Ela retorna 1 para qualquer valor de entrada v que seja maior ou igual a 0 e 0 para qualquer valor de entrada v que seja menor que 0 (HAYKIN, 2007) ;

$$\varphi(v) = \begin{cases} 1 \; se \; v \geq 0 \\ 0 \; se \; v < 0 \end{cases} \tag{4}$$

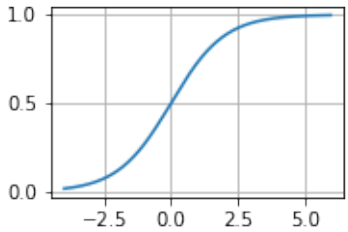
Figura 3 – Representação gráfica da função de ativação Limiar



Fonte: Autoria própria (2021).

- Função sigmoide logística, definida pela Equação 5 e representada graficamente na Figura 4. Esta função, assume valores contínuos que variam entre 0 e 1. Ela possui um parâmetro (a) responsável pela inclinação da sigmoide. Quando o parâmetro (a) se aproxima do infinito, a função sigmoide logística se torna uma função limiar (SHARMA; SHARMA; ATHAIYA, 2020);

Figura 4 – Representação linear da função de ativação Sigmoide

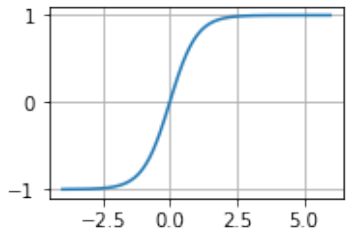


Fonte: Autoria própria (2021).

$$\varphi(v) = \frac{1}{1 + e^{-av}} \tag{5}$$

- Função tangente hiperbólica, definida pela Equação 6 e representada graficamente na Figura 5, ela varia entre -1 e 1 (WIKISTAT, 2015);

Figura 5 – Representação gráfica da função de ativação Tanh

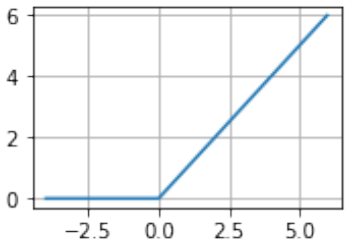


Fonte: Autoria própria (2021).

$$\varphi(v) = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}} \tag{6}$$

- *Rectified Liniear Unit* (ReLU), esta função varia entre 0 e $+\infty$ assumindo o valor de v se v for maior ou igual a 0, e 0 se v for menor que 0. Ela é definida pela Equação 7 e representada graficamente na Figura 6 (SHARMA; SHARMA; ATHAIYA, 2020);

Figura 6 – Representação gráfica da função de ativação ReLU

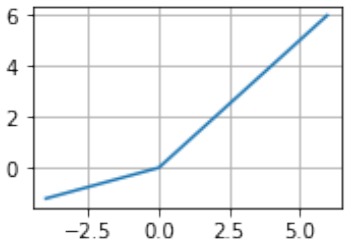


Fonte: Autoria própria (2021).

$$\varphi(v) = \begin{cases} v \; se \; v \geq 0 \\ 0 \; se \; v < 0 \end{cases} \tag{7}$$

- *Leak ReLU*, é uma versão adaptada da ReLU, varia entre $-\infty$ e $+\infty$, ela assume o valor do v se v for maior ou igual a 0 e é igual ao produto do v e do inverso de uma constante dada (a) se v for menor que 0. Ela é definida pela Equação 8 e representada graficamente pela Figura 7 (SHARMA; SHARMA; ATHAIYA, 2020).

Figura 7 – Representação gráfica da função de ativação Leak ReLU

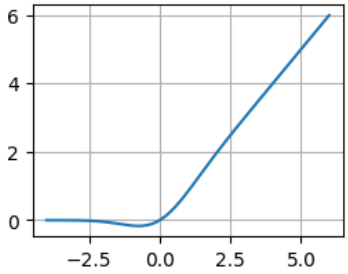


Fonte: Autoria própria (2021).

$$\varphi(v) = \begin{cases} v \; se \; v \geq 0 \\ \frac{v}{a} \; se \; v < 0 \end{cases} \tag{8}$$

- *GELU*, A função de ativação *GELU* (Unidade Linear Gaussiana de Erro) é uma função de ativação proposta para redes neurais profundas. Ela foi projetada para fornecer uma alternativa suave e diferenciável à função *ReLU*, com a vantagem de mitigar problemas de saturação que podem ocorrer com funções de ativação como a sigmoide ou a tangente hiperbólica.. Ela é definida pela Equação 9 e representada graficamente pela Figura 8 (YORK, 2023).

Figura 8 – Representação gráfica da função de ativação GELU



Fonte: Autoria própria (2021).

$$\text{Gelu}(x) = \frac{x}{2}\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}} \cdot \left(x + 0.044715x^3\right)\right)\right) \tag{9}$$

2.3.2 *Multi Layer Perceptron*

O Perceptron, projetado por Rosenblatt (1958), é a forma mais básica de classificação de padrões usando uma rede neural do tipo *feedforward* (HAYKIN, 2007). Embora o Perceptron clássico tenha bom desempenho em conjuntos de dados linearmente separáveis, como o problema dos operadores lógicos E (AND), OU (OR), $\neg E$ (NAND) e $\neg OU$ (NOR), ele se mostra ineficiente em conjuntos de dados não-linearmente separáveis, como o OU-Exclusivo (XOR). Para solucionar este problema, foram propostas as redes multi-camadas, também chamadas de *Multi Layer Perceptron* MLP, que são uma associação de pelo menos dois Perceptrons (FURTADO, 2019).

O Perceptron com n entradas, ilustrado na Figura 2, realiza a soma dos produtos das suas entradas $[x_1, x_2, \ldots, x_n]$ com os seus pesos respectivos $[w_1, w_2, \ldots, w_n]$, que são inicializados aleatoriamente, em que n é a quantidade de neurônios. É subtraído um limiar b do resultado dessa soma. O resultado obtido por esta última operação, denotado por (v), é submetido a uma função de ativação $\varphi(v)$, que é a saída (y) do neurônio. Esse processo pode ser descrito matematicamente pela Equação 10 (SILVA; SPATTI; FLAUZINO, 2010).

$$y = \begin{cases} 1 \; se \; \sum_{i=1}^{n} w_i x_i - b \geq 0 \\ -1 \; se \; \sum_{i=1}^{n} w_i x_i - b < 0 \end{cases} \tag{10}$$

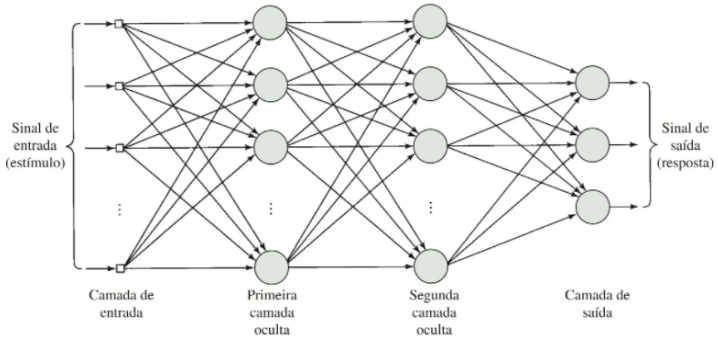
A estrutura das redes neurais MLPs representada na Figura 9, é constituída por três camadas, a camada de entrada (*input layer*) que é responsável pela geração dos sinais da rede, uma ou mais camadas ocultas (*hidden layers*) e a última camada de saída (*output layer*) (JAIN, 2016). Na última camada, a seleção da função de ativação depende da natureza do problema em questão: se é um problema de regressão, geralmente não é utilizada uma função

de ativação, enquanto em problemas de classificação, é comum empregar funções como a sigmoide para problemas binários ou a softmax para problemas multiclasse.(WIKISTAT, 2015).

2.3.3 Treinamento de uma *Multi Layer Perceptron*

As redes MLPs são treinadas utilizando vários tipos de algoritmos, mas um dos algoritmos mais utilizados é a retro-propagação do erro conhecido como o *Backpropagation* (HAYKIN, 2007).

Figura 9 – Representação de um Perceptron com duas camadas



Fonte: (HAYKIN, 2007).

O *Backpropagation* é um algoritmo clássico muito utilizado no treinamento das redes neurais MLPs. Ele é geralmente associado ao algoritmo do gradiente descendente. Neste algoritmo é aplicada a descida do gradiente para minimizar a função de custo mais conhecido como *Loss Function* (ROJAS, 1996; YAMASHITA *et al.*, 2018). Este algoritmo tem por principal finalidade os ajustes dos pesos sinápticos conforme o erro cometido pela rede durante o treinamento (GURNEY, 1997).

O *Backpropagation* é composto por duas fases principais, a primeira é conhecida como a fase de *forward* e a segunda fase é conhecida como a fase de *backward*.

Na fase de *forward*, é passado um conjunto de dados que percorre a rede camada por camada até obter uma saída, no qual a saída do neurônio anterior é a entrada do neurônio da próxima camada. A saída de um determinado neurônio j pode ser obtido pela Equação 11 e pela Equação 12, na qual, n é o n-ésimo padrão de treinamento, m representa a quantidade de neurônio na camada, w_{ji} representa o peso sináptico do neurônio , x_i a entrada do neurônio, $v_j(n)$ o potencial de ativação do neurônio e y_j a saída do neurônio j (SILVA; SPATTI; FLAUZINO, 2010).

$$v_j(n) = \sum_{i=0}^{m} w_{ji}(n)x_i(n) \tag{11}$$

$$y_j = \varphi(v_j(n)) \tag{12}$$

O erro de saída do neurônio pode ser obtido pela Equação 13 e seu erro instantâneo como $\frac{1}{2}e_j^2(n)$, sendo $d_j(n)$ a saída desejada e $y_j(n)$ a saída real do neurônio (HAYKIN, 2007).

$$e_j(n) = d_j(n) - y_j(n) \tag{13}$$

Seguindo esse raciocínio, a soma dos erros instantâneos de todos os neurônios pode ser definida pelo Erro Quadrático Médio (EQM), representado pela Equação 14, em que C denota o conjunto de neurônios nas camadas de saída (GURNEY, 1997; ROJAS, 1996).

$$E(n) = \frac{1}{2}\sum_{j \in C} e_j^2(n) \tag{14}$$

Na fase *backward*, o erro é utilizado para ajustar os pesos w_{ji} , e o cálculo é dado pela Equação 15 equivalente a Equação 16, na qual δ é o gradiente local definido pela Equação 17, η é a taxa de aprendizagem e $\varphi_j'(v_j(n))$ é a derivada da função de ativação aplicada na saída do neurônio $v_j(n)$ (HAYKIN, 2007). Ademais quanto menor a taxa de aprendizagem η, menor serão as correções efetuadas nos pesos, causando uma lenta convergência da rede. Por outro lado quanto maior a taxa de aprendizagem η maior serão as correções aplicadas nos pesos, causando uma oscilação do algoritmo que consequentemente impede a convergência da rede (HAYKIN, 2007).

$$\Delta w_{ji}(n) = -\eta \frac{\delta E(n)}{\delta w_{ji}} \tag{15}$$

$$\Delta w_{ji}(n) = -\eta \delta_i(n) y_i(n) \tag{16}$$

$$\delta_i(n) = e_j(n)\varphi_j'(v_j(n)) \tag{17}$$

O processo de ajustes de pesos se repete até minimizar-se o erro (SILVA; SPATTI; FLAUZINO, 2010).

As redes neurais MLPs são um dos mais antigos métodos de aprendizagem profunda, elas foram fundamentais para criação de novas redes mais profundas com arquiteturas complexas como as redes neurais convolucionais para reconhecimento de imagem e para séries temporais, as redes neurais recorrentes para dados sequenciais como textos e séries temporais (WIKISTAT, 2015).

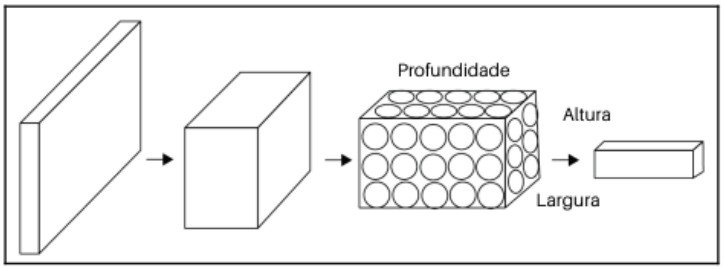
2.4 Redes Neurais Convolucionais

As Redes Neurais Convolucionais em inglês *Convolutional Neural Network* (CNNs ou ConvNets), classificadas como redes neurais profundas, são geralmente utilizadas para classi-

ficação de imagens e têm uma arquitetura parecida às redes neurais MLPs (SEWAK; KARIM; PUJARI, 2018).

