**UNIVERSIDADE FEDERAL DO CEARÁ**

**CENTRO DE CIÊNCIAS**

**DEPARTAMENTO DE COMPUTAÇÃO**

**CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

**PEDRO AMARAL FONTES DE SALES**

# AVALIAÇÃO DE MODELOS DE APRENDIZAGEM DE MÁQUINA E TÉCNICAS DE EXPLICABILIDADE PARA DESCRIÇÃO DE MICROBIOMA DE SOLO

**FORTALEZA**

**2024**

PEDRO AMARAL FONTES DE SALES

# AVALIAÇÃO DE MODELOS DE APRENDIZAGEM DE MÁQUINA E TÉCNICAS DE EXPLICABILIDADE PARA DESCRIÇÃO DE MICROBIOMA DE SOLO

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientador: Prof. Dr. César Lincoln Cavalcante Mattos.

FORTALEZA

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

---

S155a Sales, Pedro Amaral Fontes de.
Avaliação de modelos de aprendizagem de máquina e técnicas de explicabilidade para descrição de microbioma de solo / Pedro Amaral Fontes de Sales. – 2024.
43 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Ciências, Curso de Computação, Fortaleza, 2024.
Orientação: Prof. Dr. César Lincoln Cavalcante Mattos.

1. Aprendizagem de máquina. 2. Explicabilidade. 3. XAI. 4. Ecologia. 5. Metagenômica. I. Título.
CDD 005

---

PEDRO AMARAL FONTES DE SALES

# AVALIAÇÃO DE MODELOS DE APRENDIZAGEM DE MÁQUINA E TÉCNICAS DE EXPLICABILIDADE PARA DESCRIÇÃO DE MICROBIOMA DE SOLO

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovada em: 24 de Setembro de 2024.

BANCA EXAMINADORA

---

Prof. Dr. César Lincoln Cavalcante Mattos (Orientador)
Universidade Federal do Ceará (UFC)

---

Prof. Dr. João Paulo do Vale Madeiro
Universidade Federal do Ceará (UFC)

---

Prof. Msc. Pedro Hericson Machado Araújo
Instituto Federal do Ceará (IFCE)

## RESUMO

O campo da aprendizagem de máquina vem se tornando cada vez mais popular e a necessidade de compreender melhor o funcionamento dos modelos desenvolvidos por essa área se faz cada vez mais evidente. Dentre os recursos existentes para suprir essa necessidade, está a SHAP (*Shapley Additive Explanations*), uma abordagem que auxilia na interpretação de modelos de aprendizagem de máquina ao calcular a importância dos atributos que compõe os dados. Um possível uso dessa ferramenta é o auxílio na compreensão não só das decisões tomadas pelos modelos, como também da relevância dos atributos, e aquilo que eles representam, no contexto do domínio dos dados. Neste trabalho, buscamos utilizar essas ferramentas para averiguar a aplicabilidade desse tipo de análise no campo da ecologia microbiana do solo, investigando se modelos de aprendizagem de máquina são capazes de classificar de maneira satisfatória dados de metagenômica e se as análises desses modelos podem auxiliar na compreensão da diversidade microbiana das amostras, oferecendo perspectivas que análises tradicionais não conseguem oferecer. Esses dados se destacam pela alta dimensionalidade e pela dificuldade de coleta de amostras, duas características que tornam um problema bastante desafiador no contexto da aprendizagem de máquina. Portanto, também foi avaliado o efeito da projeção dos dados, para redução de dimensionalidade, nos resultados obtidos. De maneira geral, não observou-se benefício na projeção dos dados. Porém, sem a projeção, os resultados obtidos foram promissores, ainda que limitados pela quantidade reduzida de dados. Além disso, as análises de explicabilidade obtidas apresentaram certa relação com análises mais tradicionais mas também foram capazes de evidenciar certos grupos que as análises tradicionais não foi. De maneira geral, o resultados foram positivos, mas ficou clara a necessidade de mais estudos com conjunto de dados maiores e com investigações mais profundas das observações proporcionadas.



**Palavras-chave**: aprendizagem de máquina; explicabilidade; XAI; ecologia; metagenômica; microbiologia

## ABSTRACT

Machine learning has become increasingly popular and the need to better understand the inner workings of the models developed in the field makes itself clearer day by day. Among the existing resources to better understand said models its SHAP (*Shapley Additive Explanations*), an approach to explain machine learning models by computing the importance of the features that compose the data. One of the many uses of this tool its to better our understanding not only of the decision taken by the models, but also of the importance of features, and what they represent, in the context of the data's domain. In this paper, we aim to use this tools to assess the viability of this kind of approach in the field of soil microbial ecology, investigating if machine learning models are capable of classifying metagenomic data and if the analysis of these models is helpful to our understanding of microbial diversity and capable of providing insights that traditional approaches are not. This type of data is characterized by high dimensionality and by the high cost in acquiring new samples, making this problem particularly challenging in the context of machine learning. Therefore, it was also evaluated the effects of projecting the data, for dimensionality reduction, to the obtained results. The results made clear that the dimensionality reduction gave no observable benefit. On the other hand, without projection, the results were promising, albeit limited by the reduced sample size. The explainability analysis showed a small relation with more traditional ones, but were also capable of highlighting features otherwise ignored. Overall, the results were positive, but the need for bigger datasets remained clear. Also, the need to pair these analysis with more in depth investigations became apparent, if we are to validate this approach with more confidence.



**Keywords**: machine learning; explainability; XAI; ecology; metagenomic; microbiology

**SUMÁRIO**


| | | |
|---|---|---|
| **1** | **INTRODUÇÃO** | 6 |
| **1.1** | **Objetivos Gerais** | 7 |
| **1.2** | **Objetivos Específicos** | 7 |
| **2** | **TRABALHOS RELACIONADOS** | 8 |
| **3** | **FUNDAMENTAÇÃO TEÓRICA** | 9 |
| **3.1** | **Sobre o conjunto de dados** | 9 |
| **3.2** | **A projeção dos dados** | 10 |
| **3.3** | **Os modelos** | 10 |
| **3.4** | **Shapley Additive Explanations (SHAP)** | 12 |
| **4** | **METODOLOGIA** | 14 |
| **4.1** | **Normalização** | 14 |
| **4.2** | **Otimização de hiperparâmetros e avaliação dos modelos** | 14 |
| **4.3** | **Valores SHAP** | 16 |
| **5** | **RESULTADOS E DISCUSSÃO** | 18 |
| **5.1** | **Avaliação dos Modelos** | 18 |
| **5.2** | **Discussão sobres os modelos e os valores SHAP** | 20 |
| **6** | **CONCLUSÕES E TRABALHOS FUTUROS** | 34 |
| | **REFERÊNCIAS** | 35 |
| | **APÊNDICES** | 37 |
| | **APÊNDICE A –MATRIZES DE CONFUSÃO** | 38 |
| **A.1** | **Matrizes de Confusão** | 38 |
| ***A.1.1*** | ***Processos Gaussianos*** | 38 |
| ***A.1.2*** | ***Árvore de Decisão*** | 39 |
| ***A.1.3*** | ***KNN*** | 40 |
| ***A.1.4*** | ***MLP*** | 41 |
| | **ANEXOS** | 42 |
| | **ANEXO A –ANEXO A** | 43 |

# 1 INTRODUÇÃO

Modelos de aprendizagem de máquina são notoriamente complicados de interpretar e, frequentemente, são classificados como caixas-pretas. Com a crescente influência desses modelos no nosso dia-a-dia, a necessidade por ferramentas que nos auxiliem a interpretar esses modelos vem se tornando cada vez mas aparente (O'NEIL, 2016). Nesse contexto, a chamada inteligência artificial explicável (*explainable AI*, XAI) se torna cada vez mais relevante.

Em muitos contextos, compreender as decisões tomadas por um modelo, é de extrema importância para identificar possíveis vieses na modelagem do problema em questão ou até mesmo nos dados utilizados. Contudo, outra oportunidade interessante que explicar modelos de aprendizagem de máquina nos proporciona é compreender melhor a natureza dos dados analisados. No campo da XAI, uma ferramenta de explicabilidade bastante popular é a SHAP (*SHapley Additive exPlanations*). Neste trabalho, buscamos investigar a aplicabilidade das análises realizadas por essa ferramenta para descrever o conjunto de dados analisado pelos modelos testados.

Em particular, os dados avaliados são de metagenômica de solo, em um estudo de ecologia microbiana. O metagenoma é o conjunto de genomas das comunidades microbianas em um determinado ambiente. Esses dados se caracterizam por ter um número bastante reduzido de amostras e uma dimensionalidade muito elevada. É sabido que aplicações de aprendizagem de máquina são notórias por sua necessidade por uma elevada quantidade de dados (SUN *et al.*, 2017). Isso as torna particularmente desafiadoras de por em prática em domínios onde a obtenção de amostras é custosa demais e a construção de grandes conjuntos de dados é dificultada. Além disso, certos domínios são naturalmente representados por dados de alta dimensionalidade que, para muitas abordagens, podem proporcionar um custo computacional muito elevado para serem viáveis (ALTMAN; KRZYWINSKI, 2018). Em casos em que ambas essas situações são verdade, a utilização de técnicas de aprendizagem de máquina pode ser extremamente desafiadora.

Portanto, neste trabalho, avaliamos uma miríade de modelos de aprendizagem de máquina, associados a técnicas de redução de dimensionalidade, e executamos uma análise de explicabilidade neles. Essas análises foram comparadas com análises mais tradicionais deste mesmo conjunto de dados, realizadas no trabalho que originalmente o publicou (ARAUJO *et al.*, 2021). Isso foi feito para que possamos averiguar se análises de explicabilidade são similares às análises mais tradicionais ou se são capazes de captar aspectos acerca dos dados que outras análises deixam passar.

Ao fim, observamos que, apesar de possuírem certa correlação com as análises mais tradicionais, a explicação dos modelos muitas vezes fornece uma perspectiva distinta dos dados. Além disso, também vemos que essas análises geram uma quantidade muito grande de informação, proporcionando amplo material para a análise. Entretanto, também ficou claro que apesar de os modelos terem uma performance razoável com esse conjunto de dados pequeno e de alta dimensionalidade, a redução de dimensionalidade não apresentou benefícios. Assim, a construção de conjuntos de dados maiores ainda se faz extremamente necessária para a obtenção de resultados mais satisfatórios.

## 1.1 Objetivos Gerais

1. Investigar a viabilidade de técnicas de aprendizagem de máquina para a classificação de amostras de metagenoma de solo;
2. Avaliar a aplicabilidade de técnicas de explicabilidade na descrição do microbioma do solo.

## 1.2 Objetivos Específicos

1. Treinar um conjunto de modelos de aprendizagem de máquina com o conjunto de dados em questão e avaliar sua performance na classificação binária de amostras de metagenômica de solo;
2. Avaliar o impacto da redução de dimensionalidade via análise fatorial nessa classificação;
3. Computar valores SHAP locais e globais, interpretar seus resultados e comparar com as análises realizadas sobre o mesmo conjunto de dados no trabalho que o originalmente publicou.

# 2 TRABALHOS RELACIONADOS

As oportunidades e desafios relacionados a utilizar ferramentas de aprendizagem de máquina para analisar dados de metagenômica já é algo bem sedimentado (SABARIA SARAVANAN S, 2024). Entretanto, muitos dos trabalhos sobre o assunto focam na área das ciências médicas. Essa área, por sua relevância, possui uma maior diversidade de trabalhos e não é custoso encontrar trabalhos recentes, inclusive fora do escopo da metagenômica.

O trabalho "Exploratory factor analysis yields grouping of brain injury biomarkers significantly associated with outcomes in neonatal and pediatric ECMO"(HUANG *et al.*, 2024), por exemplo, utiliza dados projetados via análise fatorial e modelos de aprendizagem de máquina para classificar conjuntos de biomarcadores de pacientes, de acordo com o prognóstico em relação ao resultado de um tipo de tratamento. Já em um contexto mais próximo do presente trabalho, o estudo "Nanopore- and AI-empowered metagenomic viability inference"(UREL *et al.*, 2024) utiliza redes neurais profundas e técnicas de explicabilidade para tentar diferenciar micro-organismos viáveis de micro-organismos mortos em amostras de metagenoma.

Contudo, neste trabalho buscamos enfatizar a interpretação das análises de explicabilidade. Apesar de alguns artigos encontrados enfatizarem a importância da explicabilidade, por vezes, os resultados são abordados em termos quantitativos, com escores de interpretabilidade por parte dos usuários e hierarquização dos modelos em termos de interpretabilidade (GHANNAM; TECHTMANN, 2021). Assim, nossa abordagem foi mais similar ao trabalho "Interpretable and accurate prediction models for metagenomics data"(PRIFTI *et al.*, 2020). Apesar de propor um algoritmo e fazer uma análise em um grande número de conjuntos de dados, esse estudo faz uma análise mais aprofundada em um único conjunto de dados e elucida observações importantes acerca da composição do microbioma estudado. No entanto, é importante lembrar que o estudo previamente mencionado lida primariamente com dados relativos à saúde humana, sendo mais incomum encontrar estudos desse tipo em conjuntos de dados relativos à ecologia microbiana do solo, como no caso deste trabalho.

# 3 FUNDAMENTAÇÃO TEÓRICA

Para a total compreensão deste trabalho, é importante o entendimento acerca dos dados e do domínio da aplicação. Além disso, também é benéfico explorarmos a metodologia escolhida, os modelos avaliados e a ferramenta de explicabilidade utilizada para explorar o que os modelos aprenderam, portanto, abordamos esses assuntos a seguir.

