UNIVERSIDADE DE SÃO PAULO

DEPARTAMENTO DE ENGENHARIA MECÂNICA DA ESCOLA POLITÉCNICA DA USP

BRUNO DALEFFI DA SILVA

Ecologia Acústica e Ciência de Dados: uma análise da paisagem sonora do cerrado

Aplicabilidade e desempenho de modelos de aprendizado de máquina na identificação de formações ecológicas

São Paulo

2024

BRUNO DALEFFI DA SILVA

Ecologia Acústica e Ciência de Dados: uma análise da paisagem sonora do cerrado

Aplicabilidade e desempenho de modelos de aprendizado de máquina na identificação de formações ecológicas

Versão Corrigida

Dissertação apresentada ao Departamento de Engenharia Mecânica da Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências.

Área de Concentração: Ciência de Dados e Ecologia Acústica

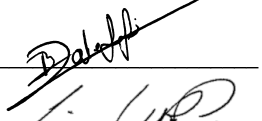
Orientador: Prof. Dr. Linilson Padovese

São Paulo

2024

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 19 de março de 2024

Assinatura do autor: 

Assinatura do orientador:

Catalogação-na-publicação

Silva, Bruno
Ecologia Acústica e Ciência de Dados: uma análise da paisagem sonora do cerrado Aplicabilidade e desempenho de modelos de aprendizado de máquina na identificação de formações ecológicas / B. Silva, L. Padovese -- versão corr. -- São Paulo, 2024.
67 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia Mecânica.

1.Aprendizado de Máquina 2.Cerrado 3.Inteligência Artificial 4.Ecologia Acústica 5.Paisagem Sonora I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia Mecânica II.t. III.Padovese, Linilson

AGRADECIMENTOS

Primeiramente, à minha mãe, Silvana Daleffi da Silva, que, com sua trajetória como professora, ensinou-me desde os primeiros passos o inestimável valor do estudo e da persistência na busca pelo conhecimento. Suas lições serão eternamente lembradas e servirão como bússola em minha jornada.

À minha irmã, Bárbara Daleffi da Silva, cuja presença constante e afetuosa serviu de ancoragem e inspiração ao longo de todos esses anos. Tua força e amor são pilares inabaláveis em minha vida.

Em memória de meu pai, Julio Cesar da Silva, que, junto à minha mãe, instalou em mim a paixão pelo aprendizado. Cada página desta dissertação carrega um pedaço do legado que você deixou.

À Amanda Alves de Souza Daleffi, minha esposa e companheira de vida, que, com amor, paciência e compreensão, esteve ao meu lado em cada decisão. Sua fé em mim e nosso amor foram a chama que iluminou os caminhos mais obscuros desta trajetória.

À equipe da Terranova e à Associação Brasileira de Jurimetria (ABJ), por sempre manterem viva a centelha nerd em mim e me incentivarem a sempre ir além, explorando novos horizontes e desafios.

Ao Prof. Dr. Linlilson Padovese, pela orientação precisa e pelo apoio constante em cada fase deste trabalho. Sua expertise e entusiasmo pela ciência foram essenciais para o desenvolvimento desta pesquisa. Sou grato por sua paciência e por sempre acreditar em meu potencial.

Por fim, mas não menos importante, a todos os meus amigos e demais familiares. Cada um de vocês, de maneira direta ou indireta, contribuiu para que eu chegasse até aqui. A todos vocês, minha eterna gratidão e reconhecimento.

RESUMO

DALEFFI, B. **Ecologia Acústica e Ciência de Dados: uma análise da paisagem sonora do cerrado: aplicabilidade e desempenho de modelos de aprendizado de máquina na identificação de formações ecológicas.** 2024. Tese (Mestrado) - Escola Politécnica da Universidade de São Paulo, São Paulo, 2024.

O Cerrado, um bioma brasileiro de significativa importância ecológica, abrange cerca de 2 milhões de quilômetros quadrados, representando quase 23% do território nacional. Além de seu papel vital na sustentação da vida e na regulação climática, abrigando aproximadamente 5% das espécies de fauna globais, o Cerrado é fundamental para a hidrologia regional. Ele forma o sistema hidrológico do Pantanal, um dos maiores complexos de água doce do mundo, e é a fonte de alguns afluentes do rio Amazonas, bem como de rios de crucial importância nacional, como o Tocantins e o São Francisco. Entretanto, o Cerrado enfrenta desafios devido ao avanço das atividades agropecuárias e ao aumento dos desastres ambientais, em decorrência das mudanças climáticas. Visando auxiliar os trabalhos de preservação e recuperação do bioma, este estudo explorou diferentes metodologias de Ciência de Dados para identificar a formação natural do Cerrado (Florestal, Savânica e Campestre) através da paisagem sonora do ambiente. Foram desenvolvidos cinco modelos estatísticos, utilizando coeficientes cepstrais de frequência mel (MFCCs) como variáveis independentes para os modelos de Gradient Boosting, Random Forest, Regressão Logística e Multilayer Perceptron, e imagens dos espectrogramas para a Rede Neural Convolucional (CNN). A análise dos resultados revelou que a CNN apresentou o melhor desempenho em todas as métricas avaliadas. Entretanto, a escolha do modelo mais adequado deve levar em consideração não apenas a performance, mas também a simplicidade do método, o tempo de resposta da predição e a capacidade de lidar com um grande número de observações em um curto período de tempo. Neste contexto, modelos mais simples, como Gradient Boosting ou Random Forest, podem ser mais apropriados em alguns casos. A aplicação da técnica de explicabilidade LIME auxiliou na compreensão dos processos de decisão do modelo de CNN,

fornecendo insights sobre como melhorar sua performance e aplicação em estudos futuros de preservação e recuperação do bioma Cerrado. A análise das frequências relevantes para cada formação do Cerrado demonstrou o potencial dessa técnica na identificação das características do espectrograma responsáveis pela classificação das paisagens acústicas. Este estudo demonstra que a combinação de Ecologia Acústica e Ciência de Dados pode ser eficiente na classificação de paisagens acústicas e na identificação do tipo de formação do Cerrado. As conclusões deste trabalho podem servir como base para futuras pesquisas na área de Ecologia Acústica e na aplicação de técnicas de aprendizado de máquina em estudos ambientais, contribuindo para a preservação da biodiversidade e a recuperação de biomas ameaçados, como o Cerrado. Além disso, fornece insumos metodológicos para outras pesquisas acerca do uso da Inteligência Artificial/Machine Learning na identificação e classificação de sinais em geral.

Palavras-Chave: Cerrado. Ecologia Acústica. Paisagens sonoras. Inteligência Artificial. Aprendizado de Máquina. Rede Neural Convolucional (CNN). Gradient Boosting. Random Forest. Preservação ambiental. Biodiversidade.

ABSTRACT

DALEFFI, B. **Acoustic Ecology and Data Science: an analysis of the Cerrado soundscape: applicability and performance of machine learning models in the identification of ecological formations**. 2024. Master's Thesis - Polytechnic School of the University of São Paulo, São Paulo, 2024.

The Cerrado, a Brazilian biome of significant ecological importance, covers about 2 million square kilometers, accounting for nearly 23% of the national territory. Besides its vital role in supporting life and in climate regulation, housing approximately 5% of the world's fauna species, the Cerrado is crucial for regional hydrology. It forms the hydrological system of the Pantanal, one of the largest freshwater wetland complexes in the world, and is the source of some tributaries of the Amazon River, as well as rivers of crucial national importance, such as the Tocantins and São Francisco. However, the Cerrado faces challenges due to the expansion of agricultural activities and the increase in environmental disasters, as a result of climate change. Aiming to support the preservation and recovery of the biome, this study explored different Data Science methodologies to identify the natural formation of the Cerrado (Forested, Savanna, and Grassland) through the soundscapes of the environment. Five statistical models were developed, using Mel Frequency Cepstral Coefficients (MFCCs) as independent variables for the Gradient Boosting, Random Forest, Logistic Regression, and Multilayer Perceptron models, and spectrogram images for the Convolutional Neural Network (CNN). The analysis of the results revealed that the CNN had the best performance in all evaluated metrics. However, the choice of the most suitable model should take into account not only performance but also the simplicity of the method, the response time of the prediction, and the ability to handle a large number of observations in a short period of time. In this context, simpler models, such as Gradient Boosting or Random Forest, may be more appropriate in some cases. The application of the LIME explainability technique helped in understanding the decision processes of the CNN model, providing insights into how to improve its performance and application in future studies of preservation and recovery of the Cerrado biome. The analysis of the relevant frequencies for each Cerrado formation demonstrated the potential of

this technique in identifying the characteristics of the spectrogram responsible for classifying the acoustic landscapes. This study demonstrates that the combination of Acoustic Ecology and Data Science can be efficient in classifying acoustic landscapes and identifying the type of Cerrado formation. The conclusions of this work can serve as a basis for future research in the field of Acoustic Ecology and the application of machine learning techniques in environmental studies, contributing to the preservation of biodiversity and the recovery of threatened biomes, such as the Cerrado. Furthermore, it provides methodological inputs for other research on the use of Artificial Intelligence/Machine Learning in the identification and classification of signals.

Keywords: Cerrado, Acoustic Ecology, Soundscapes, Machine Learning, Convolutional Neural Network (CNN), Gradient Boosting, Random Forest, Environmental Preservation, Biodiversity.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

O Cerrado é o segundo maior bioma brasileiro. É caracterizado, principalmente, por um clima semi-árido, uma vegetação arbustiva densa e abrange cerca de 2 milhões de quilômetros quadrados, o que representa cerca de 23% do território nacional, estendendo-se por 12 dos 26 estados brasileiros (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE, 2022).

Apesar de sua relevância, o Cerrado enfrenta um crescente desmatamento para dar espaço às atividades agrícolas e pecuárias. Conforme dados do Instituto Brasileiro de Geografia e Estatística (IBGE), este bioma perdeu aproximadamente 150 mil quilômetros quadrados de sua vegetação, tanto campestre quanto florestal, entre 2000 e 2018 (IBGE, 2020, Anexo 2). Essa perda resulta em profundos impactos ambientais, sociais e econômicos. Segundo Dias (1991), o Cerrado é essencial por diversos motivos:

- Abriga cerca de 5% das espécies animais conhecidas no mundo e quase um terço (30%) de toda a biodiversidade brasileira.
- Contém uma vasta gama de espécies endêmicas de plantas e animais.
- As águas do Cerrado são cruciais para o sistema hidrológico do Pantanal, um dos maiores complexos de água doce do mundo.
- É a origem de importantes rios brasileiros, incluindo o São Francisco e a bacia do Araguaia/Tocantins, além de ser o berço de alguns afluentes do rio Amazonas, desempenhando um papel vital como uma das principais fontes de água do país.
- Sua vegetação tem um papel significativo na absorção de gás carbônico, contribuindo para o equilíbrio do efeito estufa.

Diante disso, diversos estudos e projetos vêm sendo desenvolvidos com o intuito de i) preservar o bioma atual e ii) encontrar maneiras eficientes de restauração de áreas degradadas. Em ambos os casos, os processos são

longos, complexos e envolvem diversos atores, desde o poder público e o setor privado até a sociedade civil.

No caso da recuperação do ambiente, o primeiro passo para que o processo de restauração alcance os resultados desejados é entender a formação natural daquela região do Cerrado, seja ele degradado pela ação humana ou por desastres ambientais. Para isso, é necessário: i) identificar a formação original antes da degradação, ii) quantificar a perda da biodiversidade no local e iii) encontrar a formação ideal a ser recuperada. A formação ideal prevista em iii) não necessariamente corresponde ao bioma original, pois dependendo da perda da biodiversidade da região, a restauração do bioma nativo pode ser demasiadamente complexa, demorada, custosa, ou até mesmo estar comprometida (CAVA, 2018).

De acordo com a Empresa Brasileira de Pesquisa Agropecuária (Embrapa), existem 3 formações de vegetação diferentes para o Cerrado que, por sua vez, se dividem em 11 subgrupos, conforme descritos abaixo e ilustrados na Figura 1;

- Formações Florestais
    - Mata Ciliar, Mata de Galeria, Mata Seca e Cerradão
- Formações Savânicas
    - Cerrado sentido restrito (denso, típico e ralo), Parque de Cerrado, Palmeiral e Vereda
- Formações Campestres
    - Campo Sujo, Campo Limpo e Campo Rupestre

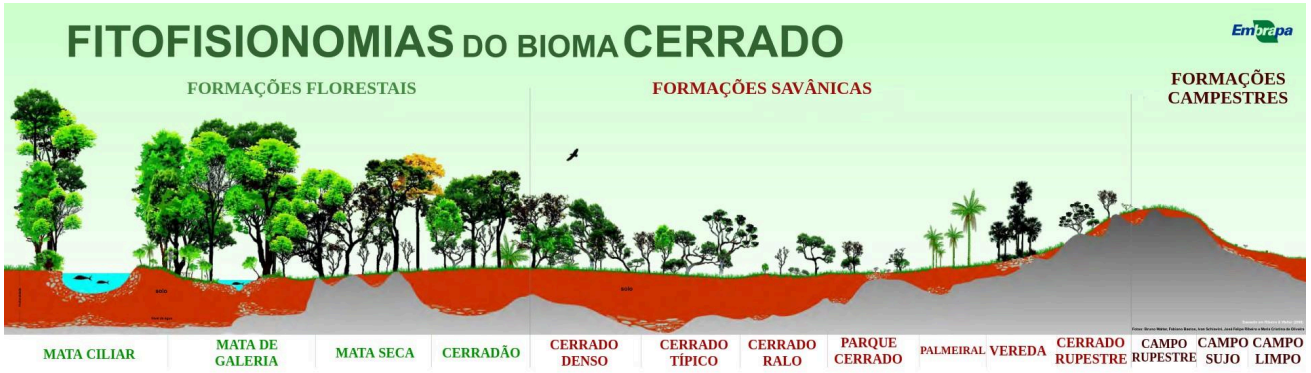


Figura 1 - Fitofisionomias do Cerrado.

Fonte: Portal Embrapa (2008)

Atualmente, o principal método para identificar a fitofisionomia de uma determinada região se baseia em análises qualitativas da flora e fauna presentes no sistema (RIBEIRO apud EMBRAPA, 2008), o que pode ser um tanto quanto complexo e custoso, uma vez que depende do deslocamento de equipes especializadas para estudos de campo.

Estudos recentes, entretanto, vêm utilizando a Ecologia Acústica no mapeamento da paisagem sonora de diversos biomas, a fim de identificar e classificar características ou espécies presentes no ambiente, reduzindo custos e mantendo, ou até melhorando, a eficiência das análises (TUCKER, 2014).

Diante do exposto sobre o Cerrado e suas diversas formações, é vital salientar a relevância das áreas de estudo selecionadas para este trabalho. As regiões de Assis, Águas de Santa Bárbara e Itirapina têm importância representativa, servindo como protótipos das formações Florestal, Savânica e Campestre do Cerrado, respectivamente. Estas classificações não são arbitrárias; ao contrário, são o resultado de rigorosos estudos de campo realizados por especialistas que analisaram as particularidades e densidades da vegetação de cada região. Ao utilizar estas áreas como referências, este estudo busca explorar a viabilidade de modelos de aprendizado de máquina no discernimento acústico de diferentes formações do Cerrado, potencializando a eficácia e precisão em tarefas de identificação e classificação.