O processo de treinamento das CNNS, assim como as redes MLPs, fazem uso do algoritmo *Backpropagation* para a atualização dos pesos (LIU *et al.*, 2017a). A diferença entre as MLPs e as CNNs é associada aos seus neurônios ocultos. As camadas em uma arquitetura CNNs tradicional são divididas em três, a saber, as camadas convolucionais, as camadas de *pooling* e as camadas totalmente conectadas ou *feedforward* (ACHARYA *et al.*, 2017). Cada neurônio é organizado em três dimensões ou seja, em altura, largura e profundidade como representado na Figura 10 (SEWAK; KARIM; PUJARI, 2018).

Figura 10 – Representação de um neurônio em CNN



Fonte: Adaptado de Sewak, Karim e Pujari (2018).

2.4.1 Camada convolucional

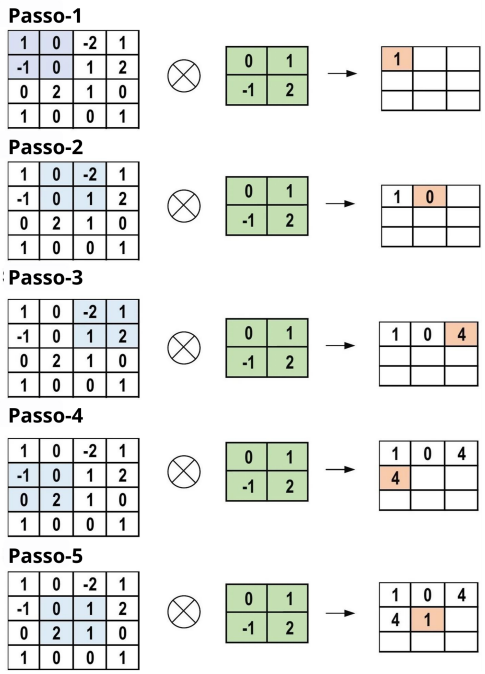
É a camada principal das CNNs, formada por uma combinação de operações lineares e não lineares (ALZUBAIDI *et al.*, 2021; YAMASHITA *et al.*, 2018). Ela é composta por um conjunto de filtros contendo núcleos, os núcleos em formato de matriz contêm valores de pesos aleatórios no início do treinamento, que são alterados durante o processo da aprendizagem para extrair características do conjunto de dados (ALZUBAIDI *et al.*, 2021; YAMASHITA *et al.*, 2018). Nesta camada o neurônio não está conectado a todos os neurônios da camada anterior, mas é conectado aos neurônios de uma determinada região especial conhecido como campo receptivo local (RANJBAR *et al.*, 2020).

Na arquitetura CNN os neurônios compartilham os pesos, o que reduz a quantidade dos mesmos, tornando consecutivamente o treinamento menos custoso em relação as redes neurais MLPs (ALZUBAIDI *et al.*, 2021; YAMASHITA *et al.*, 2018).

Nesta fase, o conjunto de dados de entrada são varridos pelos núcleos fazendo o produto de elemento por elemento que são somados no final (RANJBAR *et al.*, 2020).

Na Figura 11 está exemplificado o processo da fase de convolução de uma CNN com um conjunto de dados $4X4$ e um núcleo de $2X2$.

Figura 11 – Representação do processo na fase de convolução



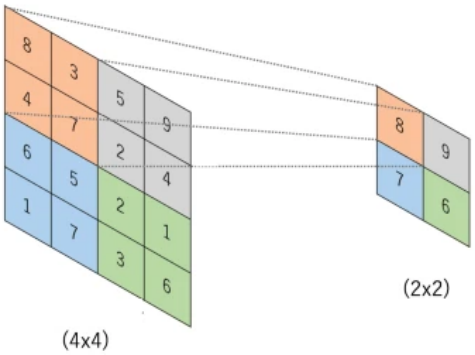
Fonte: Adaptado de Alzubaidi *et al.* (2021).

2.4.2 Camada de *Pooling*

Esta camada não efetua nenhum aprendizado e é geralmente aplicada após a fase de convolução (RANJBAR *et al.*, 2020). Ela diminui os mapas de caraterísticas geradas na fase de convolução aplicando técnicas de redução de mapa, o que ajuda a evitar o problema de *overfitting* (SEWAK; KARIM; PUJARI, 2018) .

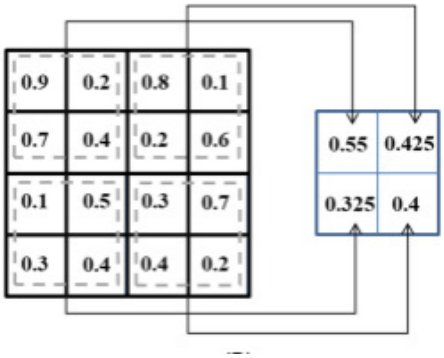
Existem vários tipos de técnicas *pooling* sendo as mais utilizadas o *Max pooling* e o *Average Pooling* (RANJBAR *et al.*, 2020; SEWAK; KARIM; PUJARI, 2018).

Figura 12 – Representação do processo de pooling usando o Max pooling



Fonte: Adaptado de Yamashita *et al.* (2018).

Figura 13 – Representação do processo de pooling usando o Average pooling



Fonte: Adaptado de Ranjbar *et al.* (2020).

O *Max pooling*, basicamente retorna o maior valor durante o processo de filtro, descartando os menores valores como está representado na Figura 12 em um conjunto de dados $4X4$ reduzido em $2X2$ (YAMASHITA *et al.*, 2018).

O *Average Pooling* retorna a média dos valores durante o processo de filtro, como está representado na Figura 13.

2.4.3 Camada Totalmente Conectada

Geralmente a camada totalmente conectada é a última camada das redes neurais CNNs, localizada após a camada de *pooling* e pode ser seguida ou não de outras camadas totalmente conectadas (RANJBAR *et al.*, 2020; SEWAK; KARIM; PUJARI, 2018).

Em uma CNN voltada para classificação, a última camada totalmente conectada utiliza funções de ativação não-lineares, como a ReLU e a Tangente Hiperbólica, para calcular as probabilidades associadas a cada classe. No entanto, é comum empregar a função *Softmax* nessa última camada para transformar as saídas em probabilidades normalizadas. A função *Softmax* converte as pontuações em uma distribuição de probabilidade, fornecendo uma interpretação mais clara das previsões do modelo para cada classe (YAMASHITA *et al.*, 2018).

2.4.4 Dropout

O *Dropout* é um método de regularização utilizado quando o modelo funciona bem em conjuntos de dados de treinos, mas é ineficiente em dados de testes, sinônimo de *overfitting* (ALZUBAIDI *et al.*, 2021; SEWAK; KARIM; PUJARI, 2018). Este método escolhe aleatoriamente os neurônios que não serão utilizados a cada fase de treino, entretanto usados na fases de testes. (NANDINI; KUMAR; K, 2021).

Um outro método muito utilizado que é parecido ao *Dropout* é o *Drop-Weights*, que ao invés de suprimir os neurônios, suprime os pesos deles, cortando assim a conexão entre eles (YAMASHITA *et al.*, 2018).

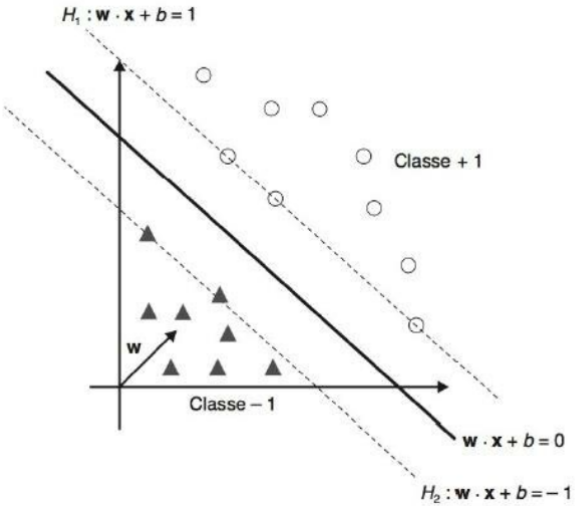
2.5 Máquinas de vetor de suporte

As Máquinas de Vetores de Suportes (SVMs) são tipos de rede *feedforward* utilizadas geralmente para a classificação com dados linearmente separáveis, mas em alguns casos são também utilizadas para dados não linearmente separáveis fazendo uso de hiperplanos (HAYKIN, 2007). Considerando um conjunto de dados de entrada X e Y a saída representada pelo conjunto $\{-1,1\}$. Um hiperplano está definido pela Equação 18 na qual $w.x$ é o produto escalar entre os vetores w e x, e b representa um número real. A Equação 18 pode ser utilizada para a divisão dos dados de entrada combinada a Equação 19 e está representada pela Figura 14 (HAYKIN, 2007). Na Figura 14 o H_1 representa a fronteira para a classe $+1$ e H_2 representa a fronteira para a classe -1, essas fronteiras são chamadas de hiperplano.

$$h(x) = w.x + b \tag{18}$$

$$y = \begin{cases} 1 \; se \; w.x + b > 0 \\ -1 \; se \; w.x + b < 0 \end{cases} \tag{19}$$

Figura 14 – Ilustração de hiperplanos canônicos e separador



Fonte: Adaptado de Faceli *et al.* (2021).

Os SVMs, ao lidar com dados não linearmente separáveis, incorporam dois conceitos essenciais. O primeiro conceito é conhecido como *One vs Rest* (Um contra Todos), o qual é empregado em problemas de classificação multiclasse, nos quais existem mais de duas classes.

Nessa abordagem, cada classificador C_i é responsável por realizar uma tarefa de classificação binária, distinguindo a classe i das demais. Quando aplicado a um novo valor x, esse valor será atribuído à classe associada ao classificador que obteve a pontuação mais alta entre os n classificadores, como representado na Equação 20. Importante notar que, para tratar dados não linearmente separáveis, é recomendado o uso de funções kernel.

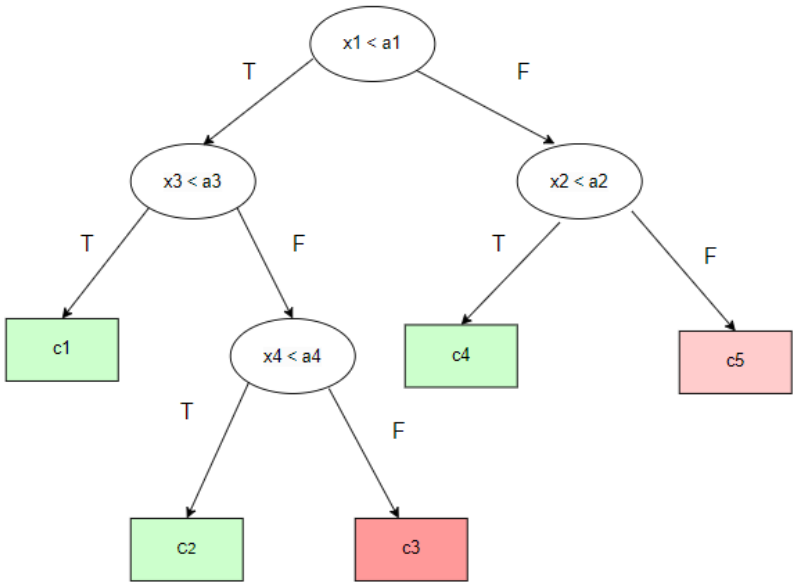
$$C(x) = \arg max_{1<i<n}(C_i(x)) \tag{20}$$

O segundo conceito é o todos contra todos, na qual o novo valor x pertence à classe com a maior quantidade de votos depois do sistema de votação (GONÇALVES, 2008).

2.6 Árvore de decisão

Uma árvore de decisão é um algoritmo de aprendizagem de máquina geralmente utilizada como método de classificação e de regressão, ela é um tipo de grafo direcionado que em cada nó gera dois ou mais nós folha. Cada ramo de um nó é definido por uma condição e as condições são testes que estão caracterizadas por um operador lógico como $>$, $<$, $=$... e um valor domínio de atributo. Na classificação o atributo é escolhido por uma regra de medida chamada *good of split*, que determina quão bem o atributo representa a classe como está ilustrado na Figura 15 (FACELI *et al.*, 2021).

Figura 15 – Uma árvore de decisão e as regiões de decisão no espaço de objetos



Fonte: Adaptado de Faceli *et al.* (2021).