## 3.1 Sobre o conjunto de dados

Primeiramente, é importante contextualizar bem a problemática envolvida. Os dados utilizados nesse trabalho foram publicados no artigo "*Distinct taxonomic composition of soil bacterial community across a native gradient of Cerrado-Ecotone-Caatinga*"(ARAUJO *et al.*, 2021) e cedidos pelos autores para uso. O trabalho em questão, compara os ecossistemas referidos com base na composição taxonômica no microbioma do solo. O microbioma é o conjunto dos táxons de micro-organismos presentes num determinado ambiente, nesse caso, no solo. Os ambientes avaliados são os solos da Caatinga, do Cerrado e do Ecótono Caatinga-Cerrado. Um ecótono é o ambiente caracterizado como o gradiente de transição entre dois ecossistemas (ODUM; BARRETT, 2005).

Os dados consistem em 16 amostras, pertencentes à três classes distintas. As classes são Cerrado (6 amostras), Caatinga (5 amostras) e Ecótono (5 amostras). Originalmente, cada amostra é definida por 7521 atributos, correspondentes a cada um dos organismos observados em todas as amostras. O valor de cada atributo corresponde a quantas observações desse organismo foram feitas na amostra. Caso um organismo não seja observado em uma amostra, o valor do atributo correspondente é 0 e, como essa ocorrência é muito comum, o conjunto de dados é altamente esparso.

A observação dessas amostras é feita via sequenciamento de amostras de DNA, extraídas das amostras de solo coletadas. Isso é feito pois estima-se que até 99% da diversidade microbiana total é não-cultivável (MARTINY, 2019), isto é, não pode ser isolada por técnicas de cultura tradicionais. Dessa forma, o sequenciamento permite uma observação mais precisa da diversidade microbiana. Entretanto, exatamente por ser capaz de observar organismos que técnicas de cultivo tradicionais não são, muitos dos organismos observados por essa técnica nunca foram completamente descritos e, portanto, não existe uma classificação taxonômica completa para todos os organismos observados.

A classificação taxonômica mais precisa conhecida para o organismo associado a cada atributo é obtida consultando-se o banco de dados SILVA (QUAST *et al.*, 2012). A classificação taxonômica permite agregar atributos de mesma classificação, a depender do nível taxonômico escolhido. Entretanto, como nem todas as classificações são completas, atributos cuja classificação é desconhecida no nível taxonômico escolhido para agregação precisam ser agrupados sob um mesmo atributo, representando grupos desconhecidos. Para esse trabalho, os atributos foram agregados sob a classificação de gênero. Assim, os 7521 atributos originais foram agrupados em 352 gêneros distintos e um atributo que agrega todos os que não possuíam gênero conhecido. Ao fim, o conjunto de dados consiste em 16 amostras com 353 atributos cada.

## 3.2 A projeção dos dados

Além disso, para reduzir o custo computacional, os dados foram projetados via análise fatorial em 14 dimensões. A análise fatorial é um modelo probabilístico com variáveis latentes contínuas (BISHOP, 2006, Capítulo 12). A motivação para esse tipo de modelo advém do fato de que, apensar de os dados observados muitas vezes serem multidimensionais e terem dimensionalidade muito elevada, as observações podem se encontrar muito mais próximas em espaços projetados de dimensionalidade inferior. A projeção dos dados em um espaço projetado de dimensionalidade menor facilita a manipulação e diminui o custo computacional de trabalhar com conjuntos de dados de dimensionalidade muito elevada, como no caso deste trabalho.

Esse modelo funciona atribuindo-se uma distribuição Gaussiana à priori ao espaço latente $p(z)$ e uma distribuição condicional Gaussiana $p(x|z)$ aos dados observados. Ambas unidas por uma relação linear. Os parâmetros do modelo são obtidos via máxima-verossimilhança, utilizando-se do algoritmo de *Expectation-Maximization*. Por fim, é possível projetar-se os dados fazendo amostragens da distribuição condicional $p(z_i|x_i)$.

## 3.3 Os modelos

Para tentar modelar esses dados, tanto projetados como em sua forma original, foi testada uma miríade de modelos com abordagens distintas. Isso foi feito para tentar buscar alguma abordagem que se adéque bem aos dados, maximizando as chances de um resultado que pudesse ser aproveitado. Dentre os modelos testados há um modelo linear, um modelo não linear, um modelo de árvore, um modelo probabilístico e uma rede neural.

O modelo linear escolhido foi a regressão logística, que é um modelo que utiliza a regressão linear para classificação binária. Os dados são modelados como uma função linear e a função logística é aplicada ao resultado. O modelo pode ser atualizado utilizando-se o algoritmo de gradiente descendente ou gradiente descendente estocástico (MURPHY, 2012, Capítulo 8). É um modelo simples e de custo computacional relativamente reduzido, em comparação com outros modelos.

Também foi testado o modelo conhecido como máquina de vetor suporte (*support vector machine*, SVM) que, assim como a regressão logística, busca separar linearmente os dados classificados. Ao separar linearmente os dados, frequentemente observamos que a fronteira de separação pode ficar mais próxima de uma das classes, devido a dispersão dos dados. Para minimizar esse efeito, pode-se focar nos dados próximos à margem de separação, chamados de vetores de suporte, e formular o modelo como uma maximização dessa margem. Além disso, esse classificador linear pode ser transformado em um classificador não-linear ao utilizar uma função de *kernel* para gerar novos atributos (MURPHY, 2012, Capítulo 14). Assim, o SVM com um *kernel* não linear foi o modelo não linear testado nesse trabalho.

A classificação via K-vizinhos mais próximos (*K-nearest neighbors*, KNN), é bastante objetiva. A distância entre a amostra a ser classificada e cada elemento do conjunto de treino é calculada e a classe escolhida é a mais comum entre os k vizinhos mais próximos. As principais escolhas a serem feitas em relação a esse modelo são a métrica de distância a ser utilizada e o valor de k. Para evitar empates, é aconselhável que k seja um valor ímpar (MURPHY, 2022, Capítulo 16). Apesar de ser bastante simples, esse tipo de modelo apresenta uma boa performance. Entretanto, seu custo computacional cresce com o tamanho do conjunto de treino. Porém, como o conjunto de dados abordado nesse trabalho possui uma quantidade relativamente reduzida de amostras, é um dos modelos avaliados.

Árvores de decisão são modelos que se baseiam no particionamento dos dados para realizar uma classificação (BISHOP, 2006, Capítudo 14). Particionamentos nos dados são criados escolhendo-se atributos e limites para os valores desses atributos. Cada particionamento pode ser recursivamente subdividido com base em novas escolhas para atributos e limites. Esse tipo de modelo é completamente dependente da heurística de particionamento escolhida para criar a árvore de decisão. O objetivo é criar a menor árvore possível com o melhor particionamento. Ao fim da construção do modelo, cada nó interno da árvore representa um particionamento com base em algum atributo e cada folha representa uma classificação, normalmente representada pela

classe mais comum entre as observações inclusas neste particionamento. Como esses modelos muitas vezes podem fazer classificação com base em poucos atributos e, em conjuntos de dados com dimensionalidade muito alta, podem auxiliar a compreender a relevância dos atributos, foi um dos incluídos na análise aqui realizada.

Além disso, foi avaliado um modelo probabilístico, os Processos Gaussianos. Estes são um tipo de modelo que, ao entender que uma observação $x_i$ é associada a uma saída $y_i$ por uma função somada a um ruído de observação, na forma $y_i = f(x_i) + \varepsilon_i$, modela os dados em termos de distribuições sobre o espaço das possíveis funções $f(x_i)$. Essa abordagem pode ser utilizada para classificação binária ao aplicar a função sigmoide à função $f(x_i)$. Essa abordagem possui uma base teórica bastante sólida, já que sua formulação é intimamente associada à distribuição Gaussiana (RASMUSSEN; WILLIAMS, 2005).

Por fim, também avaliamos um tipo de rede neural simples, o perceptron multicamadas (*multilayer perceptron*, MLP). O MLP é um modelo composto por diversas camadas, cada camada consistindo em uma determinada quantidade de unidades chamadas de perceptron (MURPHY, 2022, Capítulo 13). As camadas entre a camada de entrada, que recebe os dados, e a camada de saída, que retorna o resultado do modelo, são chamadas de camadas ocultas. Cada unidade processa os valores que recebe com uma função de ativação. Nas camadas ocultas, essa função geralmente é não linear. Nesse trabalho, a função de ativação das camadas ocultas é a função ReLU (*rectified linear unit*), que consiste em: $f(x) = max(0,x)$. Nesse tipo de modelo, a função de ativação na camada de saída depende do tipo de problema. Em problemas de classificação binária, como no caso desse trabalho, a função é utilizada é a função sigmoide: $\sigma(z) = \frac{1}{1+e^{-z}}$. Esse tipo de modelo é capaz de aproximar qualquer função contínua (HORNIK *et al.*, 1989) [1], portanto, é uma escolha natural para lidar com problemas de alta complexidade.

## 3.4 Shapley Additive Explanations (SHAP)

Uma vez decididos os modelos a serem investigados, é importante lembrar que a interpretação das decisões tomadas por um determinado modelo é muitas vezes uma tarefa complexa. Para lidar com esse aspecto da aprendizagem de máquina, o campo da explicabilidade de modelos tem se tornado cada vez mais relevante, fornecendo diversas ferramentas para auxiliar na interpretação das escolhas feitas por modelos de aprendizagem de máquina (HASSIJA *et al.*, 2024). Uma dessa ferramentas é o *framework* de explicabilidade SHAP (*Shapley Additive*

[1] Apesar de não ser necessariamente computacionalmente viável

*Explanations*) (LUNDBERG; LEE, 2017).

A forma como o SHAP funciona é imputando a cada atributo dos dados um valor de importância. Isso é feito ao treinar o modelo com e sem esse atributo e avaliar como isso afeta a saída do modelo. Utilizando-se esse valores, é possível ter uma melhor compreensão da significância de cada um dos atributos na tomada de decisões do modelo. Computar os valores SHAP de um conjunto de dados nos dá o que chamamos de SHAP global, um conjunto de valores SHAP que consiste no valor médio para cada atributo, em todas as amostras. O SHAP global nos ajuda a compreender o que o modelo foi capaz de extrair do conjunto como um todo. Ao computar os valores SHAP de uma única amostra, temos o chamado SHAP local, que nos permite uma maior compreensão sobre como os atributos afetam a predição do modelo para aquela amostra. Assim, os valores globais e locais nos permitem não só uma melhor compreensão de como o modelo interpretou os dados, mas também nos fornece informações para desenvolver uma melhor compreensão da natureza dos dados sendo estudados.

# 4 METODOLOGIA

Para serem inseridos nos modelos, os dados foram normalizados e, para cada uma das classes, foi criado um conjunto de rótulos para classificação binária. Isto é, um conjunto de rótulos que identifica todas as amostras como pertencendo à uma classe ou não, para todas as classes, totalizando três conjuntos de rótulos. Então, fez-se um processo de validação cruzada aninhada, utilizando a estratégia de divisão *Leave-One-Out* (LOO), para otimização de hiperparâmetros, avaliação dos modelos e cálculo dos valores SHAP locais. Além disso, uma validação cruzada simples, também com estratégia de divisão *Leave-One-Out*, foi realizada para otimizar hiperparâmetros e calcular os valores SHAP globais. Esse processo foi repetido para cada combinação de modelo e rótulo.

## 4.1 Normalização

Para evitar problemas numéricos, os dados foram utilizados na notação logarítmica. Como mencionado no Capítulo 3 o conjuntos de dados é esparso, então, para possibilitar o cálculo do logaritmo de todos os dados, uma constante foi adicionada a todas as amostras. Assim, os dados foram processados da seguinte forma: $X = \ln(X+1)$.

Para o treinamento dos modelos, os dados foram centralizados em torno da média e escalonados para variância unitária, de acordo com a fórmula: $X = \frac{X-\mu}{\sigma}$. Onde $\mu$ é a média dos dados e $\sigma$ o desvio padrão.

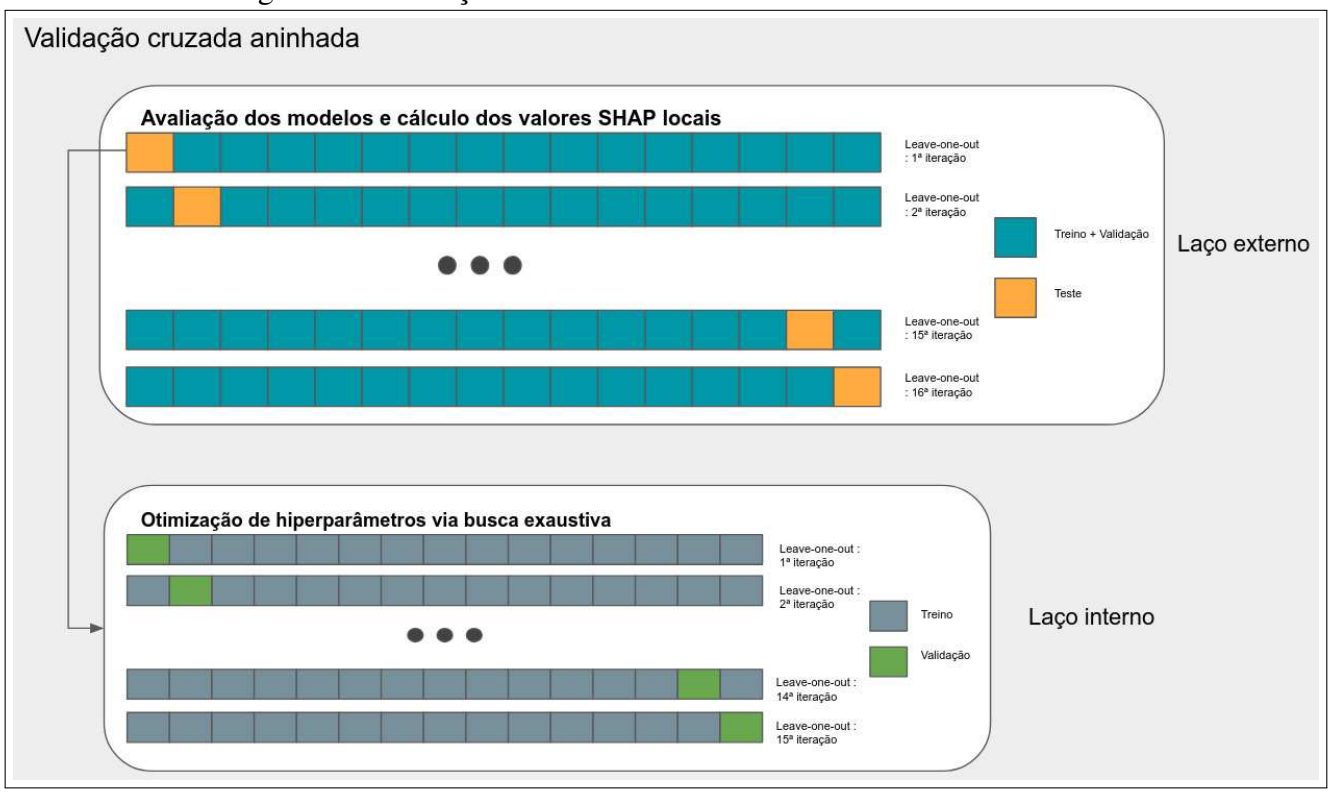
## 4.2 Otimização de hiperparâmetros e avaliação dos modelos

