



**Universidade de Brasília  
Faculdade de Tecnologia**

**Sistema de aprendizado de máquina  
para tipificação de Incêndios Florestais na  
Amazônia  
com série temporal**

Caio de Lima Saigg  
Ian Porto e Mello

PROJETO FINAL DE CURSO  
ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Brasília  
2023

**Universidade de Brasília  
Faculdade de Tecnologia**

**Sistema de aprendizado de máquina  
para tipificação de Incêndios Florestais na  
Amazônia  
com série temporal**

Caio de Lima Saigg  
Ian Porto e Mello

Projeto Final de Curso submetido como requi-  
sito parcial para obtenção do grau de Enge-  
nheiro de Controle e Automação

Orientador: Prof. Dr. Edson Mintsu Hung

Brasília  
2023

L732s Lima Saigg, Caio de.  
Sistema de aprendizado de máquina para tipificação de Incêndios Florestais na Amazônia com série temporal / Caio de Lima Saigg; Ian Porto e Mello; orientador Edson Mintsu Hung.  
-- Brasília, 2023.  
54 p.

Projeto Final de Curso (Engenharia de Controle e Automação)  
-- Universidade de Brasília, 2023.

1. Monitoramento da Amazônia. 2. Aprendizado de máquina. 3. Queimadas. 4. Sensoriamento remoto. 5. Banco de dados. 6. Censipam. 7. Censipam. I. Porto e Mello, Ian. II. Mintsu Hung, Edson, orient. III. Título

**Universidade de Brasília  
Faculdade de Tecnologia**

**Sistema de aprendizado de máquina  
para tipificação de Incêndios Florestais na Amazônia  
com série temporal**

Caio de Lima Saigg  
Ian Porto e Mello

Projeto Final de Curso submetido como requisito parcial para obtenção do grau de Engenheiro de Controle e Automação

Trabalho aprovado. Brasília, 20 de dezembro de 2023:

---

**Prof. Dr. Edson Mintsu Hung,**  
UnB/FT/ENE  
Orientador

---

**Prof. Dr. Mylène C.Q. Farias,**  
UnB/FT/ENE  
Examinador interno

---

**Dr. Henrique Bernini**  
Examinador externo

Brasília  
2023

*Este trabalho é dedicado a toda nossa equipe do projeto Curupira,  
minha família que sempre foi minha casa  
e meus amigos que me acompanharam nessa jornada.*

Caio de Lima Saigg

*Dedico deste trabalho à equipe do projeto Curupira e à minha família e amigos, cujo apoio e  
carinho foram indispensáveis e me motivaram a ir além.*

Ian Porto e Mello

# Agradecimentos

Agradecemos ao Dr. Edson Mintsu Hung por sua supervisão e orientação em nosso trabalho, aos professores Dr. Diogo Caetano Garcia e Dr. Nilson Donizete por suas contribuições e ensinamentos, ao nosso colega de projeto Bruno Scholles Soares Dias e ao Coordenador Operacional do Centro Regional de Porto Velho do Censipam Dr. Henrique Bernini.

*“If you find that you’re spending almost all your time on theory,  
start turning some attention to practical things;  
it will improve your theories.  
If you find that you’re spending almost all your time on practice,  
start turning some attention to theoretical things;  
it will improve your practice.”  
(Donald Knuth)*

# Resumo

O trabalho propõe o desenvolvimento de um algoritmo baseado em inteligência artificial para monitoramento de incêndios no território da Amazônia Legal, como ferramenta de auxílio para a tipificação diária do fogo. O uso de aprendizado de máquina foi escolhido por ser um método comprovadamente eficaz em projetos de proteção ambiental, que permite rápido reconhecimento de novos eventos com base em informações históricas. O trabalho foca na classificação de incêndios, e para isso os modelos são treinados com uma base de dados pública que não fornece informações em tempo real, mas possui a classificação para diversas queimadas ocorridas nos últimos anos

**Palavras-chave:** Monitoramento da Amazônia. Aprendizado de máquina. Queimadas. LSTM. Sensoriamento remoto. Banco de dados. Censipam.



# Abstract

This work proposes the development of an algorithm based on artificial intelligence for monitoring forest fires in the territory of the Legal Amazon, as an aid tool for daily classification of fire. The use of machine learning was chosen for being a proven effective method in environmental protection projects, which allows fast recognition of new events based on historical information. The work focuses on the classification of fires, and for this the models are trained with a public database that does not provide real-time information, but has the classification for several fires that occurred in recent years.

**Keywords:** Amazon Monitoring. Machine Learning. Fire. LSTM. Remote sensing. Dataset. Censipam.

# Lista de ilustrações

Figura 1.1	Taxas consolidadas anuais de desmatamento do PRODES (A..., 2022) .	13
Figura 2.2	Mosaico com 229 imagens Landsat em escala 1: 250.000 (COORDENACAO-GERAL..., 2019) . . . . .	17
Figura 4.3	Representação de Padding . . . . .	26
Figura 4.4	Representação visual de <i>Underfitting</i> , boa generalização e <i>Overfitting</i> (CÁRDENAS-MONTES, 2015) . . . . .	27
Figura 4.5	Ilustração exemplo de uma célula de um modelo LSTM (LSTM..., 2023)	30
Figura 4.6	Possíveis arquiteturas com LSTM (LSTM..., 2023) . . . . .	31
Figura 5.7	Mapa da América do Sul com os eventos de fogo detectados . . . . .	33
Figura 5.8	Tabela de atributos do <i>shapefile</i> . . . . .	33
Figura 5.9	Mapa no QGIS com eventos de fogo . . . . .	35
Figura 5.10	Ilustrações demonstrativas do crescimento de um evento de fogo ao longo do tempo . . . . .	35
Figura 5.11	Start day x Is new x Is active . . . . .	36
Figura 5.12	Biome x Region x Protected . . . . .	37
Figura 5.13	Biome x Region x Protected . . . . .	37
Figura 5.14	Deforestat x Tree cover x Biomass . . . . .	37
Figura 5.15	FRP x Detections x Size . . . . .	37
Figura 5.16	Persistence x Daytime x Progression . . . . .	37
Figura 5.17	Start day x Is new x Is active . . . . .	37
Figura 5.18	Esquemático do processo de obtenção de <i>features</i> . . . . .	39
Figura 5.19	Diagramas de Caixa de dados semelhantes e com valores de média e desvio padrão próximos . . . . .	41
Figura 5.20	Diagramas de Caixa de variação de área da queimada . . . . .	42
Figura 5.21	Diagramas de Caixa de variação de área da queimada corrigido em 10e6	42
Figura 6.22	<i>Overfitting</i> no treinamento . . . . .	45
Figura 6.23	Fluxograma do processo de treinamento, análise e produção do sistema em LSTM . . . . .	46
Figura 6.24	Gráfico da presença de cada classe dentro do <i>Dataset</i> . . . . .	47
Figura 6.25	Gráfico do <i>Dataset</i> após <i>data augmentation</i> e remoção aleatória de dados	48
Figura A.26	Código do modelo LSTM . . . . .	54
Figura A.27	Código do modelo de Loss Focal em pytorch (LIN et al., 2017) . . . . .	54

# Lista de tabelas

Tabela 6.1	Parâmetros do Modelo . . . . .	44
Tabela 6.2	Resultados da Avaliação . . . . .	47
Tabela 6.3	Resultados da Avaliação . . . . .	48

# Sumário

<b>1</b>	<b>Introdução</b>	<b>13</b>
1.1	Queimadas e Amazônia	13
1.2	Objetivos	14
1.3	Este documento	16
<b>2</b>	<b>Trabalhos relacionados</b>	<b>17</b>
2.1	História do Monitoramento da Amazônia por meio de Sensoriamento Remoto	17
2.2	Outros trabalhos na área de monitoramento ambiental	19
<b>3</b>	<b>Metodologia</b>	<b>21</b>
<b>4</b>	<b>Fundamentação Teórica</b>	<b>23</b>
4.1	Python	23
4.2	Scikit-Learn	23
4.3	PyTorch	23
4.4	Aprendizado de Máquina	24
4.4.1	<b>Modelo</b>	25
4.4.2	<b>Função de Custo</b>	25
4.4.3	<b>Dataset</b>	25
4.4.4	<b>Overfitting</b>	27
4.4.5	<b>Métricas de Avaliação do Modelo</b>	28
4.5	Modelos de aprendizado de máquina	29
4.6	Evento de Fogo	31
<b>5</b>	<b>Banco de dados</b>	<b>32</b>
5.1	GFED	32
5.1.1	<i>Shapefile</i>	32
5.1.2	<i>Dataframe</i>	34
5.2	Censipam	34
5.3	Análise das <i>features</i> do GFED	36
5.4	Adição de <i>features</i>	38
5.5	Dados nos Servidores do Censipam	40
5.5.1	Dados de Treinamento e Dados em Produção	40
5.5.2	Comparando dados do Treinamento com os dados em Produção	40
<b>6</b>	<b>Resultados</b>	<b>44</b>
6.1	LSTM	44

6.1.1	Modelo . . . . .	44
6.1.2	Lidando com <i>Overfitting</i> . . . . .	44
6.1.3	Organização em Série Temporal . . . . .	45
6.1.4	Fluxograma do Desenvolvimento . . . . .	45
6.1.5	Desempenho . . . . .	46
6.1.6	Balanceamento . . . . .	46
<b>7</b>	<b>Conclusão . . . . .</b>	<b>49</b>
	<b>Referências . . . . .</b>	<b>50</b>
	<b>Apêndice A Códigos . . . . .</b>	<b>54</b>

# 1 Introdução

## 1.1 Queimadas e Amazônia

O Brasil é o país com a maior presença da floresta Amazônica na América, 60% do total da floresta está dentro do território brasileiro, o que equivale a aproximadamente 5 milhões de quilômetros quadrados ou 59% do território nacional (AMAZONIA..., 2023). Por consequência disto, possuímos uma enorme responsabilidade com ela, existente em diversas formas, como na defesa do território nacional ou na proteção de um enorme patrimônio da humanidade.