2.7 *Random Forest*

O algoritmo Random Forest é uma técnica de aprendizado de máquina que combina múltiplas árvores de decisão para construir um modelo preditivo mais robusto e preciso. Cada árvore de decisão é treinada em uma amostra aleatória do conjunto de dados original e é utilizada para fazer uma previsão. A predição final é feita pela média das predições de todas as árvores individuais (BREIMAN, 2001).

A construção das árvores de decisão no Random Forest é feita através do método de *bootstrap*, que consiste em amostrar aleatoriamente o conjunto de dados original com reposição, gerando novos subconjuntos de dados para treinamento das árvores individuais. Além disso, em cada nó da árvore, uma amostra aleatória de atributos é selecionada para determinar qual atributo será utilizado para dividir os dados nesse nó (ALI *et al.*, 2012).

A seleção aleatória de atributos e a utilização de múltiplas árvores reduzem a correlação entre as predições e aumentam a generalização do modelo para novos dados. O Random Forest também permite medir a importância de cada atributo para a predição, o que pode ser útil para entender o comportamento do modelo e selecionar atributos relevantes para a análise.

2.8 Métricas

Há várias métricas para avaliar o desempenho de classificador, pode-se citar alguns mais populares como a Acurácia (*Accuracy*), a perda (*Loss*), Recall e Precisão (*Recall and Precision*), *F1-Score*.

- **Acurácia (*Accuracy*)**:

  A Acurácia é uma métrica comum usada para avaliar o desempenho de modelos de classificação. Ela é definida como a proporção de classificações corretas em relação ao número total de amostras (MAXWELL; WARNER; GUILLéN, 2021). A fórmula para calcular a acurácia é representada na Equação 21:

$$\text{Acurácia} = \frac{\text{Número de amostras classificadas corretamente}}{\text{Número total de amostras}} \tag{21}$$

  A precisão é uma métrica útil para avaliar a eficácia geral de um modelo de classificação.

- **Perda (*Loss*)**:

  A perda é uma métrica comum usada para avaliar o desempenho de modelos de regressão. Ela é definida como a discrepância entre a saída prevista pela rede neural e a saída real. A fórmula para calcular a perda depende do problema em questão e do tipo

de função de perda usada na rede neural. Em geral, o objetivo é minimizar a perda ao treinar a rede neural (SOWELL, 2021).

- ***Recall* e Precisão (*Recall and Precision*)**:

  O *recall* e a precisão são métricas comuns usadas para avaliar o desempenho de modelos de classificação binária. O recall mede a proporção de instâncias positivas que foram corretamente identificadas em relação ao número total de instâncias positivas. A precisão mede a proporção de instâncias positivas que foram corretamente identificadas em relação ao número total de instâncias identificadas como positivas (SOWELL, 2021; MAXWELL; WARNER; GUILLéN, 2021). Nas Equações 22 e 23 são apresentadas as formulas de *recall* e precisao, respectivamente:

$$\textit{Recall} = \frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos + Falsos negativos}} \tag{22}$$

$$\text{Precisão} = \frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos + Falsos positivos}} \tag{23}$$

  O *recall* e a precisão são importantes para avaliar o desempenho da rede neural em problemas de classificação binária.

- ***F1-Score***:

  O F1-Score é uma métrica que combina *recall* e precisão em uma única medida de desempenho. É a média harmônica entre *recall* e precisão e é calculado pela Equação 24:

$$\textit{F1-Score} = 2 \times \frac{\text{Precisão} \times \textit{Recall}}{\text{Precisão} + \textit{Recall}} \tag{24}$$

  O F1-Score é uma métrica útil para avaliar a eficácia geral de um modelo de classificação, equilibrando recall e precisão (GHORI *et al.*, 2020).

- ***F2-Score* ou *F2-Measure*** : *F2-Score* ou *F2-Measure*, é uma métrica de avaliação comumente utilizada em problemas de classificação binária. Ela é uma variação da pontuação F1, que é a média harmônica entre precisão e *recall*. A *F2-Measure* dá mais peso ao *recall* do que à precisão, tornando-se útil em situações em que é mais crucial identificar corretamente os verdadeiros positivos, mesmo que isso resulte em mais falsos positivos.

  A fórmula da *F2-Measure* é dada pela Equação 25:

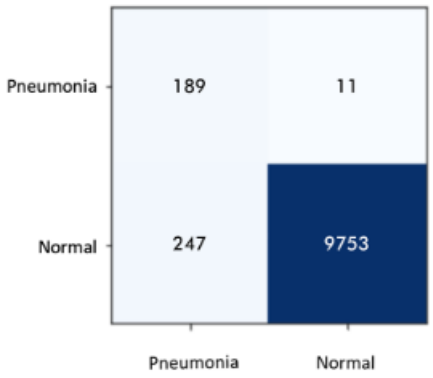
$$\textit{F2-Measure} = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \textit{Recall}}{(\beta^2 \cdot \text{Precision}) + \textit{Recall}} \tag{25}$$

O parâmetro β controla o equilíbrio entre precisão e *recall*. No caso da *F2-Measure*, β é definido como 2, dando mais peso ao *recall*. Se $\beta = 1$, a fórmula se reduz à *F1-Score* (GHORI *et al.*, 2020).

- **Matriz de Confusão (*Confusion Matrix*)**:

  A matriz de confusão é uma tabela que mostra o número de instâncias classificadas corretamente e incorretamente para cada classe. Ela é comumente usada para avaliar o desempenho de modelos de classificação. Uma matriz de confusão para um problema (SOWELL, 2021; MAXWELL; WARNER; GUILLéN, 2021). A Figura 16 apresenta um exemplo de matriz de confusão, na qual é possível observar que, de um total de 200 pacientes diagnosticados com pneumonia, o classificador classificou corretamente 189, cometendo erro em 11 pacientes. Em relação aos 10.000 pacientes saudáveis, o classificador obteve uma taxa de acerto de 9.753, com 247 casos erroneamente classificados.

Figura 16 – Matriz de confusão de pacientes



Fonte: Sowell (2021).

- **Curva *Receiver Operating Characteristic***

  A curva *Receiver Operating Characteristic* (ROC) é uma ferramenta fundamental na avaliação de desempenho de modelos de classificação, fornecendo uma representação visual da taxa de verdadeiros positivos (TPR ou sensibilidade) em relação à taxa de falsos positivos (FPR ou $1-$especificidade). Essa análise é crucial em situações em que a discriminação entre duas classes é de grande importância, como em problemas médicos, detecção de fraudes e diagnósticos (SOWELL, 2021).

  A representação gráfica da curva ROC é obtida variando o limiar de classificação do modelo. O limiar determina a probabilidade mínima para atribuir uma instância à classe positiva. Seja $P(c|x)$ a probabilidade predita da classe positiva dado um exemplo x, a classificação c ocorre quando $P(c|x) >$ limiar.

A curva ROC é construída ao plotar a TPR em função da FPR para diferentes valores do limiar de classificação. A sensibilidade (TPR) e a especificidade (TNR ou $1 - \text{FPR}$) são definidas respectivamente pelas Equações 26 e 27.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{26}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{27}$$

Sendo TP (True Positives), FN (False Negatives), FP (False Positives) e TN (True Negatives) representam as contagens de instâncias corretamente classificadas como positivas, erroneamente classificadas como negativas, erroneamente classificadas como positivas e corretamente classificadas como negativas, respectivamente.

A área sob a curva ROC (AUC-ROC) é uma métrica que resume a eficácia global do modelo. A fórmula para calcular a AUC-ROC é dada pela Equação 28.

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR})\, \text{dFPR} \tag{28}$$

A AUC-ROC varia de 0 a 1, sendo que 1 representa um desempenho perfeito e 0.5 indica um desempenho equivalente ao acaso (SOWELL, 2021; MAXWELL; WARNER; GUILLéN, 2021).

A crescente complexidade dos modelos de aprendizagem de máquina, aliada à sua proliferação em uma ampla gama de aplicações, tem colocado um desafio crucial diante dos cientistas de dados e pesquisadores: como interpretar e compreender o comportamento desses modelos? A simples busca pela acurácia não é mais suficiente. À medida que modelos mais complexos são desenvolvidos para tarefas críticas, como diagnóstico médico, detecção de fraudes financeiras e muito mais, é imperativo que possamos entender como e por que esses modelos tomam suas decisões.

A interpretabilidade de modelos de aprendizagem de máquina é essencial, não apenas para cumprir requisitos regulatórios, mas também para construir confiança e aceitação pública dessas tecnologias. No entanto, os modelos modernos muitas vezes são “caixas-pretas” devido à complexidade de seus algoritmos, o que torna desafiador entender as razões subjacentes a suas previsões.

Neste contexto, o *Shapley Additive Explanation* (SHAP) emerge como uma ferramenta fundamental para a interpretabilidade de modelos.

2.8.1 *Shapley Additive Explanation* (SHAP)

O SHAP baseia-se na sólida teoria dos jogos cooperativos, especificamente na Solução de Shapley, para fornecer uma abordagem sistemática e teoricamente embasada na explicação das previsões do modelo.

Uma das características distintivas do SHAP é a capacidade de atribuir valores SHAP a cada variável, o que permite quantificar a contribuição de cada variável para uma previsão específica. Isso é realizado por meio de uma fórmula matemática precisa, que é representada na Equação 29:

$$\text{Valor SHAP}(x_i) = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N| - |S| - 1)!}{|N|!}[f(S \cup i) - f(S)] \tag{29}$$

- x_i é a característica de interesse.
- S é um subconjunto de N que exclui a característica x_i.
- $f(S \cup i)$ é a previsão do modelo quando todas as características em S e a característica x_i estão presentes.
- f(S) é a previsão do modelo quando apenas as características em S estão presentes.
- |S| denota o número de elementos no conjunto S.
- |N| denota o número de elementos no conjunto N.

Assim, o SHAP proporciona uma interpretação transparente, possibilitando uma análise mais profunda das previsões dos modelos de aprendizagem de máquina.

2.9 Estado da arte

São escassos os estudos que utilizam Redes Neurais Convolucionais (CNNs) para classificar organismos baseando-se em seus microbiomas. A maioria destes trabalhos são aplicados ao microbioma humano (SHARMA; XU; XU, 2021), sendo poucos os trabalhos nesta área aplicados a predição de saúde ou doença no contexto agrícola.

Em um estudo realizado por Wang *et al.* (2021), um método de classificação baseado em microbioma foi apresentado e comparado com outros métodos existentes. Esse novo método utiliza uma CNN regularizadora filogenética chamada *MDeep*, que recebe como entrada dados taxonômicos, tais como unidades taxonômicas operacionais (OTUs) que representam classificações de seres vivos em diferentes níveis, desde o reino até a espécie. Os dados passam por diversas camadas de convolução para identificar correlações filogenéticas entre os taxa[2],

[2] Plural de um taxón

seguidas de várias camadas totalmente conectadas. Para evitar o *overfitting*, é utilizado o método de *dropout*. Nesse estudo foi possível fazer a predição da idade com base no microbioma intestinal de pessoas no Estados Unidos, também foi utilizado o método de classificação binária para a predição do gênero de gêmeos com base no microbioma intestinal. Nesse estudo, *MDeep* obteve resultados significativos em torno de 95% para os dados de classificação binaria de gênero de gêmeos e 75% para os dados de regressão prevendo a idade cronológica com base no microbioma intestinal.

A utilização de CNNs é muito comum para previsão e classificação de doenças como é no caso do artigo de Sharma *et al.* (2020), que apresenta uma nova abordagem de aprendizado de máquina que incorpora uma estratégia estratificada para agrupar unidades taxonômicas operacionais (OTUs) em grupos de filos e usa redes neurais convolucionais (CNNs) para treinar cada cluster individualmente. Essa abordagem melhorou a precisão de previsão em comparação com o uso de uma única CNN que ignora as correlações entre OTUs. A nova abordagem foi testada em conjuntos de dados simulados e em estudos de microbioma humano de cirrose e diabetes tipo 2, produzindo resultados encorajadores com os valores médios de AUC de 0,88, 0,92, 0,75.