A otimização de hiperparâmetros foi feita utilizando a estratégia de busca exaustiva, em que todas as possíveis combinações dos hiperparâmetros considerados são testadas. A otimização e o treinamento dos modelos foram realizados utilizando a validação cruzada aninhada (Figura 1). A estratégia de divisão escolhida foi a *Leave-One-Out* (LOO), onde o modelo é treinado com todas as amostras exceto uma, que é utilizada para teste. O processo é repetido com todas as amostras como conjunto de teste. Assim, os dados são divididos em um conjunto de treino e um de teste. Esse conjunto de treino é utilizado para treinar e validar o modelo, utilizando a mesma estratégia de divisão, uma vez para cada conjunto de hiperparâmetros. Os hiperparâmetros com melhor métrica de avaliação são utilizados para treinar e avaliar o modelo, considerando a divisão inicial dos dados, e o processo é repetido com todas as amostras como

conjunto de teste.

Devido à estratégia de divisão escolhida para a validação cruzada, cada iteração do laço de treino só é avaliada com uma predição. Dessa forma, não é possível calcular métricas como precisão, revocação e F1-score de maneira convencional. Por isso, a métrica escolhida para a avaliação foi a acurácia. Apesar de não terem validade para avaliar a capacidade de generalização dos modelos, também foram calculadas versões adaptadas da precisão, revocação e F1-Score, para auxiliar na interpretação do desempenho dos modelos. Para esse cálculo, a contagem das predições feitas foi realizada ao longo das iterações do laço externo da validação cruzada aninhada. Cada um desses modelos, treinados em cada iteração do laço externo, foi salvo utilizando o *MLflow* (ZAHARIA *et al.*, 2018). Além dos modelos treinados serem salvos, também foram armazenados os conjuntos de hiperparâmetros utilizados.

Figura 1 – Validação cruzada aninhada




Fonte: elaborada pelo autor.

Além dessa validação cruzada aninhada, feita para avaliar os modelos, também foi feita uma busca exaustiva de hiperparâmetros ótimos, em uma validação cruzada simples, para calcular os valores SHAP globais dos modelos. Esses modelos também foram salvos utilizando o *MLflow*.

Apesar de o conjunto de dados ter três classes, o problema foi convertido em três classificações binárias, gerando-se um conjunto de rótulos para cada uma das classes. Portanto, cada modelo teve os hiperparâmetros otimizados, foi treinado e foi avaliado para cada uma das classificações binárias. Além disso, todo o processo foi repetido tanto com os dados originais quanto com os dados projetados via análise fatorial.

Os modelos testados foram a Regressão Logística, K-Vizinho Mais Próximos (KNN), Máquinas de Vetor Suporte (SVM), Árvores de Decisão, um Perceptron Multicamadas (MLP) e Processos Gaussianos. O gerenciamento e armazenamento dos modelos treinados e dos resultados obtidos foi feito utilizando o *MLflow* , uma plataforma para a facilitação do ciclo de vida de projetos de aprendizagem de máquina.

Tabela 1 – Hiperparâmetros investigados na busca exaustiva

| Modelo | Parâmetros testados |
|---|---|
| Regressão Logística | C={0,01; 0,1; 1; 1,25; 1,5; 1,75};tol={$10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$} |
| SVM | C={0,01; 0,1; 1; 1,25; 1,5; 1.75};kernel={"rbf", "sigmoid"} |
| KNN | n_neighbors={1, 3, 5};weights={"uniform", "distance"};p={1, 2, 3} |
| Árvore de Decisão | criterion={"gini", "entropy", "log_loss"}; splitter={"best", "random"} |
| Processos Gaussianos | kernel={"rbf", "matern5/2", "matern3/2", "matern1/2"} |
| MLP | hidden_layer_sizes={(75,),(100,), (125,)};alpha={0,0001; 0,01; 0,1; 1} |

Fonte: elaborada pelo autor.

### 4.3 Valores SHAP

Como o conjunto de dados avaliado é pequeno, os valores SHAP locais foram calculados para cada uma das iterações do laço externo da validação cruzada aninhada. Isto é, para cada conjunto de hiperparâmetros ótimo encontrado, o modelo foi treinado com o conjunto de treino e os valores SHAP foram computados para o conjunto de teste.

Já os valores SHAP globais, foram calculados utilizando os hiperparâmetros encontrados na segunda busca exaustiva, utilizando o conjunto de dados inteiro. É importante ressaltar que os valores de SHAP globais não foram computados para avaliar a capacidade de generalização do modelo, mas sim para explicar o que os modelos foram capazes de extrair dos dados.

Os valores SHAP calculados foram avaliados tanto em sua forma original, um valor para cada atributo, como agregando os atributos com base no filo a qual o gênero que o atributo representa pertence, facilitando a interpretação dos resultados. Além disso, também para facilitar a interpretabilidade dos gráficos, os valores das amostras, indicados como valores de referencia

nos gráficos gerados, foram desnormalizados.

# 5 RESULTADOS E DISCUSSÃO

Nas seções a seguir, apresentamos e discutimos as métricas dos modelos testados. Também apresentamos as imagens geradas tanto a partir dos modelos e suas classificações como a partir dos valores SHAP computados. Em razão do grande número de imagens geradas, nem todas estão inclusas abaixo.

## 5.1 Avaliação dos Modelos

Como podemos ver na Tabela 2, com exceção da Árvore de Decisão, todos os modelos apresentaram acurácia abaixo de 0.75 na maioria das classificações, quando analisando o conjunto de dados projetado por análise fatorial. Além disso, para uma completa interpretação dos resultados dessa análise, seria necessário compreender as relações de cada um dos fatores da análise fatorial com todos os atributos originais. Considerando a baixa performance dos modelos associada à essa camada adicional de complexidade, optou-se por não dar prosseguimento nas análises dos dados projetados com análise fatorial.

Tabela 2 – Métricas dos modelos treinados com os dados projetados via análise fatorial evidenciando a má performance da abordagem

| Modelo | Rótulo | Acurácia | Precisão | Revocação | F1-Score |
|---|---|---|---|---|---|
| Regressão Logística | Cerrado | 0,62 | 0,00 | 0,00 | 0,00 |
| SVM | Cerrado | 0,62 | 0,00 | 0,00 | 0,00 |
| KNN | Cerrado | 0,69 | 0,60 | 0,50 | 0,54 |
| Árvore de Decisão | Cerrado | **0,75** | 0,75 | 0,50 | 0,60 |
| Processos Gaussianos | Cerrado | 0,62 | 0,00 | 0,00 | 0,00 |
| MLP | Cerrado | 0,56 | 0,00 | 0,00 | 0,00 |
| Regressão Logística | Ecótono | 0,62 | 0,00 | 0,00 | 0,00 |
| SVM | Ecótono | 0,69 | 0,00 | 0,00 | 0,00 |
| KNN | Ecótono | 0,31 | 0,00 | 0,00 | 0,00 |
| Árvore de Decisão | Ecótono | **0,81** | 0,67 | 0,80 | 0,73 |
| Processos Gaussianos | Ecótono | 0,19 | 0,10 | 0,20 | 0,13 |
| MLP | Ecótono | 0,69 | 0,50 | 0,60 | 0,54 |
| Regressão Logística | Caatinga | 0,69 | 0,00 | 0,00 | 0,00 |
| SVM | Caatinga | **0,75** | 1,00 | 0,20 | 0,33 |
| KNN | Caatinga | 0,69 | 0,50 | 0,40 | 0,44 |
| Árvore de Decisão | Caatinga | 0,50 | 0,20 | 0,20 | 0,20 |
| Processos Gaussianos | Caatinga | 0,69 | 0,00 | 0,00 | 0,00 |
| MLP | Caatinga | 0,69 | 0,50 | 0,40 | 0,44 |

Fonte: elaborada pelo autor.

Ao avaliar a performance dos modelos treinados utilizando o conjunto de dados sem projeção, é importante lembrar que, como descrito na seção 4.2, a precisão, a revocação e o

f1-score foram calculados de uma maneira não tradicional. Por isso, deu-se preferência a avaliar a performance dos modelos e tomar decisões a cerca da continuação dos experimentos com ênfase na acurácia dos modelos. Não obstante, vale ressaltar que as outras métricas ainda são informativas. Assim, considerando as métricas relatadas na Tabela 3, optou-se por desconsiderar a regressão logística e o SVM das análises de valores SHAP.

Tabela 3 – Métricas dos modelos treinados com os dados sem projeção com as melhores acurácias por classe em destaque

| Modelo | Rótulo | Acurácia | Precisão | Revocação | F1-Score |
|---|---|---|---|---|---|
| Regressão Logística | Cerrado | 0,75 | 1,00 | 0,33 | 0,50 |
| SVM | Cerrado | 0,75 | 0,75 | 0,50 | 0,60 |
| KNN | Cerrado | 0,75 | 0,62 | 0,83 | 0,71 |
| Árvore de Decisão | Cerrado | **0,94** | 0,86 | 1,00 | 0,92 |
| Processos Gaussianos | Cerrado | 0,81 | 0,71 | 0,83 | 0,77 |
| MLP | Cerrado | 0,81 | 0,71 | 0,83 | 0,77 |
| Regressão Logística | Ecótono | 0,81 | 1,00 | 0,40 | 0,57 |
| SVM | Ecótono | 0,69 | 0,00 | 0,00 | 0,00 |
| KNN | Ecótono | 0,75 | 0,57 | 0,80 | 0,67 |
| Árvore de Decisão | Ecótono | 0,69 | 0,50 | 0,40 | 0,44 |
| Processos Gaussianos | Ecótono | 0,75 | 0,57 | 0,80 | 0,67 |
| MLP | Ecótono | **0,87** | 1,00 | 0,60 | 0,75 |
| Regressão Logística | Caatinga | 0,62 | 0,00 | 0,00 | 0,00 |
| SVM | Caatinga | 0,62 | 0,00 | 0,00 | 0,00 |
| KNN | Caatinga | 0,87 | 1,00 | 0,60 | 0,75 |
| Árvore de Decisão | Caatinga | **0,94** | 0,83 | 1,00 | 0,91 |
| Processos Gaussianos | Caatinga | 0,81 | 1,00 | 0,40 | 0,57 |
| MLP | Caatinga | 0,81 | 0,75 | 0,60 | 0,67 |

Fonte: elaborada pelo autor.

Observando a acurácia dos modelos restantes, podemos ver que a árvore de decisão e os processos gaussianos apresentaram uma acurácia menor na classificação binária com o rótulo de Ecótono. Isso é intuitivo e condizente com a natureza dos dados, pois o ecótono estudado é a região de transição entre o Cerrado e a Caatinga. Assim, é de se esperar que seja mais fácil diferenciar as amostras do Cerrado ou as da Caatinga das demais do que diferenciar as amostras do Ecótono das amostras das outras áreas. Além disso, excetuando o KNN, podemos observar que a acurácia das classificações binárias relativas ao Cerrado e à Caatinga são iguais. Isso indica uma adequação similar dos modelos a ambas as classificações. Vale ressaltar que a árvore de decisão e para o MLP exibiram um comportamento similar nas outras métricas.

Essa relativa consistência entre o comportamento dos modelos é positiva, pois parece indicar que os modelos, de certa forma, concordam na interpretação dos dados. Entretanto, é importante ressaltar que de maneira geral a performance dos modelos não foi muito boa. As

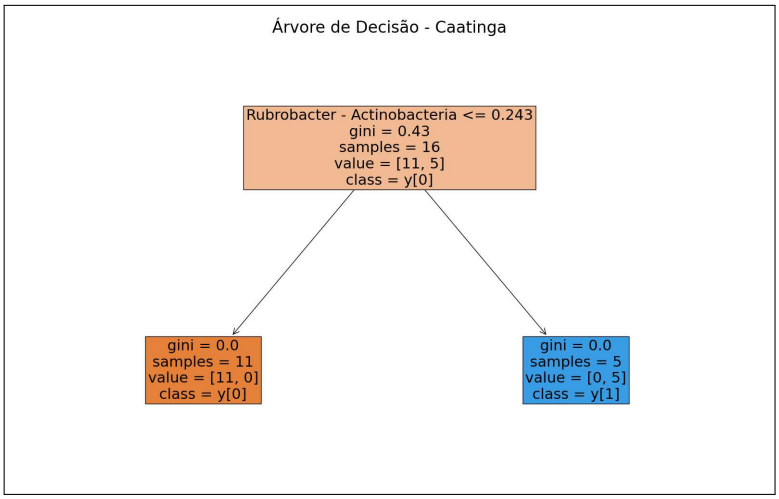
notáveis exceções são a árvore de decisão, nas classificações relativas ao Cerrado e à Caatinga, o KNN, na classificação relativa à Caatinga, e o MLP na classificação relativa ao ecótono.