A compreensão das diferenças entre as formações Florestal, Savânica e Campestre no Cerrado é enriquecida pela inclusão de um conjunto de imagens fotográficas. A primeira imagem, representando a região de Assis, mostra a formação Florestal, caracterizada por uma densa cobertura arbórea. A segunda fotografia, de Águas de Santa Bárbara, ilustra a formação Savânica, com uma combinação de árvores dispersas e vegetação rasteira. Por fim, a imagem de Itirapina exemplifica a formação Campestre, destacando-se por suas extensas áreas de campo aberto e vegetação esparsa.

Figura 2 - Formações Vegetais do Cerrado: Florestal, Savânica e Campestre.
Fonte: Assis: Catálogo de Plantas das Unidades de Conservação do Brasil[1]
Santa Bárbara: Instituto de Pesquisas Ambientais[2]
Itirapina: Scielo Brasil[3]

[1] Disponível em: https://catalogo-ucs-brasil.jbrj.gov.br/descr_areas.php?area=EEAssis

[2] Disponível em: https://www.infraestruturameioambiente.sp.gov.br/institutoflorestal/areas-protegidas/estacoes-ecologicas/santa-barbara/

[3] Disponível em: https://doi.org/10.1590/S1676-06032008000300019

Este trabalho introduz avanços notáveis no campo do reconhecimento acústico das formações do Cerrado. Primeiramente, destaca-se a possibilidade da classificação de sons com base em Inteligência Artificial, uma abordagem inovadora no contexto atual. Em segundo lugar, leva-se em consideração a ponderação entre a complexidade do método e sua praticidade; quando um evento é observado várias vezes rapidamente, pode-se usar um modelo de desempenho médio para classificá-lo várias vezes, sendo a classificação final a média das probabilidades dessas tentativas. Por fim, o uso do LIME nas CNN's, baseadas em espectrogramas, representa um avanço significativo. Esse método permite identificar as regiões do espectro e faixas de frequência que caracterizam cada formação, fornecendo insights valiosos sobre a composição sonora específica de cada região do Cerrado.

2. OBJETIVOS

O principal objetivo deste trabalho é desenvolver e comparar modelos de aprendizado de máquina para classificar diferentes formações naturais do Cerrado com base nas paisagens sonoras das respectivas regiões. Esta análise é, essencialmente, uma avaliação da hipótese central de que **formações naturais com identidades ecológicas semelhantes apresentam paisagens acústicas similares**. A eficiência em identificar e classificar essas formações por meio de seus perfis sonoros funcionará como uma validação prática dessa hipótese.

As áreas de Assis, Águas de Santa Bárbara e Itirapina foram preliminarmente identificadas como amostras das Formações Florestal, Savânica e Campestre, respectivamente. Essa categorização foi baseada em estudos de campo e na densidade da vegetação onde os equipamentos de gravação foram instalados, seguindo as fitofisionomias do Cerrado definidas pelo Portal Embrapa (ver Figura 1).

Para atingir o objetivo, o estudo será estruturado em torno das seguintes questões norteadoras:

1. Quais modelos de aprendizado de máquina, abrangendo desde métodos estatísticos tradicionais até redes neurais avançadas, são mais eficientes e eficazes na classificação dos tipos de formações naturais do Cerrado com base em suas paisagens sonoras?
2. De que forma a complexidade e o custo computacional dos modelos impactam na escolha do método mais apropriado para a classificação de paisagens acústicas?
3. Quais atributos dos sinais sonoros são mais relevantes para a classificação das paisagens acústicas do Cerrado?

Neste sentido, e, conforme detalhado no capítulo 4, o trabalho buscará analisar a eficácia de modelos que utilizam **Coeficientes Cepstrais de Frequência Mel (MFCCs) e espectrogramas** como variáveis de entrada na classificação de paisagens acústicas. Para isso, serão utilizados modelos de

Machine Learning e Deep Learning, como Gradient Boosting, Random Forest, Regressão Logística, Multilayer Perceptron e Redes Neurais Convolucionais.

Além disso, será analisado o impacto da quantidade de observações na performance dos modelos e a importância de se ponderar a simplicidade do método, o tempo de treinamento e resposta da predição e a capacidade de lidar com um grande número de observações em um curto período de tempo. Por fim, este estudo buscará entender a importância das principais características do espectrograma na classificação das paisagens acústicas do Cerrado.

3. REVISÃO DE LITERATURA

A presente seção tem como objetivo investigar a convergência entre a Ecologia Acústica, a Ciência de Dados e a Inteligência Artificial, com um enfoque especial nas Redes Neurais e suas aplicações na análise de paisagens sonoras. Essa revisão abrangerá uma visão geral dessas áreas, culminando na união de todas em estudos recentes que buscam entender e conservar a biodiversidade através de um novo prisma: o acústico.

3.1. ANÁLISE DE PAISAGEM SONORA

A história da análise de paisagem sonora começa na segunda metade do século XX. Nesse período, Raymond Murray Schafer, um músico, compositor e ex-professor de Estudos de Comunicação na Simon Fraser University, propôs uma maneira inovadora de interagir com o mundo sonoro ao nosso redor. Ele sugeriu que as pessoas deveriam ouvir a paisagem acústica do dia a dia como uma composição musical (SCHAFER, 1977). Essa perspectiva, embora parecesse incomum na época, deu origem à área de Ecologia Acústica (WRIGHTSON, 2000).

Ao longo dos anos, o conceito de paisagem sonora expandiu-se e começou a ser usado para estudar vários aspectos do ambiente. Barry Truax, um acadêmico canadense que começou a trabalhar significativamente com ecologia acústica e paisagens sonoras na década de 1970, foi um dos primeiros estudiosos na área. Em um de seus trabalhos, Truax utilizou a análise de paisagem sonora para estudar os efeitos da poluição sonora na vida urbana. Seu trabalho pioneiro, realizado no início do século XXI, destacou a importância dos sons no nosso ambiente e como eles afetam a nossa percepção do mundo (TRAUX, 2001).

A criação do Fórum Mundial de Ecologia Acústica (WFAE), em 1993, marcou outro momento significativo para o campo, promovendo o reconhecimento internacional da ecologia acústica e facilitando a divulgação de estudos relacionados (WRIGHTSON, 2000).

Além dessas contribuições, o trabalho de Bernie Krause, autor de 'The Great Animal Orchestra: Finding the Origins of Music in the World's Wild Places' (KRAUSE, 2012), também merece destaque. Krause expandiu o campo da ecologia acústica, explorando as conexões entre os sons da natureza e a música, argumentando que os sons da vida selvagem têm uma estrutura musical intrínseca. Seu trabalho não apenas enriqueceu a compreensão da ecologia acústica, mas também inspirou uma nova forma de pensar sobre a conservação da biodiversidade e a importância de preservar os ambientes sonoros naturais.

Na era moderna, com os avanços da Ciência de Dados e da Inteligência Artificial, a análise de paisagem sonora adquiriu uma nova e poderosa ferramenta. Segundo Kate Priestman (2017), as paisagens sonoras são compostas por uma trama de sons, cada um fornecendo informações valiosas sobre o ambiente, divididas em **biofonia**, **geofonia** e **antropofonia**.

Através da inteligência artificial, cada uma dessas categorias de som pode ser analisada separadamente, permitindo uma compreensão muito mais profunda da composição da paisagem sonora. O casamento da análise de paisagem sonora com a inteligência artificial está abrindo um mundo de novas oportunidades. Agora é possível estudar com precisão a biodiversidade de uma região, monitorar mudanças no ambiente sonoro e entender de forma mais detalhada como nós, humanos, nos relacionamos e influenciamos o ambiente sonoro à nossa volta.

Diversas ferramentas foram desenvolvidas graças à intersecção entre a análise de paisagem sonora e a Inteligência Artificial. Entre elas, estão a medição da **saturação da paisagem** (quantidade de fontes sonoras distintas que é correlacionada com a quantidade de espécies no local), a avaliação da **similaridade entre espécies**, o **monitoramento de mudanças na composição das espécies** e o **reconhecimento de espécies específicas** com base em suas identidades sonoras (PRIESTMAN, 2017).

Nesse contexto, estudos recentes têm explorado o uso de métodos de Machine Learning e Deep Learning para a classificação de espécies através de sons. Um exemplo notável é o estudo de Nahian Ibn Hasan (2022), que

apresenta um híbrido de processamento de sinais tradicionais e abordagens de deep learning para identificar espécies de aves a partir de gravações de áudio, alcançando uma precisão de 90,45% para um conjunto de 10 classes de aves.

Assim, a história da análise de paisagem sonora, desde sua concepção até a sua intersecção com a Inteligência Artificial, ilustra a constante evolução e o potencial deste campo. Hoje, é possível não apenas ouvir o mundo ao nosso redor, mas também compreendê-lo e protegê-lo de uma forma nunca antes imaginada. A próxima seção tratará de uma ferramenta chave neste processo.

3.2. AS REDES NEURAIS

As Redes Neurais Artificiais (RNAs) figuram como uma importante e impactante abordagem no campo da Inteligência Artificial. Desde os anos 1950, tais estruturas têm sido objeto de inúmeras discussões e avanços na literatura científica. O conceito central que permeia as RNAs é a tentativa de emular, através de estruturas computacionais, o funcionamento do sistema nervoso biológico, principalmente no que tange o processamento e o aprendizado de informações (DATA SCIENCE ACADEMY, 2019).

A inspiração biológica para a concepção dessas estruturas é evidente na arquitetura das redes, que consistem em neurônios interconectados organizados em diferentes camadas. Esses neurônios processam informações, modificam-se através do aprendizado e contribuem para a realização de tarefas complexas, como a classificação de dados.

Dentre as várias arquiteturas de redes neurais que foram propostas ao longo dos anos, o Perceptron Multicamada (Multilayer Perceptron, MLP) é uma das mais notórias e historicamente relevantes. O MLP foi introduzido ao mundo científico por Frank Rosenblatt em 1958, no artigo "The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain". O trabalho pioneiro de Rosenblatt é considerado uma das obras mais influentes da história da Inteligência Artificial e marcou o início dos estudos de redes neurais artificiais. O autor propôs uma rede com múltiplas camadas de

neurônios interconectados, onde cada neurônio possuía seus próprios pesos e saídas, e a saída de cada neurônio era utilizada como entrada para o próximo (ROSENBLATT, 1958).

A Figura 3, originária do Asimov Institute, apresenta uma visão geral das principais arquiteturas de redes neurais utilizadas atualmente. Neste espectro de técnicas e arquiteturas, é notável a diversidade e a especialização de cada modelo para lidar com diferentes tipos de problemas e estruturas de dados. Tal diversidade permite a seleção da abordagem mais adequada para cada caso de uso específico, o que resulta em aplicações mais eficientes e precisas das redes neurais.

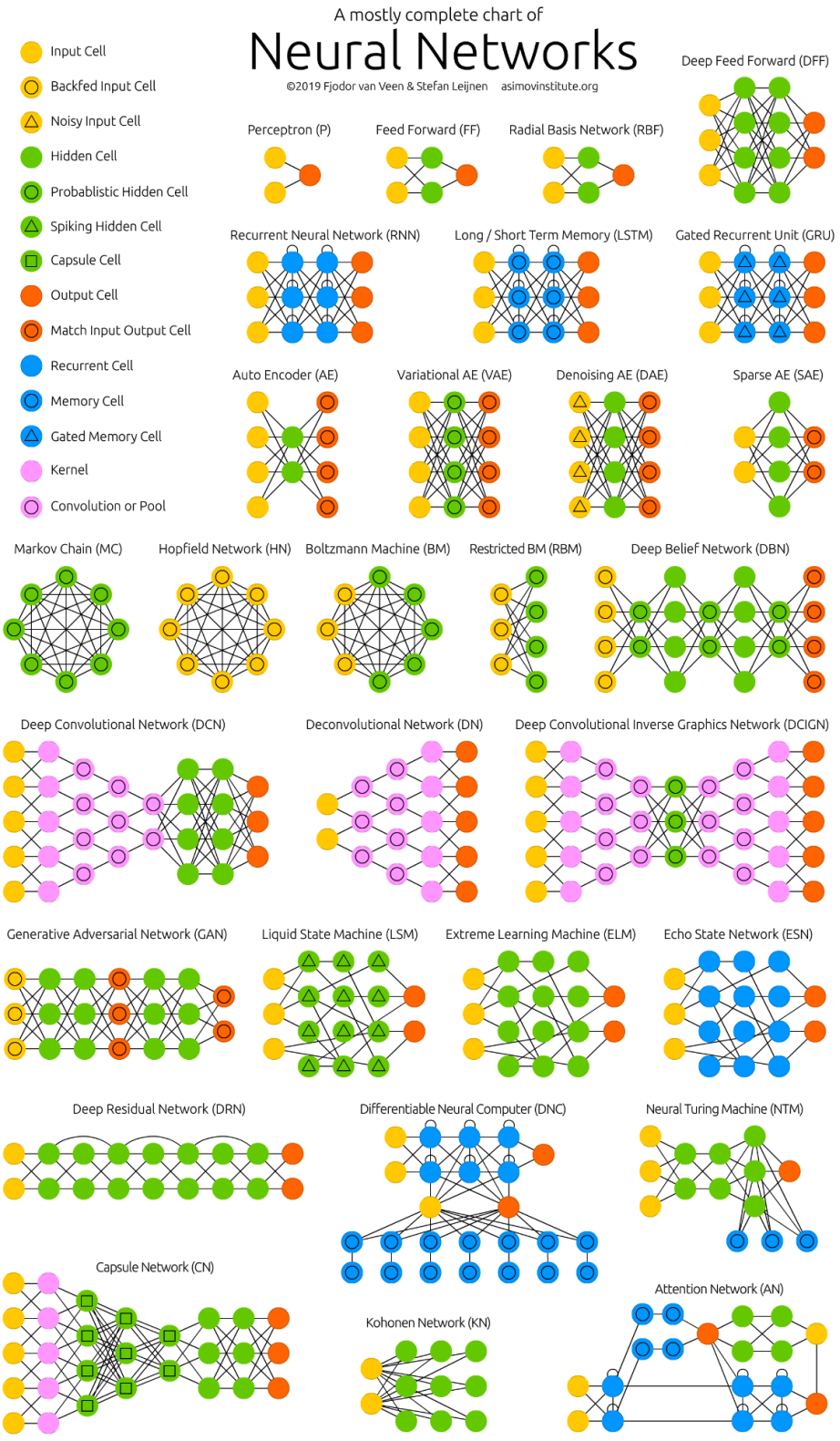


Figura 3 - Principais arquiteturas das redes neurais artificiais utilizadas atualmente.

Fonte: Asimov Institute (2016)

Cabe destacar que a evolução computacional observada, especialmente na primeira década do século XXI, foi decisiva para o crescimento exponencial do interesse e aplicações práticas do Deep Learning, um subcampo da Inteligência Artificial focado no desenvolvimento e aplicação de redes neurais profundas. Hoje, é possível encontrar soluções comerciais e acadêmicas baseadas nesta tecnologia em uma variedade de setores.

3.3. AS REDES NEURAIS CONVOLUCIONAIS

As Redes Neurais Convolucionais (CNNs - Convolutional Neural Networks) constituem um marco na evolução das redes neurais artificiais e tornaram-se a arquitetura de escolha para tarefas de processamento de imagem. As CNNs são caracterizadas por sua capacidade de aprender automaticamente e generalizar características das imagens, sem a necessidade de engenharia manual de recursos.