No artigo de Sharma, Xu e Xu (2021), é proposto um novo sistema de aprendizado profundo chamado “phyLoSTM” para análise de dados longitudinais de sequenciamento do microbioma humano e fatores ambientais do hospedeiro para previsão de doenças. O *framework* utiliza uma combinação de redes neurais convolucionais e redes neurais de memória de longo prazo (LSTM) para extração de recursos e análise de dependência temporal. O método também lida com pontos de tempo variáveis em indivíduos e equilibra o peso entre casos e controles desbalanceados. Os autores simularam 100 conjuntos de dados para testar o modelo e o aplicaram em dois estudos reais de microbioma humano longitudinal: DIABIMMUNE e DiGiulio. Os resultados mostraram que o modelo proposto superou o desempenho do método Random Forest, com um aumento de 5% na AUC nas simulações e aumentos de 19% e 8% nas AUCs dos estudos DIABIMMUNE e DiGiulio, respectivamente. O *framework* proposto melhora a acurácia preditiva em estudos longitudinais de microbioma humano que contêm dados espacialmente correlacionados e avalia a mudança na composição do microbioma que contribui para a previsão de resultados. É importante ressaltar que a utilização neste estudo de classificadores como a SVM, *Random Forest*, foram também utilizados no artigo Namkung (2020), no qual os autores realizam uma síntese dos métodos de aprendizado de máquina para investigar a correlação entre microbiomas e características do hospedeiro, ressaltando a influência do ambiente bacteriano intestinal humano sobre o sistema imunológico, condições psicológicas, câncer, obesidade e doenças metabólicas. Com o avanço da tecnologia de sequenciamento, estudos de microbiomas com amostras volumosas tornaram-se viáveis a um custo acessível. O texto explora diversas abordagens de aprendizado de máquina, como regressão penalizada, máquina de vetores de suporte (SVM), floresta aleatória e redes neurais artificiais (ANN), incluindo métodos de redes neurais profundas. Todo o processo de análise, desde a configuração

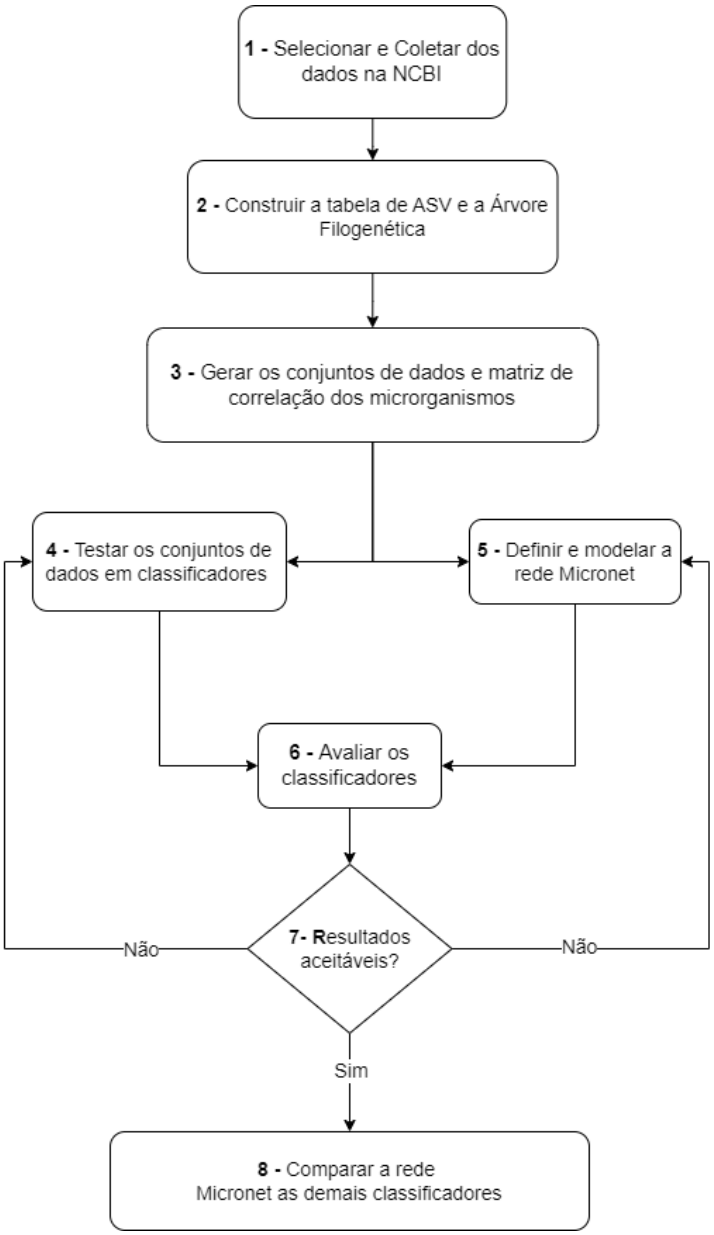
do ambiente até a extração dos resultados, é detalhado utilizando a linguagem de programação Python.

Em um outro estudo realizado por Barque (2021), explorou-se a previsão da saúde do coral com base em seu microbioma, utilizando a rede neural *MDeep*, juntamente com algoritmos clássicos de aprendizado de máquina como SMO e J48 da *Weka*. Com dados reais, o *MDeep* alcançou uma precisão de 85,11% nos dados de treino e aproximadamente 72% nos dados de validação, superando SMO e J48, que obtiveram 66,17% e 57,35%, respectivamente. A performance superior do *MDeep* foi atribuída à sua exploração de informações filogenéticas e ao agrupamento de OTUs com base na correlação filogenética. Neste estudo, destacou-se bactérias importantes para a saúde do coral.

3 MATERIAL E MÉTODOS

Neste capítulo é explanada a metodologia aplicada para a predição do microbioma saudável em uma cultura utilizando as redes neurais. O metódo aplicado para está pesquisa esta resumido em 8 etapas como pode ser observado na Figura 17.

Figura 17 – Diagrama da metodologia



Fonte: Autoria própria (2022).

3.1 Coleta e Pré-processamento dos Dados

No início do estudo, conduziu-se uma investigação sobre os dados disponíveis na NCBI[1] para selecionar a cultura apropriada para esta pesquisa. Após avaliações criteriosas, levando em conta o volume de amostras e as metodologias de sequenciamento para identificação de micro-organismos, decidiu-se utilizar o conjunto de dados referente à cultura de arroz, analisado no estudo de Bez *et al.* (2021).

O referido estudo focou na avaliação de culturas de arroz em duas regiões da Itália (Barone e Sole), investigando a doença causada por *Dickeya zeae*, que induz a podridão do pé nas plantações. Neste estudo, os autores examinaram patobiomas em plantas de arroz, tanto saudáveis quanto afetadas, e constataram que a infecção por *Dickeya zeae* altera a comunidade bacteriana em termos de composição, abundância e diversidade de espécies.

Essas variações estão ligadas à configuração de consórcios microbianos associados à condição patológica. Vários tipos bacterianos foram identificados em coocorrência significativa com o agente patogênico, sugerindo uma vinculação ao processo da doença.

Os dados de sequenciamento brutos empregados nesta pesquisa estão armazenados no *Sequence Read Archive* (SRA) da NCBI, acessíveis pelo *Bioproject* PRJNA602829. Os dados incluem sequências de amostras de arroz saudável e doente, abrangendo 110 amostras (63 doentes e 47 saudáveis) das duas localidades italianas (Barone e Sole), identificando aproximadamente 3.816 ASVs.

Para a obtenção desses dados, foi necessário o download do *SRA toolkit*[2], uma ferramenta do NCBI para baixar dados de sua plataforma. Subsequentemente, os dados para cada amostra foram descarregados utilizando seus códigos SRA[3], nos formatos fastq: R1 (*forward*) e R2 (*reverse*).

No estágio da pesquisa, correspondente às etapas 2 e 3 da Figura 17, foram criadas a árvore filogenética e a tabela ASV a partir dos arquivos fastq (R1 e R2) através do uso do programa QIIME2, seguindo os parâmetros descritos em Bez *et al.* (2021).

O QIIME é uma ferramenta reconhecida para análise de dados de microbioma, oferecendo um diversas ferramentas e algoritmos para análise de sequências de DNA, sendo amplamente utilizado em pesquisas relacionadas à microbiologia, ecologia microbiana, biologia ambiental, entre outras áreas. O QIIME2 destaca-se como uma plataforma imprescindível para estudos nesses campos. Subsequentemente, com base na árvore filogenética, realizou-se um processamento para gerar a matriz de correlação (3816×3816) entre ASVs utilizando a função *cophenetic* no software R, que é uma linguagem multi-paradigma direcionada para manipulação e análise de dados. Para essas análises, empregou-se o RStudio[4] na versão 4.1.0, que fornece um ambiente integrado para a linguagem R.

[1] Centro Nacional de Informações sobre Biotecnologia
[2] https://github.com/ncbi/sra-tools/wiki/02.-Installing-SRA-Toolkit
[3] *Sequence Read Archive*
[4] https://www.rstudio.com/

Após o processamento bioinformático, analisando os dados, foi percebido um grande números de colunas de ASVs com dados nulos. Para resolver, foi aplicado uma técnica de limpeza que faz a exclusão dos ASVs (micro-organismos) ausentes em 20%, 30%, 50%, 70% e 100% das amostras.

Após essa etapa, os dados foram distribuídos em dois conjuntos: um com 77 amostras alocadas para o treinamento dos classificadores, incluindo 43 amostras de plantas doentes e 34 de plantas saudáveis; e outro com 33 amostras destinadas aos testes, compostas por 20 amostras de plantas doentes e 13 saudáveis. Os conjuntos de dados foram normalizados para um intervalo de 0 a 1 usando a função *MinMaxScaler* da biblioteca *Scikit Learn*.

3.2 Implementação e Avaliação dos Classificadores

A seleção dos classificadores utilizados na etapa 4 da Figura 17, como a MLP e SVM, foi baseada em sua eficácia comprovada na classificação de conjuntos de dados binários, sendo amplamente reconhecidos na literatura científica (WILHELM; ES; BUCKLEY, 2021a). Da mesma forma, a escolha de outros algoritmos, como a árvore de decisão, foi fundamentada em sua aplicabilidade consagrada na área da microbiologia para a classificação de classes. O uso do *Random Forest* também foi adotado devido à sua robustez em lidar com problemas de classificação binária, sendo uma associação de árvores de decisão (LIU *et al.*, 2017b). Além disso, a aplicação do *MDeep*, uma rede neural convolucional mencionada no artigo de Wang *et al.* (2021), para a classificação do gênero humano com base no microbioma intestinal, assim como sua utilização por Barque *et al.* (2024) para a classificação da saúde de amostras de corais com base nos micro-organismos presentes, foi motivada pela sua eficácia demonstrada em estudos anteriores.

Para a implementação dos classificadores clássicos com MLP, SVM, árvore de decisão e *Random Forest*, foi utilizada a biblioteca Scikit-learn[5] na versão 0.20.3.

Para o classificador *MDeep*, foi utilizada a biblioteca TensorFlow[6] na versão 1.12.0, que suporta processamento com GPU através de CUDA. Também foi utilizada a biblioteca Scipy[7] na versão 1.2.1, que serve para ilustrar agrupamentos de micro-organismos por meio de dendrogramas.

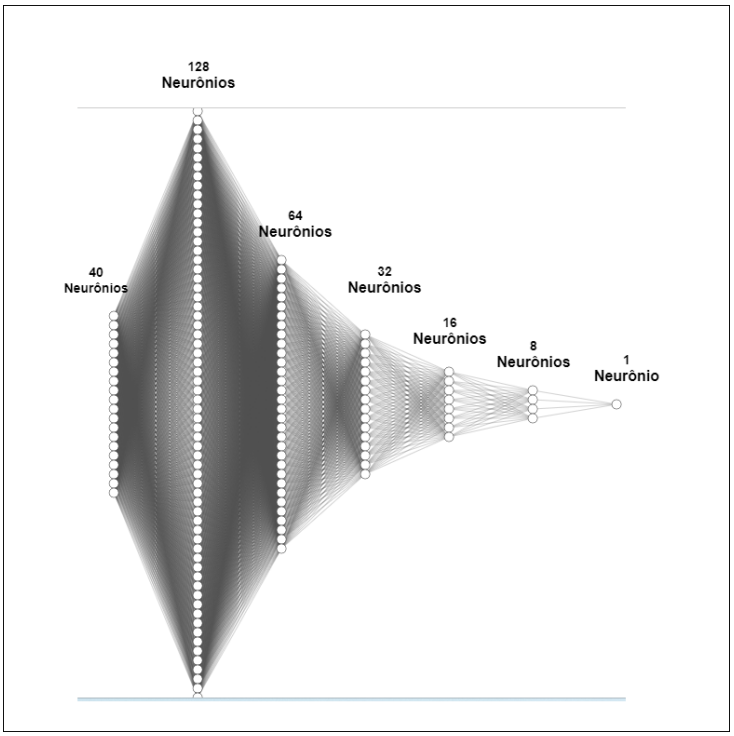
O Matplotlib na versão 3.1.0 foi utilizado para a criação de gráficos, além da linguagem Python na versão 3.

Em seguida, na etapa 5 da Figura 17 foi construido um novo classificador, conforme listado nos objetivos deste projeto. A arquitetura do classificador elaborado foi uma rede neural completamente conectada, denominada Micronet, representada simbolicamente pela Figura 18.