As matrizes de confusão das predições utilizadas para calcular as métricas dos modelos na Tabela 3 podem ser encontradas no Apêndice A

## 5.2 Discussão sobres os modelos e os valores SHAP

Antes mesmo de prosseguirmos para a análise dos valores SHAP, já podemos fazer observações interessantes a partir dos modelos. As árvores de decisão, por exemplo, obtiveram os melhores resultados quando classificando as amostras utilizando os rótulos de Caatinga e de Cerrado.

Figura 2 – Árvore de Decisão - Caatinga




Fonte: elaborada pelo autor.

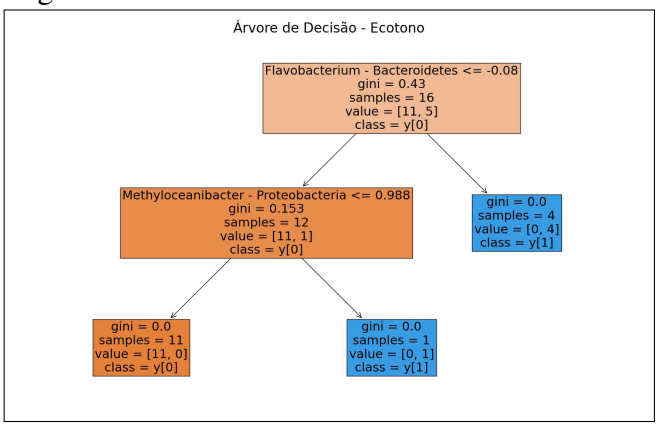
Na Figura 2, observamos a árvore de decisão que classificou as amostras como pertencendo à Caatinga ou não. Como podemos ver, com apenas um atributo, em particular, o gênero *Rubrobacter*, essa árvore de decisão é capaz de classificar as amostras. A indicação desse gênero é interessante, pois o gênero *Rubrobacter* é notório por ter membros com alta resistência à radiação (CHEN *et al.*, 2004), uma adaptação condizente com a alta incidência de radiação ultra-violeta no solo da Caatinga (SILVA *et al.*, 2024). Essa observação é importante para nosso propósito pois, como vemos que o modelo é capaz de fazer a classificação com um atributo, e esse atributo representa um gênero que é particularmente adaptado para o contexto do ecossistema em questão, podemos ter certa confiança de que a árvore de decisão é capaz de identificar grupos relevantes na descrição do microbioma da Caatinga.

É importante ressaltar que, como o modelo ilustrado acima utiliza apenas um atributo

para fazer a classificação, o gráfico de SHAP global não é particularmente informativo, uma vez que o único atributo capaz de influenciar na classificação é o atributo indicado pela árvore.

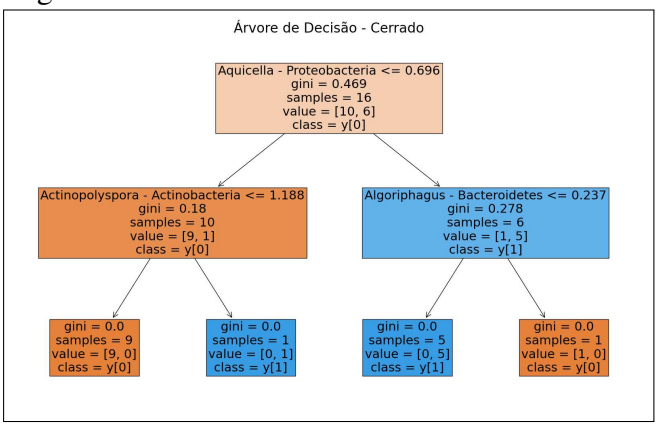
Se observarmos as árvores de decisão relativas ao Ecótono e ao Cerrado, nas figuras a seguir, vemos que, de forma similar à Figura 2, poucos atributos são utilizados na classificação. Entretanto, a relevância dos gêneros indicados não é tão imediatamente evidente nesse caso. Dessa forma, os resultados fornecidos pelas árvore de decisão podem servir como um ponto de partida para um ecólogo buscando investigar grupos importantes nos microbiomas do Cerrado e do Ecótono Caatinga-Cerrado. Vale lembrar que, enquanto as métricas de avaliação da árvore de decisão que realizou a classificação binária com Cerrado como classe positiva foram muito boas, as métricas de avaliação da árvore de decisão que realizou a mesma tarefa com o Ecótono como classe positiva foram menos satisfatórias.

Figura 3 – Árvore de Decisão - Ecótono




Fonte: elaborada pelo autor.

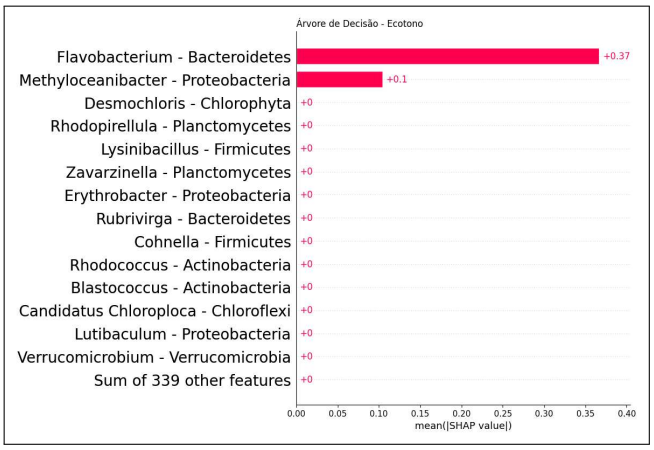
Figura 4 – Árvore de Decisão - Cerrado




Fonte: elaborada pelo autor.

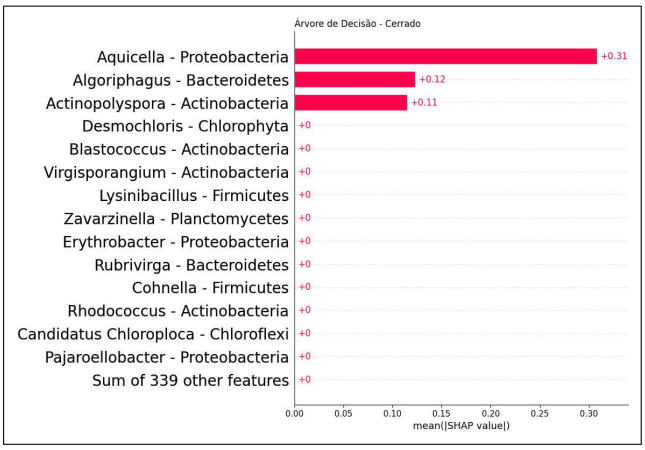
Nesses casos, o gráficos do SHAP global podem nos ajudar a hierarquizar os grupos indicados pelas árvores, como vemos nas imagens a seguir.

Figura 5 – SHAP Global - Árvore de Decisão - Ecótono




Fonte: elaborada pelo autor.

Figura 6 – SHAP Global - Árvore de Decisão - Cerrado

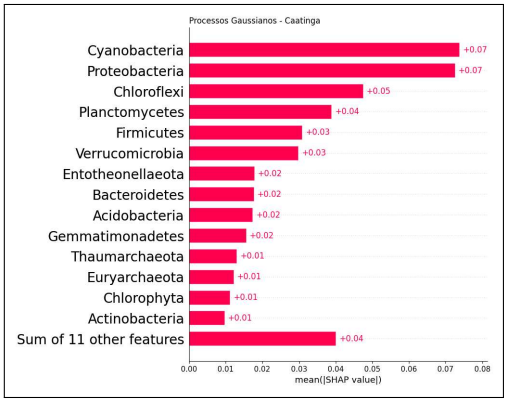



Fonte: elaborada pelo autor.

Como vimos, as árvores de decisão, por conseguirem fazer classificações com poucos atributos, podem ajudar a identificar atributos relevantes sem mesmo ser necessária uma análise de valores SHAP. Essa dependência em poucos atributos, entretanto, faz com que as análises dos valores SHAP não sejam tão ricas quanto as de outros modelos, já que poucos atributos são capazes de influenciar na decisão do modelo. Quando consideramos a análise dos valores SHAP, os outros modelos que investigamos são mais interessantes.

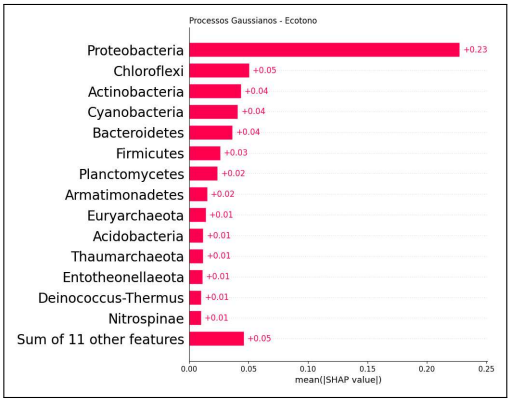
O trabalho que originalmente publicou os dados elegeu os filos Firmicutes, Proteobacteria, Acidobacteria e Chloroflexi como os quatro principais responsáveis pela dissimilaridade entre as classes, em um teste SIMPER (CLARKE, 1993). Como podemos ver nos gráficos do SHAP global para todas as classificações feitas pelos modelos remanescentes, frequentemente estes filos estão entre os dez atributos mais relevantes para a tomada de decisão. Mesmo que seja difícil traçar uma relação direta entre essas duas análises, a aparente consistência entre o teste SIMPER e a análise dos valores SHAP parece indicar uma correlação entre a contribuição de um atributo para a dissimilaridade entre as classes e a importância dele na classificação da amostra.

Figura 7 – SHAP Global - Processos Gaussianos - Caatinga



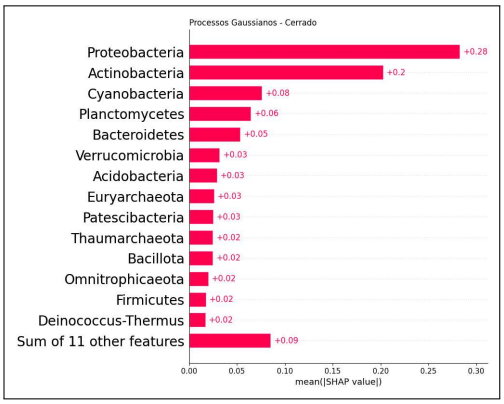

Fonte: elaborada pelo autor.

Figura 8 – SHAP Global - Processos Gaussianos - Ecótono



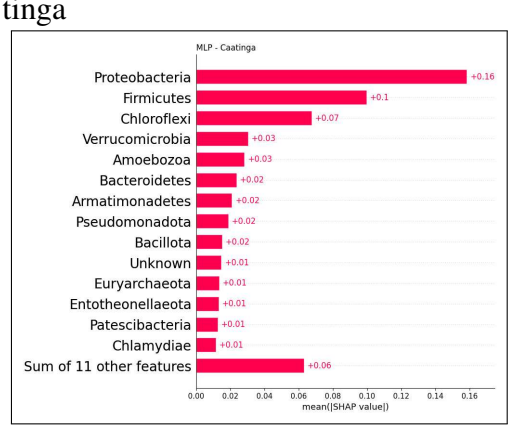

Fonte: elaborada pelo autor.

Figura 9 – SHAP Global - Processos Gaussianos - Cerrado



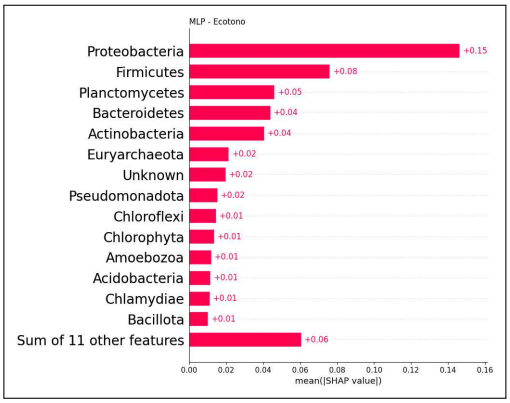

Fonte: elaborada pelo autor.

Figura 10 – SHAP Global - MLP - Caatinga



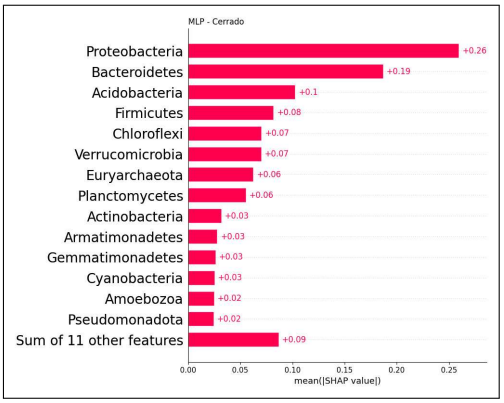

Fonte: elaborada pelo autor.

Figura 11 – SHAP Global - MLP - Ecótono



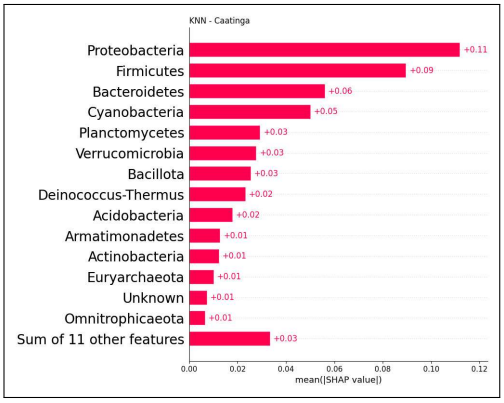

Fonte: elaborada pelo autor.