Tal patrimônio vem requerendo cada vez mais atenção nos últimos tempos, de acordo com dados gerados pelo PRODES (A..., 2022), entre agosto de 2021 e julho de 2022 a taxa de desmatamento da Amazônia Legal foi cerca de  $11594\text{km}^2$ . Como podemos ver na Figura 1.1, tal taxa vem acompanhando uma alta nos últimos 10 anos do desmatamento no bioma. O INPE (INPE..., 2023) estima que, desde 1988, 17% do bioma já tenha sido desmatado. Tais dados reforçam a necessidade de métodos de monitoramento e controle que reduzam as perdas do meio ambiente.

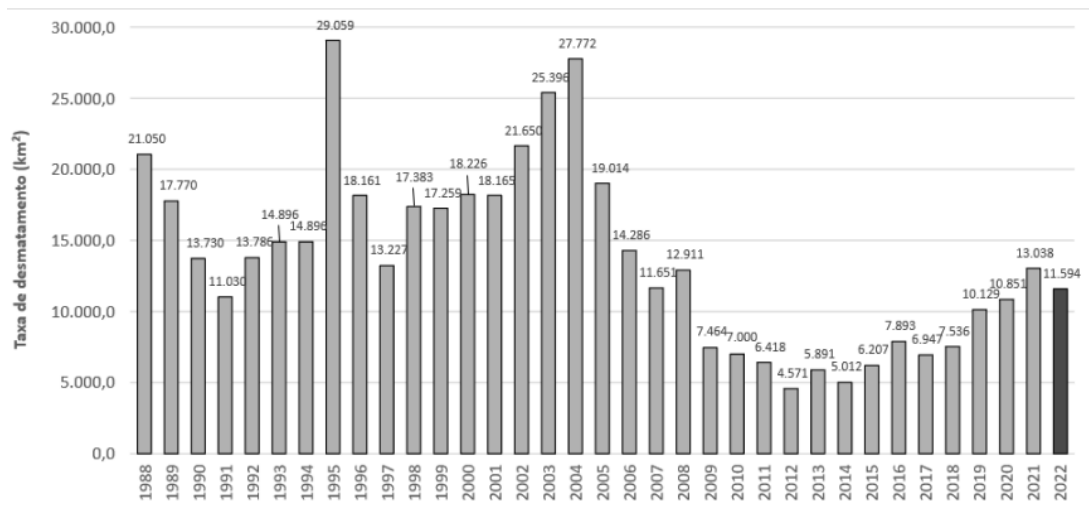


Figura 1.1 – Taxas consolidadas anuais de desmatamento do PRODES (A..., 2022)

Uma importante área de atuação na preservação da floresta se trata do combate à incêndios. Todos os anos, milhares de quilômetros da floresta amazônicas são queimados por incêndios florestais, causando impactos ecológicos e econômicos, redução da biomassa florestal, mudanças na composição de espécies de árvores e esgotamento do solo (PENHA, 2018).

Os incêndios podem ocorrer naturalmente ou podem estar atrelados a atividades humanas. No entanto, queimadas resultantes da ação humana constituem quase que a

totalidade dos incêndios registrados nos territórios no Brasil, que ocupa o 5º lugar entre os países mais poluidores (INPE... , 2022).

Já as queimadas naturais, por sua vez, costumam ter sua ignição dada por raios, tornando-as mais comuns em períodos de secas prolongadas quando as queimadas costumam atingir seu pico. São mais frequentes no bioma Cerrado que é naturalmente mais seco e adaptado a queimadas naturais recorrentes, normalmente no início e fim das estações chuvosas. Em contraste, o bioma Amazônia é extremamente úmido e apresenta baixos índices de queimadas naturais.

Conforme pesquisas recentes (INPE, 2022), a queima e o desflorestamento excessivo na floresta Amazônica estão afetando a capacidade de absorção de CO<sub>2</sub> da atmosfera, ao mesmo tempo que estas atividades emitem grandes quantidades de carbono por meio da queima de madeira. Além disso, estas ações apresentam um impacto climático na região, tornando a estação seca mais seca, quente e longa, agredindo ainda mais o bioma.

## 1.2 Objetivos

Os incêndios florestais de origem antropológica são predominantemente e principalmente ligados ao desmatamento e a agricultura de manutenção (BARLOW; CARMENTA; FRANÇA, 2020). Nem todo fogo iniciado por atividade humana resultará em tragédia, mas a frouxidão das leis de proteção ambiental estimula pessoas mal intencionadas a queimar grandes áreas públicas para grilagem (VIEIRA; SÁ; PIEDADE, 2021), então ser capaz de controlar queimadas é importante tanto para proteger os bens naturais quanto para garantir a aplicação da legislação.

Dessa forma, este projeto visa criar uma ferramenta de **tipificação automática de fogo** no território amazônico para auxiliar no combate ao fogo na região. De acordo com o *EU Horizon 2020 Work Programme* (EU... , 2020), o ciclo de gestão do fogo pode ser amplamente segmentado em três etapas: (1) prevenção e preparação (pré-fogo); (2) detecção e resposta (gestão de incêndios florestais ativos); (3) atividades de restauração e adaptação (pós-incêndio). Este trabalho estará focado em desenvolver uma ferramenta que sirva de auxílio para a segunda etapa, na resposta dos órgãos de defesa para com incêndios ativos.

Este trabalho visou o desenvolvimento de um **algoritmo** projetado para ser incluído ao Painel do Fogo, uma plataforma Web que disponibiliza informações sobre incêndios e queimadas no Brasil, desenvolvido pelo Centro Gestor e Operacional do Sistema de Proteção da Amazônia (CENSIPAM, 2022). No entanto, diferente de outras plataformas de monitoramento, o Painel do Fogo trabalha ao lado de batalhões e brigadas no combate ao fogo. Através desta plataforma se é possível obter o "perímetro" e o "status" mais recente sobre uma queimada ou incêndio de modo que um evento esteja associado a uma ocorrência ou acionamentos de equipes.

O modelo finalizado é capaz de **classificar queimadas** dentro das quatro classificações utilizadas pelo Censipam para tipificar fogos florestais na Amazônia (ANDELA et al., 2018):

- **Savana e pastagem:** Em regiões onde a vegetação natural é composta principalmente de savana, cerrado ou pastagens, os incêndios frequentemente ocorrem durante a estação seca devido à combinação de condições climáticas secas e atividade humana, como a queima de pastagens para melhorar o crescimento da grama. Vegetação de porte baixo, com pouca presença de biomassa, longa duração dias-semanas;
- **Pequenas clareiras:** Esse tipo de incêndio ocorre em áreas florestais onde as árvores foram cortadas recentemente, criando clareiras. O objetivo pode ser a remoção de resíduos de madeira ou a preparação do solo para o plantio. Pequenos incêndios em sistemas florestais (cobertura de árvores > 50% e igual ou inferior a 5 detecções de fogo), curta duração, área menor que 100 ha;
- **Sub-bosque:** As queimadas no sub-bosque ocorrem em florestas maduras, onde o fogo queima principalmente sob a copa das árvores, afetando a vegetação e o solo abaixo da copa das árvores. Esses fogos podem ser causados por raios, atividades humanas ou ser usados como ferramenta de manejo florestal. Vegetação de porte arbustivo alto, apresenta altas taxas de biomassa. Ocorre muito em borda de áreas desmatadas com floresta. Área variável, longa duração, poucos focos de calor por detecção, mesma região geográfica do desmatamento;
- **Desmatamento:** Essas queimadas ocorrem principalmente devido à atividade humana, como a conversão de áreas florestais em terrenos agrícolas, pastagens ou áreas urbanas. São frequentemente associados ao desmatamento ilegal. As queimadas de desmatamento normalmente têm alto poder radiativo inicial do fogo, pois os detritos lenhosos empilhados levam a uma maior liberação de energia e longa persistência do fogo, uma vez que essas pilhas podem arder por dias. Maior poder radiativo no início, maior duração, área variável.

A classificação de queimadas florestais desempenha um papel crucial na gestão ambiental, pois fornece informações essenciais para o desenvolvimento de estratégias de prevenção e combate ao fogo. Ao tipificar as queimadas dentro dessas categorias, as autoridades podem avaliar a ameaça que representam para a biodiversidade, comunidades humanas e ecossistemas. Essa análise permite a alocação eficiente de recursos, o planejamento de evacuações quando necessário e a mobilização de equipes de combate a incêndios. Além disso, a classificação de queimadas fornece dados valiosos para estudos e pesquisas que visam entender as causas subjacentes dos fogos florestais e desenvolver medidas preventivas mais eficazes para proteger as florestas e o meio ambiente.



O Pannel do Fogo já possui atualmente capacidade de tipificar seu banco de dados com as quatro classificações de fogo, porém o tempo e pessoal necessário para realizar tão avaliação não permite que estas informações possam ser incluídas na plataforma em tempo real junto com o atual produto da plataforma que é a visualização diária de novos eventos de fogo. O projeto é responsável por realizar tal trabalho de forma ligeira e com um bom nível de acurácia.

Para este fim, a plataforma fará uso de sistemas de **aprendizado de máquinas** para tipificar queimadas em curso, integrando estas informações à plataforma Web. Através dela, as informações são disponibilizadas para todos que tenham interesse e alertas serão emitidos para os agentes competentes conforme a necessidade.

O uso de inteligência artificial para cuidados durante o ciclo do fogo não são uma novidade, em "*A systematic review of applications of machine learning techniques for wildfire management decision support*" (BOT; BORGES, 2022) foi feita uma revisão de diversos artigos publicados entre 2019 e 2022 que tratavam do uso de métodos de aprendizado de máquina no controle de queimadas florestais. O que se pode concluir da revisão é que existe um potencial para a adoção de métodos de aprendizado de máquina para previsão e classificação e para melhorar o suporte às decisões gerais no monitoramento de danos ambientais relacionados ao fogo.