[5] https://scikit-learn.org/
[6] https://www.tensorflow.org/
[7] https://www.scipy.org/

A estrutura da Micronet foi composta por um número equivalente de ASVs em relação à quantidade de neurônios na camada de entrada, os quais passaram pela função de ativação *GELU*, 128 neurônios na primeira camada oculta, seguida por 64 neurônios na segunda camada oculta, 32 na terceira, 16 na quarta, e 8 na quinta. A camada de saída foi configurada para indicar 0 para amostras saudáveis e 1 para amostras doentes, após passar por uma função de *SoftMax* e *Sigmoid*.

Figura 18 – Representação simbólica da Micronet



Fonte: Autoria própria (2024).

Para implementar a Micronet, utilizou-se a linguagem Python na versão 3.6 e as bibliotecas Pytorch[8] na versão 2.0.1.

Após a fase de implementação, os classificadores foram treinados e testados, comparando-os por meio de métricas como Precisão, *Recall*, Acurácia, *F1-Score*, e *F2-Measure*. Além das métricas, utilizou-se a matriz de confusão e a biblioteca *Shap values* para aprimorar a avaliação dos micro-organismos mais relevantes identificados por cada classificador, e por último, a visualização gráfica das árvores de decisões e do *Random Forest*.

[8] https://pytorch.org/

4 RESULTADOS E DISCUSSÃO

Neste capítulo, serão apresentados a configuração experimental, os resultados obtidos após os experimentos e serão discutidas as análises relacionadas a esses resultados.

4.1 Configuração experimental

No experimento, foram comparadas as métricas dos classificadores ao remover as ASVs (micro-organismos) ausentes nas amostras em diferentes proporções: 100% resultando a 3816 ASVs, , representada na Tabela 1; 70% com 144 ASVs, representada na Tabela 2; 50% com 40 ASVs, representada na Tabela 3; 30% com 7 ASVs, representada na Tabela 4 e 20% com 4 ASVs, representada na Tabela 5. Observou-se que o melhor resultado foi obtido baseado na média das métricas obtidas dos classificadores, ao remover 50% das ASVs nulas. Assim, foi decidido seguir com essa taxa de remoção, resultando em 40 ASVs em vez de 3816, o que representa uma redução de aproximadamente 99% das ASVs.

Tabela 1 – Tabela comparativa das métricas dos classificadores utilizados nesse estudo com a remoção de 100% dos ASVs nulos.

Classificador	Precisão (%)	Recall (%)	Acurácia (%)	F1-Score (%)	F2-Measure (%)
Micronet	100	90,91	94,59	95,24	92,59
MDeep	100	95,45	97,30	97,67	96,33
Árvore de Decisão	100	77,27	86,49	87,18	80,95
Random Forest	100	77,27	86,49	87,18	80,95
SVM	100	45,45	67,57	62,50	51,02
MLP	100	68,18	81,08	81,08	72,82
Média	100	75,76	85,59	85,14	79,11

Fonte: Elaborado pelo autor (2024).

Tabela 2 – Tabela comparativa das métricas dos classificadores utilizados nesse estudo com a remoção de 70% dos ASVs nulos.

Classificador	Precisão (%)	Recall (%)	Acurácia (%)	F1-Score (%)	F2-Measure (%)
Micronet	100	63,64	78,38	77,78	68,63
MDeep	94,74	81,82	86,49	87,80	84,11
Árvore de Decisão	81,82	81,82	78,38	81,82	81,82
Random Forest	100	86,36	91,89	92,68	88,79
SVM	100	68,18	81,08	81,08	72,82
MLP	94,44	77,27	83,78	85,00	80,19
Média	95,17	76,52	83,33	84,36	79,39

Fonte: Elaborado pelo autor (2024).

Durante o experimento, todos os classificadores foram treinados com a mesma semente aleatória (*random seed*) estabelecida em 42. Os parâmetros utilizados para a rede *MDeep* foram consistentes com os parâmetros empregados por Barque (2021), incluindo um tamanho de lote (*batch size*) de 32, execução por 500 épocas, taxa de aprendizagem de $0{,}0001$ e uma taxa de

Tabela 3 – Tabela comparativa das métricas dos classificadores utilizados nesse estudo com a remoção de 50% dos ASVs nulos.

Classificador	Precisão (%)	Recall (%)	Acurácia (%)	F1-Score (%)	F2-Measure (%)
Micronet	95,45	**95,45**	**94,59**	**95,45**	**95,45**
MDeep	95,24	90,91	91,89	93,02	31,74
Random Forest	**100**	81,82	89,19	90	84,91
SVM	**100**	81,82	89,19	90	84,91
MLP	94,74	81,82	86,49	87,80	84,11
Árvore de Decisão	94,12	72,73	81,08	82,05	76,19
Média	96,59	84,09	88,74	89,72	76,21

Fonte: Elaborado pelo autor (2024).

Tabela 4 – Tabela comparativa das métricas dos classificadores utilizados nesse estudo com a remoção de 30% dos ASVs nulos.

Classificador	Precisão (%)	Recall (%)	Acurácia (%)	F1-Score (%)	F2-Measure (%)
Micronet	79,17	86,36	78,38	82,61	84,82
MDeep	60	95,45	59,46	73,68	85,37
Árvore de Decisão	83,33	68,18	72,97	75,00	70,75
Random Forest	90	81,82	83,78	85,71	83,33
SVM	81,82	81,82	78,38	81,82	81,82
MLP	79,17	86,36	78,38	82,61	84,82
Média	78,92	83,33	75,23	80,24	81,82

Fonte: Elaborado pelo autor (2024).

Tabela 5 – Tabela comparativa das métricas dos classificadores utilizados nesse estudo com a remoção de 20% dos ASVs nulos.

Classificador	Precisão (%)	Recall (%)	Acurácia (%)	F1-Score (%)	F2-Measure (%)
Micronet	59,46	100	59,46	74,58	88,00
MDeep	0	0	40,54	0	0
Árvore de Decisão	76,19	72,73	70,27	74,42	73,39
Random Forest	73,91	77,27	70,27	75,56	76,58
SVM	64,52	90,91	64,86	75,47	84,03
MLP	0	0	40,54	0	0
Média	45,68	56,82	57,66	50,01	53,67

Fonte: Elaborado pelo autor (2024).

dropout de 0.2. Já para a Micronet, os parâmetros adotados foram os seguintes: tamanho de lote de 32, execução por 10 épocas e taxa de aprendizagem de $0{,}001$. Os demais classificadores foram configurados com os parâmetros padrões fornecidos pela biblioteca *Scikit*.

Alguns parâmetros da árvore de decisão utilizados foram a função *Gini*, como função para medir a qualidade da divisão da árvore. Como estratégia para a divisão em cada nó foi utilizado a propriedade *best* que tem por objetivo escolher a melhor divisão. O número mínimo de amostras necessárias para dividir um nó interno é 2. As outras propriedades foram inicializadas como 0 ou $None$.

Para o *Random Forest*, foi também utilizado os parâmetros padrões do *Scitkit*, sendo 100 como o numero de árvores da floresta representado pela propriedade $n\_estimators$. Para medir a qualidade da divisão foi utilizado a função *Gini*. Para a profundidade máxima da árvore

foi alterado para 3 sendo 2 o padrão. As outras propriedades foram inicializados como 0 ou $None$.

Para o *SVM* foi utilizado a função sigmoide na propriedade *Kernel*, a propriedade *Gamma* foi inicializada como $auto$. Para as outras propriedades foram utilizados os parâmetros padrões do *Scikit*.

4.2 Comparação do desempenho preditivo

Após o treinamento dos classificadores, diversas métricas foram avaliadas, incluindo Acurácia, Precisão, Recall, F1-Score e F2-Measure, as quais estão apresentadas na Tabela 3. Observa-se que, em relação à métrica de Precisão, os melhores desempenhos foram obtidos pelos classificadores *Random Forest* e SVM, atingindo 100%. Quanto à Acurácia, a Micronet obteve a melhor taxa com $94{,}59$%, e em relação à Recall e outras métricas, mais uma vez a Micronet alcançou o melhor resultado, atingindo $95{,}45$%. Além dessas métricas, foram geradas matrizes de confusão para cada classificador, conforme apresentado na Figura 19.

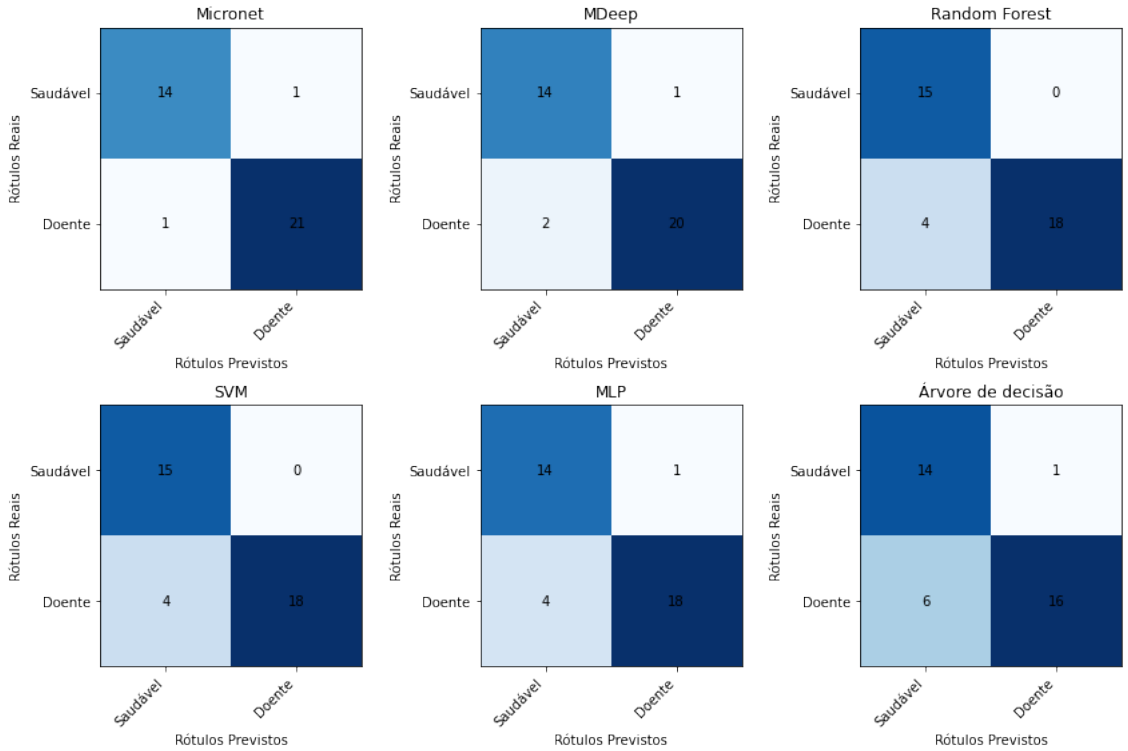
Na Figura 19, pode-se observar as classificações realizadas pelas diferentes técnicas. A Micronet, por exemplo, acertou 14 das 15 amostras saudáveis e 21 das 22 amostras doentes. Já a MDeep obteve um desempenho semelhante, com 14 acertos nas amostras saudáveis e 20 nas amostras doentes. Tanto o *Random Forest* quanto a SVM classificaram corretamente todas as amostras saudáveis, mas alcançaram 18 acertos e 4 erros nas amostras doentes. A MLP, por sua vez, acertou 14 amostras saudáveis e 18 amostras doentes. Por fim, a Árvore de Decisão obteve 14 acertos nas amostras saudáveis e 16 acertos nas amostras doentes.

Para adicionar outra avaliação dos classificadores, foi gerada a curva ROC ilustrada na Figura 20. Neste gráfico, é possível perceber que a Micronet obteve o melhor desempenho, com uma taxa de AUC de 94%. Em seguida, o MDeep alcançou 91%, seguido pelo *Random Forest* e SVM, ambos com 88% de AUC, e por fim, a Árvore de Decisão com 83% de AUC.

Tendo os resultados dos classificadores em presentes, buscou-se entender os resultados obtidos utilizando o *framework SHAP values* para verificar quais micro-organismos cada classificador baseou-se mais para efetuar suas previsões, destacando os micro-organismos chave identificados para cada classificador. A Figura 21 apresenta os 10 micro-organismos identificados pelos classificadores Micronet, *MDeep*, *Random Forest*, Árvore de decisão, SVM e MLP como tendo o maior impacto.