Figura 12 – SHAP Global - MLP - Cerrado



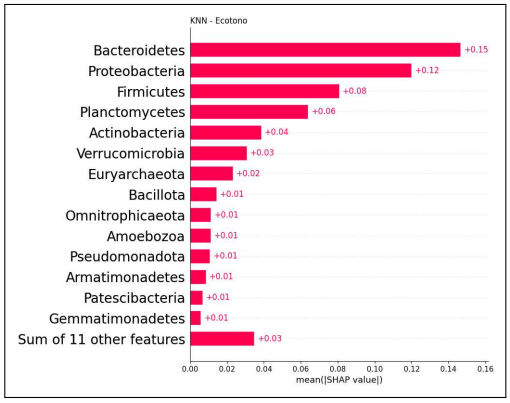

Fonte: elaborada pelo autor.

Figura 13 – SHAP Global - KNN - Caatinga



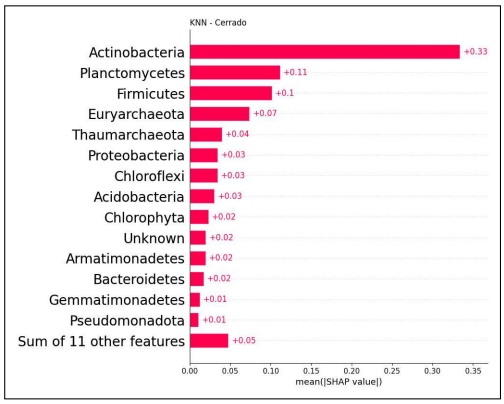

Fonte: elaborada pelo autor.

Figura 14 – SHAP Global - KNN - Ecótono




Fonte: elaborada pelo autor.

Figura 15 – SHAP Global - KNN - Cerrado

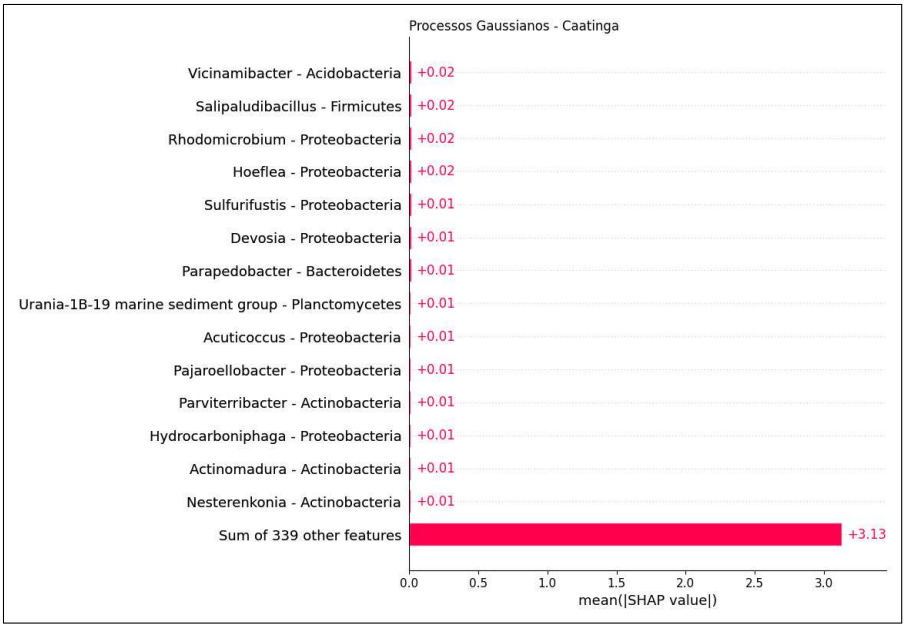



Fonte: elaborada pelo autor.

É importante ressaltar que a análise SIMPER faz uma análise da dissimilaridade entre classes, indicando quais grupos são os principais responsáveis e em qual classe eles são mais abundantes. Apesar de tentador, não é correto assumir que, por um determinado grupo ser mais abundante em uma classe e ter influência na dissimilaridade dessa classe com outras, esse atributo é importante para classificar um amostra como sendo dessa mesma classe, uma vez que a análise SIMPER é pertinente apenas para os dados, sem ter relação nenhuma com os modelos testados, que de fato classificam as amostras.

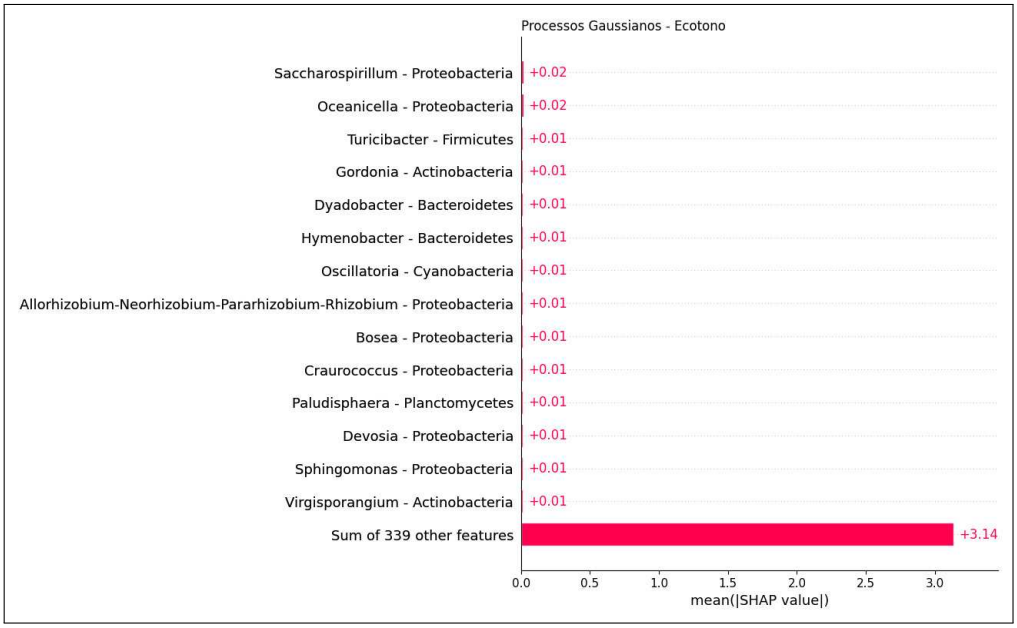
Observando os valores SHAP globais dos modelos de processos gaussianos e comparando com a análise SIMPER feitas pelo trabalho que originalmente publicou os dados (Figura 42) vemos alguns exemplos disso.

Figura 16 – SHAP Global - Processos Gaussianos - Caatinga




Fonte: elaborada pelo autor.

Figura 17 – SHAP Global - Processos Gaussianos - Ecótono

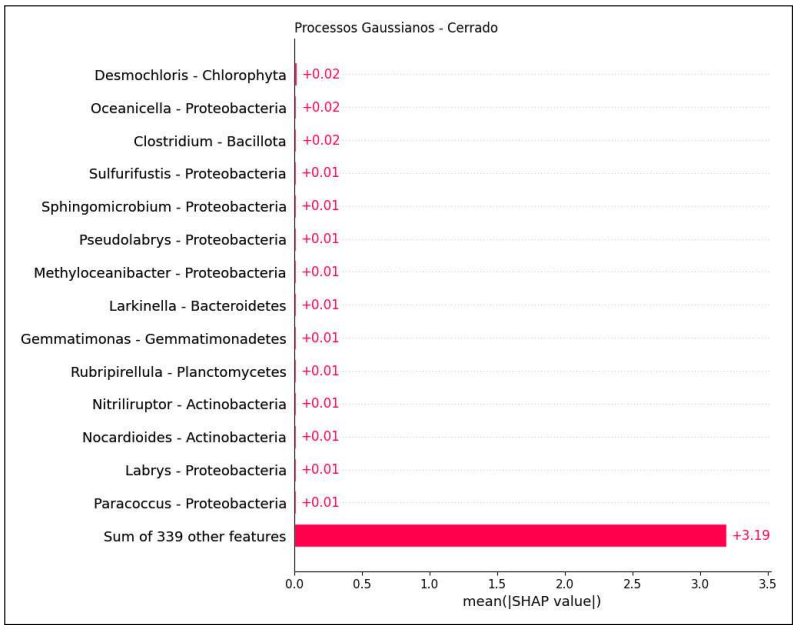



Fonte: elaborada pelo autor.

No teste SIMPER, a família *Sphingomonodaceae* é tida como um dos principais responsáveis pela dissimilaridade entre as amostras da Caatinga e as do Ecótono, sendo mais abundante nas amostras da Caatinga. No entanto, os valores SHAP globais indicam o gênero *Sphingomonas* (membro da família *Sphingomonodaceae*) como um dos atributos mais importan-

tes para a classificação de amostras como pertencentes ao Ecótono, enquanto nenhum membro dessa família é indicado como relevante pela análise SHAP na classificação binária relativa às amostras da Caatinga. O mesmo ocorre com o gênero *Sphingomicrobium* (também membro da família *Sphingomonodaceae*), quando consideramos o Cerrado e a Caatinga. Outro exemplo é o gênero *Vicinamibacter*, indicado como o atributo mais importante na classificação das amostras como pertencentes à Caatinga, pelo modelo de processos gaussianos, ao mesmo tempo que, apesar de também ser importante de acordo com o teste SIMPER, é mais abundante na amostras de solo do Cerrado.

Figura 18 – SHAP Global - Processos Gaussianos - Cerrado



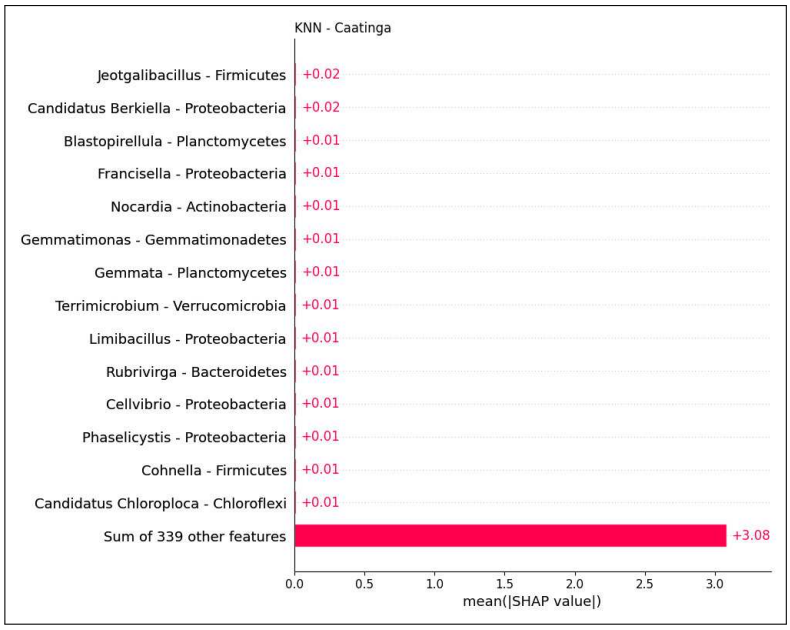

Fonte: elaborada pelo autor.

Desta forma, apesar de vermos que os grupos identificados como relevantes pelo teste SIMPER muitas vezes também apresentam valores SHAP médios elevados, fica claro que é difícil compreender uma relação clara entre essas métricas. O fato de a análise dos valores SHAP não ser de todo consistente com a análise SIMPER não a torna necessariamente inválida. Porém, sem uma base experimental sólida e com a escassez de trabalhos similares, é difícil ser assertivo sobre a validade das conclusões extraídas dessa análise.

Apesar disso, isoladamente, as análises dos valores SHAP fornecem uma grande variedade de informação, que, se válida, podem auxiliar bastante na compreensão da composição do microbioma do solo e na identificação de grupos importantes para os ecossistemas. Observando os valores SHAP globais das classificações das amostras como pertencentes à Caatinga

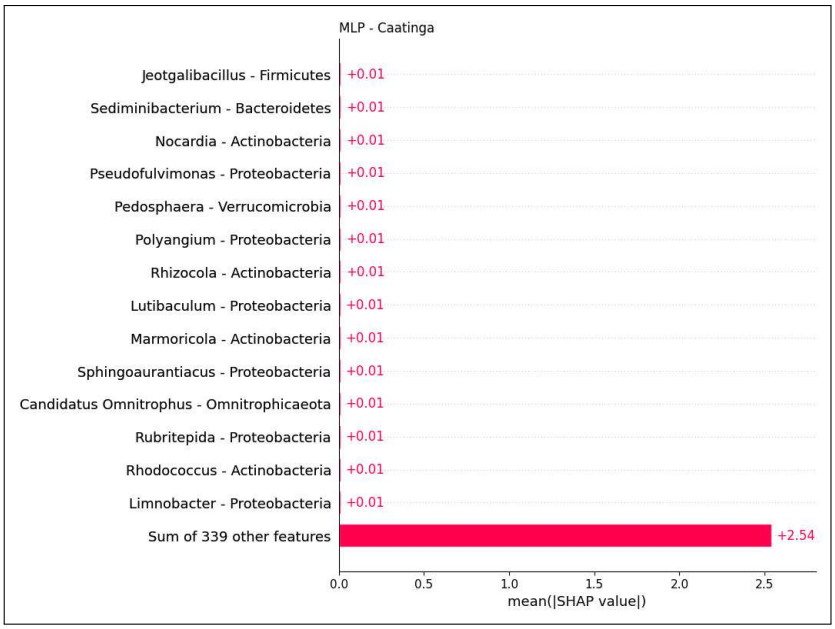
ou não, feitas pelo KNN e pelo MLP, por exemplo, pode-se ver que ambos identificaram como grupo mais importante para a classificação o gênero *Jeotgalibacillus*.

Figura 19 – SHAP Global - KNN - Caatinga




Fonte: elaborada pelo autor.