Em conjunto com um algoritmo de monitoramento, é necessário que acoplado a ele exista um método de fornecimento de dados. A maneira mais comum atualmente de se obter informações do estado mais recente do vasto território da floresta amazônica é por meio do **sensoriamento remoto**. Compreendendo os tipos de dados que podem ser obtidos por meio do sensoriamento remoto, a relevância deles e sua frequência de obtenção são de grande importância para que uma ferramenta de monitoramento possua um bom êxito.

### 1.3 Este documento

Ao longo deste documento, revisitaremos contribuições significativas no campo do sensoriamento remoto e contribuições de outros trabalhos na área de monitoramento de queimadas e da Amazônia, oferecendo uma perspectiva abrangente das abordagens e descobertas preexistentes. A metodologia adotada é detalhada no capítulo subsequente, delineando o plano de ação que orientou a condução da pesquisa. No âmbito da "Fundamentação Teórica", são apresentados os conceitos essenciais que sustentam a nossa metodologia, proporcionando um embasamento sólido para a análise dos resultados obtidos. Uma explanação minuciosa sobre os bancos de dados utilizados é oferecida em um capítulo dedicado, delineando as fontes de informações que fundamentaram a pesquisa. Por fim, os "Resultados" são apresentados, consolidando as descobertas e conclusões oriundas do projeto.

## 2 Trabalhos relacionados

### 2.1 História do Monitoramento da Amazônia por meio de Sensoriamento Remoto

Devido ao valor da Amazônia e da necessidade em protegê-la, é natural o interesse em se desenvolver tecnologias de monitoramento visando preservar este grande patrimônio natural. A história do sensoriamento remoto na região data desde o ano 1975, quando se foram produzidas as primeiras imagens de satélite Landsat, por meio de um sensor MSS ou sensor multi-espectral. Estas imagens compuseram o primeiro grande mapeamento do território para se verificar o potencial desta tecnologia no monitoramento de desflorestamento. Um mosaico construído a partir de diversas imagens Landsat da amazônia encontra-se na Figura 2.2.

A experiência contemplou uma área de 55 milhões de hectares referentes a uma região crítica ligada a grandes projetos agropecuários, revelando que 10% do território já estava desflorestado. Diante da indignação internacional oriunda deste conhecimento, as autoridades brasileiras instauraram uma Comissão Parlamentar de Inquérito, ou CPI, para investigar o tema (COORDENACAOO-GERAL..., 2019).

Em contrapartida, as imagens coletadas para este estudo contemplavam apenas uma parte crítica do território, levantando a dúvida sobre o estado do resto do território. Em 1988, teve origem o projeto PRODES Analógico (Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite), produzindo métricas desmatamento anuais para todo o território amazônico até 2003.

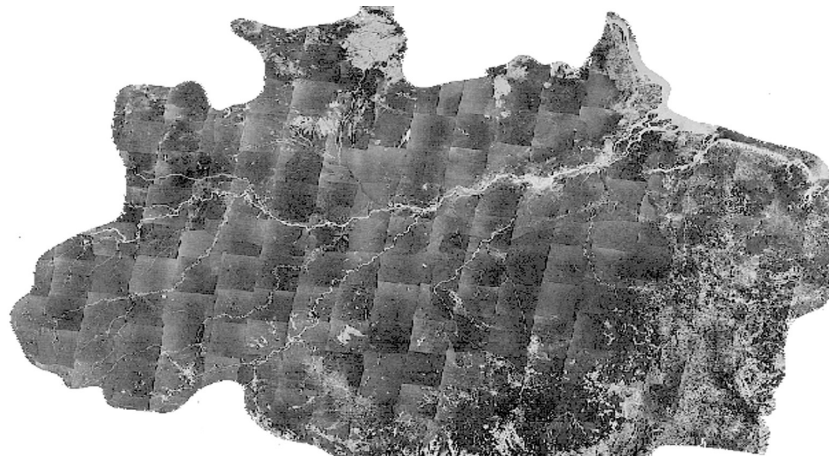


Figura 2.2 – Mosaico com 229 imagens Landsat em escala 1: 250.000 (COORDENACAOO-GERAL..., 2019)

---

Em sequência, há o nascimento do projeto PANAMAZÔNIA I (INPE, 1992) em 1992, operando de forma semelhante ao PRODES Analógico. Este novo projeto também faria o monitoramento da Amazônia de forma analógica com imagens do satélite Landsat de resolução 1:250.000, visando aumentar a escala e fazer o monitoramento para toda a América do Sul.

A próxima etapa desta história se dá em 1997 por meio do PRODES Digital. Este representou um grande avanço tecnológico na forma da criação de um banco de dados digital, porém, foi limitado a operar dentro da complexidade do projeto PRODES Analógico. Por consequência disto, várias melhorias ao sistema de monitoramento e classificação não foram aplicadas.

A partir de 2003 tem-se a origem do projeto DETER (Detecção do Desmatamento em Tempo Real) (INPE..., 2004). Esta nova empreitada utilizou imagens do sensor MODIS, abordo do satélite Terra. Apesar da baixa resolução das imagens, tinha-se uma alta taxa de amostragem, gerando-se imagens diárias. A partir disto, o sistema fazia a sobreposição destas imagens e polígonos de desmatamento detectados. Caso um polígono de desmatamento fosse sobreposto a uma imagem de vegetação anteriormente intocada, era gerado um alerta de alteração da cobertura vegetal para os órgãos de fiscalização de desmatamento oficiais. Este foi o primeiro sistema de monitoramento em tempo quase real do território amazônico e opera até hoje.

Em 2020 foi lançado pelo *Global Fire Emission Database* o *Amazon Dashboard*, uma ferramenta que rastreia queimadas individuais na região da Amazônia usando uma abordagem para agrupar e classificar as detecções ativas em diferentes tipos de fogo. O GFED é um banco de dados focado em informações de fogos obtidos por meio de satélites da NASA como NOAA-20, SUOMI NPP e MODIS (GLOBAL..., 2022). No atual estado da arte, o GFED é o único dado a nível mundial com distribuição gratuita na internet que faz referência sobre tipos de fogo.

Em 2021, o Centro Gestor e Operacional do Sistema de Proteção da Amazônia desenvolve o Painel do Fogo, uma plataforma Web que permite visualizar informações em tempo real sobre queimadas florestais na região da Amazônia com o intuito de subsidiar o acionamento de brigadas ou batalhões durante o combate ao fogo. A novidade se baseia em trabalhar agrupamentos de focos de calor para entender se tal situação é um evento individual de queimada, e disponibilizá-los no Mapa Interativo de Incêndios e Queimadas.

## 2.2 Outros trabalhos na área de monitoramento ambiental

Muitos trabalhos abordam a detecção de incêndios florestais, especialmente usando aprendizado de máquina. Vários sensores podem ser usados, como câmeras IP (TRAN; HOA, 2017), sensores sem fio para emissão de monóxido de carbono e temperatura (DAMAGE et al., 2022) e dados de satélite.

Abid (ABID, 2021) fornece uma visão geral dos sistemas de detecção e previsão de incêndios florestais com base em algoritmos de aprendizado de máquina, separados por tipo de arquitetura como redes neurais, modelos logísticos, árvores de decisão, etc. Os estudos que avaliam os fatores que afetam a ocorrência e o risco de incêndio também são discutidos, juntamente com as principais questões e resultados de cada estudo. Este trabalho demonstra que vários trabalhos chegam em conclusões semelhantes quanto aos principais fatores influenciadores de queimadas ou incêndios em florestas. Os fatores mais importantes para a ignição e propagação do fogo são meteorológicos como chuva e vento, além de fatores topológicos, de vegetação e ação humana que são mais importantes para o cálculo de risco de incêndio.

Estudos feitos para o bioma Cerrado mostram grande precisão no mapeamento de cicatrizes de queimaduras. O estudo de Arruda (ARRUDA et al., 2021) teve como objetivo fazer uma metodologia semiautomática para mapeamento de áreas queimadas no Cerrado, utilizando imagens de satélite e algoritmo *Deep Learning* nas plataformas *Google Earth Engine* e *Google Cloud Storage*. O trabalho empregou um algoritmo MLP, ou *Multi-Layer Perceptron*, em imagens de satélite Landsat para mapear áreas de cicatriz de fogo no Cerrado e depois comparou seus resultados com dados disponíveis pelo MODIS (MODIS..., 2023) e pelo INPE (INPE..., 2023). O estudo mostrou um potencial de dados de sensoriamento remoto com séries temporais Landsat utilizadas no projeto MapBiomias (Projeto MapBiomias, 2022).

Os estudos para o bioma amazônico são mais escassos. No entanto, Lima et al. (LIMA et al., 2009) realizou um trabalho focado no mapeamento de cicatrizes de queimaduras na Amazônia a partir de Modelo Linear de Mistura Espectral de imagens de sensores MODIS (MOD09), estes dados são estáo segmentados, classificados de forma não supervisionada e rotulação dos grupos de *pixels* dos mapas. O estudo foi baseado no DETER (Projeto Detecção de Áreas Desmatadas em Tempo Real) (INPE..., 2004) usando um algoritmo de classificação não supervisionado baseado em regiões. Os resultados mostraram que cerca de 50.000km<sup>2</sup> da superfície amazônica foram queimados e que os dados diários do sensor MODIS são uma importante fonte de informação para mapeamento de áreas queimadas.

Utilizando dados de sensoriamento remoto, Faria et al. (FARIA et al., 2022) desenvolveu um indicador quase em tempo real da severidade do fogo. São usadas informações dos satélites NOAA-20 e Suomi NPP para monitorar eventos de fogo com base no nível de

atenção que eles exigem e aumentar a eficiência do combate ao fogo. O índice de severidade leva em consideração quatro critérios: tamanho geral, duração, extensão e intensidade. A extensão espacial deste indicador inclui os biomas Amazônia e Pantanal brasileiros e foi demonstrado em um caso de estudo em Porto Velho.