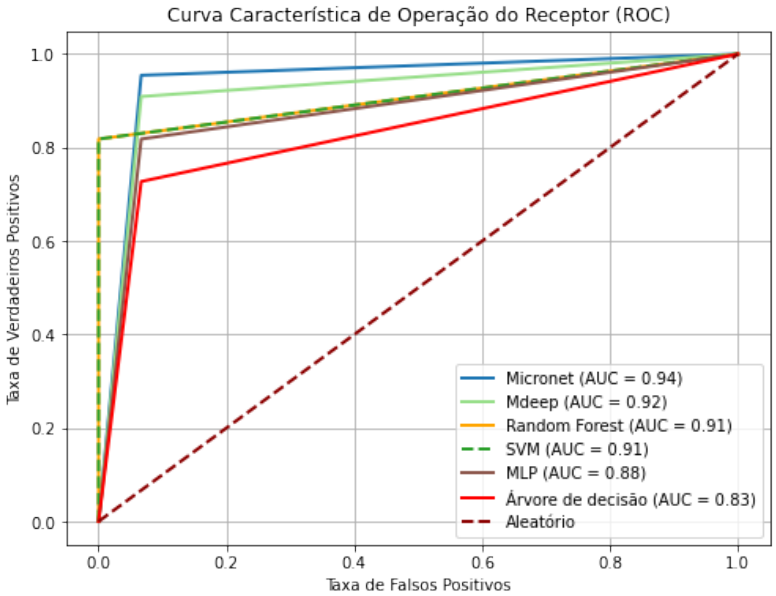
A análise comparativa dos resultados obtidos através dos classificadores Micronet, *MDeep*, *Random Forest*, Árvore de Decisão, SVM e MLP evidencia a influência diferenciada da presença de certos micro-organismos nas decisões de classificação das amostras analisadas. Detalhadamente, observam-se os seguintes padrões relacionados à classificação das amostras como “doentes” ou “saudáveis”:

Figura 19 – Matriz de confusão do conjunto de dados utilizando os classificadores desse estudo



Fonte: Elaborado pelo autor (2024).

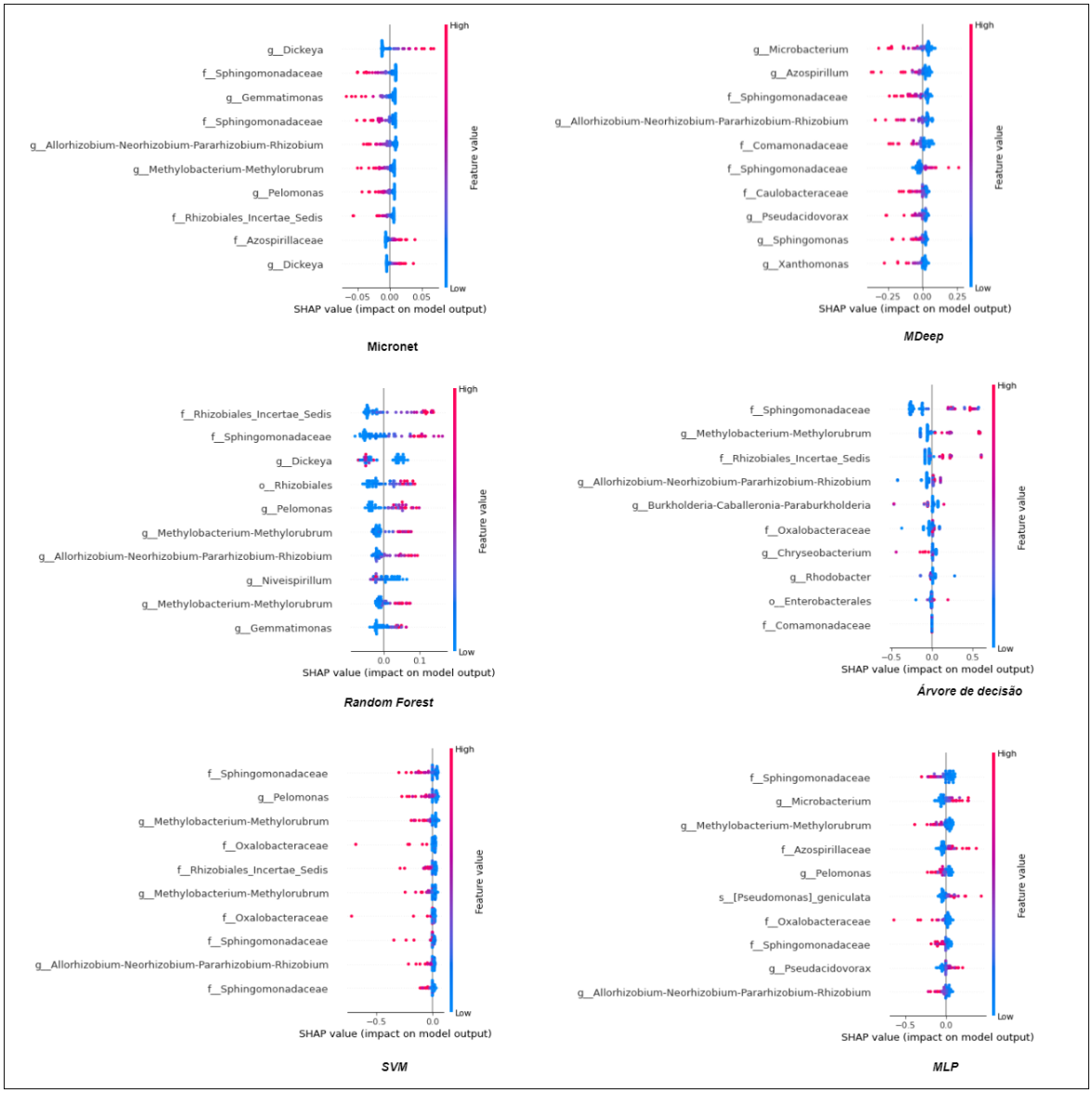
Figura 20 – Curva ROC demostrando o desempenho dos modelos de predição deste estudo



Fonte: Elaborado pelo autor (2024).

- **Micronet:** Demonstra uma tendência de classificar as amostras como “doentes” quando o *gênero Dickeya* e a *família Azospirillaceae* são prevalentes. Em contrapartida, identifica amostras como “saudáveis” frente à predominância da *família Sphingomonadaceae* e do *gênero Gemmatimonas*.

Figura 21 – SHAP values dos classificadores utilizados neste estudo



Fonte: Elaborado pelo autor (2024).

- **MDeep:** Este classificador é propenso a categorizar as amostras como “doentes” na presença em abundância da *família Sphingomonadaceae* e do *gênero Allorhizobium-Neorhizobium-Pararhizobium,* enquanto sinaliza como “saudáveis” amostras ricas em *Microbacterium* e *Azospirillum*.

- **Random Forest:** Revela a tendência de indicar amostras como “doentes” com a alta representatividade da *família Rhizobiales Incertae Sedis* e da *família Sphingomonadaceae*, diferentemente, aponta como “saudáveis” aquelas com maior ocorrência de *Chryseobacterium* e da *família Azospirillaceae*.

- **Árvore de Decisão:** Classifica como “doentes” as amostras que apresentam elevada concentração da *família Sphingomonadaceae*, do *gênero Methylobacterium-Methylorubrum,* e da *família Rhizobiales Incertae Sedis,* em contraste, identifica como

"saudáveis" as que detêm abundância de *Burkholderia-Caballeronia-Paraburkholderia* e *Chryseobacterium*.

- **SVM:** Tem a disposição de marcar como "doentes" as amostras com grande quantitativo da *família Azospirillaceae* e do *gênero Pseudacidovorax*, enquanto tende a reconhecer como "saudáveis" aquelas que possuem a *família Sphingomonadaceae* e o *gênero Pelomonas* em destaque.
- **MLP:** Mostra propensão a rotular as amostras como "doentes" quando existe significativa presença de *Microbacterium* e da *família Azospirillaceae,* contraposto a isso, classifica como "saudáveis" as amostras com predominância da *família Sphingomonadaceae,* do *gênero Methylobacterium-Methylorubrum,* e do *gênero Pelomonas*.

Por fim, foram gerados, as árvores dos classificadores árvore de decisão e o *Random Forest* ilustrado repesctivamente na Figura 24 e na Figura 23

4.3 Discussão

Os resultados alcançados a partir da análise dos dados e do desempenho dos classificadores mostram métricas variadas, proporcionando uma visão abrangente sobre a eficácia desses modelos na diferenciação entre amostras saudáveis e doentes. A distribuição dos dados treinados e testados em termos de amostras saudáveis e doentes foi estrategicamente organizada para refletir uma representação equilibrada dos dois grupos, garantindo robustez na avaliação dos classificadores.

Inicialmente, a avaliação quantitativa dos modelos permitiu identificar que a Micronet demonstrou desempenho superior, alcançando uma taxa de acurácia de $94{,}59\%$, juntamente com outros indicadores favoráveis, como precisão, recall e F1-Score. Esse resultado corrobora com a capacidade do modelo de Micronet em distinguir adequadamente as amostras doentes das saudáveis. Por outro lado, a *MDeep*, embora tenha obtido resultados de precisão e recall elevados, mostrou uma ligeira inferioridade em comparação com a Micronet.

É relevante notar que a árvore de decisão e o *Random Forest* apresentaram resultados divergentes em comparação com a Micronet e a *MDeep*. Ambos os classificadores demonstraram precisões perfeitas de 100% ao identificar amostras saudáveis. No entanto, esses modelos mostraram uma precisão relativamente inferior na identificação de amostras doentes. O SVM e MLP exibiram desempenho intermediário, com resultados de precisão e recall oscilando entre $94{,}74\%$ a $81{,}82\%$ e acurácia entre $89{,}19\%$ a $86{,}49\%$.

Além disso, a análise das matrizes de confusão, conforme refletido na Figura 19, destaca a capacidade de cada classificador em corretamente identificar amostras saudáveis e doentes. A Micronet e a MDeep, mais uma vez, mostraram um bom desempenho, corretamente classificando a maioria das amostras, enquanto o *Random Forest* e SVM tiveram um bom desem-

penho na identificação de amostras saudáveis, mas mostraram dificuldades na identificação de amostras doentes. A MLP e a Árvore de Decisão também mostraram algumas limitações, especialmente na identificação de amostras doentes.

A curva de Característica de Operação do Receptor proporcionou uma avaliação mais detalhada do desempenho dos classificadores. A Micronet obteve o melhor desempenho com uma área sob a curva de 94%, seguida pela MDeep com 91%. O *Random Forest* e SVM apresentaram desempenho semelhante com 88% de AUC, enquanto a Árvore de Decisão obteve 83% de AUC. Esse resultado é consistente com as métricas quantitativas anteriores, corroborando a superioridade da Micronet e a relativa inferioridade do *Random Forest* e da Árvore de Decisão na diferenciação entre amostras saudáveis e doentes.

A análise dos SHAP values revelou os micro-organismos que mais influenciaram as previsões de cada classificador. Os resultados indicaram que para a Micronet e a MDeep, os micro-organismos *gênero Dickeya, família Azospirillaceae*, entre outros, demonstraram ser mais relevantes na classificação de amostras como doentes, enquanto os micro-organismos *família Sphingomonadaceae, gênero Gemmatimonas*, entre outros, tiveram mais influência na classificação de amostras como saudáveis.

A identificação de micro-organismos específicos, como *Dickeyea*, *Sphingomonadaceae*, e outros, tanto nos classificadores com mais taxa de acerto (Micronet e Mdeep) quanto no artigo científico de Bez *et al.* (2021), apresentado na Figura 22, indica uma convergência nas conclusões, mesmo tendo uma redução de 50% dos ASVs realizado nesse estudo. Esta convergência é valiosa, pois sugere uma robustez nas associações feitas por diferentes métodos.