Figura 20 – SHAP Global - MLP - Caatinga

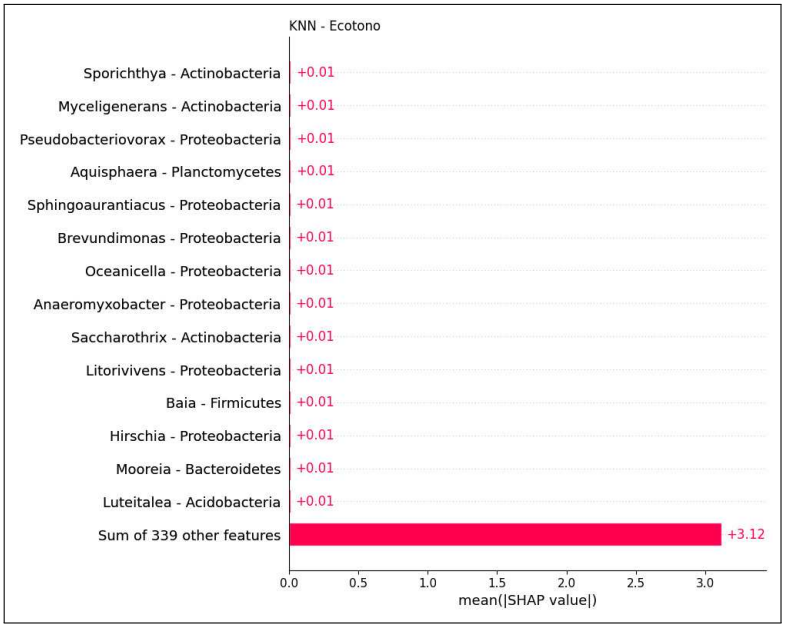



Fonte: elaborada pelo autor.

As mesmas análises também indicam o gênero *Nocardia* como relevante. Esse gênero, em particular, contém bactérias patogênicas, responsáveis pela doença nocardiose. Outro

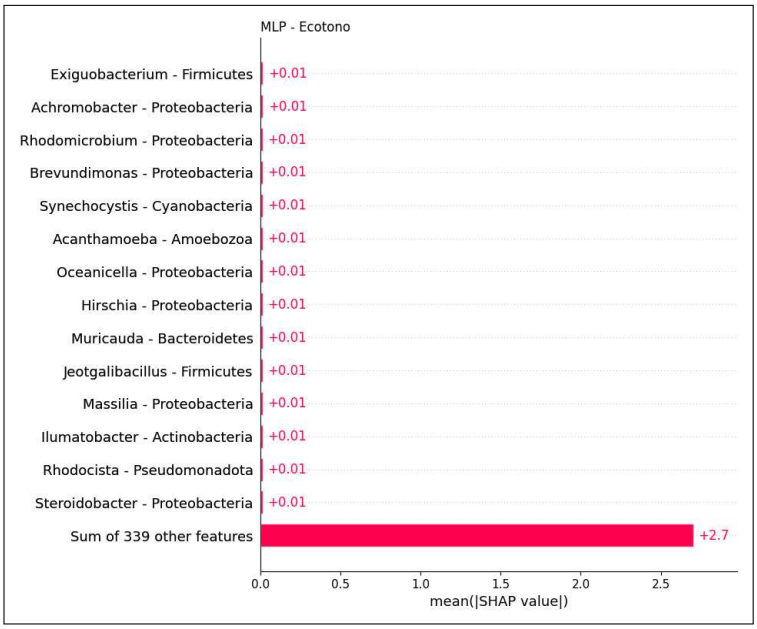
exemplo que podemos observar, está presente nas classificações relativas ao Ecótono (Figura 17, Figura 21, Figura 22). Nos gráficos, podemos ver que os processos gaussianos, o KNN e o MLP indicaram alguma relevância do gênero *Oceanicella*.

Figura 21 – SHAP Global - KNN - Ecótono




Fonte: elaborada pelo autor.

Figura 22 – SHAP Global - MLP - Ecótono

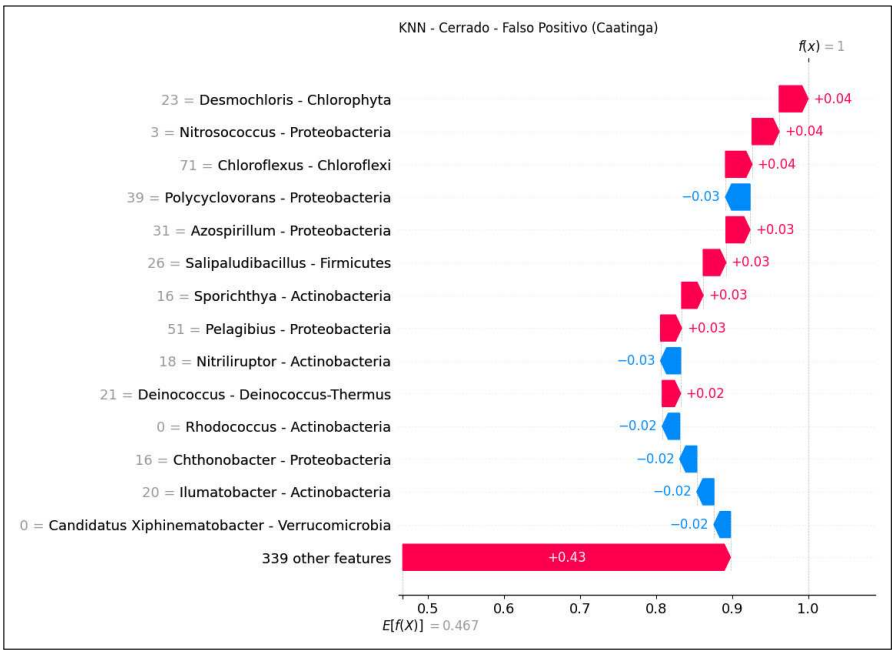



Fonte: elaborada pelo autor.

Contudo, também vale ressaltar, que o grande volume de análises não é necessaria-

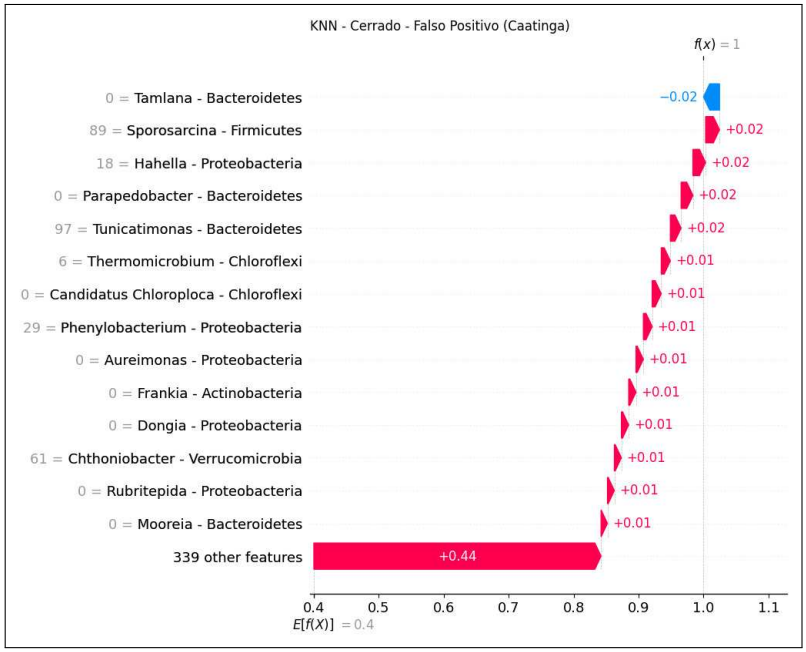
mente algo produtivo. Observando os valores SHAP locais dos falsos positivos cometidos ao tentar classificar as amostras como Cerrado ou não, poderia se esperar observar algum padrão que pudesse indicar algum conjunto de gêneros responsável pelas classificações errôneas. Entretanto, como podemos ver, há pouquíssimos grupos indicados como relvantes paras as predições que estão presentes em mais de uma predição.

Figura 23 – SHAP Local - KNN - Cerrado - Falso Positivo



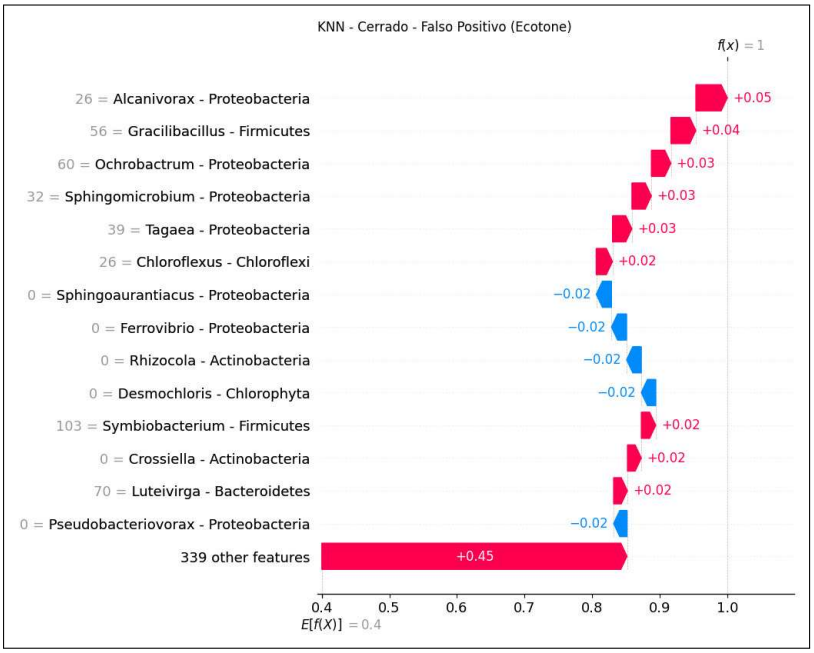

Fonte: elaborada pelo autor.

Figura 24 – SHAP Local - KNN - Cerrado - Falso Positivo



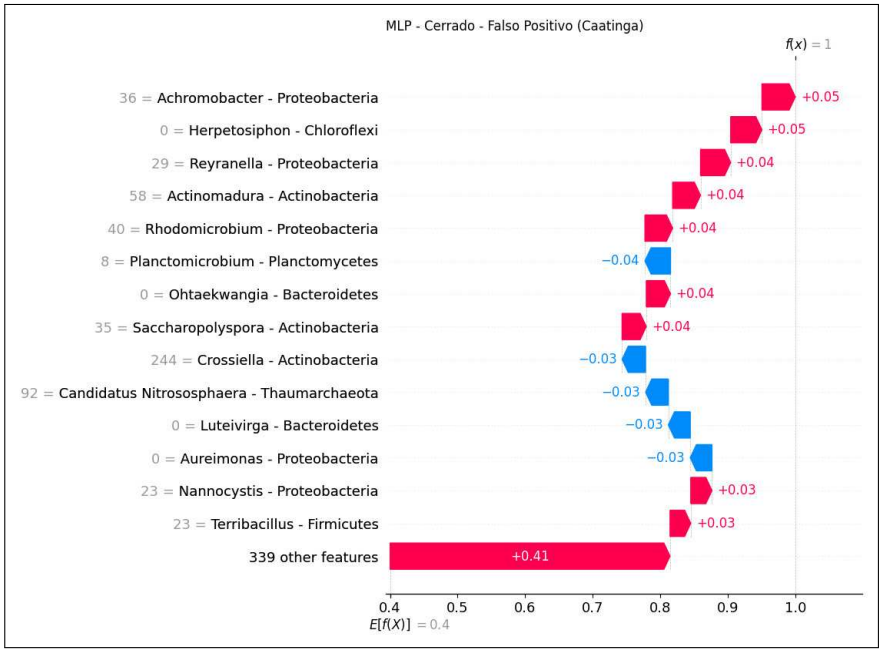

Fonte: elaborada pelo autor.

Figura 25 – SHAP Local - KNN - Cerrado - Falso Positivo



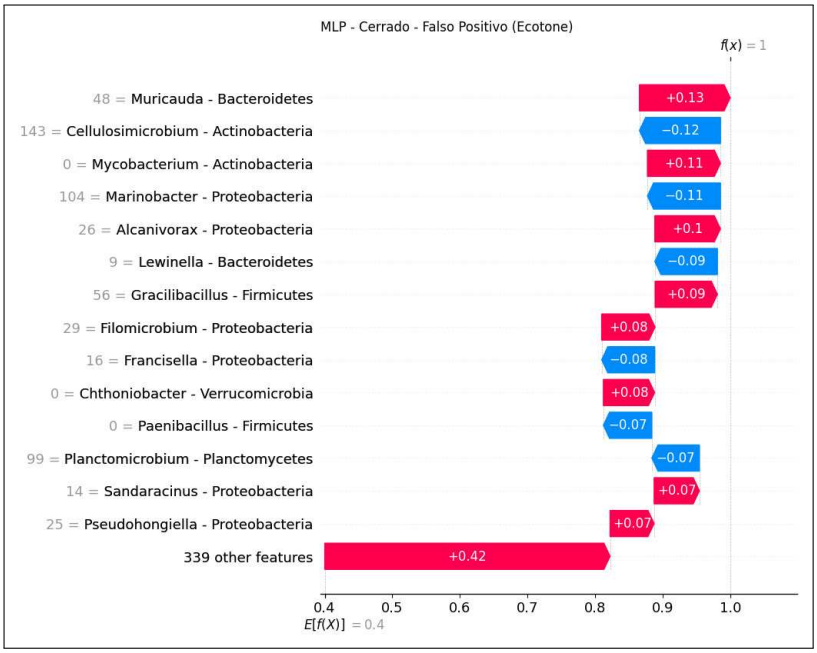

Fonte: elaborada pelo autor.