A inspiração para as CNNs vem da organização do córtex visual de animais, sugerindo que talvez possamos replicar esta eficiência computacional em uma rede neural artificial (HUBEL & WIESEL, 1962). Este conceito foi primeiro implementado na forma de uma rede neural convolucional por Kunihiko Fukushima em 1980 com a proposta da arquitetura "Neocognitron", mas ainda não continha o método de retropropagação para o treinamento (FUKUSHIMA, 1980).

O avanço veio com Yann LeCun e colaboradores em 1998, no que é considerado o trabalho seminal que popularizou as CNNs. LeCun introduziu o algoritmo LeNet-5 para classificação de dígitos manuscritos, utilizando a técnica de retropropagação para treinamento. Seu trabalho serviu como a base para o rápido desenvolvimento e popularização de CNNs nos anos seguintes (LECUN et al., 1998).

Os anos 2010s foram um período de crescimento rápido e inovação em redes neurais convolucionais. Em 2012, Alex Krizhevsky, Ilya Sutskever e Geoffrey Hinton introduziram a rede AlexNet, que ganhou a competição ImageNet Large Scale Visual Recognition Challenge (ILSVRC) por uma

margem significativa. Isso demonstrou o poder das CNNs em reconhecer e classificar imagens em larga escala, popularizando ainda mais seu uso em uma variedade de aplicações de visão computacional (KRIZHEVSKY et al., 2012).

A partir daí, uma série de outras arquiteturas notáveis foram introduzidas, como a VGGNet, GoogLeNet, e ResNet, cada uma contribuindo para o progresso do campo e expandindo o alcance de aplicações de CNNs (SIMONYAN & ZISSERMAN, 2014; SZEGEDY et al., 2015; HE et al., 2016).

A figura abaixo contém uma representação visual da arquitetura por trás de uma rede neural convolucional tradicional.

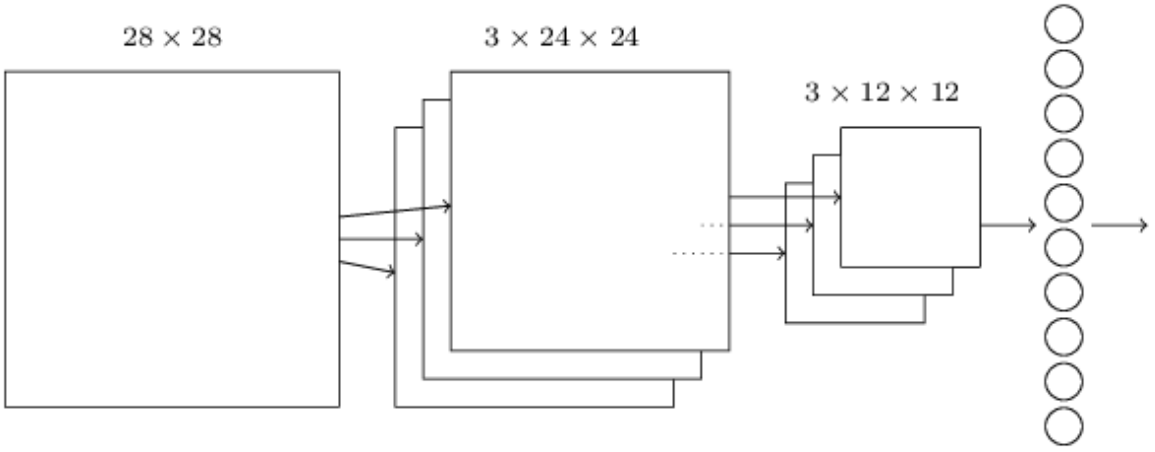


Figura 4 - Representação da estrutura de uma rede neural convolucional.

Fonte: Data Science Academy (2019, p. 43)

Quando aplicadas à Ecologia Acústica, as CNNs têm mostrado grande potencial. A capacidade dessas redes de extrair automaticamente características dos dados de entrada as torna particularmente adequadas para a análise de paisagens sonoras. As CNNs podem ser usadas para analisar espectrogramas - representações visuais de frequências de som ao longo do tempo - e classificar ou identificar os sons específicos dentro deles. Isso permite a identificação de espécies ou eventos sonoros particulares dentro de uma gravação de áudio, o que pode ser de grande importância para

o monitoramento da biodiversidade e estudos de ecologia em geral (STOWEL et al., 2014; PELLEGRINI et al., 2020).

4. MÉTODOS

Um artigo publicado no Journal of Applied Ecology em 2017 lança luz sobre as complexidades associadas à regeneração do Cerrado, especificamente destacando as diferenças notáveis na biodiversidade entre diferentes tipos de formações do bioma (CAVA, 2017). Esta pesquisa sublinha a importância de compreender de forma precisa e eficaz a natureza dessas formações, especialmente quando se trata de estratégias de conservação e restauração.

Dada a biodiversidade única e o papel ecológico crucial do Cerrado, tal compreensão é fundamental. A relevância deste ponto para o presente estudo se manifesta nos objetivos centrados em desenvolver modelos de aprendizado de máquina eficazes para classificar essas formações com base em paisagens sonoras. Ao fazer isso, é possível contribuir para estratégias mais direcionadas e eficazes de preservação e recuperação, que consideram as características e necessidades específicas de cada tipo de formação do Cerrado.

Para demonstrar a eficiência e sinergia da interdisciplinaridade entre a Ciência de Dados e a Ecologia Acústica, foi desenvolvida uma série de modelos de Machine Learning/Deep Learning treinados com uma grande quantidade de arquivos de áudio que representam a paisagem acústica das três principais formações do Cerrado.

Para fortalecer a robustez e a precisão de nosso estudo, a metodologia foi intencionalmente estruturada em quatro etapas fundamentais que não só refletem o ciclo completo da ciência de dados (SHEARER, 2000), mas também estão alinhadas com as fases iterativas do fluxo de uma análise de dados rigorosa.

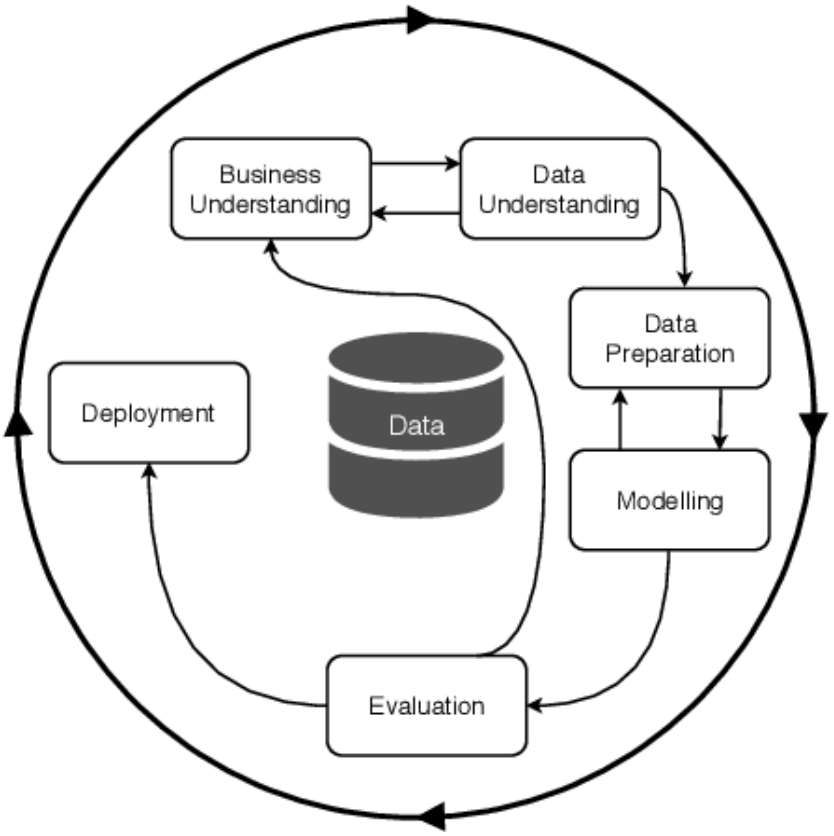


Figura 5 - Ciclo da ciência de dados, de acordo com Shearer (2000).
Fonte: Shearer (2000)

Essa divisão foi concebida para garantir que cada etapa da investigação seja tratada com a devida atenção aos detalhes e rigor científico. São elas:

1. Coleta dos dados
2. Pré-processamento
3. Modelagem
4. Explicabilidade

4.1. COLETA DOS DADOS

A coleta de dados foi realizada a partir da implantação de equipamentos de registro de paisagem acústica, em diferentes regiões do Cerrado. Estes dispositivos foram desenvolvidos pelo Laboratório de Acústica e Meio Ambiente (LACMAM) do Departamento de Engenharia Mecânica da Escola Politécnica da Universidade de São Paulo, por meio de uma iniciativa

que visava registrar e armazenar a paisagem sonora do Cerrado para uso futuro como banco de dados para investigações científicas.

Neste estudo, foram selecionadas três Estações Ecológicas (ESECs) para representar as diferentes formações naturais do Cerrado: a **ESEC de Águas de Santa Bárbara** (savânica), a **ESEC de Assis** (florestal) e a **ESEC de Itirapina** (campestre). Estas ESECs foram escolhidas por serem representações das três formações do bioma Cerrado, bem como por serem próximas ao Laboratório de Acústica e Meio Ambiente (LACMAM) do Departamento de Engenharia Mecânica da POLI-USP.

Com os equipamentos de captura de paisagem sonora alocados, foram realizadas gravações sonoras em diferentes trechos de vegetação em cada uma dessas ESEC, abrangendo dois pontos na ESEC de Águas de Santa Bárbara e em Assis, e um ponto na ESEC de Itirapina.

Os dados de áudios referentes à ESEC de Itirapina foram gravados em arquivos de 5 minutos, com taxa de amostragem de 8 KHz e profundidade de 16 bits. Já as paisagens acústicas das ESECs de Santa Bárbara e Assis foram gravadas em arquivos de 3 minutos, com taxa de amostragem de 32 KHz e profundidade de 16 bits. Ao todo, foram armazenados cerca de 3 TB de informação.

As gravações utilizadas neste trabalho para a ESEC de Itirapina abrangem os meses de Novembro a Dezembro, enquanto para as ESECs de Santa Bárbara e Assis, o intervalo vai de Setembro a Novembro. Tal especificidade temporal indica que os modelos de Machine Learning e Deep Learning desenvolvidos neste estudo focam exclusivamente nessas janelas temporais e, portanto, não foram avaliados para sua aplicabilidade em outras épocas do ano.

Os dados sonoros das ESECs de Santa Bárbara e Assis apresentam uma elevada taxa de amostragem. Se submetidos à Transformada de Fourier para conversão em uma base de dados, resultariam em uma matriz de dimensões excepcionais (526 bilhões x 3). Tal escala requereria um sistema computacional de alta capacidade para processamento e modelagem. Com o

objetivo de otimizar as análises, o estudo concentra-se em desenvolver e avaliar modelos de Machine Learning para classificar diferentes formações do Cerrado, utilizando uma seleção aleatória de arquivos de áudio do conjunto de dados disponível.

Foram selecionados aleatoriamente 7.500 arquivos, totalizando 70 GB de dados, assegurando uma representação diversificada tanto em termos de condições climáticas quanto de horários de gravação. Essa variedade inclui diferentes cenários, como ventos, chuvas e variações diurnas e noturnas, permitindo aos modelos aprender a reconhecer os padrões das paisagens acústicas do Cerrado em diferentes contextos.

O software R foi empregado para conduzir as análises estatísticas. Para o desenvolvimento dos modelos de Deep Learning, foram utilizadas as bibliotecas Tensorflow (GOOGLE, 2023) e Keras (CHOLLET et al., 2023), do Python.

O ambiente computacional utilizado para estruturar os dados, projetar e treinar os modelos apresentava as seguintes especificações técnicas: uma capacidade de 64GB de memória RAM, um processador dodeca-core e uma placa de vídeo NVIDIA Ge-Force RTX 3090.

4.2. PRÉ-PROCESSAMENTO

Para o treinamento eficaz de modelos de Machine Learning/Deep Learning, os arquivos de áudio precisam ser pré-processados e adaptados em uma estrutura de dados apropriada para a modelagem. Depois de coletar, randomizar e organizar 7.500 gravações, foram realizadas as seguintes etapas de processamento a fim de consolidar as informações em uma base estruturada:

Tempo de gravação

Foi considerado apenas o primeiro minuto (0-59 segundos) de cada arquivo de áudio. Esta escolha deve-se ao objetivo de padronizar a

quantidade de dados extraídos de cada gravação e reduzir a possibilidade de variância decorrente de diferenças de duração entre os arquivos de áudio.

Reamostragem

Por conta da diferença na taxa de amostragem entre os dados da ESEC de Itirapina (8kHz) e as Esecs Assis e Santa Bárbara , (16kHz), estes últimos foram subamostrados. O processo foi feito através da função downsample da biblioteca tuneR[4] do software R. A função ***downsample*** reduz a taxa de amostragem de um arquivo de áudio particionando o número de amostras do arquivo original em n partes iguais, onde

$$n = Taxa\ de\ amostragem\ final * \frac{Número\ de\ amostras\ inicial}{Taxa\ de\ amostragem\ inicial}$$

STFT

Cada arquivo de áudio foi convertido em uma base de dados tridimensional através da função "**spectro**"[5] da biblioteca **seewave**[6] no software R. Essa base de dados é organizada em três colunas distintas: tempo, frequência e amplitude. A primeira coluna representa o tempo ***t*** da gravação, a segunda denota a frequência ***f*** observada nesse instante (janela) ***t***, e a terceira coluna indica a amplitude correspondente ao mesmo instante (janela) ***t*** e frequência ***f***. Desse modo, a base de dados fornece uma representação detalhada das frequências sonoras ao longo do tempo e suas amplitudes associadas.

Existem alguns detalhes cruciais a serem destacados sobre o processo de cálculo da STFT:

A função "spectro" segmenta o sinal de áudio em janelas temporais de tamanho fixo, realiza a transformada de Fourier para cada janela e organiza os resultados em uma matriz em que cada coluna corresponde ao instante

[4] Boa parte da biblioteca ***tuneR*** foi inspirada no pacote ***rastamat*** do Matlab (Hermansky, 1990 e Hermansky et al., 1994).
[5] A função **spectro** fornece uma representação bidimensional do espectro de sinal de onda, correspondente à Transformada de Fourier de Curto Termo (STFT, na sigla em inglês).
[6] A biblioteca **seewave** se baseia no livro "Animal Acoustic Communication: Sound Analysis and Research Methods" (HOPP et al. 1998).

(janela) correspondente à janela e cada linha às amplitudes do espectro de frequência.

Os parâmetros definidos na função "spectro" regulam o tamanho das janelas e o grau de sobreposição entre elas. Nesta análise, foram utilizados os valores padrão da função: o tamanho da janela (parâmetro 'wl') foi fixado em 512 pontos, enquanto a sobreposição (parâmetro 'ovlp') foi configurada como 0, indicando a ausência de sobreposição entre as janelas.

Com esses parâmetros, cada janela de 512 pontos do sinal de áudio corresponde a um vetor de 256 pontos na operação STFT. Assim, a matriz resultante contém 256 linhas.

Dado um tamanho de janela de 512 pontos e uma taxa de amostragem de 8kHz, cada janela representa 512 / 8000 = 0,064 segundos de áudio. Para cobrir 59 segundos de áudio, requer-se 59 / 0,064 ≈ 921 janelas.