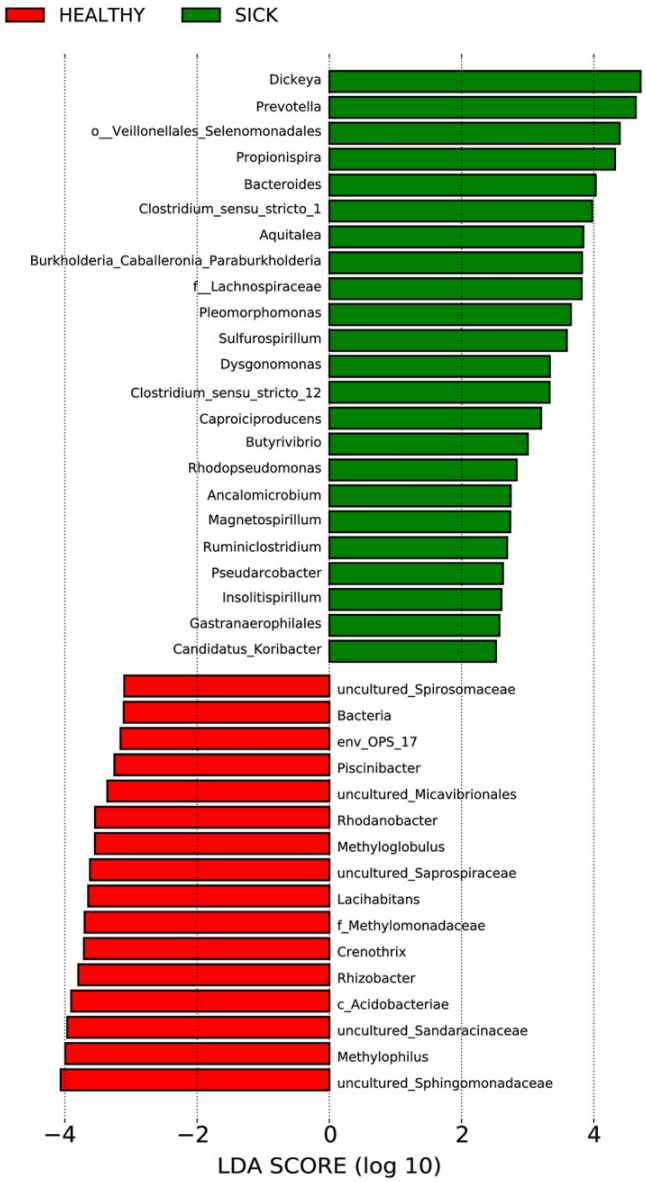
Essa técnica de remoção de micro-organismos menos abundantes também foi aplicada e destacada no artigo de referência Janssen *et al.* (2019), no qual foi identificada uma melhoria considerável no desempenho dos classificadores.

A coerência na identificação da *Dickeya* como o agente causador da podridão do pé de arroz entre os classificadores Micronet, Mdeep e o artigo de Bez *et al.* (2021) é notável e demonstra uma concordância crucial na caracterização dessa bactéria como patógeno associado a essa doença específica. Essa consistência aumenta a confiabilidade das conclusões e reforça a validade da associação.

Em estudos adicionais, foram confirmados os benefícios dos micro-organismos no crescimento das plantas, especialmente no caso da *Sphingomonadaceae* e da *Gemmatimonas* (WANG *et al.*, 2022; BARNAWAL; SINGH; SINGH, 2018; LIU *et al.*, 2022).

Em suma, os resultados destacam a Micronet como o classificador mais eficaz no contexto da diferenciação entre amostras saudáveis e doentes. No entanto, há nuances na performance dos classificadores, com diferenças significativas nas métricas de avaliação. Essa análise fornece uma visão abrangente sobre a influência dos micro-organismos na classificação e ressalta a necessidade de considerar as particularidades de cada modelo ao aplicá-los em cenários reais.

Figura 22 – Diferenças nas Abundâncias Bacterianas entre amostras Saudáveis e Doentes.



Fonte: Bez *et al.* (2021).

Esse estudo abre portas para investigações futuras, como a exploração de técnicas de otimização ou o desenvolvimento de modelos híbridos que possam melhorar ainda mais a precisão na identificação de amostras doentes, contribuindo assim para avanços significativos na agricultura.

Essa avaliação detalhada oferece uma compreensão abrangente do desempenho e dos impactos dos micro-organismos nos modelos de classificação, contribuindo para a área da bioinformática e da agricultura ao explorar melhor os dados genômicos para fins diagnósticos.

Por fim, a aplicação de técnicas de aprendizado de máquina supervisionado neste estudo e em outros estudos (JANSSEN *et al.*, 2019; WILHELM; ES; BUCKLEY, 2021b) demonstrou a facilidade de analisar, classificar e identificar micro-organismos chave que determinam a saúde das amostras, em comparação com o artigo de Bez *et al.* (2021), que utilizou métodos estatísticos mais tradicionais.

5 CONCLUSÕES E PERSPECTIVAS

Em face dos resultados obtidos da análise dos classificadores aplicados para diferenciar amostras de plantas de arroz saudáveis e doentes, a Micronet emerge como o modelo mais eficaz, exibindo um desempenho robusto e consistente em relação aos outros classificadores. Seus resultados de alta acurácia, precisão, recall, F1-Score e área sob a curva ROC (AUC) ilustram sua capacidade de identificar com sucesso amostras saudáveis e doentes. No entanto, ao observar mais detalhadamente os SHAP values, é evidente que a Micronet, juntamente com a MDeep, prioriza certos micro-organismos para fazer tais distinções. Essa descoberta abre caminho para compreender as bases genéticas subjacentes à diferenciação entre amostras de plantas de arroz.

Embora a Micronet tenha se destacado, os demais classificadores também apresentaram desempenho digno. O Random Forest e o SVM, por exemplo, obtiveram precisão perfeita na identificação de amostras saudáveis, mas enfrentaram mais dificuldades na diferenciação de amostras doentes. A análise das matrizes de confusão forneceu percepções valiosos sobre o desempenho individual de cada classificador na identificação de amostras saudáveis e doentes.

Além disso, a identificação consistente da *Dickeya* como a bactéria causadora da podridão do pé de arroz, fundamentada em múltiplos classificadores e evidências científicas, destaca a importância dessa bactéria como um patógeno significativo para a agricultura. Este conhecimento serve como base para a implementação de medidas práticas e cientificamente fundamentadas para o controle eficaz da doença e ressalta a necessidade contínua de pesquisas para aprimorar nossa compreensão das interações microbiológicas no contexto agrícola.

Portanto, este estudo ressalta a importância de uma abordagem multivariada, que combine a aplicação de diferentes modelos de classificação com técnicas que permitam identificar e interpretar as influências genômicas nas decisões de classificação. Essa compreensão tem relevância significativa no contexto da agricultura e bioinformática, fornecendo percepções para futuras pesquisas na identificação e diagnóstico de doenças em plantas de arroz.

Em suma, os resultados obtidos oferecem uma perspectiva abrangente sobre a eficácia dos modelos de classificação na diferenciação de amostras de plantas de arroz saudáveis e doentes. Esses achados são promissores e representam um passo significativo para a aplicação prática de métodos computacionais avançados na agricultura, contribuindo para melhorias na saúde das plantas e na produtividade agrícola.

REFERÊNCIAS

ACHARYA, U. R. *et al.* A deep convolutional neural network model to classify heartbeats. **Computers in Biology and Medicine**, Elsevier Ltd, v. 89, p. 389–396, 2017. ISSN 18790534. Disponível em: http://dx.doi.org/10.1016/j.compbiomed.2017.08.022.

ALI, J. *et al.* Random forests and decision trees. **null**, 2012.

ALZUBAIDI, L. *et al.* **Review of deep learning: concepts, CNN architectures, challenges, applications, future directions**. Springer International Publishing, 2021. v. 8. ISSN 21961115. ISBN 4053702100444. Disponível em: https://doi.org/10.1186/s40537-021-00444-8.

AMTHOR, F. **Neuroscience For Dummies**. Wiley, 2016. (For dummies). ISBN 9781119224891. Disponível em: https://books.google.bj/books?id=WkjhCgAAQBAJ.

BARNAWAL, D.; SINGH, R.; SINGH, R. P. **Role of Plant Growth Promoting Rhizobacteria in Drought Tolerance: Regulating Growth Hormones and Osmolytes**. [*S.l.*: *s.n.*], 2018.

BARQUE, B. M. B.S. thesis, **Predição de microbioma saudável baseada em micro-organismos presentes no coral Mussismilia hispida, utilizando uma rede neural profunda**. 2021.

BARQUE, B. M. *et al.* Prediction of health of corals mussismilia hispida based on the microorganisms present in their microbiome. *In*: PEREIRA, A. I. *et al.* (Ed.). **Optimization, Learning Algorithms and Applications**. Cham: Springer Nature Switzerland, 2024. p. 409–423. ISBN 978-3-031-53025-8.

BEZ, C. *et al.* The rice foot rot pathogen Dickeya zeae alters the in-field plant microbiome. **Environmental Microbiology**, v. 23, n. 12, p. 7671–7687, 2021. ISSN 14622920.

BREIMAN, L. Random forests. **null**, 2001.

CHANDRAN, H.; MEENA, M.; SWAPNIL, P. Plant growth-promoting rhizobacteria as a green alternative for sustainable agriculture. **Sustainability (Switzerland)**, v. 13, n. 19, p. 1–30, 2021. ISSN 20711050.

FACELI *et al.* **Inteligência artificial: uma abordagem de aprendizado de máquina**. Grupo GEN, 2021. ISBN 9788521637509. Disponível em: https://integrada.minhabiblioteca.com.br/books/9788521637509.

FURTADO, M. I. V. **Redes Neurais Artificiais: Uma Abordagem Para Sala de Aula**. [*S.l.*: *s.n.*], 2019. ISBN 9788572473262.

GHORI, K. M. *et al.* Performance analysis of machine learning classifiers for non - technical loss detection. **Journal of Ambient Intelligence and Humanized Computing**, Springer Berlin Heidelberg, n. 0123456789, 2020. ISSN 1868-5145. Disponível em: https://doi.org/10.1007/s12652-019-01649-9.

GONÇALVES, A. R. Fundamentos e Aplicações de Técnicas de Aprendizado de Máquina. 2008. Disponível em: https://andreric.github.io/posts/2018/05/blog-post-2/.

GURNEY, K. **An Introduction to Neural Networks**. USA: Taylor Francis, Inc., 1997. ISBN 1857286731.

HAYKIN, S. **Redes Neurais: Princípios e Prática**. Artmed, 2007. ISBN 9788577800865. Disponível em: https://books.google.com.br/books?id=bhMwDwAAQBAJ.

JAIN, A. Fundamentals of deep learning – starting with artificial neural network. 2016. Disponível em: https://www.analyticsvidhya.com/blog/2016/03/introduction-deep-learning-fundamentals-neural-networks/.

JAIN, A. *et al.* A review of plant leaf fungal diseases and its environment speciation. **Bioengineered**, Taylor Francis, v. 10, n. 1, p. 409–424, 2019. ISSN 21655987. Disponível em: https://doi.org/10.1080/21655979.2019.1649520.

JANSSEN, R. *et al.* An artificial neural network identifies glyphosate-impacted brackish communities based on 16s rrna amplicon miseq read counts. **bioRxiv**, Cold Spring Harbor Laboratory, p. 711309–, 2019.

JESKE, J. T.; GALLERT, C. Microbiome Analysis via OTU and ASV-Based Pipelines—A Comparative Interpretation of Ecological Data in WWTP Systems. **Bioengineering**, v. 9, n. 4, 2022. ISSN 23065354.

JOOS, L. *et al.* Daring to be differential: metabarcoding analysis of soil and plant-related microbial communities using amplicon sequence variants and operational taxonomical units. **BMC Genomics**, BMC Genomics, v. 21, n. 1, p. 1–17, 2020. ISSN 14712164.

LIU, C. *et al.* Soil bacterial communities of three types of plants from ecological restoration areas and plant-growth promotional benefits of microbacterium invictum (strain x-18). **Frontiers in Microbiology**, 2022.

LIU, W. *et al.* A survey of deep neural network architectures and their applications. **Neurocomputing**, Elsevier B.V., v. 234, n. November 2016, p. 11–26, 2017. ISSN 18728286. Disponível em: http://dx.doi.org/10.1016/j.neucom.2016.12.038.

LIU, Y. *et al.* Experimental study and random forest prediction model of microbiome cell surface hydrophobicity. **Expert Systems With Applications**, Pergamon, v. 72, p. 306–316, 2017.

MANNAA, M.; SEO, Y.-S. Plants under the Attack of Allies : Moving towards the Plant. **Plants**, v. 10, n. 125, p. 1–15, 2021.

MAXWELL, A. E.; WARNER, T. A.; GUILLéN, L. A. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—part 1: Literature review. **Remote Sensing**, v. 13, n. 13, 2021. ISSN 2072-4292. Disponível em: https://www.mdpi.com/2072-4292/13/13/2450.

MONTESINOS, E. Plant-associated microorganisms: A view from the scope of microbiology. **International Microbiology**, v. 6, n. 4, p. 221–223, 2003. ISSN 11396709.

NAMKUNG, J. Machine learning methods for microbiome studies. **Journal of Microbiology**, 2020.

NANDINI, G. S.; KUMAR, A. S.; K, C. Dropout technique for image classification based on extreme learning machine. **Global Transitions Proceedings**, v. 2, n. 1, p. 111–116, 2021. ISSN 2666-285X. 1st International Conference on Advances in Information, Computing and Trends in Data Engineering (AICDE - 2020). Disponível em: https://www.sciencedirect.com/science/article/pii/S2666285X21000157.