Figura 26 – SHAP Local - MLP - Cerrado - Falso Positivo



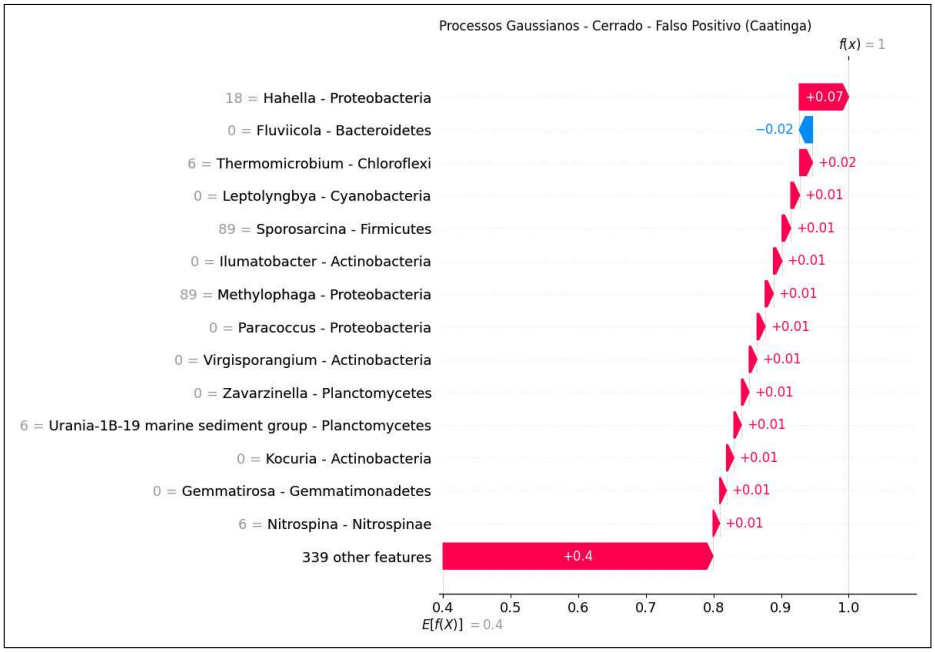

Fonte: elaborada pelo autor.

Figura 27 – SHAP Local - MLP - Cerrado - Falso Positivo



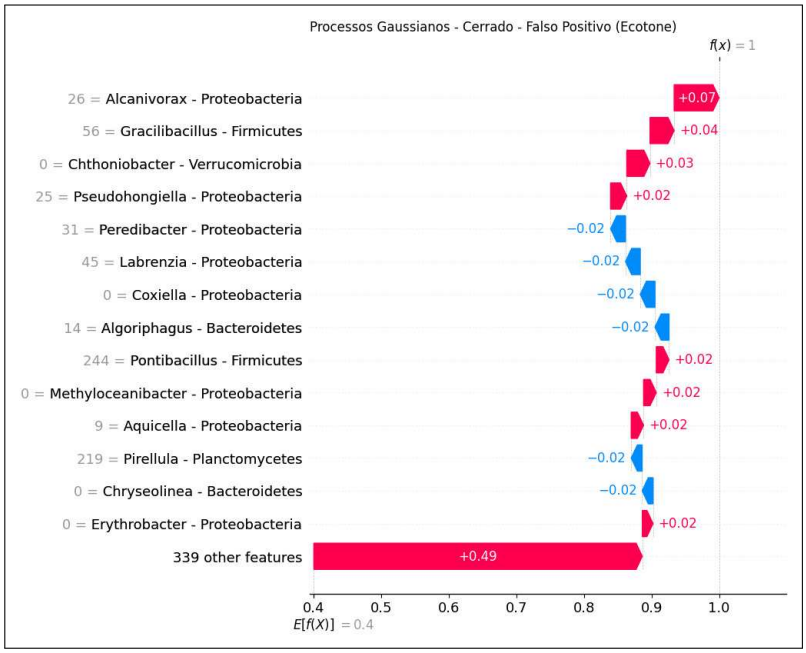

Fonte: elaborada pelo autor.

Figura 28 – SHAP Local - Processos Gaussianos - Cerrado - Falso Positivo




Fonte: elaborada pelo autor.

Figura 29 – SHAP Local - Processos Gaussianos - Cerrado - Falso Positivo




Fonte: elaborada pelo autor.

Isso provavelmente acontece pois, como discutido na seção 4.2, cada predição é feita em uma iteração do laço da validação cruzada aninhada. Assim, cada predição é feita por um modelo com hiperparâmetros diferentes treinado em um conjunto de dados ligeiramente

diferente. Consequentemente, aquilo que o modelo foi capaz de extrair do conjunto de dados é ligeiramente diferente para cada iteração, o que dificulta a análise conjunta dos resultados. Isso evidencia que, apesar de serem interessantes, as análises são prejudicadas pela falta de dados.

# 6 CONCLUSÕES E TRABALHOS FUTUROS

De maneira geral, ficou evidente que a projeção dos dados via análise fatorial, apesar de oferecer uma diminuição do custo computacional devida à redução na dimensionalidade das amostras, prejudicou os desempenhos dos modelos. Também ficou claro que a avaliação dos modelos indicou que, apesar do número reduzido de amostras, é possível ter performance razoável com esse tipo de dado. Entretanto, para ter maior confiança nas análises, modelos com um melhor desempenho são desejáveis. Por isso, faz-se necessário a realização de estudos com conjuntos de dados maiores, para determinar se a performance observada deu-se pela natureza dos dados ou pela quantidade reduzida de amostras.

Além disso, outra dificuldade encontrada na validação dessas análises, foi a comparação com as análises mais tradicionais, feitas no trabalho que originalmente publicou os dados. As análises SIMPER foram realizadas em uma configuração par a par, como mencionado no Capítulo 5, enquanto as análises aqui realizadas foram feitas em uma configuração um contra todos (*one versus rest*, ovr). A principal motivação para isso é, novamente, a quantidade reduzida de amostras, uma vez que realizar uma classificação binária par a par envolveria descartar um terço do conjunto de dados, fazendo o conjunto a ser avaliado ter apenas 10 amostras. Dessa forma, buscar conjuntos de dados maiores também conferiria uma liberdade maior para configurações experimentais que pudessem facilitar a comparação dos resultados.

Contudo, é importante lembrar que a motivação original desse trabalho é averiguar se modelos de aprendizagem de máquina treinados com esses dados, e as suas respectivas análises, são capazes de captar nuances nas relações entre as amostras e na importância de seus atributos, que as análises tradicionais não são. Dessa forma, mesmo em uma situação ideal (nesse caso, com uma maior disponibilidade de dados), uma correspondência clara entre a análise SIMPER e a análise dos valores SHAP dos modelos não seria necessariamente algo desejado. Nesse aspecto, ficou claro que os resultados obtidos são distintos dos observados na análise SIMPER e podem ser um importante complemento na descrição dos microbiomas estudados.

A análise SIMPER é bastante tradicional e, por isso, apresenta amplo respaldo teórico e prático. Já a avaliação dos valores SHAP, nesse contexto, é algo relativamente inédito e, consequentemente, sem uma base teórico-prática sedimentada. Dessa forma, dificilmente pode-se fazer alguma afirmação conclusiva sobre sua validade. Portanto, para esse tipo de análise poder ser validada ou descartada, mais trabalhos, incluindo experimentos práticos, são de extrema importância.

# REFERÊNCIAS

ALTMAN, N.; KRZYWINSKI, M. The curse (s) of dimensionality. **Nat Methods**, v. 15, n. 6, p. 399–400, 2018.

ARAUJO, A. S. F.; OLIVEIRA, L. M. de S.; MELO, V. M. M.; ANTUNES, J. E. L.; ARAUJO, F. F.; MENDES, L. W. Distinct taxonomic composition of soil bacterial community across a native gradient of cerrado-ecotone-caatinga. **Applied Soil Ecology**, v. 161, p. 103874, 2021. ISSN 0929-1393. Disponível em: https://www.sciencedirect.com/science/article/pii/S0929139320308039.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738.

CHEN, M.-Y.; WU, S.-H.; LIN, G.-H.; LU, C.-P.; LIN, Y.-T.; CHANG, W.-C.; TSAY, S.-S. Rubrobacter taiwanensis sp. nov., a novel thermophilic, radiation-resistant species isolated from hot springs. **International Journal of Systematic and Evolutionary Microbiology**, Microbiology Society, v. 54, n. 5, p. 1849–1855, 2004. ISSN 1466-5034. Disponível em: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.63109-0.

CLARKE, K. R. Non-parametric multivariate analyses of changes in community structure. **Australian Journal of Ecology**, v. 18, n. 1, p. 117–143, 1993. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1442-9993.1993.tb00438.x.

GHANNAM, R. B.; TECHTMANN, S. M. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. **Computational and Structural Biotechnology Journal**, Elsevier, v. 19, p. 1092–1107, 2021.

HASSIJA, V.; CHAMOLA, V.; MAHAPATRA, A.; SINGAL, A.; GOEL, D.; HUANG, K.; SCARDAPANE, S.; SPINELLI, I.; MAHMUD, M.; HUSSAIN, A. Interpreting black-box models: a review on explainable artificial intelligence. **Cognitive Computation**, Springer, v. 16, n. 1, p. 45–74, 2024.

HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. **Neural Networks**, v. 2, n. 5, p. 359–366, 1989. ISSN 0893-6080. Disponível em: https://www.sciencedirect.com/science/article/pii/0893608089900208.

HUANG, V.; ROEM, J.; NG, D. K.; SCHWARTZ, J. M.; EVERETT, A. D.; PADMANABHAN, N.; ROMERO, D.; JOE, J.; CAMPBELL, C.; SIGAL, G. B. *et al.* Exploratory factor analysis yields grouping of brain injury biomarkers significantly associated with outcomes in neonatal and pediatric ecmo. **Scientific reports**, Nature Publishing Group UK London, v. 14, n. 1, p. 10790, 2024.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems 30**. Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

MARTINY, A. C. High proportions of bacteria are culturable across major biomes. **The ISME Journal**, v. 13, n. 8, p. 2125–2128, 04 2019. ISSN 1751-7362. Disponível em: https://doi.org/10.1038/s41396-019-0410-3.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. MIT Press, 2012. Disponível em: https://probml.github.io/pml-book/book0.html.

MURPHY, K. P. **Probabilistic Machine Learning: An introduction**. MIT Press, 2022. Disponível em: probml.ai.

ODUM, E.; BARRETT, G. **Fundamentals of Ecology**. Thomson Brooks/Cole, 2005. ISBN 9780534420666. Disponível em: https://books.google.com.br/books?id=vC9FAQAAIAAJ.

O'NEIL, C. **Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy**. USA: Crown Publishing Group, 2016. ISBN 0553418815.

PRIFTI, E.; CHEVALEYRE, Y.; HANCZAR, B.; BELDA, E.; DANCHIN, A.; CLÉMENT, K.; ZUCKER, J.-D. Interpretable and accurate prediction models for metagenomics data. **GigaScience**, Oxford University Press, v. 9, n. 3, p. giaa010, 2020.

QUAST, C.; PRUESSE, E.; YILMAZ, P.; GERKEN, J.; SCHWEER, T.; YARZA, P.; PEPLIES, J.; GLöCKNER, F. O. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. **Nucleic Acids Research**, v. 41, n. D1, p. D590–D596, 11 2012. ISSN 0305-1048. Disponível em: https://doi.org/10.1093/nar/gks1219.

RASMUSSEN, C. E.; WILLIAMS, C. K. I. **Gaussian Processes for Machine Learning**. The MIT Press, 2005. ISBN 9780262256834. Disponível em: https://doi.org/10.7551/mitpress/3206.001.0001.

SABARIA SARAVANAN S, M. K. D. T. S. A. P. R. S. P. P. M. S. Harnessing machine learning for metagenomics: Discovering the invisible microbial world. **Communications on Applied Nonlinear Analysis**, v. 31, n. 8, p. 147–164, 2024.

SILVA, D. J.; SILVA, T. R.; de Oliveira, M. L.; de Oliveira, G.; MISHRA, M.; SANTOS, C. A. G.; SILVA, R. M. da; dos Santos, C. A. Analysis of surface radiation fluxes and environmental variables over caatinga vegetation with different densities. **Journal of Arid Environments**, v. 222, p. 105163, 2024. ISSN 0140-1963. Disponível em: https://www.sciencedirect.com/science/article/pii/S0140196324000430.

SUN, C.; SHRIVASTAVA, A.; SINGH, S.; GUPTA, A. Revisiting unreasonable effectiveness of data in deep learning era. In: **Proceedings of the IEEE International Conference on Computer Vision (ICCV)**. [*S. l.*: *s. n.*], 2017.

UREL, H.; BENASSOU, S.; RESKA, T.; MARTI, H.; RAYO, E.; MARTIN, E. J.; SCHLOTER, M.; FERGUSON, J. M.; KESSELHEIM, S.; BOREL, N.; URBAN, L. Nanopore- and ai-empowered metagenomic viability inference. **bioRxiv**, Cold Spring Harbor Laboratory, 2024. Disponível em: https://www.biorxiv.org/content/early/2024/06/11/2024.06.10.598221.

ZAHARIA, M.; CHEN, A.; DAVIDSON, A.; GHODSI, A.; HONG, S. A.; KONWINSKI, A.; MURCHING, S.; NYKODYM, T.; OGILVIE, P.; PARKHE, M. *et al.* Accelerating the machine learning lifecycle with mlflow. **IEEE Data Eng. Bull.**, v. 41, n. 4, p. 39–45, 2018.

# APÊNDICES

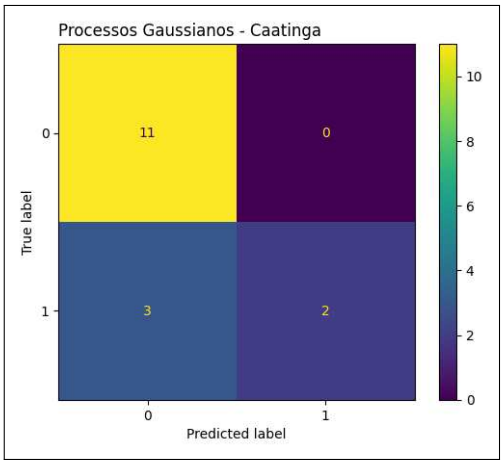
# APÊNDICE A – MATRIZES DE CONFUSÃO

Na seção a seguir estão as matrizes de confusão utilizadas para calcular as métricas de todos os modelos treinados utilizando o conjunto de dados sem projeção.