Bruno Scholles ([DIAS, 2023](#)), em conjunto com os autores e o conteúdo apresentado neste trabalho, desenvolveu um sistema de classificação de queimadas dentro da Amazônia Legal utilizando dados de sensoriamento remoto e um método de aprendizado de máquina conhecido como *Random Forest* para integralização com o Painel do Fogo.

## 3 Metodologia

A velocidade e a eficiência na detecção e monitoramento dos incêndios florestais são fundamentais para a viabilização do controle do fogo, e para isso é necessário possuir informações confiáveis da localidade e da área da queimada (BATISTA, 2004). Visto a importância da preservação do território amazônico, este projeto se dispõe a desenvolver um sistema capaz de classificar por meio de algoritmos de aprendizado de máquina e de dados de satélites esses incêndios. Compondo um sistema inteligente de monitoramento para todo o território Amazônico.

De acordo com Naqa (NAQA; MURPHY, 2015), “aprendizado de máquina é um ramo em evolução de algoritmos computacionais que são projetados para emular a inteligência humana aprendendo com o ambiente circundante.” Além disso, “pode melhorar automaticamente através da experiência” (JAIN et al., 2020). Ao escolhermos desenvolver a ferramenta por meio de algoritmos de aprendizado de máquina, temos o objetivo de que ela possa fazer de maneira autônoma um trabalho de monitoramento que até então precisava da supervisão de um ser humano.

O trabalho aqui apresentado melhor se enquadra em um modelo de pesquisa experimental, onde diversas técnicas de *machine learning*, processamento de dados, e engenharia de *features* foram testadas com o objetivo de comparar a eficiência entre elas e como cada elemento é capaz de influenciar no desempenho do algoritmo.

Neste estudo, dados públicos disponíveis gratuitamente na internet foram utilizados como amostras e *features* para o treinamento da rede. Os dados são disponibilizados por fontes confiáveis e renomadas, utilizadas mundialmente por pesquisadores e órgãos governamentais, como FIRMS e VIIRS, e classificado de acordo com o próprio Censipam e dados do GFED Amazon Dashboard. Para treinamento foram utilizadas amostras e informações referentes ao ano de 2020, porém as fontes dos dados são capazes de liberar novos dados frequentemente de forma que o produto final sempre se mantenha atualizado com os conhecimentos mais recentes, podendo entregar suas previsões em tempo real sem que elas percam relevância com o passar do tempo.

É importante salientar que um diferencial dos dados oferecidos pelo Painel do Fogo é a capacidade de monitorar o que chamam de eventos de fogo, observando a evolução de uma queimada ao longo do tempo através de várias detecções de focos de fogo em proximidade. Dessarte, se é possível analisar um fogo florestal como um evento variante no tempo.

Como mencionado anteriormente, a qualidade e significância de cada dado dentro do projeto foi analisada experimentalmente. Normalmente ao trabalhar com *machine learning* e redes neurais é possível medir o quanto que cada informação está sendo importante para

as tomadas de decisões do modelo. Realizando tal análise foi-se capaz de definir quais informações deveríamos priorizar a obtenção, e quais não possuíam impacto, assim como quais modelos de aprendizado de máquina apresentavam melhor desempenho.

Com os resultados preliminares, a próxima etapa é investir em um modelo promissor e criar um código funcional que seja capaz de receber um banco de dados, povoá-lo com as *features* necessárias e então retornar a saída de forma que ela possa ser incorporada de volta ao banco de dados com o máximo de independência.

A eficácia final do trabalho foi medida tomando como referência o padrão estabelecido pelo Censipam para tipificação de queimadas. Ele também está responsável pela revisão do produto e averiguação de sua utilidade.

O modelo atual foi treinado com uma base de dados experimental, e que possui diversas informações obtidas de diferentes fontes que possuem resoluções e frequências de amostragem diferentes. Temos a expectativa que uma colaboração conjunta futura maior com o Censipam e outros órgãos públicos possa aumentar a qualidade das informações com que trabalhamos. Tal colaboração é interessante devido a inovação que este projeto traz no ramo do monitoramento de queimadas, possuindo grande espaço para crescimento e investimento.

## 4 Fundamentação Teórica

### 4.1 Python

Python é uma linguagem de programação de alto nível, orientada a objetos e interpretada (ROSSUM; JR, 1995). Possui uma presença muito forte nos meios acadêmicos e profissionais devido a sua filosofia de priorizar o esforço do programador sobre o esforço da máquina. Em outras palavras, Python é conhecido por sua fácil leitura, aprendizado e rapidez de desenvolvimento de aplicações, além de possuir código aberto e conexões com os mais diversos módulos e *frameworks*.

A linguagem é talvez a mais utilizada no campo de ciência da informação pois, além dos motivos já mencionados, possui diversos módulos de visualização de dados e que, somados à facilidade de leitura do código, proporciona uma visão clara dos dados, até para os não iniciados. Dessa forma, a linguagem é excelente para o desenvolvimento de aplicações de aprendizado de máquina, graças à facilidade da manipulação e visualização de dados e estatísticas e da alta integração com bibliotecas de desenvolvimento de modelos de inteligência artificial.

### 4.2 Scikit-Learn

Uma biblioteca para Python muito conhecida na área de ciência de dados e de aprendizado de máquinas (BUITINCK et al., 2013). O projeto de código aberto vem desenvolvendo algoritmos desde 2010, focando na manipulação e tratamento de dados, assim como no desenvolvimento de algoritmos de aprendizado de máquina (PEDREGOSA et al., 2011).

O scikit-learn é uma ferramenta valiosa para cientistas de dados, engenheiros de aprendizado de máquina e pesquisadores que desejam explorar e aplicar técnicas de aprendizado de máquina e mineração de dados em seus projetos. Ele permite que os usuários construam, avaliem e ajustem modelos de *machine learning* de maneira eficaz e eficiente, tornando-o uma escolha popular no campo do processamento de dados e aprendizado de máquina.

### 4.3 PyTorch

Trata-se de uma biblioteca para programação em Python de código aberto amplamente utilizada para o aprendizado de máquina e desenvolvimento de modelos de aprendizado profundo (*deep learning*) (PASZKE et al., 2019). Ele foi desenvolvido principalmente



pelo *Facebook's AI Research lab* (FAIR) e é conhecido por sua flexibilidade e facilidade de uso, o que o tornou uma escolha popular entre pesquisadores e desenvolvedores de aprendizado profundo.

O PyTorch fornece uma estrutura eficiente para realizar operações com tensores, que são essencialmente matrizes multidimensionais, semelhantes às matrizes NumPy. Ele oferece suporte para operações de álgebra linear e cálculo em GPUs, o que o torna adequado para tarefas de aprendizado profundo que envolvem grandes volumes de dados.

Além disso, devido a sua comunidade grande e ativa e a sua flexibilidade, o PyTorch é uma ferramenta poderosa para produção de programas de aprendizado de máquina para as mais variadas aplicações.

## 4.4 Aprendizado de Máquina

No âmbito de inteligência artificial (IA), isto é, no desenvolvimento de sistemas que simulem a inteligência humana em máquinas ou computadores, existe o campo do Aprendizado de Máquina. Este campo também se dispõe a simular uma inteligência, porém a partir de um processo de treinamento de um algoritmo com um conjunto de dados de forma que o mesmo possa servir para interpretar novos conjuntos.

Aprendizado de Máquina se trata de uma forma de estatística aplicada com a intenção de aproximar funções muito complexas (GOODFELLOW; BENGIO; COURVILLE, 2016). Ou seja, é um campo do conhecimento focado no desenvolvimento de algoritmos que possuem uma alta capacidade de generalização após serem treinados a partir de um conjunto de dados. Algumas aplicações destes algoritmos incluem: classificação de conjuntos de dados, detecção de anomalias e regressão (modelamento matemático e previsão de valores).

O campo do aprendizado de máquina teve origem nas décadas de 50 e 60, a partir do *perceptron*, o primeiro modelo de aprendizado de máquina, modelado à imagem de um neurônio (MINSKY; PAPER, 1969). Desde então, esta área da ciência vem evoluindo e passou por uma série de inovações significativas que reverberaram pela humanidade. Nesta linha, pode-se citar pontos marcantes em sua evolução como o aprendizado profundo, as redes convolucionais e as redes recorrentes, que baseiam o algoritmo desenvolvido neste trabalho.

Algoritmos de aprendizado são separados entre algoritmos supervisionados e não-supervisionados (BISHOP, 1995). Os primeiros se referem aos modelos cujo os dados fornecidos são previamente rotulados, possuem sua classificação ou regressão conhecida e estes dados são computados na função de custo no treinamento. Os segundos não possuem dados rotulados e operam tentando agrupar os dados de forma autônoma, formando grupos ou vizinhanças. Estes algoritmos em questão possuem alguns componentes integrais para seu

---

funcionamento: Modelo, Função de Custo e *Dataset*.

#### 4.4.1 Modelo

O modelo é um termo utilizado para descrever a arquitetura geral de um algoritmo de aprendizado de máquina. Diferentes modelos podem ser vastamente diferentes uns dos outros, tanto na forma como processam os dados quanto em sua finalidade mais apropriada. Diferentes características intrínsecas a cada modelo podem torna-lo melhor ou pior para interpretar certos tipos de dados ou objetivos.

#### 4.4.2 Função de Custo

Todo modelo de aprendizado de máquina faz uso de uma função de custo ou função de erro. Este termo se trata de uma equação que computa a diferença entre a saída do modelo e o resultado esperado, e a partir deste valor os parâmetros do modelo são ajustados de forma a minimizar esta função.