A multiplicação do número de linhas (frequências) pelo número de colunas (janelas) resulta no número total de células na matriz, que corresponde ao número total de pontos na representação STFT do sinal de áudio. Neste caso, tem-se 256x921 = 235.776 pontos.

Ao concluir esse processo, cada arquivo de áudio é transformado em uma base de dados composta por 235.776 linhas. Este conjunto de dados não apenas encapsula informações espectrais detalhadas do sinal de áudio, mas também inclui metadados relevantes, como o local e a data da gravação. Assim, essa base de dados multidimensional e abrangente serve como alicerce para o treinamento subsequente e a avaliação dos modelos de Machine Learning/Deep Learning.

Normalização das amplitudes

Para minimizar possíveis vieses devido ao equipamento de gravação utilizado, as amplitudes de todas as bases de dados foram normalizadas.

1. As amplitudes nos instantes ***j*** de cada base de dados ***i*** foram normalizadas através do seguinte princípio:

$$Z_{ij} = \frac{amplitude_{ij} - média(amplitude_i)}{desvio\ padrão(amplitude_i)}$$ garantindo que

Z tenha média igual a 0 e variância igual a 1.

Este processo de pré-processamento transforma as gravações originais em uma matriz essencial para desenvolvimentos futuros. Ela servirá como alicerce fundamental para a construção dos modelos de classificação, sendo utilizada tanto para a extração de características através dos Coeficientes Cepstrais de Frequência de Mel (MFCC's) quanto na criação de espectrogramas.

MFCC

Após os passos de pré-processamento anteriores, a função **melfcc** do pacote **tuneR** foi empregada para calcular os coeficientes Mel-frequency Cepstral (MFCCs) a partir dos sinais de áudio. O MFCC é uma técnica sofisticada que extrai características acústicas, emulando a percepção auditiva humana (DAVIS et. al. 1980). O processo por trás da função melfcc começa com a aplicação de um filtro de pré-ênfase, que amplifica as altas frequências. Em seguida, utiliza-se a Transformada de Fourier de Curto Prazo (STFT) com uma janela de Hamming para obter o espectro de curto prazo. Esse espectro é mapeado em uma escala Mel, que aproxima a percepção humana das frequências. Posteriormente, toma-se o logaritmo desse espectro e, sobre ele, aplica-se a Transformada Cosseno Discreta (DCT) para se obter os coeficientes cepstrais. Finalmente, os primeiros 'ncep' coeficientes são extraídos e representam os MFCCs. Para este estudo, foram considerados os primeiros 12 coeficientes cepstrais. Após a obtenção destes coeficientes, foi realizada uma normalização, com o intuito de eliminar potenciais ruídos provenientes do aparelho utilizado e assegurar a robustez do método.

Espectrogramas

Um espectrograma é uma representação visual do espectro de frequências de um sinal ao longo do tempo.

O primeiro passo do processo de construção dos espectrogramas envolveu a leitura dos arquivos de dados, provenientes da aplicação da STFT, que resultou em uma matriz que contém informações sobre o instante/janela de tempo (time), a frequência (freq) no respectivo instante/janela, e sua amplitude (amp), assim como uma amplitude normalizada (amp_norm).

Durante o processo, utilizou-se o pacote **ggplot2** para criar a imagem visual dos espectrogramas. O processo mapeia o instante/janela de tempo no eixo x, a frequência no eixo y e usa a amplitude normalizada para colorir cada ponto, produzindo uma representação gradiente das amplitudes. Uma escala de cores que varia do violeta ao vermelho foi empregada, representando a variação da amplitude. Os limites do eixo x foram definidos de 0 a 60 e os do eixo y de 0 a 4000, garantindo consistência na representação visual dos dados. Para melhor visualização, ajustes estéticos foram feitos para minimizar distrações visuais, como remover a legenda e ajustar margens. Após este processo, as imagens foram armazenadas como arquivos JPEG de tamanho padrão 1000x1000 pixels.

4.3. MODELAGEM

Nesta etapa da metodologia, serão discutidas a seleção e a implementação dos modelos de aprendizado de máquina utilizados para classificar as paisagens acústicas. A escolha dos modelos foi baseada na capacidade de capturar e aprender padrões complexos nas características acústicas extraídas, bem como na eficácia e eficiência para realizar previsões em novos dados.

Os modelos selecionados foram divididos em dois grupos principais, de acordo com as variáveis utilizadas como entrada para a classificação das paisagens acústicas. O primeiro grupo é composto pelos modelos que utilizam um vetor de características espectrais, baseados nas MFCCs (Davis et. al. 1980) como variáveis de entrada (inputs), que são coeficientes que representam as características espectrais das gravações. O segundo grupo é formado pelos modelos baseados nos espectrogramas, que são imagens bidimensionais que descrevem a energia espectral das gravações ao longo do tempo.

A divisão dos modelos em dois grupos permite a avaliação e comparação do desempenho dos diferentes algoritmos de aprendizado de máquina ao lidar com diferentes tipos de variáveis. Além disso, esta abordagem possibilita a identificação do conjunto de variáveis e do modelo mais adequado para classificar com precisão as paisagens acústicas nas diferentes formações naturais estudadas.

4.3.1. Modelagem com MFCC's

A análise com MFCCs é uma abordagem que utiliza os coeficientes Mel-frequency Cepstral para extrair características discriminativas das paisagens sonoras. Nesta seção, serão detalhados os modelos escolhidos, a construção e seleção das variáveis de entrada e saída, o processo de treinamento e validação dos modelos, visando analisar a eficácia dessa metodologia na classificação das formações do bioma com base em suas características acústicas.

Antes de mergulhar nos detalhes da modelagem utilizando MFCCs, houve uma busca rigorosa e extensa por literaturas relevantes e estudos no campo de paisagens sonoras. O objetivo era compreender, consolidar e sintetizar as melhores práticas, metodologias e insights que poderiam ser aproveitados para este projeto. Esta investigação bibliográfica não apenas embasa as decisões tomadas ao longo do processo de modelagem, mas também garante que o trabalho atual esteja alinhado com as mais recentes inovações e descobertas na área.

4.3.1.1. Modelos

Na modelagem com MFCCs, foram selecionados quatro modelos com base na sua relevância e sucesso em estudos anteriores na área de classificação de paisagens sonoras.

Gradient Boosting (FRIEDMAN, 2001): um exemplo de aplicação desse modelo pode ser encontrado no estudo de Fonseca et al. (2017), que propõe um sistema de classificação de cenas acústicas utilizando duas abordagens: engenharia de características e aprendizado de representações.

O sistema, composto por uma Gradient Boosting Machine e uma Rede Neural Convolucional, melhora o desempenho da linha de base fornecida em 8,2%.

Random Forest (BREIMAN, 2001): este modelo foi utilizado por Grama et al. (2017) em um sistema de classificação de sinais de áudio baseado em Código de Previsão Linear e Florestas Aleatórias. O estudo aborda um problema de classificação multiclasse com conjuntos de dados desequilibrados, classificando sinais relacionados à detecção de intrusos na vida selvagem, como sons de pássaros, tiros, motosserras, vozes humanas e tratores. A taxa de classificação correta geral obtida foi de 99,25%.

Logistic Regression (COX, 1958): o estudo de Noviyanti et al. (2019) empregou regressão logística binária para analisar a composição da paisagem sonora urbana, usando parâmetros de ecologia acústica e coeficientes MFCC. Os resultados mostraram que os coeficientes MFCC são mais eficazes na previsão da percepção sonora, com índices de Classificação Correta (CCR) de até 88,3%, superando as métricas tradicionais da ecologia acústica.

Multilayer Perceptron (ROSENBLATT, 1958): Zhang et al. (2016) utilizaram um Perceptron de Múltiplas Camadas em uma abordagem de classificação multi-rótulo baseada em relevância binária para reconhecer padrões acústicos simultâneos em clipes de áudio de um minuto. O uso de índices acústicos como características globais e o Multilayer Perceptron como classificador base resultou em uma boa performance de classificação com dados de campo. Adicionalmente, a abordagem de classificação multi-rótulo forneceu informações mais detalhadas sobre a distribuição de vários padrões acústicos em gravações de longa duração.

4.3.1.2. Variáveis

A seleção adequada de variáveis é essencial para garantir a precisão e a robustez dos resultados obtidos em uma análise. Neste estudo, a variável resposta é a formação do bioma Cerrado, que será classificada com base na paisagem acústica. Essa variável corresponde à região onde a amostragem foi realizada, representando a formação do bioma em que a gravação sonora

foi feita. A formação do Cerrado é classificada em três níveis: florestal, savânico e campestre.

O MFCC, como um conjunto de características, é derivado da Transformada de Fourier de Curto Prazo (STFT) do sinal de áudio. A STFT é usada para decompor o sinal de áudio em suas componentes de frequência ao longo do tempo. Para obter os MFCCs, o espectro da STFT é primeiramente mapeado em uma escala de frequência Mel, que simula a percepção de frequência do ouvido humano. Esta escala coloca mais resolução em frequências mais baixas e menos nas frequências mais altas. Após a conversão para a escala Mel, o logaritmo da potência em cada banda é calculado. A Transformada de Cosseno Discreta (DCT) é, então, aplicada a esta sequência logarítmica para obter os coeficientes Mel-frequency Cepstral, os MFCCs. Estes coeficientes capturam as características espectrais do sinal de áudio, sendo amplamente utilizados para a análise e modelagem de áudio devido à sua capacidade de representar eficientemente as propriedades sonoras.

Para a modelagem com MFCCs, foram consideradas variáveis explicativas baseadas nos Mel-frequency Cepstral Coefficients (MFCC) com 12 coeficientes. O cálculo dos MFCC foi realizado utilizando a função melfcc da biblioteca tuneR, no software R. Essa função é inspirada na função melfcc.m, desenvolvida na biblioteca rastamat do Matlab (HERMANSKY, 1990 e HERMANSKY et al., 1994). A função melfcc foi aplicada após o processo de downsample (para as regiões de Assis e Santa Bárbara) e antes da padronização da amplitude. Os coeficientes obtidos foram, então, normalizados, seguindo a mesma estratégia adotada para a amplitude. Esta estratégia consistiu em subtrair a média dos coeficientes e dividir pelo desvio padrão de cada conjunto de dados, resultando em dados com média 0 e variância 1. Este processo de normalização assegura a comparabilidade entre diferentes conjuntos de dados e elimina possíveis viéses introduzidos por variações na escala dos coeficientes brutos.

4.3.1.3. Treinamento

A hipótese central deste estudo é que **formações naturais com identidades ecológicas semelhantes apresentam paisagens acústicas similares**. Essa hipótese, baseada na obra de Priestman (2017), fundamenta, ainda que indiretamente, a maior parte dos estudos de Ecologia Acústica e permite a associação de Assis como uma amostra da Formação Florestal, Santa Bárbara como uma amostra da Formação Savânica e Itirapina como uma amostra da Formação Campestre.

Os modelos de aprendizado de máquina Gradient Boosting, Random Forest, Regressão Logística e Perceptron Multicamadas (MLP) foram treinados utilizando os MFCCs como variáveis explicativas (de entrada). Com exceção do MLP, os modelos foram implementados com a biblioteca tidymodels do R, que fornece uma abordagem integrada e consistente para modelagem estatística e aprendizado de máquina (KUHN, 2022), seguindo os princípios do tidyverse.

Entre os pacotes oferecidos por tidymodels, estão o "tune" e o "parsnip". O pacote "tune" desempenha um papel fundamental na otimização dos hiperparâmetros dos modelos de aprendizado de máquina. Hiperparâmetros são configurações do modelo que precisam ser especificadas antes do treinamento e que não são aprendidas a partir dos dados. Eles podem influenciar a eficácia do modelo, e sua otimização permite a avaliação e comparação de diferentes modelos, melhorando seu desempenho.

Por outro lado, o pacote "parsnip" é utilizado para a especificação de modelos. Ele fornece uma interface consistente para a definição e configuração de modelos de aprendizado de máquina, independentemente do motor de computação subjacente que o implementa. Este pacote foi usado extensivamente neste estudo para definir e ajustar os hiperparâmetros dos modelos de Gradient Boosting, Random Forest e Regressão Logística.

Para o modelo Gradient Boosting, utilizando a função boost_tree do pacote parsnip, foram ajustados os seguintes hiperparâmetros: mtry, min_n,

tree_depth, sample_size, learn_rate e loss_reduction. Aqui, mtry se refere ao número de variáveis disponíveis para divisão em cada nó da árvore, min_n é o número mínimo de observações nos nós, tree_depth é a profundidade máxima de qualquer árvore, sample_size é a fração dos dados usados para construir cada árvore, learn_rate é a taxa de aprendizagem e loss_reduction refere-se à redução mínima da perda necessária para fazer uma nova divisão na árvore.

O modelo Random Forest, implementado através da função rand_forest do pacote parsnip, teve os hiperparâmetros mtry e min_n ajustados. O mtry, como no modelo anterior, é o número de variáveis disponíveis para a divisão em cada nó da árvore, enquanto min_n é o número mínimo de observações nos nós.

Na Regressão Logística, usando a função multinom_reg do pacote parsnip, foram ajustados os hiperparâmetros penalty e mixture. O penalty se refere à quantidade de regularização aplicada, que ajuda a evitar o sobreajuste, enquanto mixture determina o tipo de regularização aplicada (L1, L2 ou uma mistura de ambas).

No processo de treinamento dos modelos, os hiperparâmetros foram cuidadosamente ajustados, assim como os pesos associados às variáveis explicativas (MFCCs). Os algoritmos específicos utilizados no treinamento dos modelos, incluindo a seleção e ajuste dos hiperparâmetros, assim como as arquiteturas detalhadas utilizadas para as redes neurais, foram documentados de maneira rigorosa para garantir a reprodutibilidade e a transparência deste estudo. Todos esses detalhes, incluindo o código-fonte usado para implementar e treinar os modelos, podem ser encontrados no Apêndice A deste trabalho.

A arquitetura do MLP foi desenvolvida com base nos pacotes Keras e Tensorflow do R (advindas do Python). Essa arquitetura apresentou seis camadas intermediárias, incluindo duas camadas densas com 30 e 18 neurônios, respectivamente, além de uma camada de dropout e uma de normalização, dispostas entre as camadas densas. O número total de

parâmetros ajustados no modelo MLP foi de 1.101. A arquitetura completa da rede está apresentada no Apêndice A.

Os tempos de treinamento dos modelos variaram significativamente, refletindo as diferenças em complexidade e abordagem de cada algoritmo. O modelo de Gradient Boosting levou 5,86 minutos para ser treinado, enquanto o Random Forest teve um tempo de treinamento de 5,31 minutos. A Regressão Logística apresentou o menor tempo de treinamento, com apenas 9,89 segundos. O Multilayer Perceptron demandou um tempo intermediário, com 2,33 minutos para o treinamento.

Essas diferenças nos tempos de treinamento são relevantes para avaliar a **eficiência** e a **aplicabilidade** de cada modelo no contexto de classificação de paisagens acústicas.

Os tempos de treinamento indicam a **eficiência** computacional de cada algoritmo. Modelos mais rápidos, como a Regressão Logística, são computacionalmente eficientes, o que pode ser especialmente valioso ao trabalhar com grandes conjuntos de dados ou quando se deseja explorar uma variedade de hiperparâmetros ou pré-processamentos de dados. Por outro lado, modelos mais complexos, que demoram mais tempo para treinar, como uma Rede Neural Convolucional, ou, no caso, um Gradient Boosting, podem ser menos eficientes do ponto de vista computacional, mas muitas vezes oferecem uma performance preditiva superior.