## A.1 Matrizes de Confusão

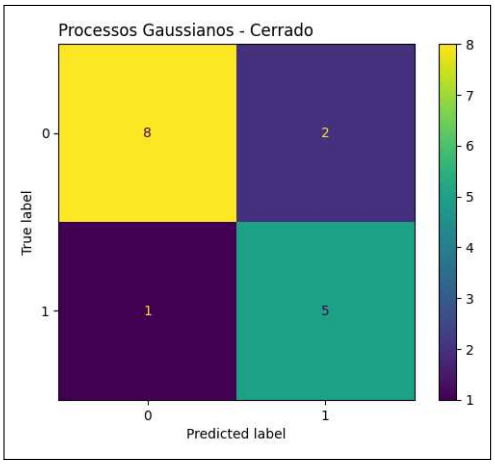
### *A.1.1 Processos Gaussianos*

Figura 30 – Caatinga



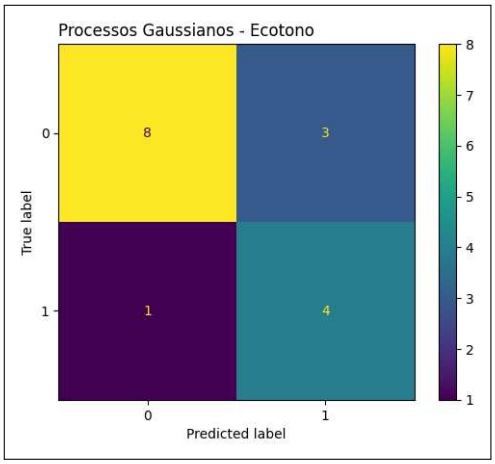

Fonte: elaborada pelo autor.

Figura 31 – Cerrado




Fonte: elaborada pelo autor.

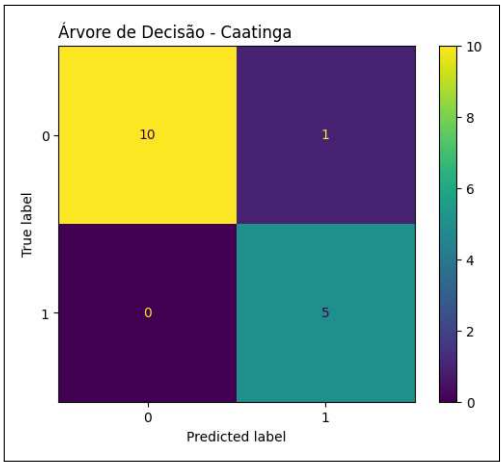
Figura 32 – Ecótono




Fonte: elaborada pelo autor.

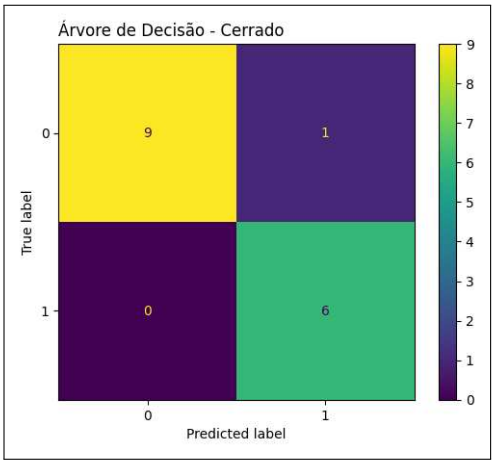
### *A.1.2 Árvore de Decisão*

Figura 33 – Caatinga



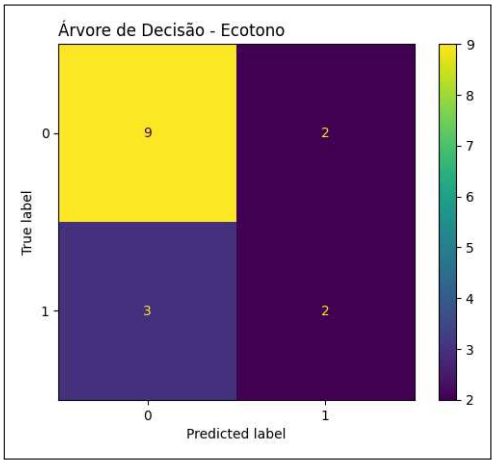

Fonte: elaborada pelo autor.

Figura 34 – Cerrado




Fonte: elaborada pelo autor.

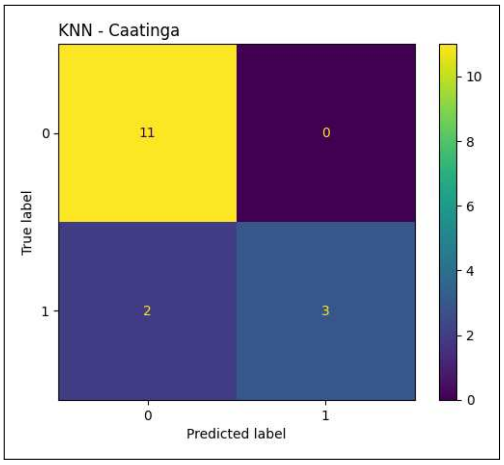
Figura 35 – Ecótono




Fonte: elaborada pelo autor.

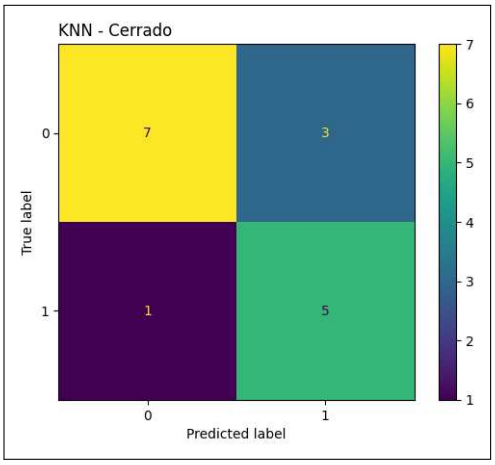
### *A.1.3 KNN*

Figura 36 – Caatinga



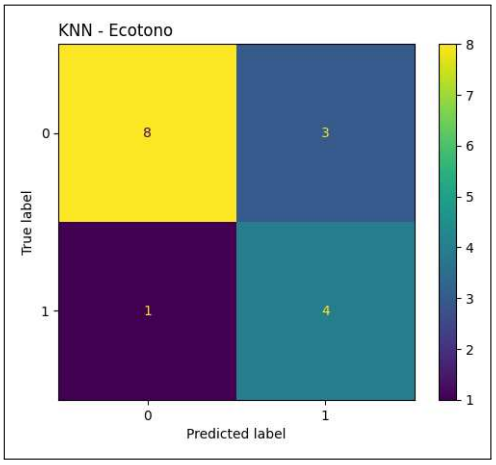

Fonte: elaborada pelo autor.

Figura 37 – Cerrado




Fonte: elaborada pelo autor.

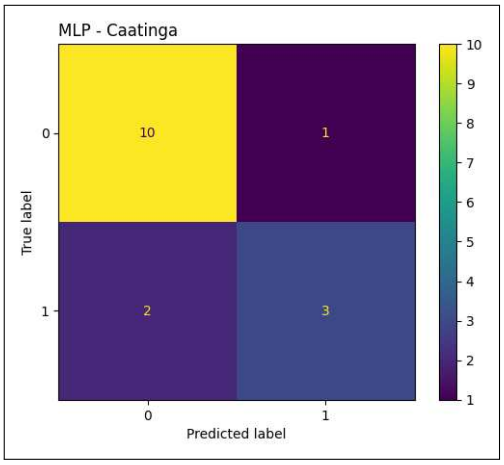
Figura 38 – Ecótono




Fonte: elaborada pelo autor.

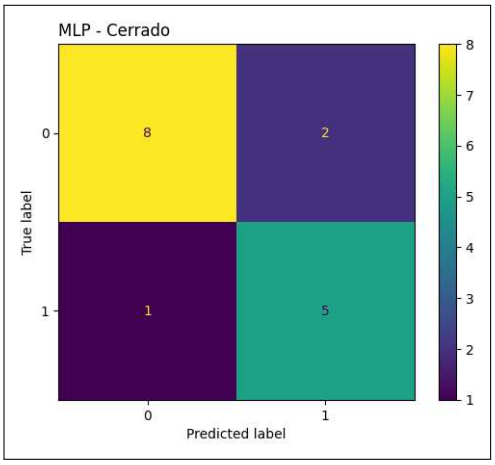
### *A.1.4 MLP*

Figura 39 – Caatinga



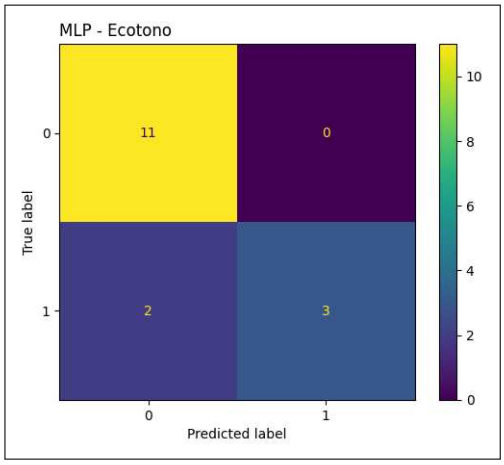

Fonte: elaborada pelo autor.

Figura 40 – Cerrado




Fonte: elaborada pelo autor.

Figura 41 – Ecótono



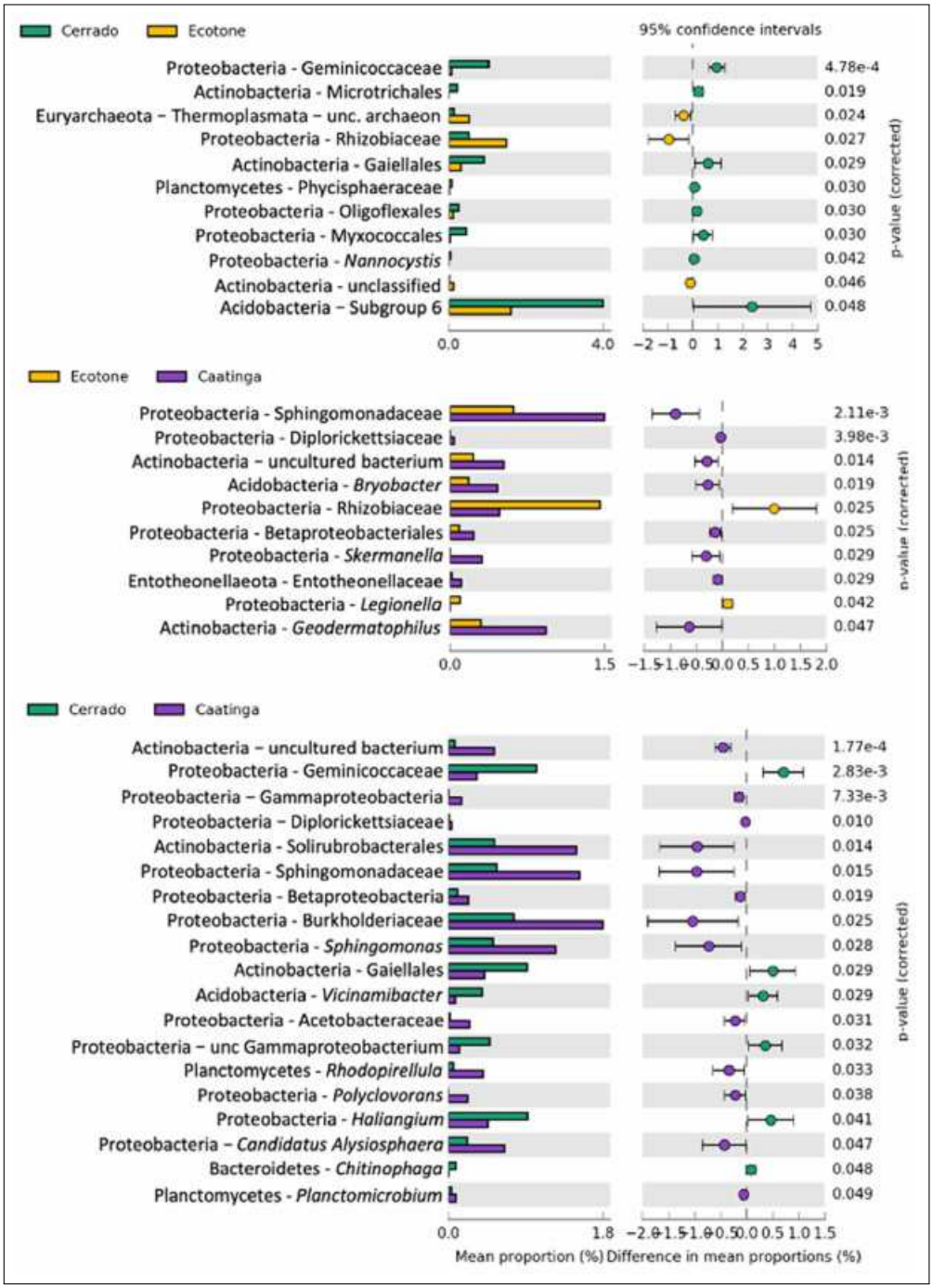

Fonte: elaborada pelo autor.

# ANEXOS

# ANEXO A – ANEXO A

Anexo abaixo, está o gráfico sumarizando as análises SIMPER realizadas no conjunto de dados estudado nesse trabalho, publicado no mesmo trabalho que originalmente publicou os dados (ARAUJO *et al.*, 2021).

Figura 42 – Abundância Diferencial




Fonte: (ARAUJO *et al.*, 2021)