Utilizando uma linguagem mais técnica, uma estratégia comum de se encontrar um mínimo de uma função se dá pelo uso de sua derivada, que fornece o ângulo da curva função em um dado ponto. Ou seja, a partir disto é possível incrementar a função em pequenos valores, encaminhando-a a uma direção onde sua derivada seja nula, isto é, um ponto mínimo. Esta técnica é conhecida como *Gradient Descent* (CAUCHY, 1847), ou Gradiente Descendente.

#### 4.4.3 Dataset

O conjunto de dados é vital para o funcionamento do algoritmo de aprendizado. Em geral, o primeiro passo no desenvolvimento de um algoritmo de aprendizado de máquina é a criação de um *Dataset*. Para um modelo tabular, os dados são dispostos na forma de uma tabela enquanto para um modelo convolucional, os dados são fornecidos na forma de imagens. De qualquer forma, uma quantidade suficiente de dados é necessária para um bom treinamento do modelo.

Os dados normalmente são separados em conjuntos de dados de Treinamento e dados de Validação. Dados de treinamento são auto explicativos, isto é, são utilizados para treinar o modelo, enquanto dados de validação são utilizados rotineiramente no treinamento para se checar a acurácia do modelo com um conjunto de dados diferente do utilizado no treinamento. Também é comum de se utilizar um terceiro conjunto, um conjunto de Teste onde o modelo com treinamento concluído processa um novo conjunto diferente para fins de métricas de performance.

#### 4.4.3.1 Série Temporal

Um conjunto de dados de série temporal (ou "*dataset*" de série temporal) é uma coleção de informações **organizadas de maneira sequencial**, onde cada ponto de dados está associado a um período de tempo específico. Esse tipo de conjunto de dados é usado para registrar observações, medições ou valores ao longo do tempo. Os *datasets* de séries temporais são frequentemente usados em análises estatísticas, modelagem preditiva e aprendizado de máquina para entender tendências, padrões sazonais e comportamentos ao longo do tempo, isso devido a presença de **dependência** temporal entre seus dados.

A dependência temporal refere-se à relação ou influência que existe entre eventos ou observações em momentos distintos no tempo. Em outras palavras, para cada informação contida em uma série temporal, o seu estado é impactado pelos eventos que a precedem. Cada ocorrência de um evento de uma série temporal também pode levar a ocorrência de novos eventos futuros.

#### 4.4.3.2 Padding

Para treinamento no PyTorch com séries temporais, é necessário que todas as amostras possuam o mesmo tamanho. O desafio em questão é transformar dados de comprimento variado em dados padronizados com o mesmo comprimento, para que possam ser organizados como um tensor tridimensional. Dessarte, é necessário de se empregar um algoritmo capaz de preencher lacunas e eliminar excessos para garantir que todos os dados tenham as mesmas proporções.

"*Padding*" não se trata de um algoritmo específico, mas de qualquer algoritmo capaz de preencher lacunas para garantir regularidade entre dados. Tratando-se de séries temporais, os algoritmos de *Padding* mais comuns envolvem preencher com zeros, com médias ou repetindo valores. Neste trabalho, optou-se por preencher as lacunas repetindo-se a última detecção do evento.

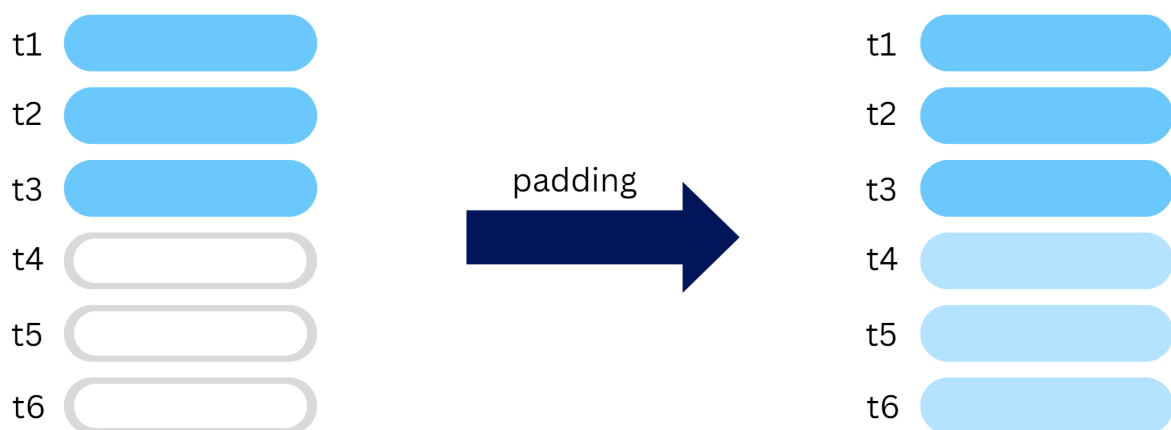


Figura 4.3 – Representação de Padding

A escolha do método utilizado para o *Padding* se deu pela observação de que ao processar os dados que estamos utilizando em sequência temporal, a rede neural tende a convergir para um rótulo dentre as 5 primeiras iterações. Ou seja, foram-se evitados métodos de *Padding* que interfiram nas primeiras instâncias de dados de uma série temporal, como por exemplo, inserir dados intermediários antes ou entre os dados originais de uma sequência.

#### 4.4.3.3 Normalização do Dataset

Normalizar os dados é uma etapa crucial de pré-processamento que consiste em ajustar as características para uma faixa comum, evitando que valores numericamente mais elevados de algumas características dominem sobre os valores numericamente menores. O principal propósito é reduzir o viés daquelas características cuja contribuição numérica é mais significativa na distinção entre as classes (SINGH; SINGH, 2020).

#### 4.4.4 Overfitting

*Overfitting*, ou sobreajuste em português, ocorre quando um modelo de *machine learning* se ajusta excessivamente aos dados de treinamento de maneira que ele passe a falsa impressão de que possui um alto rendimento, porém ao receber dados novos de entrada, a precisão da saída cai consideravelmente (CÁRDENAS-MONTES, 2015).

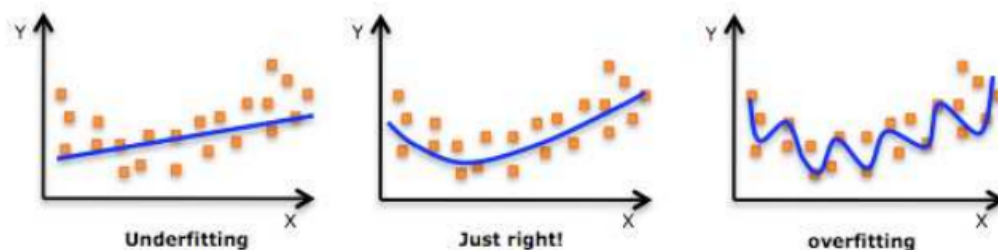


Figura 4.4 – Representação visual de *Underfitting*, boa generalização e *Overfitting* (CÁRDENAS-MONTES, 2015)

Na Figura 4.4 temos um exemplo de sobre-ajuste para um modelo de regressão, podemos deduzir do gráfico que a função encontrada é a que melhor minimiza a função de erro do modelo, porém ela certamente não é capaz de ser generalizada para novos dados.

O *overfitting* ocorre quando um modelo é treinado excessivamente nos dados de treinamento, capturando padrões específicos desses dados que não são generalizáveis para novos dados. O modelo acaba por incorporar o ruído dos dados de entrada, levando a previsões imprecisas quando o modelo é exposto a novos dados com ruído diferente.

#### 4.4.5 Métricas de Avaliação do Modelo

A avaliação de modelos de aprendizado de máquina é crucial para desenvolver um bom algoritmo de aprendizado de máquina. Métricas de avaliação fornecem uma medida objetiva do desempenho do modelo, permitindo comparações entre diferentes abordagens.

Métricas são essenciais no ajuste fino dos hiper parâmetros do modelo de aprendizado de máquina. A partir deles, é possível diagnosticar problemas como *overfitting* ou explosão de gradiente. Os dados gerados são acurácia, precisão, *recall* e *f-score* para cada classe e no total de forma ponderada. Além disso também são registrados os valores dos pesos dos nós da rede neural de forma a se observar explosões ou desaparecimentos de gradiente.

Abaixo estão as métricas utilizadas. Onde "tp" é quantidade de verdadeiros positivos, "fp" falsos positivos e "fn" falsos negativos:

$$\text{Acurácia} = \frac{\text{Número de previsões corretas}}{\text{Número total de previsões}} \quad (4.1)$$

A acurácia 4.1 é uma métrica que reflete o desempenho geral de um modelo e representa a proporção de instâncias corretamente classificadas em relação ao total de instâncias. Em termos simples, é a habilidade do modelo de acertar as previsões.

$$\text{Precisão} = \frac{tp}{tp + fp} \quad (4.2)$$

A precisão 4.2 mede a proporção de instâncias positivas identificadas corretamente em relação ao total de instâncias identificadas como positivas (verdadeiras positivas mais falsos positivos). Em outras palavras, ela expressa a qualidade das previsões positivas do modelo.

$$F_1 = \frac{2tp}{2tp + fp + fn} \quad (4.3)$$

A equação do *F-score* 4.3 é uma métrica útil para medir o aprendizado da rede pois é uma forma de levar em consideração o desequilíbrio entre classes. Por exemplo, em uma tarefa de classificação cujo a classe 1 representa apenas 1% dos dados teria uma acurácia de 99% caso exibisse apenas a classe 0 como resultado, porém o *F-score* traria um resultado péssimo, informando a falha da rede.

$$\text{Recall} = \frac{tp}{tp + fn} \quad (4.4)$$

Já a equação do *Recall* 4.4, também conhecido como sensibilidade ou taxa de verdadeiros positivos, é utilizado no cálculo do *F-score* e mede a capacidade de um modelo identificar corretamente todos os exemplos positivos em um conjunto de dados.