A **aplicabilidade** dos modelos, por outro lado, refere-se à sua capacidade de se ajustar aos dados e fazer previsões úteis. Embora todos os quatro modelos sejam aplicáveis à classificação de paisagens acústicas, as diferenças em seus tempos de treinamento e suas performances preditivas podem afetar sua viabilidade em diferentes contextos. Por exemplo, se o tempo é uma consideração crítica, pode-se optar por usar um modelo mais rápido como a Regressão Logística, mesmo que a sua performance preditiva seja um pouco inferior. Da mesma forma, se a performance preditiva é a prioridade e os recursos computacionais e o tempo não são um problema, um modelo que leva mais tempo para treinar, como o Gradient Boosting, ou uma Rede Neural Convolucional, pode ser o mais apropriado.

A escolha final do modelo depende de um equilíbrio entre **eficiência** e **aplicabilidade**.

4.3.1.4. Validação

A validação dos modelos elaborados por meio das bibliotecas tidymodels, TensorFlow e Keras foi efetuada mediante a divisão sistemática dos dados em três conjuntos distintos: treinamento, validação e teste cego. Inicialmente, foram alocadas 1.000 observações para o conjunto de teste cego, enquanto as demais 6.500 observações foram partilhadas em 70% para treinamento e 30% para validação.

A base de treinamento é o conjunto de dados usado para instruir o modelo a realizar previsões. Já a base de validação é usada para testar o desempenho do modelo durante a fase de treinamento. Este processo fornece um feedback valioso para ajustar o modelo e evitar o sobreajuste - quando o modelo “decora” os dados de treinamento, ao invés de “aprender” com eles.

O teste cego, por sua vez, também conhecido como teste de validação externa, é uma prática comum em ciência de dados para avaliar o desempenho dos modelos treinados com dados nunca antes vistos por eles. Trata-se de um teste robusto de generalização, pois verifica se os modelos são capazes de fazer previsões corretas em dados novos e desconhecidos.

Os modelos foram posteriormente avaliados com base em três métricas derivadas da matriz de confusão: **acurácia, precisão** e **revocação** (FAWCETT, 2006). A acurácia é a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões. A precisão, por sua vez, é a proporção de verdadeiros positivos (previsões corretas da classe positiva) em relação ao total de previsões positivas realizadas pelo modelo. Já a revocação, também conhecida como sensibilidade, mede a proporção de verdadeiros positivos em relação ao total de observações realmente positivas no conjunto de dados.

Essas métricas proporcionam uma visão detalhada e precisa da eficácia das previsões do modelo, complementando-se mutuamente para

avaliar tanto a proporção de acertos quanto a habilidade do modelo de identificar corretamente as diferentes classes.

Em suma, a tripartição dos dados em treinamento, validação e teste cego configura uma estratégia efetiva e robusta para assegurar a precisão e eficiência das previsões, permitindo a obtenção de uma visão abrangente sobre o desempenho dos modelos desenvolvidos e sua capacidade de generalização.

4.3.2. Modelagem com Espectrogramas

Após a construção do modelo que utiliza os Coeficientes Cepstrais de Frequência Mel (MFCCs) como variáveis explicativas, o estudo prossegue com a "Análise com Espectrogramas". Esta seção se concentra nas imagens dos espectrogramas das paisagens sonoras, que serão adotadas como variáveis explicativas em uma Rede Neural Convolucional (CNN). Estas representações bidimensionais, que exibem a distribuição de frequências e amplitudes ao longo do tempo, cujo seu desenvolvimento fora detalhado na etapa de pré-processamento, possibilitam o uso da CNN para a classificação das formações do bioma Cerrado. Nesta parte do estudo, serão detalhados o modelo, as variáveis e a técnica de treinamento/validação. O objetivo desta seção é avaliar a efetividade da CNN na detecção precisa das diferentes formações do bioma, com base em suas características acústicas através de representações visuais.

4.3.2.1. Modelos

Nesta seção, será abordado o modelo selecionado para a análise de espectrogramas:

Analogamente à modelagem com MFCCs, a escolha do modelo de Rede Neural Convolucional (CNN) para a análise de espectrogramas foi precedida por uma busca extensa e criteriosa na literatura especializada. A intenção era identificar, compreender e consolidar as melhores práticas, metodologias e insights que poderiam ser aproveitados neste estudo. O uso da CNN, conforme documentado em pesquisas anteriores, demonstra grande

potencial na classificação de sons ambientais com base em espectrogramas. Assim, considerando a riqueza de informações contidas nos espectrogramas e os resultados promissores obtidos em estudos anteriores, optou-se por essa abordagem na tentativa de alcançar uma classificação precisa das paisagens sonoras do bioma Cerrado.

Convolutional Neural Network (CNN) (LECUN, 1998): A CNN tem sido amplamente aplicada em estudos de classificação de sons ambientais com base em espectrogramas. Um exemplo de pesquisa nessa área é o trabalho de Khamparia et al. (2019). Os autores exploraram a classificação de sons ambientais utilizando redes neurais profundas, especificamente Redes Neurais Convolucionais (CNN) e Redes de Empilhamento Profundo Tensor (TDSN). A metodologia envolveu a geração de espectrogramas dos sons ambientais, que são imagens bidimensionais representando a frequência, intensidade e tempo dos sons, e utilizando esses espectrogramas como entrada para as redes de aprendizado profundo. Essas redes foram treinadas com dois conjuntos de dados de sons ambientais amplamente reconhecidos: o ESC-10 e o ESC-50. O ESC-10 é composto por 10 classes de sons, enquanto o ESC-50 contém 50 classes diferentes de sons ambientais, como praias, chuva, canto de pássaros etc (PICZAK, 2015). Ambos os conjuntos de dados foram criados para avaliar a capacidade de algoritmos de aprendizado de máquina em classificar sons ambientais. Os principais resultados demonstraram que a CNN alcançou uma precisão de 77% no conjunto de dados ESC-10 e 49% no ESC-50, enquanto a TDSN obteve uma precisão de 56% no conjunto de dados ESC-10. Os autores concluíram que o uso de espectrogramas de sons como entrada para redes de aprendizado profundo pode ser um método eficiente para o desenvolvimento de sistemas de classificação e reconhecimento de sons.

4.3.2.2. Variáveis

Para a construção dos modelos, foram consideradas duas categorias de variáveis explicativas: o primeiro grupo, já abordado na seção 3.1, é composto pelos 12 coeficientes Mel-frequency Cepstral Coefficients (MFCC).

O segundo grupo consiste nos espectrogramas (imagens) das paisagens sonoras.

Os espectrogramas foram gerados a partir das bases de tempo, frequência e amplitude normalizada, conforme descrito na seção de "Pré-processamento". Esses espectrogramas foram construídos com especificações de 250 px de largura, 250 px de altura, limite temporal de 0 a 59 segundos, limite de frequência de 0 a 4 kHz, e remoção de eixos, legendas e títulos para preservar apenas a imagem resultante.

A seguir, são apresentados exemplos aleatórios de espectrogramas das três regiões analisadas:

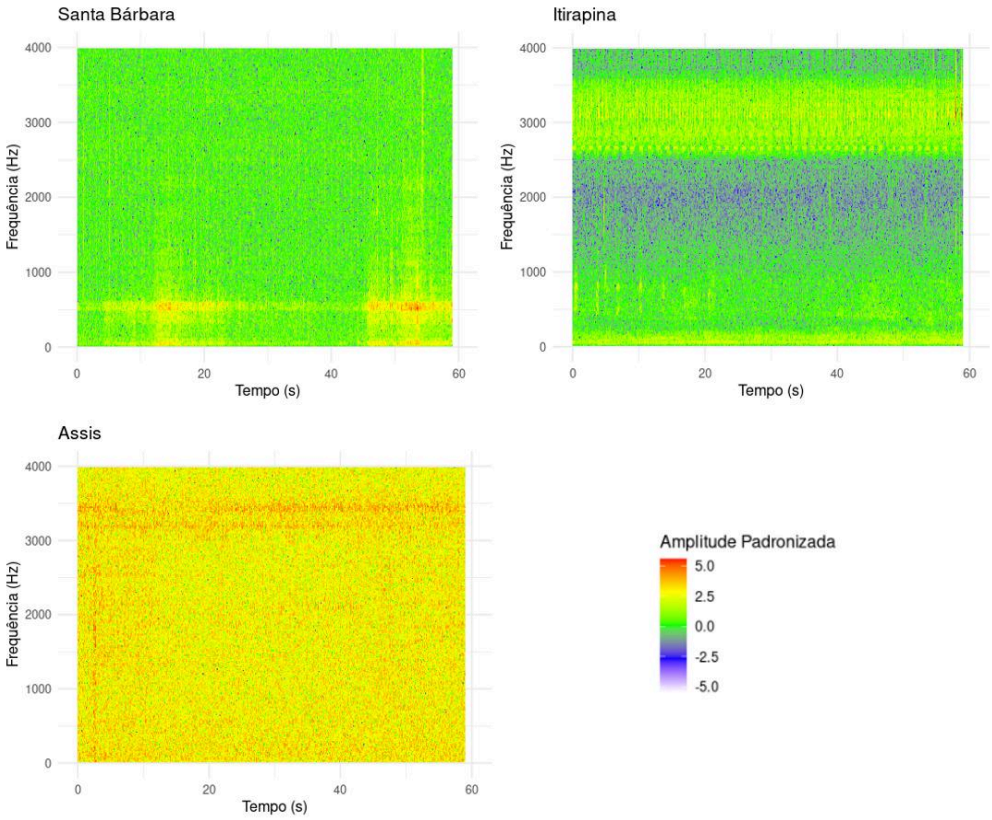


Figura 6 - Exemplo de espectrogramas das três regiões/formações

4.3.2.3. Treinamento

A Rede Neural Convolucional (CNN) foi desenvolvida utilizando a linguagem de programação R e as bibliotecas Keras e TensorFlow. A base de dados foi dividida em três partes: treinamento (6000 observações), validação

(500 observações) e teste (1000 observações), visando uma avaliação adequada do modelo durante o treinamento e, posteriormente, testar sua capacidade de generalização.

A arquitetura da CNN é composta por várias camadas, incluindo camadas convolucionais, camadas de normalização em lote (Batch Normalization), camadas de pooling máximo (Max Pooling) e camadas de dropout espacial (Spatial Dropout). A rede inicia com uma camada de entrada com dimensões 250x250x3, que representam as dimensões das imagens de entrada. A arquitetura completa da rede está apresentada no Apêndice A.

O modelo possui três blocos convolucionais, cada um com duas camadas convolucionais, seguidas por uma camada de normalização em lote, uma camada de pooling máximo e uma camada de dropout espacial. O número de filtros nas camadas convolucionais aumenta progressivamente a cada bloco, iniciando com 32, depois 64 e, por fim, 128. Todas as camadas convolucionais utilizam funções de ativação ReLU e possuem tamanho de kernel de 5x5, com preenchimento ('padding') do tipo 'same' para garantir que a saída tenha a mesma dimensão que a entrada.

Após os três blocos convolucionais, a rede possui uma camada de global average pooling, seguida por uma camada de achatamento (Flatten) e uma camada densa (Dense) com 128 unidades e função de ativação ReLU. Uma camada de dropout com taxa de 0,4 é aplicada antes da camada de saída, que possui três unidades e função de ativação softmax, correspondente às três categorias de classificação.

O modelo tem um total de 815.139 parâmetros ajustáveis, dos quais 814.243 são treináveis e 896 não são treináveis. Durante o treinamento, o modelo foi compilado utilizando o otimizador Adam com taxa de aprendizado de 0,001, função de perda 'categorical_crossentropy' e métrica de avaliação 'accuracy'. Foram usadas duas funções de callback: uma para reduzir a taxa de aprendizado quando a perda de validação parasse de melhorar (callback_reduce_lr_on_plateau) e outra para parar o treinamento quando a acurácia de validação superasse 98% após 10 épocas consecutivas acima de 97% (CustomEarlyStopping).

O treinamento do modelo envolveu até 300 épocas (eixo x), com um lote de 64 unidades. Os dados de validação foram essenciais para monitoramento e ajuste da taxa de aprendizado do modelo. A performance é demonstrada na figura subsequente, onde o eixo y exibe a diminuição da função de perda ('loss') e o aumento da acurácia, tanto no conjunto de validação quanto no de treinamento.

A aplicação de 'callbacks' mostrou-se uma técnica eficiente, contribuindo para a prevenção de 'overfitting' e economia de tempo computacional. Com essa estratégia, o treinamento da Rede Neural Convolucional (CNN) foi completado em 79 épocas, durando um total de 8 dias, sem o uso de uma Unidade de Processamento Gráfico (GPU).

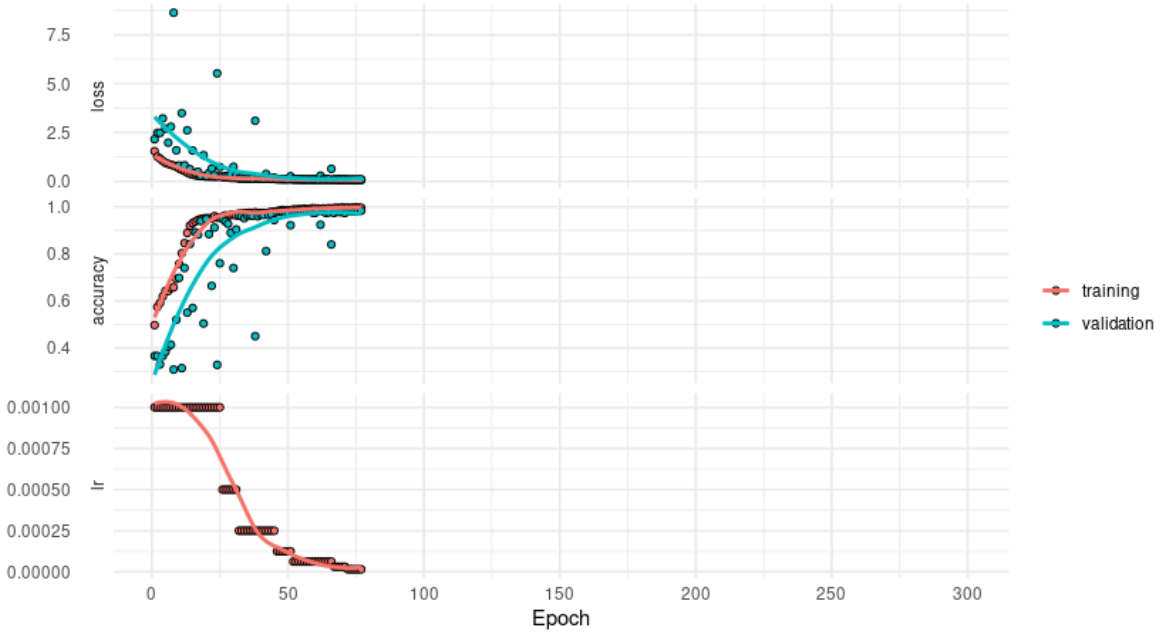


Figura 7 - Treinamento da CNN.

As diferenças nos tempos de treinamento devem ser consideradas ao avaliar a eficiência e aplicabilidade de cada modelo no contexto de classificação de paisagens acústicas.