QUAIL, B. W. Thematic Review. **Media Management in the Age of Lyndon B. Johnson**, v. 21, n. 4, p. 13–68, 2021.

RANJBAR, S. *et al.* 3 - computational intelligence for modeling of asphalt pavement surface distress. *In*: SAMUI, P. *et al.* (Ed.). **New Materials in Civil Engineering**. Butterworth-Heinemann, 2020. p. 79–116. ISBN 978-0-12-818961-0. Disponível em: https://www.sciencedirect.com/science/article/pii/B978012818961000003X.

ROJAS, R. **Neural Networks: A Systematic Introduction**. Berlin, Heidelberg: Springer-Verlag, 1996. ISBN 3540605053.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 3. ed. [*S.l.*]: Prentice Hall, 2010.

SCHLOSS, P. D. Amplicon sequence variants artificially split bacterial genomes into separate clusters. **mSphere**, v. 6, n. 4, p. e00191–21, 2021. Disponível em: https://journals.asm.org/doi/abs/10.1128/mSphere.00191-21.

SEWAK, M.; KARIM, R.; PUJARI, P. **Practical Convolutional Neural Networks**. [*S.l.*: *s.n.*], 2018. 199 p. ISBN 9781788392303.

SHADE, A.; STOPNISEK, N. Abundance-occupancy distributions to prioritize plant core microbiome membership. **Current Opinion in Microbiology**, Elsevier Ltd, v. 49, p. 50–58, 2019. ISSN 18790364. Disponível em: https://doi.org/10.1016/j.mib.2019.09.008.

SHARMA, D. *et al.* Taxonn: ensemble of neural networks on stratified microbiome data for disease prediction. **Bioinformatics**, 2020.

SHARMA, D.; XU, W.; XU, W. phylostm: a novel deep learning model on disease prediction from longitudinal microbiome data. **Bioinformatics**, 2021.

SHARMA, S.; SHARMA, S.; ATHAIYA, A. Activation Functions in Neural Networks. **International Journal of Engineering Applied Sciences and Technology**, v. 04, n. 12, p. 310–316, 2020. ISSN 2455-2143.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. Redes Neurais Artificiais Para Engenharia e Ciências Aplicadas. **São Paulo: Artliber**, 2010.

SOWELL, T. Magician's Corner: 9. Performance Metrics for Machine Learning Models. v. 55905, p. 1–7, 2021.

STEWART, J. E. *et al.* Pathobiome and microbial communities associated with forest tree root diseases. **Forest Microbiology: Volume 1: Tree Microbiome: Phyllosphere, Endosphere and Rhizosphere**, p. 277–292, 2021.

WANG, F. *et al.* Sphingomonas sp. hbc-6 alters physiological metabolism and recruits beneficial rhizosphere bacteria to improve plant growth and drought tolerance. **Frontiers in Plant Science**, 2022.

WANG, Y. *et al.* A novel deep learning method for predictive modeling of microbiome data. **Briefings in bioinformatics**, v. 22, n. 3, p. 1–14, 2021. ISSN 14774054.

WIKISTAT. Neural Networks and Introduction to Deep Learning. p. 1–17, 2015. Disponível em: http://klab.tch.harvard.edu/academia/classes/BAI/pdfs/intro-deep-learning.pdf.

WILHELM, R. C.; ES, H. M. van; BUCKLEY, D. H. Predicting measures of soil health using the microbiome and supervised machine learning. **Soil Biology Biochemistry**, Pergamon, v. 164, p. 108472–, 2021.

WILHELM, R. C.; ES, H. M. van; BUCKLEY, D. H. Predicting measures of soil health using the microbiome and supervised machine learning. **Soil Biology Biochemistry**, Pergamon, v. 164, p. 108472–, 2021.

WILHELM, R. C.; ES, H. M. van; BUCKLEY, D. H. Predicting measures of soil health using the microbiome and supervised machine learning. **Soil Biology and Biochemistry**, Elsevier Ltd, v. 164, n. August 2021, p. 108472, 2022. ISSN 00380717. Disponível em: https://doi.org/10.1016/j.soilbio.2021.108472.

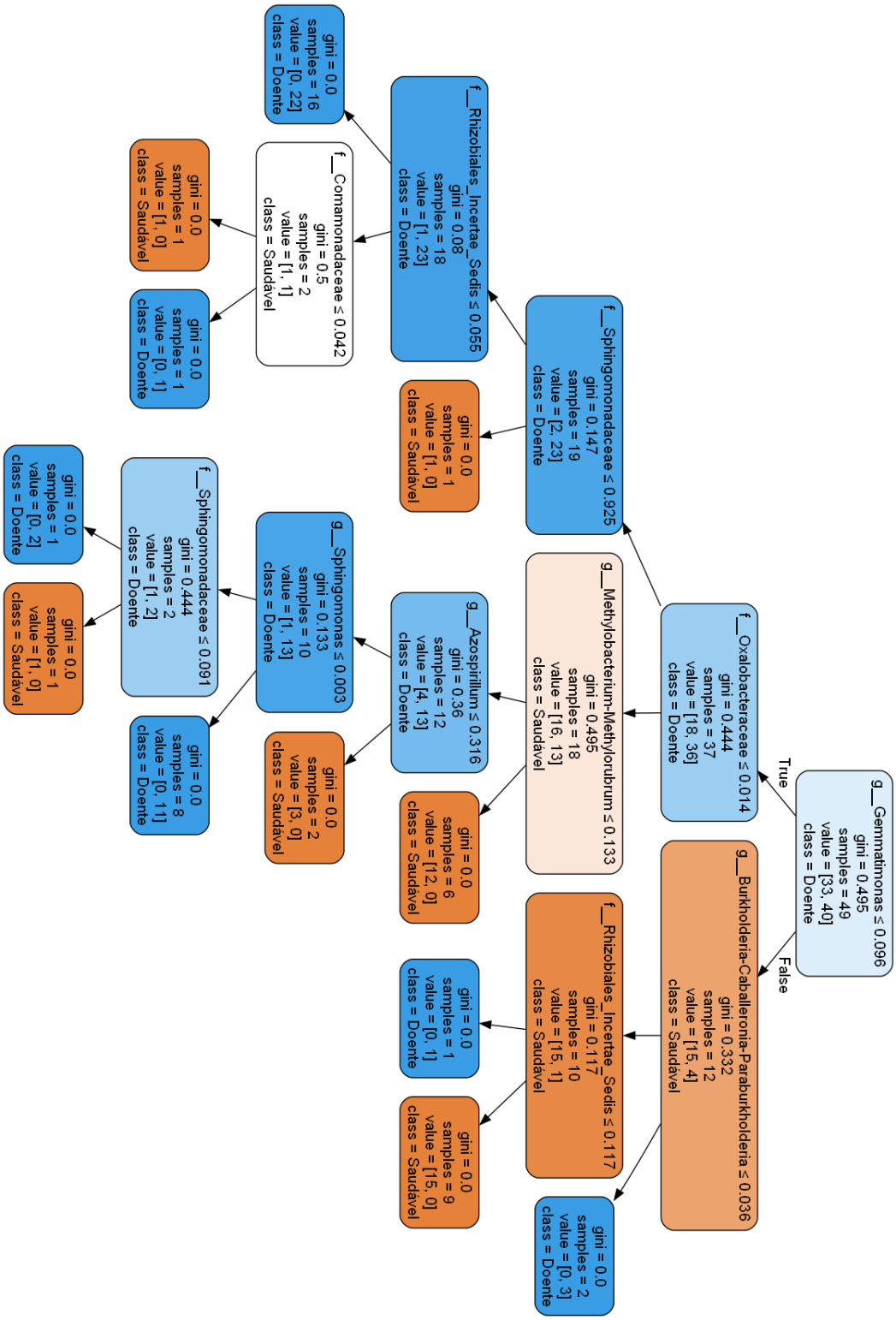
YAMASHITA, R. *et al.* Convolutional neural networks: an overview and application in radiology. **Smart Innovation, Systems and Technologies**, Insights into Imaging, v. 195, p. 21–30, 2018. ISSN 21903026.

YORK, R. O. Gelu activation function in deep learning: A comprehensive mathematical analysis and performance. 2023.

YUAN, J. *et al.* Predicting disease occurrence with high accuracy based on soil macroecological patterns of Fusarium wilt. **ISME Journal**, Springer US, v. 14, n. 12, p. 2936–2950, 2020. ISSN 17517370. Disponível em: http://dx.doi.org/10.1038/s41396-020-0720-5.

APÊNDICE A – Figura da árvore gerada pelo *Random Forest*

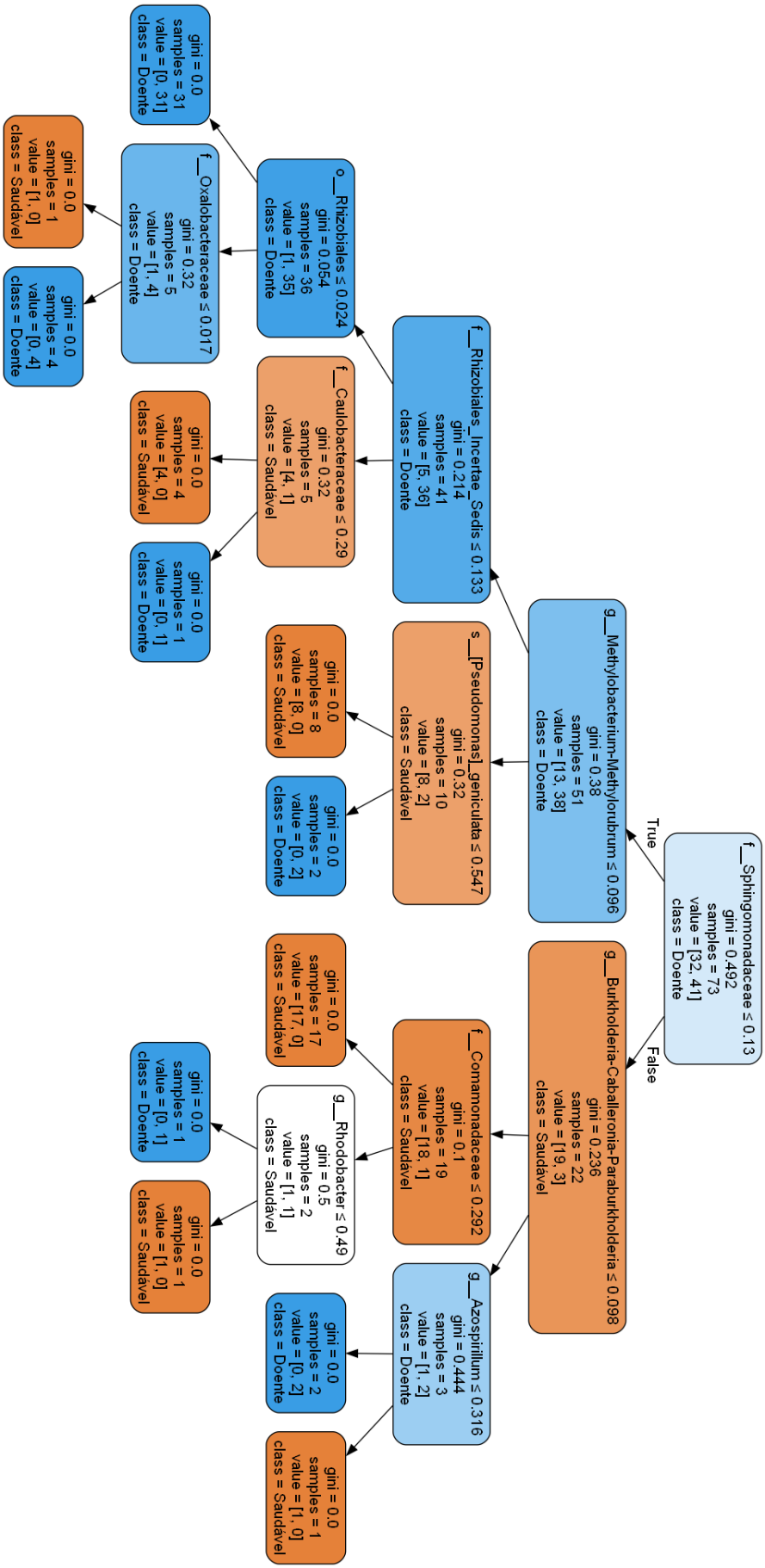
Figura 23 – Radom forest *Tree*



Fonte: Autoria própria (2023).

APÊNDICE B – Figura da árvore gerada pela árvore de decisão

Figura 24 – Árvore de decisão *Tree*



Fonte: Autoria própria (2023).