## 4.5 Modelos de aprendizado de máquina

### LSTM (*Long Short-Term Memory*)

LSTM (*Long Short-Term Memory*) é um tipo de arquitetura de rede neural recorrente (RNN) que é projetada para lidar com sequências de dados e é especialmente eficaz em capturar dependências de longo prazo em sequências temporais (HOCHREITER; SCHMIDHUBER, 1997). Sua característica fundamental é a capacidade de "lembrar" informações importantes de longo prazo enquanto processam sequências de dados. Isso é alcançado por meio de unidades de memória especiais chamadas células de memória, representadas pela ilustração na Figura 4.5.

- Porta de Esquecimento (*Forget Gate*): Esta porta decide quais informações da célula de memória anterior devem ser esquecidas ou mantidas com base no conteúdo atual da sequência. Ela ajuda a lidar com a questão do esquecimento de informações em sequências longas.
- Porta de Entrada (*Input Gate*): A porta de entrada permite que novas informações sejam adicionadas à célula de memória com base na sequência atual. Ela decide quais informações são relevantes para adicionar.
- Porta de Saída (*Output Gate*): A porta de saída determina qual parte da célula de memória será a saída da LSTM com base no conteúdo atual da sequência.

Em questões práticas, as células de memória da rede neural LSTM recebem três entradas: *cell state* ou estado de célula, *hidden state* ou estado oculto e os dados de entrada. A entrada *cell state* é projetada para capturar dependências de longo prazo em sequências, armazenando e transportando informações relevantes ao longo do tempo e é alterada pelas três portas da rede. Já a *hidden state* é uma representação resumida das informações e pode ser usado como saída da LSTM ou alimentado às camadas subsequentes. O tamanho do *hidden state* é o *hidden size*, esse valor nos diz o quanto de informação que a célula de memória pode armazenar e manter ao longo do tempo.

As etapas que acontecem dentro de uma célula de memória são (GERS; SCHMIDHUBER, 2001):

1. Entrada: a LSTM recebe uma entrada que é combinada com o *hidden state* anterior;
2. Porta de esquecimento: decide-se quais informações da célula de memória anterior devem ser mantidas ou esquecidas;
3. Portão de entrada: decide quais novas informações devem ser adicionadas à célula de memória;

4. Atualização da célula de memória: o *cell state* é atualizado usando os valores do portão de esquecimento e do portão de entrada;
5. Porta de saída: decide-se qual parte da célula de memória atualizada será emitida como saída;
6. Saída e próximo *hidden state*: a saída da célula de memória é gerada multiplicando o *cell state* pelo valor do portão de saída. Essa saída também é usada como o próximo estado oculto.

Para cada célula de memória que adicionamos, estamos aumentando a quantidade de camadas que a rede possui. A presença de uma única célula de memória é suficiente para termos uma LSTM funcional. O que acontece é que a célula de memória, após gerar uma saída para cada elemento da série temporal, mantém uma conexão recorrente consigo mesma. Essa conexão recorrente permite que a célula de memória receba informações do seu estado anterior em cada passo de tempo subsequente, permitindo assim a captura de dependências temporais ao longo da sequência. Quando temos mais de uma camada, essas informações são passadas de uma célula para a próxima.

Esta arquitetura de rede neural foi projetada para lidar com um problema comum em redes recorrentes, a explosão ou o desaparecimento do gradiente, a partir do uso das portas e células de memória. Trata-se de uma rede com a capacidade de separar informações irrelevantes, permitindo com que o gradiente não varie caso não seja necessário.

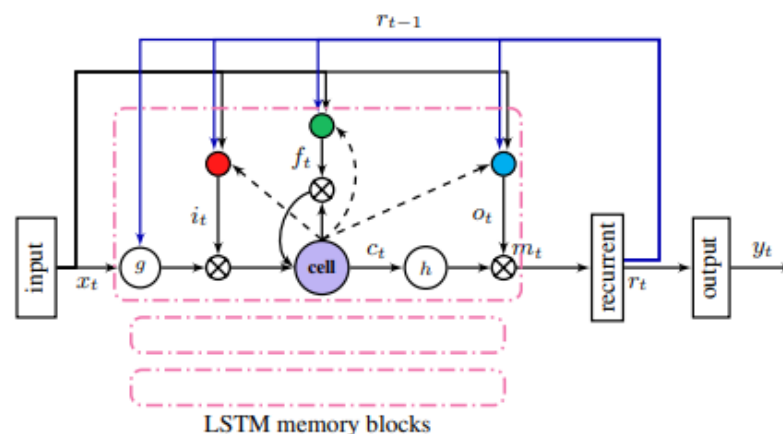


Figura 4.5 – Ilustração exemplo de uma célula de um modelo LSTM (LSTM..., 2023)

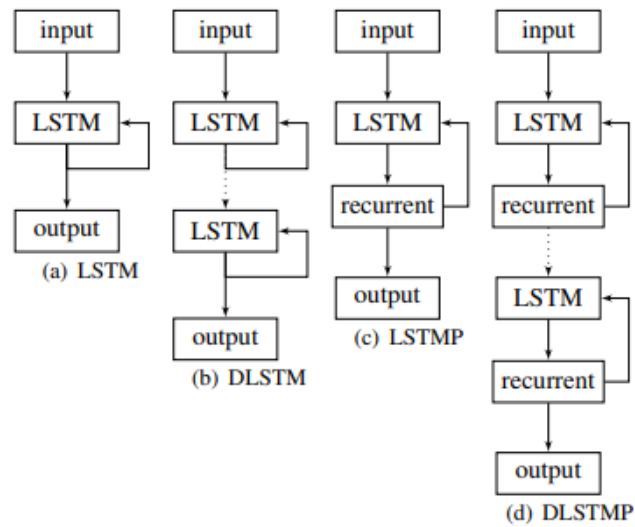


Figura 4.6 – Possíveis arquiteturas com LSTM (LSTM..., 2023)

## 4.6 Evento de Fogo

Os Eventos de Fogo podem ser acessados por meio do Painel do Fogo (CENSIPAM, 2022), uma plataforma gerida pelo Censipam (CENTRO..., 2023). Este órgão é subordinado ao ministério da defesa e é responsável por articular ações de governo em prol da proteção ambiental e do desenvolvimento sustentável da Amazônia Legal e da Amazônia Azul.

Os eventos são gerados a partir de observações dos sensores orbitais VIIRS, por meio dos satélites NOAA-20 e SUOMI NPP, e MODIS, por meio dos satélites Aqua e Terra. Estes sensores detectam fogos ativos, formando uma camada vetorial. Em outras palavras, eventos detectados delimitam uma região e quando esta se intersecta com outra durante a mesma passagem do satélite, é gerado um vetor, polígono, que contém todos os eventos detectados na vizinhança e sua evolução ao longo do tempo.



## 5 Banco de dados

Como mencionado na subseção 4.4.3, a criação de um banco de dados ou *dataset* é fundamental para o treinamento de um algoritmo de aprendizado de máquina. No conjunto de dados devemos especificar quais são os dados de entrada que o algoritmo irá receber e qual a saída esperada para aquela entrada, a partir disso o programa deve fazer uso da função de custo para ajustar os parâmetros do modelo de forma a conseguir acertar a saída com a maior precisão.

Neste trabalho, o banco de dados está no formato tabular, isto é, as informações estão dispostas em uma tabela. Cada linha da tabela possui um conjunto de atributos (os dados de entrada/*features*) e um rótulo (*label* ou saída desejada). Este tipo de conjunto de dados é chamado de *dataframe*, uma estrutura de dados bidimensional com os dados alinhados de forma tabular em linhas e colunas.

Dois bancos de dados foram construídos para a experimentação com diferentes códigos de *machine learning*, o primeiro construído utilizando dados obtidos e rotulados pelo Amazon Dashboard do GFED, e o segundo com dados obtidos e rotulados pelo Censipam. A partir do treinamento com o banco do GFED, foi possível estipular as *features* mais importantes na hora de realizar a tipificação do fogo e que não poderiam ser deixadas de lado na hora de treinar o modelo para trabalhar com o banco do Censipam.

### 5.1 GFED

Pelo site do Amazon Dashboard é possível fazer o download de um arquivo *.zip* com diferentes tipos de arquivos que contém um mapa da região da floresta amazônica na América do Sul e os diversos eventos de fogo que aconteceram na região desde 2020.

#### 5.1.1 *Shapefile*

O arquivo obtido é do tipo *shapefile*, um formato de armazenamento de dados de vetor para armazenar a posição, forma e atributos de feições geográficas. Este tipo de arquivo pode ser aberto por meio de *softwares* gratuitos como o QGIS, utilizado para visualização de dados GIS (Sistema de Informação Geográfica), para captura de dados, para análise avançada de GIS e para apresentações na forma de mapas, atlas e relatórios sofisticados (GUIA..., 2023). Um GIS integra informações descritivas a um mapa, eles incluem imagens, feições e mapas base vinculados a planilhas e tabelas.