Levando em conta os modelos construídos na seção Análise com MFCC, os tempos de treinamento variaram significativamente, demonstrando as diferenças na complexidade e abordagem de cada algoritmo. Por exemplo,

recapitulando o que foi discutido na no subtópico Treinamento da seção anterior, o tempo de treinamento dos modelos que utilizaram MFCCs como variáveis explicativas variou de aproximadamente 10 segundos (Regressão Logística) até 6 minutos (Gradient Boosting).

Por outro lado, a Rede Neural Convolucional (CNN) empregada na análise com espectrogramas exigiu o maior tempo de treinamento entre os modelos, destacando a complexidade dessa abordagem. Essas diferenças nos tempos de treinamento são cruciais ao avaliar a eficiência e a aplicabilidade de cada modelo no contexto de classificação de paisagens sonoras, permitindo a escolha da melhor opção conforme a situação.

4.3.2.4. Validação

No caso da Rede Neural Convolucional (CNN) aplicada na análise com espectrogramas, a divisão entre a base de treinamento, a base de validação e a base de teste foi estruturada com 6.000 observações para o treinamento do modelo e 500 para a validação, além de outras 1.000 observações reservadas para o teste cego.

Essa abordagem de teste cego, utilizando uma base separada de 1.000 observações, permitiu uma avaliação robusta do desempenho da CNN, incluindo acurácia, precisão e revocação, ao mesmo tempo que proporcionou insights valiosos sobre a capacidade de generalização do modelo. Ao testar a CNN com dados desconhecidos, foi possível verificar se as previsões eram consistentes e eficazes mesmo fora do conjunto de treinamento e validação.

Assim, a estratégia de divisão entre treinamento, validação e teste foi crucial para assegurar a confiabilidade e a precisão das previsões feitas pela CNN, bem como para proporcionar uma análise comparativa mais justa com os resultados obtidos na seção anterior, que focou na análise com os MFCCs.

4.4. EXPLICABILIDADE

A capacidade de entender e interpretar modelos de aprendizado de máquina tornou-se um campo emergente de pesquisa conhecido como

Explainable AI (IA Explicável) (RIBEIRO et al., 2016). A transparência e a confiabilidade dos modelos de aprendizado de máquina são essenciais, principalmente em aplicações que têm repercussões significativas ou que lidam com informações sensíveis. Dessa maneira, a IA Explicável desempenha um papel crucial, permitindo que as decisões tomadas pelos algoritmos sejam justificadas e compreendidas tanto por especialistas como pelo público geral.

Os estudos na área da IA Explicável buscam estabelecer métodos, métricas e taxonomias que auxiliem na interpretação de modelos complexos (DOSHI-VELEZ & KIM, 2017; CARVALHO et al., 2019). Além disso, a pesquisa tem se estendido para integrar insights das ciências sociais para melhorar o entendimento humano sobre as decisões tomadas por algoritmos (MILLER, 2019). Vários métodos e abordagens têm sido propostos para tornar os modelos de aprendizado de máquina mais interpretáveis, como o uso de "explicações locais" (RIBEIRO et al., 2016) e a aplicação de teorias da informação para uma melhor compreensão dos modelos (CHEN et al., 2018). Tais avanços mostram que a IA Explicável é uma área de pesquisa multifacetada, focada em tornar as decisões de algoritmos transparentes, justificáveis e compreensíveis.

Dentro desse contexto, o LIME (Local Interpretable Model-agnostic Explanations) surge como uma ferramenta influente para a explicabilidade dos modelos. Proposto por Ribeiro et al. em 2016, o LIME busca criar explicações aproximadas e interpretações humanamente compreensíveis para qualquer modelo de aprendizado de máquina, independentemente de sua complexidade. Utilizando a abordagem de perturbar os dados de entrada e aprender um modelo linear interpretável no contexto local, consegue elucidar e traduzir os processos de decisão do modelo original.

O LIME é fundamentado na ideia de que, mesmo que o modelo global seja complexo e difícil de interpretar, como é o caso das Redes Neurais Convolucionais (CNNs), ainda é possível aproximá-lo com um modelo linear mais simples em uma pequena vizinhança em torno de uma instância de interesse. Este modelo linear pode ser facilmente interpretado e fornece

insights sobre as decisões do modelo mais complexo nesta vizinhança específica.

Para atingir esse objetivo, o LIME começa por perturbar a instância de entrada (neste caso, um espectrograma) múltiplas vezes, gerando um conjunto de dados "perturbado". Em seguida, faz-se predições com o modelo complexo (a CNN) para cada uma dessas instâncias perturbadas. O próximo passo é aprender um modelo linear simples (por exemplo, uma regressão linear) usando os dados perturbados e suas respectivas predições. As características (ou superpixels, no caso de imagens) que são mais influentes na decisão são representadas pelos pesos maiores no modelo linear. Assim, o LIME consegue traduzir de forma eficaz o que acontece no espaço intrincado e de alta dimensionalidade do modelo original para um espaço de interpretação mais compreensível.

Um ponto-chave do LIME é que ele não busca explicar o funcionamento global do modelo, mas sim suas decisões em torno de instâncias específicas. Isso o torna altamente valioso para verificar se um modelo está fazendo escolhas por razões justificáveis. Por exemplo, no contexto desta pesquisa, o LIME poderia revelar se a CNN está identificando uma formação do Cerrado com base em características sonoras genuínas ou se está sendo influenciada por artefatos ou ruídos nos espectrogramas.

Além disso, o LIME é versátil. Ele foi projetado para ser "agnostic" em relação ao modelo, o que significa que ele pode ser aplicado a uma variedade de modelos de aprendizado de máquina, desde árvores de decisão até redes neurais. Esta flexibilidade torna o LIME uma ferramenta de escolha para muitos pesquisadores que buscam entender e confiar mais nas decisões de seus modelos complexos.

Em resumo, o LIME oferece uma ponte entre o mundo complexo e muitas vezes opaco dos modelos de aprendizado de máquina avançados e a necessidade humana de compreensão e transparência. Ao fazer isso, ele ajuda a garantir que os insights derivados desses modelos sejam não apenas precisos, mas também confiáveis e aplicáveis de maneira responsável em contextos práticos, como a preservação do bioma Cerrado.

Além do LIME, o Grad-CAM, apresentado por Selvaraju et al. (2017), também se destaca em explicabilidade. Ele é particularmente útil com CNNs, permitindo visualizar as áreas de uma imagem que uma CNN considera importante durante a classificação. De forma semelhante ao LIME, Selvaraju e sua equipe usaram o Grad-CAM para entender como modelos identificavam diferentes usos da terra em imagens de satélite. Os mapas de calor produzidos por essa técnica revelaram partes cruciais da imagem para a decisão do modelo, oferecendo perspectivas valiosas aos ecologistas.

Neste trabalho, foi o LIME que foi adotado para a Rede Neural Convolucional (CNN) no exame das paisagens acústicas do Cerrado. Ao se debruçar sobre 12 características e 50 superpixels presentes nos espectrogramas, foi possível discernir quais áreas e frequências eram mais relevantes para classificar as formações do Cerrado. Optando pela CNN, pudemos contornar as intrincadas dificuldades de interpretação associadas ao MFCC, que intrinsecamente não oferece uma claridade interpretativa semelhante.

Portanto, a inclusão do LIME neste estudo solidificou a competência da CNN na análise de paisagens acústicas, fornecendo um arcabouço rigoroso para a pesquisa em explicabilidade. Estes insights e metodologias podem ser determinantes para pesquisas subsequentes em Ecologia Acústica, servindo como fundamento para melhorar a eficácia dos modelos e aperfeiçoar os esforços de conservação do bioma Cerrado.

5. RESULTADOS

Ao todo, cinco modelos estatísticos foram desenvolvidos com o intuito de classificar a Formação do Cerrado (Florestal, Savânica e Campestre) com base nas paisagens sonoras das respectivas regiões. As variáveis independentes utilizadas consistiram nos coeficientes cepstrais de frequência mel (MFCCs) para os modelos de Gradient Boosting, Random Forest, Regressão Logística e Multilayer Perceptron, e nas imagens dos espectrogramas para a Rede Neural Convolucional (CNN).

Os resultados obtidos por cada modelo, considerando as métricas de acurácia, precisão e revocação, estão apresentados na Tabela 1.

Tabela 1 - Desempenho dos modelos desenvolvidos conforme as métricas analisadas.

Modelo	**Acurácia**	**Precisão**	**Revocação**	**Tempo de treinamento**
Gradient Boosting	93%	93%	93%	5,86 minutos
Random Forest	92%	92%	92%	5,31 minutos
Regressão Logística	82%	83%	82%	9,89 segundos
Multilayer Perceptron	83%	79%	69%	2,33 minutos
CNN	99%	98%	98%	~ 8 dias

A análise dos dados apresentados na Tabela 1 revela que a Rede Neural Convolucional (CNN) exibiu o melhor desempenho em todas as métricas, alcançando uma acurácia de 99%, precisão de 98% e revocação de 98%. Esses resultados evidenciam que a CNN foi a metodologia mais eficiente no que se refere à identificação do tipo de formação do Cerrado a partir da paisagem sonora, dentre os modelos avaliados.

Abaixo, a Tabela 2 ilustra a matriz de confusão da Rede Neural Convolucional (CNN) aplicada ao conjunto do teste cego, com 1.000 observações. Esta tabela proporciona uma visão detalhada sobre o

desempenho do modelo em classificar corretamente as diferentes categorias de formação do Cerrado.

Tabela 2 - Matriz de confusão da CNN aplicada à base do teste cego.

	Previsto		
Observado	Florestal	Campestre	Savânica
Florestal	324	0	5
Campestre	0	347	1
Savânica	6	0	317

Entretanto, é importante considerar que, em alguns casos práticos, o ganho de desempenho proporcionado pela CNN pode não ser tão relevante (LECUN et al., 1998; CHANG et al., 2015). Para a classificação de paisagens acústicas, fatores como uma acurácia razoável, simplicidade do método de modelagem e tempo de resposta da predição podem ser determinantes na escolha do modelo mais adequado.

Quando o objetivo é classificar um evento que pode ser observado várias vezes em um curto período de tempo, uma abordagem inspirada em Breiman (1996) pode ser considerada. Embora o conceito original de Breiman envolva a criação de múltiplos modelos a partir de um conjunto de dados através de amostragens bootstrap para melhorar a acurácia do modelo, a ideia central de combinar múltiplas previsões para obter um resultado final se mantém. Neste contexto, sugere-se a utilização de um modelo com desempenho razoável para classificar repetidamente o evento em questão. A classificação final para o evento é então dada pela média das probabilidades encontradas nas diversas classificações.

Essa abordagem adaptada pode ser aplicada a diversos processos de classificação de eventos que ocorrem ao longo do tempo, como imagens capturadas em vídeos (classificação e reconhecimento de imagens), sons

registrados em gravações (classificação e reconhecimento de sons), e outras observações baseadas em sinais que variam com o tempo.

Assim, é possível utilizar modelos mais simples, como o Gradient Boosting ou o Random Forest, para realizar múltiplas classificações em amostras curtas de paisagens acústicas para melhorar a eficiência do método.

Dessa forma, ainda que a CNN apresente a melhor performance entre os modelos avaliados, é importante ponderar se essa diferença de desempenho justifica a sua adoção em detrimento de modelos mais simples, como Gradient Boosting ou Random Forest. A escolha do modelo mais adequado deve levar em consideração não apenas a performance, mas também a simplicidade do método, o tempo de resposta da predição e a capacidade de lidar com um grande número de observações em um curto período de tempo.

Entretanto, apesar da complexidade da CNN, a aplicação do LIME auxiliou na compreensão das principais características do espectrograma que classificam a paisagem acústica em uma determinada formação do Cerrado. Os resultados do LIME, aplicado em um conjunto de três exemplos aleatórios de espectrogramas, um de cada região, mostraram que a formação florestal, representada por Assis, é explicada pela faixa de frequência que vai até cerca de 1.1 KHz. A formação savânica, representada por Santa Bárbara, é explicada pela faixa de frequência que vai até cerca de 0.6 KHz. Por fim, a formação campestre, representada por Itirapina, é explicada pela faixa de frequência que vai até cerca de 0.4 KHz. Estes resultados podem ser verificados nos espectrogramas de exemplos explicados pelo LIME, na figura abaixo.

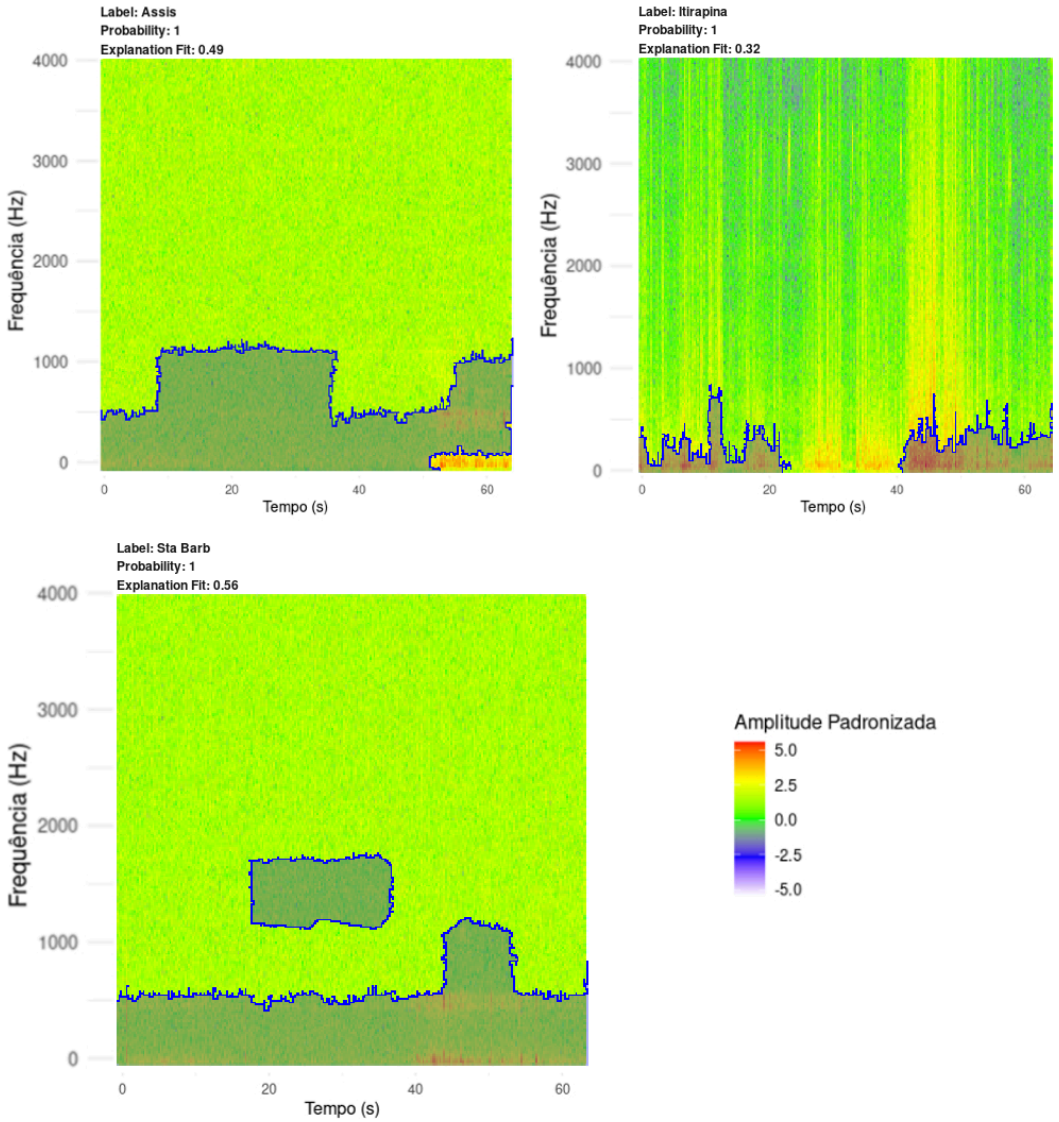


Figura 8 - Aplicação do LIME para explicar a CNN em exemplos de espectrogramas.