Ao abrirmos o *shapefile* no QGIS, primeiro nos deparamos com uma imagem das queimadas que foram analisadas pelo GFED na região que contém a floresta amazônica desde

2020 até 31 de maio de 2022, como mostrado na Figura 5.7. Com o projeto aberto, podemos escolher visualizar a tabela de atributos que acompanha o mapa, que está representada na Figura 5.8. A tabela de atributos contém as feições relativas a cada polígono que compõe a imagem do mapa da América do Sul. Os atributos que fazem parte da tabela são:

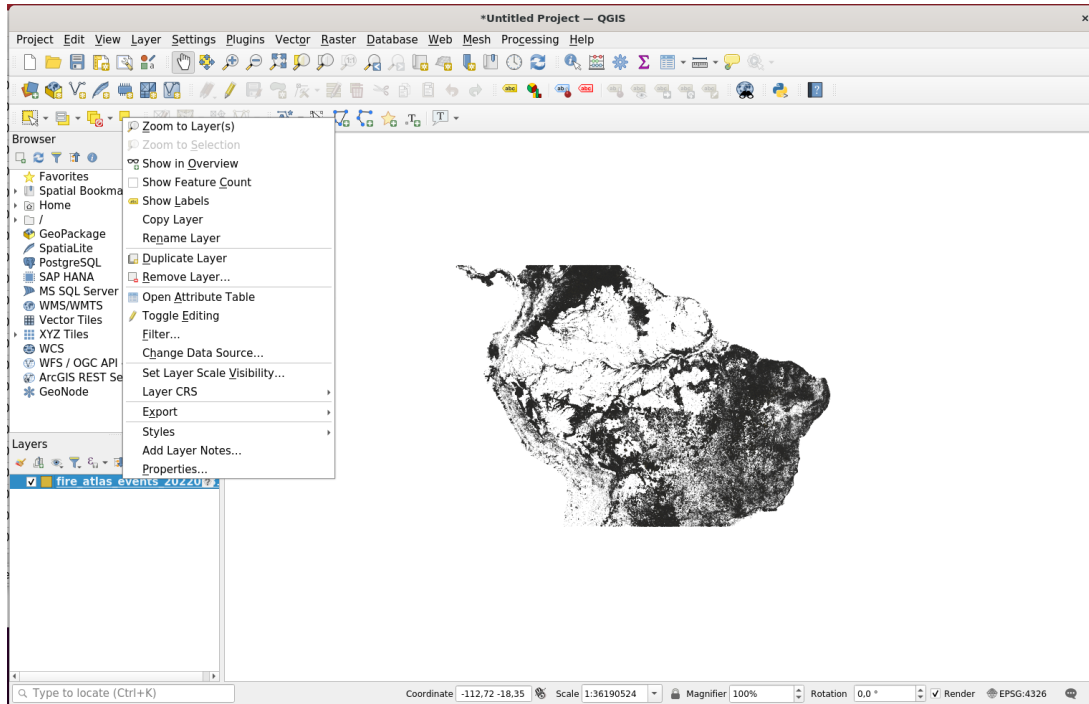


Figura 5.7 – Mapa da América do Sul com os eventos de fogo detectados

id	confidence	deforestation	tree_cover	biomass	frp	fire_count	size	persistenc	daytime	progressio	biome	region	protected	start_DOY	last_DOY	is_new	is_active
1	2	3	0	55,25	38,4	8,0	6	1,24	1,00	0	0	0	0	72	81	0	0
2	1	2	0	21,00	33,4	2,4	2	0,62	1,00	1,00	0	0	0	34	36	0	0
3	1	3	0	8,00	25,7	6,7	2	0,31	1,00	1,00	0	0	0	2	45	0	0
4	1	3	0	15,00	24,4	4,0	3	0,31	1,00	0	0	0	0	51	62	0	0
5	1	2	0	26,00	27,1	7,6	5	0,93	1,00	1,00	0	0	0	23	51	0	0
6	2	3	0	67,00	35,5	1,8	1	0,31	1,00	1,00	0	0	0	79	79	0	0
7	1	1	0	49,50	22,0	1,1	3	0,62	1,00	0	0	0	0	75	75	0	0
8	2	3	0	57,00	24,0	16,3	1	0,31	1,00	1,00	0	0	0	94	94	0	0
9	1	3	0	17,00	24,3	6,0	3	0,93	1,00	1,00	0	0	0	71	71	0	0
10	1	3	0	3,60	26,0	5,5	10	1,55	1,00	0	0,04	0	0	55	56	0	0
11	1	3	0	7,00	29,9	3,6	1	0,31	1,00	1,00	0	0	0	32	32	0	0
12	1	2	0	23,00	31,8	1,2	1	0,31	1,00	0	0	0	0	88	88	0	0
13	1	3	0	3,00	42,2	5,3	2	0,31	1,00	1,00	0	0	0	55	67	0	0
14	1	3	0	4,00	0	1,8	12	0,31	1,00	0	0	0	0	17	322	1	1
15	1	3	0	7,00	0	0,5	1	0,31	1,00	0	0	0	0	322	322	1	1
16	1	3	0	12,50	0	4,5	7	0,62	1,00	0	0	0	0	31	87	0	0
17	1	3	0	8,00	0	2,0	1	0,31	1,00	0	0	0	0	67	67	0	0
18	2	3	0	74,00	46,9	2,8	1	0,31	1,00	1,00	0	0	0	32	32	0	0
19	1	3	0	9,50	0	4,5	3	0,62	1,00	1,00	0	0	0	86	86	0	0
20	1	3	0	1,00	12,5	2,3	1	0,31	1,00	1,00	0	0	0	81	81	0	0
21	1	2	0	20,00	25,1	6,3	1	0,31	1,00	1,00	0	0	0	39	39	0	0
22	1	1	0	46,00	122,8	5,5	1	0,31	1,00	1,00	0	0	0	72	72	0	0
23	2	3	0	79,00	201,7	7,7	2	0,62	1,00	1,00	0	0	0	78	82	0	0
24	2	3	0	88,50	195,8	25,3	4	0,62	1,00	1,00	0	0	0	45	83	0	0
25	2	3	0	57,00	125,5	2,2	1	0,31	1,00	1,00	0	0	0	42	42	0	0
26	2	3	0	58,00	83,3	6,2	1	0,31	1,00	1,00	0	0	0	38	38	0	0
27	2	3	0	73,00	120,9	2,9	1	0,31	1,00	1,00	0	0	0	80	80	0	0
28	2	3	0	53,00	110,7	4,7	1	0,31	1,00	1,00	0	0	0	85	85	0	0
29	3	1	0	72,75	91,4	10,7	6	1,24	1,00	1,00	0	0	0	65	68	0	0

Figura 5.8 – Tabela de atributos do *shapefile*

- **fire\_type:** tipo de fogo - (1) savana e pastagem, (2) pequenas clareiras, (3) sub-bosque, (4) desmatamento;
- **confidence:** confiança - (1) baixa, (2) moderada, (3) alta;
- **tree\_cover:** fração média de cobertura arbórea dentro do perímetro (%);
- **biomass:** biomassa média dentro do perímetro do fogo (ton/ha);
- **deforestation:** fração (0-1) de quadrante de 550m com desmatamento histórico (2015-2019);
- **size:** tamanho do fogo em  $km^2$ ;
- **detections:** número total de detecções dentro do período;
- **frp:** potência radiativa média de fogo (FRP) em megawatts (MW). Trata-se de uma técnica para quantizar biomassa queimada medindo a radiação de energia infravermelha (MODIS..., 2023);
- **persistence:** persistência em dias;
- **progression:** fração média de progressão do fogo em quadrantes de 550m (0-1);
- **daytime:** (0) não de dia, (1) de dia;
- **start\_DOY:** dia do início do novo fogo a partir do dia do ano (1-366);
- **last\_DOY:** detecção de fogo ativo mais recente a partir do dia do ano (1-366);
- **is\_new:** (1) começou nas últimas 24 horas, (0) queimada já existente;
- **is\_active:** (1) o fogo estava ativo nos últimos 10 dias (0) ou não;
- **biome:** (1) o fogo está dentro (0) ou fora do bioma Amazônia;
- **protected:** fração (0-1) de fogo ocorrendo dentro de área protegida ou terra indígena;
- **region:** (0-21) qual a região do evento do fogo.

### 5.1.2 Dataframe

Para converter o conteúdo da tabela de atributos em um *dataframe*, faz-se necessário o uso das bibliotecas *geopandas* e *pandas* do *python*. *Pandas* é uma ferramenta de análise e manipulação de dados, e *geopandas* estende os tipos de dados usados pelos *pandas* para permitir operações espaciais em tipos geométricos. O que fazemos então é primeiro abrir o arquivo do tipo *shapefile* usando *geopandas* para criar um *GeoDataFrame* e então usar uma função da biblioteca *pandas* para convertê-lo em um *dataframe* normal com as *features* que iremos utilizar como entrada para os algoritmos de aprendizado de máquina.

## 5.2 Censipam

Assim como o *Amazon Dashboard*, o banco de dados obtido pelo Censipam está em formato *shapefile*. Tal banco pode ser acessado através do recurso de adição de camadas Web

Feature Service (WFS) no QGIS, utilizando uma URL específica que aponta para um serviço remoto contendo dados vetoriais geoespaciais. O WFS é um padrão da Open Geospatial Consortium (OGC) que define interfaces para servir e consumir dados geoespaciais vetoriais pela web.

Na Figura 5.9 temos um exemplo do que podemos visualizar ao acessar o banco de dados do Censipam pelo QGIS:

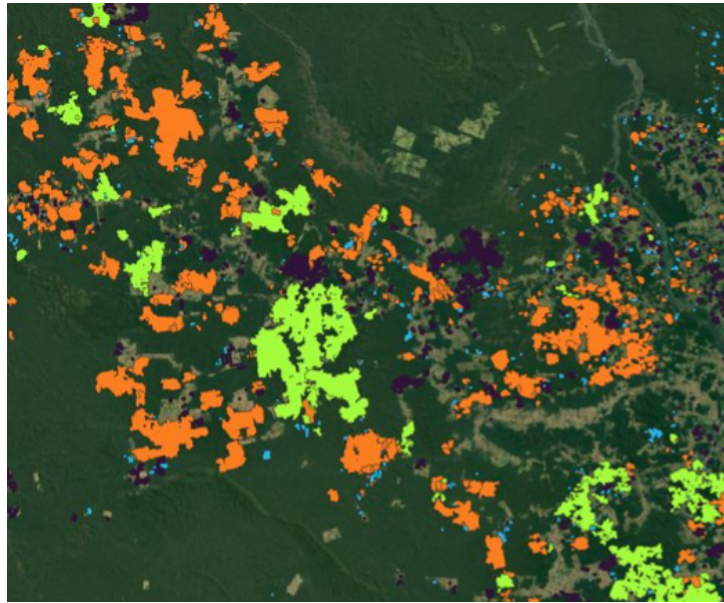


Figura 5.9 – Mapa no QGIS com eventos de fogo

Uma adição interessante presente nos dados do Censipam se comparados com os do GFED, é a capacidade de adquirirmos diversas detecções de um mesmo evento de fogo como demonstrado na Figura 5.10. Com o GFED tínhamos acesso apenas a classificação e *features* mais recente da queimada, referente a sua última detecção realizada, porém com o Censipam podemos obter ambos para detecções passadas, o que abre a possibilidade de incluir as características dinâmicas e temporais da queimada na sua classificação.