Nesta pesquisa, o LIME foi empregado especificamente para as Redes Neurais Convolucionais (CNN's) devido à sua capacidade de desvendar a importância de variáveis em modelos altamente complexos como estes. A escolha por não aplicar o LIME aos modelos que utilizaram os MFCC's como variáveis explicativas baseou-se em duas razões principais. Primeiramente, os modelos que incorporam os MFCC's como variáveis explicativas, como Random Forest e Gradient Boosting, possuem seus próprios mecanismos para avaliar a importância das variáveis, não necessitando do LIME para tal fim. Em segundo lugar, os MFCC's, por sua natureza, não são facilmente interpretáveis. Assim, mesmo que técnicas explicativas apontassem a relevância de determinados coeficientes cepstrais, a contextualização e

compreensão desses coeficientes no domínio da ecologia acústica seriam desafiadoras. Dessa forma, o foco do uso do LIME restringiu-se às CNN's, buscando maximizar a clareza e aplicabilidade dos insights gerados.

De todo modo, seja através de Redes Neurais Convolucionais aplicadas sob os espectrogramas, ou através de modelos menos complexos, como Random Forest e Gradient Boosting, este estudo demonstra que a combinação de Ecologia Acústica e Ciência de Dados pode ser eficiente na classificação de paisagens acústicas.

6. DISCUSSÃO

Os resultados obtidos neste estudo demonstram o potencial e os benefícios de encontrar um modelo adequado para a classificação de paisagens acústicas, combinando Ecologia Acústica e Ciência de Dados. A aplicação destes modelos possibilita uma melhor compreensão das características distintivas de diferentes formações do Cerrado, contribuindo para estudos ecológicos e para a preservação destes ambientes.

Um aspecto importante a ser destacado é a facilidade na coleta dos dados utilizados nos modelos. Como os modelos foram treinados com amostras aleatórias de trechos das paisagens acústicas, eles aprenderam os padrões de cada região independentemente do horário da gravação e da condição climática, como ventos e chuvas. Isso significa que, mesmo com poucos minutos de gravação na primavera, em qualquer período do dia, os modelos são capazes de gerar resultados confiáveis. No entanto, é importante ressaltar que os dados analisados neste estudo referem-se apenas à primavera, portanto, deve-se ter cautela ao generalizar os resultados para outras estações do ano.

Os resultados deste estudo destacaram-se não apenas pela performance excepcional da Rede Neural Convolucional (CNN), que atingiu uma acurácia de 99%, mas também por duas observações relevantes: primeiro, os modelos mais simples, como Gradient Boosting e Random Forest, que utilizam MFCCs, também demonstraram um desempenho notável na classificação das paisagens acústicas do Cerrado; segundo, a possibilidade de adoção de uma estratégia eficaz de “classificação repetitiva” para eventos de alta exposição, como é o caso das paisagens acústicas, tende a ser promissora.

Neste método, registros contínuos do ambiente são segmentados em subamostras menores, cada uma sendo classificada de forma independente. Conforme o número de subamostras classificadas aumenta e todas originam-se do mesmo contexto ambiental, a probabilidade de os modelos menos complexos e com performance moderada classificarem corretamente o evento em análise é amplificada, graças ao princípio de voto majoritário. Esta

abordagem retira sua eficácia da consistência e convergência evidenciada na identificação acurada da formação natural do Cerrado ao longo de diversas observações, ecoando os fundamentos da Lei dos Grandes Números (BERNOULLI, 1713).

De todo modo, embora a CNN possa não ser o modelo mais adequado para todas as aplicações práticas devido à sua alta complexidade, a utilização do LIME em conjunto com a CNN revelou nuances e características nos sinais de áudio que explicam a classificação do modelo. Com o auxílio do LIME, foi possível identificar as faixas de frequência que caracterizam cada região, permitindo uma melhor compreensão dos fatores que influenciam a classificação das paisagens acústicas em diferentes formações do Cerrado. Essa abordagem também destaca a importância de se explorar diferentes técnicas e modelos para extrair insights valiosos dos dados analisados.

Em conclusão, este estudo demonstra que a aplicação de diferentes modelos de classificação de paisagens acústicas pode fornecer insights valiosos para a compreensão das características das formações do Cerrado. Os resultados obtidos destacam a importância de se escolher o modelo mais adequado para cada situação, levando em consideração aspectos como performance, complexidade e custo computacional.

7. CONCLUSÃO

Neste estudo, foi investigada a aplicação de diferentes modelos de aprendizado de máquina para classificar a formação do Cerrado (Florestal, Savânica e Campestre) com base nas paisagens sonoras das respectivas regiões. Foram desenvolvidos cinco modelos estatísticos, utilizando coeficientes cepstrais de frequência mel (MFCCs) como variáveis independentes para os modelos de Gradient Boosting, Random Forest, Regressão Logística e Multilayer Perceptron, e imagens dos espectrogramas para a Rede Neural Convolucional (CNN).

Os resultados obtidos revelaram que a CNN apresentou o melhor desempenho em todas as métricas avaliadas, alcançando uma acurácia de 99%, precisão de 98% e revocação de 98%. Entretanto, é importante ponderar se a diferença de desempenho entre a CNN e os modelos mais simples, como Gradient Boosting ou Random Forest, justifica a sua adoção, considerando fatores como simplicidade do método, tempo de resposta da predição e capacidade de lidar com um grande número de observações em um curto período de tempo.

Embora a eficiência e a aplicabilidade de cada modelo variem, todos eles apresentam potencial para classificar paisagens acústicas de forma efetiva. O modelo escolhido para qualquer projeto ou análise específica dependerá do balanceamento entre o tempo disponível, os recursos computacionais e a performance desejada.

Foi destacado que, quando o objetivo é classificar eventos que ocorrem várias vezes em um curto período, a utilização de modelos com desempenho razoável, como o Gradient Boosting, Random Forest ou a Regressão Logística, pode ser mais apropriada. Uma vez que o modelo esteja bem ajustado e possuir uma performance razoável, a média das probabilidades encontradas em múltiplas classificações pode fornecer uma classificação final precisa e convergir para a classificação verdadeira da região, com base em elementos existentes no princípio da Lei dos Grandes Números.

A aplicação da técnica de explicabilidade LIME contribuiu para a compreensão dos processos de decisão do modelo de CNN, fornecendo insights sobre como melhorar sua performance e aplicação em estudos futuros de preservação e recuperação do bioma Cerrado. A análise das frequências relevantes para cada formação do Cerrado, obtidas a partir da aplicação do LIME, demonstrou o potencial dessa técnica na identificação das características do espectrograma responsáveis pela classificação das paisagens acústicas.

Este estudo demonstra que a combinação de Ecologia Acústica e Ciência de Dados pode ser eficiente na classificação de paisagens acústicas e na identificação do tipo de formação do Cerrado. Os resultados obtidos ressaltam a importância de selecionar o modelo mais adequado para cada situação, levando em consideração não apenas a performance, mas também a simplicidade do método, o tempo de resposta da predição e a capacidade de processar um grande volume de observações.

As conclusões deste trabalho podem servir como base para futuras pesquisas na área de Ecologia Acústica, bem como na aplicação de técnicas de aprendizado de máquina em estudos ambientais. A metodologia apresentada pode ser adaptada e aprimorada para abordar outras questões relacionadas à conservação e monitoramento de ecossistemas, contribuindo para a preservação da biodiversidade e a recuperação de biomas ameaçados, como o Cerrado.

8. TRABALHOS FUTUROS

A presente dissertação estabelece um marco no estudo da modelagem e classificação de paisagens sonoras com base em atributos acústicos. A seguir, delineiam-se diretrizes para futuras investigações inspiradas por este trabalho:

Aplicação em Ecologia Acústica e Preservação Ambiental: A investigação presente serve de fundação para pesquisadores da ecologia acústica e preservação ambiental. A técnica de classificação baseada em sons oferece meios para monitorar ambientes, identificar mudanças em ecossistemas e formular estratégias de conservação.

Adoção de Modelos Simplificados para Eventos de Alta Exposição: A complexidade de um modelo nem sempre determina sua eficiência. Para eventos de alta exposição, como classificações de imagens ou sons, um modelo mais simples pode atender às necessidades. Se o objetivo se concentra em uma classificação generalizada, o pesquisador pode empregar modelos de performance razoável. Ao aplicá-los repetidamente, a estratégia de voto majoritário fortalece a decisão final, potencializando sua precisão.

Emprego do LIME em CNN para Sinais Sonoros: O estudo sugere a combinação do método LIME (Local Interpretable Model-Agnostic Explanations) com uma CNN adaptada para espectrogramas. Este método permite ao pesquisador discernir as faixas de frequência cruciais para as decisões de classificação do modelo, aprimorando sua compreensão dos determinantes sonoros.

Extensão para Diferentes Modalidades de Sinais: A técnica discutida nesta dissertação tem potencial de aplicação em áreas diversificadas, como sismologia, análise de tráfego em redes e estudos eletrocardiográficos. Em cada contexto, a identificação dos determinantes chave das classificações se revela essencial.

Integração com Outras Fontes de Dados: A pesquisa futura pode contemplar a fusão dos espectrogramas com outros conjuntos de dados,

como registros meteorológicos ou visuais. Esta abordagem integrada pode fornecer insights mais holísticos dos ambientes sob estudo e refinar a precisão da classificação.

Ao avançar neste campo de estudo, espera-se que a comunidade acadêmica não apenas refine a precisão das classificações, mas também expanda sua compreensão das inter-relações entre diferentes modalidades de dados e os fenômenos que representam.

9. REFERÊNCIAS

BERNOULLI, Jakob. Ars Conjectandi: Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis. Basel, Suiça: Thurneysen Brothers, 1713.

BREIMAN, L. Bagging Predictors. Machine Learning, v. 24, p. 123-140, 1996. Boston: Kluwer Academic Publishers. Disponível em: https://sci2s.ugr.es/keel/pdf/algorithm/articulo/1996-ML-Breiman-Bagging%20Predictors.pdf. Acesso em: 28 abr. 2023.

BREIMAN, L. Random Forests. Machine Learning, v. 45, p. 5-32, 2001. Disponível em: https://doi.org/10.1023/A:1010933404324. Acesso em: 03 fev. 2023.

CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics, v. 8, n. 8, p. 832, 2019. Disponível em: https://doi.org/10.3390/electronics8080832. Acesso em: 26 ago. 2023.

CAVA, M.G.B.; PILON, N.A.L.; RIBEIRO, M.C.; DURIGAN, G. Abandoned pastures cannot spontaneously recover the attributes of old-growth savannas. Journal of Applied Ecology, v. 55, p. 1164–1172, 2018. Disponível em: https://doi.org/10.1111/1365-2664.13046. Acesso em: 23 ago. 2022.

CHEN, J.; SONG, L.; WAINWRIGHT, M. J.; JORDAN, M. I. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. Disponível em: https://doi.org/10.48550/arXiv.1802.07814. Acesso em: 26 ago. 2023.

CHOLLET, François. Keras: The Python Deep Learning library. Astrophysics Source Code Library, record ascl:1806.022, Jun. 2018. Disponível em: https://keras.io/. Acesso em: 6 jun. 2023.

COX, D. R. The Regression Analysis of Binary Sequences (with Discussion). Journal of the Royal Statistical Society, Series B, v. 20, n. 2, p. 215-242, 1958.

DATA SCIENCE ACADEMY. Deep Learning Book. 2019. Disponível em: http://www.deeplearningbook.com.br/. Acesso em: 23 ago. 2022.

DAVIS, S. B.; MALMBERG, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, v. 28, n. 4, p. 357-366, ago. 1980.

DIAS, B.F.S. Alternativas de Desenvolvimento do Cerrado: Manejo e Conservação dos Recursos Naturais Renováveis. Brasília: Fundação Pró-Natureza (FUNATURA), 1991.

DOSHI-VELEZ, F.; KIM, B. Towards A Rigorous Science of Interpretable Machine Learning. ArXiv: Machine Learning, 2017. Disponível em: https://doi.org/10.48550/arXiv.1702.08608. Acesso em: 26 ago. 2023.

FAWCETT, T. An introduction to ROC analysis. Pattern Recognition Letters, v. 27, n. 8, p. 861-874, 2006.

FONSECA, E. et al. Acoustic scene classification by ensembling gradient boosting machine and convolutional neural networks. DETECTION AND CLASSIFICATION OF ACOUSTIC SCENES AND EVENTS WORKSHOP DCASE2017, Munique, 2017.

FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, v. 29, n. 5, p. 1189-1232, 2001.

FUKUSHIMA, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, v. 36, n. 4, p. 193-202, 1980.

GILPIN, L. H. et al. Explaining Explanations: An Overview of Interpretability of Machine Learning. In: PROCEEDINGS OF THE 5TH IEEE INTERNATIONAL CONFERENCE ON DATA SCIENCE AND ADVANCED ANALYTICS (DSAA 2018), 2018. Disponível em: https://doi.org/10.48550/arXiv.1806.00069. Acesso em: 26 ago. 2023.

GOOGLE BRAIN TEAM. TensorFlow: An end-to-end open source machine learning platform. Disponível em: https://www.tensorflow.org/. Acesso em: 6 jun. 2023.

GRAMA, L.; RUSU, C. Audio signal classification using Linear Predictive Coding and Random Forests. International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucareste, Romêmia, 2017.

HE, K. et al. Deep Residual Learning for Image Recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, p. 770-778, 2016.

HEBB, D. O. The organization of behavior. New York: Wiley & Sons, 1949.

HERMANSKY, H. Perceptual Linear Predictive (PLP) Analysis of Speech. The Journal of the Acoustical Society of America, vol. 87, nº. 4, p. 1738-1752, abril de 1990.

HERMANSKY, H.; MORGAN, N. Rasta Processing Of Speech. IEEE Transactions on Speech and Audio Processing, vol. 2, nº. 4, p. 578-589, outubro de 1994.

HILASACA, L.H. et al. Visual Active Learning for Labeling: A Case for Soundscape Ecology Data. Information, v. 12, n. 7, 2021. Disponível em: https://doi.org/10.3390/info12070265. Acesso em: 23 ago. 2022.

HOPP, S. L. et al. Animal Acoustic Communication: Sound Analysis and Research Methods. Berlin: Springer International, 1998.

HUBEL, D. H.; WIESEL, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, v. 160, n. 1, p. 106-154, 1962.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. Banco de Dados de Informações Ambientais. Disponível em: https://bdiaweb.ibge.gov.br/#/consulta/pesquisa. Acesso em: 23 ago. 2022.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. Contas De Ecossistemas - O Uso da Terra nos Biomas Brasileiros. 2020. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101753.pdf. Acesso em: 23 ago. 2022.

KHAMPARIA, A. et al. Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. IEEE, v. 7, p. 7717 - 7727, 2019. Disponível em: https://ieeexplore.ieee.org/abstract/document/8605515. Acesso em: 23 ago. 2022.

KRAUSE, B. The Great Animal Orchestra: Finding the Origins of Music in the World's Wild Places. New York: Little, Brown and Company, 2012.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - NEURIPS'12, v. 1, p. 1097-1105, 2012.

KUHN, M.; SILGE, J. Tidy Modeling with R. California: O'Reilly Media, 2022.

LECUN, Yann. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, vol. 86, no. 11, p. 2278-2324, 1998.

HASAN, Nahian Ibn. Bird Species Classification And Acoustic Features Selection Based on Distributed Neural Network with Two Stage Windowing of Short-Term Features. 2022. Disponível em: https://doi.org/10.48550/arXiv.2201.00124. Acesso em: 15 mar. 2024.

MILLER, Tim. Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence, v. 267, 2017. DOI: 10.1016/j.artint.2018.07.007. Disponível em: https://doi.org/10.48550/arXiv.1706.07269. Acesso em: 26 ago. 2023.

NOVIYANTI, A; SUDARSONO, A. S.; KUSUMANINGRUM, D. Urban soundscape prediction based on acoustic ecology and MFCC parameters. AIP Conference Proceedings, v. 2187, 2019. Disponível em: https://doi.org/10.1063/1.5138335. Acesso em: 03 fev. 2023.

PELLEGRINI, A. F. A. et al. Birdsong and anthropogenic noise: implications for conservation. Landscape Ecology, v. 35, n. 5, p. 1161-1179, 2020.

PICZAK, Karol J. Environmental Sound Classification with Convolutional Neural Networks. In: Proceedings of the International Conference on Computational Intelligence in Music, Sound, Art and Design (CIM'15), 2015.

PRIESTMAN, K. The Science of Soundscapes. Inside Ecology, Set. 2017. Disponível em: https://insideecology.com/2017/09/28/the-science-of-soundscapes/. Acesso em: 23 ago. 2022.

RIBEIRO, J. F.; WALTER, B.M.T. As Principais Fitofisionomias do Bioma Cerrado. In: SANO, S. M.; ALMEIDA, S. P. de; RIBEIRO, J. F. (Eds.). Cerrado: ecologia e flora, v. 2. Brasília: EMBRAPA-CERRADOS, 2008.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, Califórnia, p. 97-101, 2016. Disponível em https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf. Acesso em: 22 abr. 2023.

ROSENBLATT, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review, v. 65, n. 6, p. 386-408, 1958.

SCHAFER, R. M. The Soundscape: Our Sonic Environment and the Tuning of the World. Rochester: Destiny Books, 1977.

SELVARAJU, R. R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 2017. Disponível em: https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf. Acesso em: 21 ago. 2023.

SHEARER, C. The CRISP-DM model: The new blueprint for data mining. Journal of Data Mining, v. 5, n. 4, 2000. Disponível em: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf. Acesso em: 26 ago. 2023.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (ICLR), 2015.

STOWELL, D. et al. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ, v. 2, e488, 2014.

SZEGEDY, C. et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 1-9, 2015.

THE ASIMOV INSTITUTE. The Neural Network Zoo, 2016. Disponível em: https://www.asimovinstitute.org/neural-network-zoo/. Acesso em: 23 ago. 2022.

TRUAX, B. Acoustic Communication. 2. ed. Westport: Ablex Publishing, 2001.

TUCKER, D.; GAGE, S.H.; WILLIAMSON, I. et al. Linking ecological condition and the soundscape in fragmented Australian forests. Landscape Ecology, v. 29, p. 745–758, mar. 2014. Disponível em: https://doi.org/10.1007/s10980-014-0015-1. Acesso em: 23 ago. 2022.

WRIGHTSON, K. An Introduction to Acoustic Ecology. Soundscape: The Journal of Acoustic Ecology, p. 10-13, 2000. Disponível em:

http://www.econtact.ca/5_3/wrightson_acousticecology.html. Acesso em: 23 ago. 2022.

ZHANG, L.; TOWSEY, M.; XIE, J.; ZHANG, J.; ROE, P. Using multi-label classification for acoustic pattern detection and assisting bird species surveys. Applied Acoustics, v. 110, p. 91–98, set. 2016. Disponível em: https://doi.org/10.1016/j.apacoust.2016.03.027. Acesso em: 03 fev. 2023.

ZHANG, X.; ZHU, M.; WANG, T.; SUN, J. Deep learning for remote sensing data: A technical tutorial on the state of the art. IEEE Geoscience and Remote Sensing Magazine, v. 4, n. 2, p. 22-40, 2015. Disponível em: https://ieeexplore.ieee.org/document/7486259. Acesso em: 21 abr. 2023.

ZNIDERSIC, E.; WATSON, D.M. Acoustic Restoration: Using Soundscapes to Benchmark and Fast-Track Recovery of Ecological Communities. Ecology Letters, v. 25, p. 1597 – 1603, abr. 2022. Disponível em: https://doi.org/10.1111/ele.14015. Acesso em: 23 ago. 2022.

APÊNDICE A

— Funções de treinamento dos modelos —

```
library(tidyverse)
library(tidymodels)
library(purrr)

# Função para otimização de hiperparâmetros de modelos de classificação
meu_tune_grid <- function(workflow,rsamples_cv, grid = 5) {
  tune_grid(
    workflow,
    resamples = rsamples_cv,
    grid = grid,
    metrics = metric_set(accuracy, precision, recall, roc_auc),
    control = control_grid(verbose = TRUE, allow_par = TRUE)
  )
}

# Função para ajustar o modelo usando validação cruzada e otimização de hiperparâmetros
ajusta_modelo <- function(recip,nomes = ' ',modelo,rsample_cv,tipo = 'class'){

  wf <- workflows::workflow() %>%
    workflows::add_recipe(recip) %>%
    workflows::add_model(modelo)

  # Registrando o tempo de execução
  tictoc::tic(nomes)
  tune_grid <- meu_tune_grid(wf,rsample_cv)
  tictoc::toc()

  resp <- meu_fit(tune_grid, modelo, wf)

  return(resp)
}

# Definindo modelos para ajustar (XGBoost, Random Forest e Regressão Logística)
modelos = list(
  xgb_model = parsnip::boost_tree(
    mtry = tune::tune(), min_n = tune::tune(),
```

```
    tree_depth = tune::tune(), trees = 1500,
    sample_size = 0.75, learn_rate = tune::tune(),
    loss_reduction = tune::tune()
  ) %>%
    parsnip::set_mode("classification") %>%
    parsnip::set_engine("xgboost"),

      rf_model = parsnip::rand_forest(mtry = tune::tune(), min_n =
tune::tune(), trees = 1000) %>%
    parsnip::set_mode("classification") %>%
    parsnip::set_engine("ranger"),

    lr_model = parsnip::multinom_reg(penalty = tune::tune(), mixture =
tune::tune()) %>%
    parsnip::set_mode("classification") %>%
    parsnip::set_engine("glmnet")
)

# Ajustando os modelos
tictoc::tic('Tempo total')
ft <- purrr::map2(
  modelos,
  names(modelos),
   ~ajusta_modelo(recip = recip, modelo = .x, nomes = .y, rsample_cv =
rsample)
)
tictoc::toc()

## KERAS - MLP ##

# Definindo a arquitetura da rede neural
input_x <- keras::layer_input(shape = ncol(x_train))
output <- input_x %>%
  # Definindo as camadas da rede neural
  keras::layer_dense(units = 30, activation = 'relu') %>%
  keras::layer_dropout(0.5) %>%
  keras::layer_layer_normalization() %>%
  keras::layer_dense(units = 18, activation = 'relu') %>%
  keras::layer_dropout(0.5) %>%
  keras::layer_layer_normalization() %>%
  keras::layer_dense(units = 3, activation = 'softmax')

model <- keras::keras_model(inputs = input_x, outputs = output)
```

— Arquitetura do MLP —

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 12)]              0
 dense_2 (Dense)             (None, 30)                390
 dropout_1 (Dropout)         (None, 30)                0
 layer_normalization_1 (LayerNormalization)  (None, 30)  60
 dense_1 (Dense)             (None, 18)                558
 dropout (Dropout)           (None, 18)                0
 layer_normalization (LayerNormalization)  (None, 18)  36
 dense (Dense)               (None, 3)                 57
=================================================================
Total params: 1101 (4.30 KB)
Trainable params: 1101 (4.30 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

— Treinamento da CNN —

```
# Definindo a camada de entrada com imagens de tamanho 250x250 e 3 canais (RGB)
input <- keras::layer_input(shape = c(250,250,3))

# Definindo o número inicial de filtros para a CNN
fts <- 32

# Construindo a arquitetura da CNN
output <- input %>%

  # Primeiro bloco de convolução com regularização L2
    keras::layer_conv_2d(filters = fts, kernel_size = c(5,5), activation = 'relu', padding = 'same', kernel_regularizer = regularizer_l2(0.001)) %>%
  keras::layer_batch_normalization() %>%
    keras::layer_conv_2d(filters = fts, kernel_size = c(5,5), activation = 'relu', padding = 'same', kernel_regularizer = regularizer_l2(0.001)) %>%
  keras::layer_batch_normalization() %>%
  keras::layer_max_pooling_2d(pool_size = c(2,2)) %>%
  keras::layer_spatial_dropout_2d(0.3) %>%

   # Segundo bloco de convolução com regularização L2 e o dobro de filtros
    keras::layer_conv_2d(filters = 2*fts, kernel_size = c(5,5), activation = 'relu', padding = 'same', kernel_regularizer = regularizer_l2(0.001)) %>%
  keras::layer_batch_normalization() %>%
    keras::layer_conv_2d(filters = 2*fts, kernel_size = c(5,5), activation = 'relu', padding = 'same', kernel_regularizer = regularizer_l2(0.001)) %>%
```

```
  keras::layer_batch_normalization() %>%
  keras::layer_max_pooling_2d(pool_size = c(2,2)) %>%
  keras::layer_spatial_dropout_2d(0.3) %>%

  # Terceiro bloco de convolução com regularização L2 e quatro vezes o número inicial de filtros
   keras::layer_conv_2d(filters = 4*fts, kernel_size = c(5,5), activation = 'relu', padding = 'same', kernel_regularizer = regularizer_l2(0.001)) %>%
  keras::layer_batch_normalization() %>%
   keras::layer_conv_2d(filters = 4*fts, kernel_size = c(5,5), activation = 'relu', padding = 'same', kernel_regularizer = regularizer_l2(0.001)) %>%
  keras::layer_batch_normalization() %>%
  keras::layer_max_pooling_2d(pool_size = c(2,2)) %>%
  keras::layer_spatial_dropout_2d(0.3) %>%

   # Camadas finais: agrupamento global, achatamento e camadas densas
  keras::layer_global_average_pooling_2d() %>%
  keras::layer_flatten() %>%
   keras::layer_dense(units = 4*fts, activation = 'relu', kernel_regularizer = regularizer_l2(0.001)) %>%
  keras::layer_dropout(0.4) %>%
  keras::layer_dense(units = 3, activation = 'softmax') # Três unidades de saída, supondo 3 classes

# Definindo o modelo com entrada e saída
model <- keras::keras_model(
  input,
  output)

# Definindo o callback para reduzir a taxa de aprendizado se a perda de validação estagnar
reduce_lr <- keras::callback_reduce_lr_on_plateau(
  monitor = "val_loss",
  factor = 0.5,
  patience = 5,
  verbose = 1,
  mode = "auto",
  min_delta = 0.0001,
  cooldown = 0,
  min_lr = 0
)

# Compilando o modelo com otimizador Adam e perda de entropia cruzada categórica
model %>%
```

```
  keras::compile(
    optimizer = keras::optimizer_adam(learning_rate = 0.001),
    loss = 'categorical_crossentropy',
    metrics = 'accuracy'
  )

# Definindo um callback personalizado para interromper o treinamento se a acurácia de validação exceder 98% por 10 épocas consecutivas
CustomEarlyStopping <- R6::R6Class(
  "CustomEarlyStopping",
  inherit = KerasCallback,

  public = list(
    epochs_above_97 = 0,
    consecutive_epochs = 10,
    target_accuracy = 0.982,

    on_epoch_end = function(epoch, logs = NULL) {
      val_acc <- logs[["val_accuracy"]]

      if (val_acc > 0.97) {
        self$epochs_above_97 <- self$epochs_above_97 + 1
      } else {
        self$epochs_above_97 <- 0
      }

      if (self$epochs_above_97 >= self$consecutive_epochs && val_acc > self$target_accuracy) {
                          cat("Atingiu a acurácia alvo de 98% após", self$consecutive_epochs, "épocas consecutivas acima de 97%. Parando o treinamento...\n")
        self$model$stop_training <- TRUE
      }
    }
  )
)

stop_training <- CustomEarlyStopping$new()

# Treinando o modelo com os dados, usando um tamanho de lote de 64 e 300 épocas máximas
hist <- model %>%
  keras::fit(
    data_cnn$train$x,
    data_cnn$train$y,
    batch_size = 64,
```

```
    epochs = 300,
    validation_data = list(data_cnn$valid$x,data_cnn$valid$y),
    view_metrics = TRUE,
      callbacks = list(stop_training, reduce_lr) # Adicionando os callbacks definidos anteriormente
  )
```

— Arquitetura da CNN —

```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 250, 250, 3)]     0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 250, 250, 32)      2432
_________________________________________________________________
batch_normalization_5 (BatchNormalization) (None, 250, 250, 32)      128
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 250, 250, 32)      25632
_________________________________________________________________
batch_normalization_4 (BatchNormalization) (None, 250, 250, 32)      128
_________________________________________________________________
max_pooling2d_2 (MaxPooling2D) (None, 125, 125, 32)      0
_________________________________________________________________
spatial_dropout2d_2 (SpatialDropout2D) (None, 125, 125, 32)      0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 125, 125, 64)      51264
_________________________________________________________________
batch_normalization_3 (BatchNormalization) (None, 125, 125, 64)      256
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 125, 125, 64)      102464
_________________________________________________________________
batch_normalization_2 (BatchNormalization) (None, 125, 125, 64)      256
_________________________________________________________________
max_pooling2d_1 (MaxPooling2D) (None, 62, 62, 64)       0
_________________________________________________________________
spatial_dropout2d_1 (SpatialDropout2D) (None, 62, 62, 64)       0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 62, 62, 128)       204928
_________________________________________________________________
batch_normalization_1 (BatchNormalization) (None, 62, 62, 128)       512
_________________________________________________________________
conv2d (Conv2D)              (None, 62, 62, 128)       409728
_________________________________________________________________
batch_normalization (BatchNormalization) (None, 62, 62, 128)       512
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 31, 31, 128)       0
_________________________________________________________________
spatial_dropout2d (SpatialDropout2D) (None, 31, 31, 128)       0
_________________________________________________________________
global_average_pooling2d (GlobalAveragePooling2D) (None, 128)              0
_________________________________________________________________
flatten (Flatten)            (None, 128)               0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               16512
_________________________________________________________________
dropout (Dropout)            (None, 128)               0
_________________________________________________________________
dense (Dense)                (None, 3)                 387
=================================================================
Total params: 815,139
Trainable params: 814,243
Non-trainable params: 896
_________________________________________________________________
```