Figura 5.10 – Ilustrações demonstrativas do crescimento de um evento de fogo ao longo do tempo

Da tabela de atributos destes *shapefiles*, utilizamos as seguintes colunas para o treinamento:

- **id\_evento**: Representa o código de identificação interno de um evento de fogo catalogado pelo sistema. Detecções distintas, mas consideradas como um mesmo evento recebem a mesma identificação.
- **delta\_area**: Variação da área coberta pela queimada em quilômetros quadrados.
- **q\_focos**: Quantidade de focos de fogo detectados pelo sistema como pertencendo ao mesmo evento de fogo.
- **qtd\_passagens\_com\_deteccao**: Quantas passagens de satélites ocorreram por este evento. Como se trata de 4 pares de satélite em órbitas polares, pode-se ter até 8 detecções por dia (CENSIPAM, 2022).
- **frp\_med**: *Fire Radiative Power* ou FRP, um dado relativo ao sensor MODIS (*Moderate Resolution Imaging Spectroradiometer*).
- **area\_passagem**: Área coberta pela queimada no momento da detecção;
- **delta\_t\_horas**: Tempo acumulado em horas entre detecções.
- **ve\_expansao**: Velocidade estimada de expansão da queimada.

### 5.3 Análise das *features* do GFED

Ao todo são utilizados 17 atributos diferentes para realizarmos a classificação do tipo de fogo, que estão descritas na subseção 5.1.1 (`cluster_ID` não é utilizado). Nem sempre todas as informações que entregamos para um algoritmo de aprendizado de máquina são úteis para a previsão, o que explica não utilizarmos o `cluster_ID`, pois se cada evento de fogo possui um número de identificação diferente, esse número nada diz a respeito sobre qual o tipo de queimada que ocorreu. Podemos fazer uma correlação entre cada tipo de dado fornecido e os tipos de fogo, observando assim quais informações são mais descritivas para o algoritmo, quais delas são mais úteis para separar os 4 tipos.

Para podermos observar a relação que cada *feature* possui com a saída, foram feitos diversos gráficos de dispersão em 3 dimensões que combinam o efeito que até 3 *features* em conjunto podem ter para separar cada tipo de queimada em nuvens. Os gráficos estão dispostos nas Figuras a seguir, a perspectiva de cada um deles é alterada de forma a melhorar a visibilidade:

Figura 5.11 – Start day x Is new x Is active

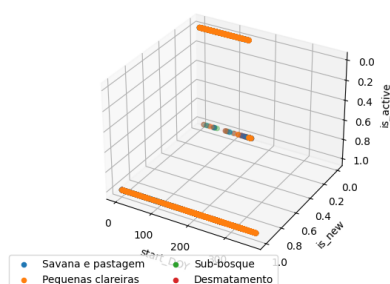


Figura 5.12 – Biome x Region x Protected

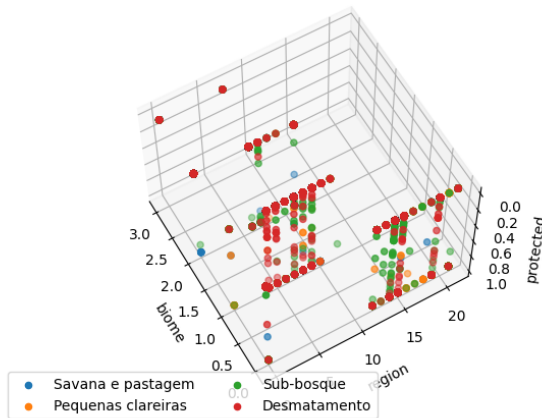


Figura 5.15 – FRP x Detections x Size

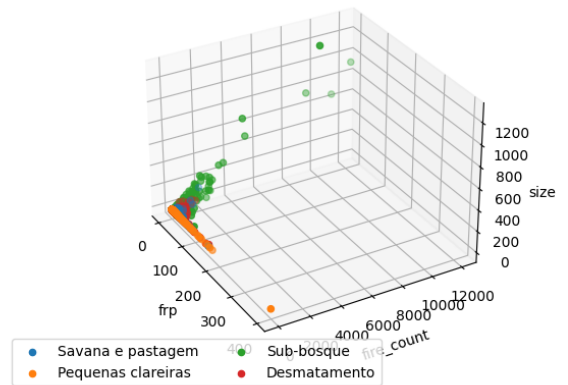


Figura 5.13 – Biome x Region x Protected

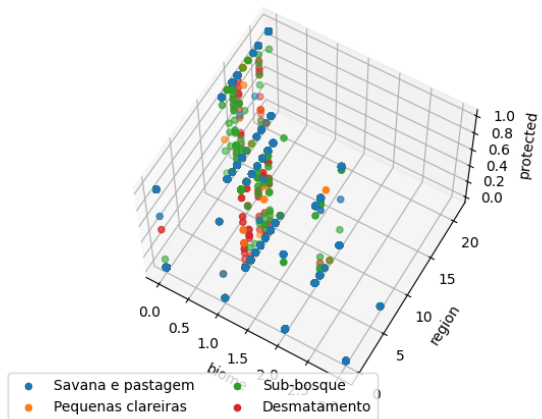


Figura 5.16 – Persistence x Daytime x Progression

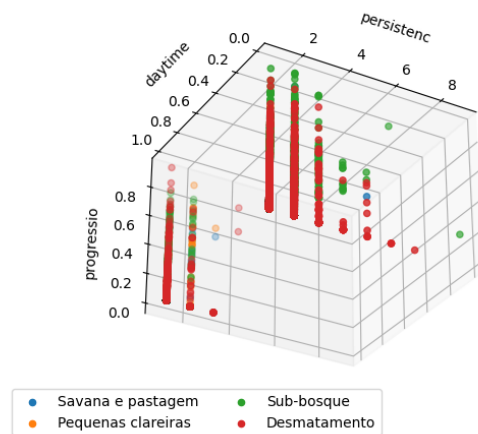


Figura 5.14 – Deforestat x Tree cover x Biomass

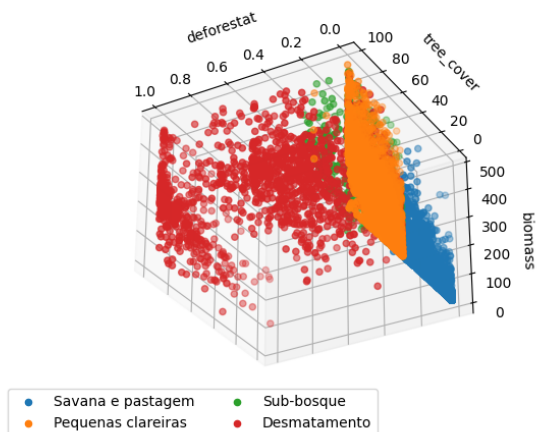
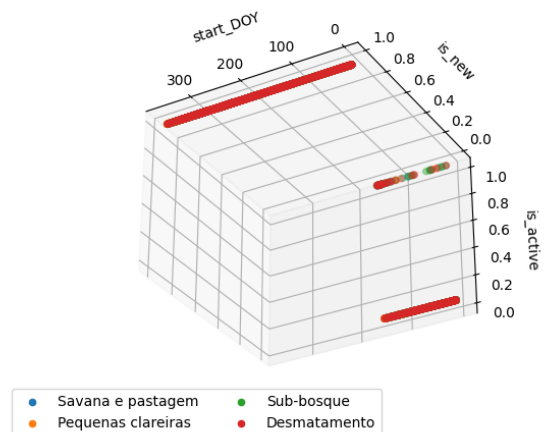


Figura 5.17 – Start day x Is new x Is active



Este tipo de gráfico pode oferecer uma perspectiva útil de análise de dados. Algumas combinações de dados, quando representadas dessa maneira, podem sugerir quais atributos são bons para se discretizar cada classe. A Figura 5.14, por exemplo, mostra uma clara divisão entre as classes "Pequenas Clareiras" e "Savana e Pastagem" através do atributo Cobertura de Árvore (*tree cover*), e outra divisão para a classe "Desmatamento" a partir do atributo de Desflorestamentos.

Podemos observar através dos gráficos que os tipos de fogo de savana (0) e pequenas clareiras (1) são mais fáceis de se tipificar, dados como *tree cover*, *deforestat*, *frp*, *fire count* e *biomass* em conjunto são capazes de delimitar fortemente esses dois tipos. O desmatamento (4) e sub-bosque (3) apresentam maior diluição um com o outro e com os outros tipos, o que explica a maior dificuldade dos algoritmos em prevê-los corretamente, como será visto no capítulo 6.

Como podemos ver na seção 5.2, o banco de dados do Censipam não conta com boa parte das *features* presentes no banco do GFED, e como podemos ver algumas delas se mostraram essenciais para acusação do tipo do fogo. Torna-se necessário então buscar outras fontes que forneçam tais informações.

## 5.4 Adição de *features*

A produção destes dados para a adição ao banco de dados do Censipam envolve o cruzamento com dados de diversas outras entidades. Sua produção se dá pelo cálculo da intersecção geográfica dos eventos de fogo e diversos mapas disponibilizadas pelas entidades Terrabrasilis (ASSIS et al., 2019), MapBiomass (Projeto MapBiomass, 2022) e GFED (GLOBAL..., 2022).

A partir das coordenadas dos focos de queimada, é possível de se fazer uma comparação com outros mapas para se encontrar intersecções com áreas de interesse. Por exemplo, dados relativos à proximidade à áreas de conservação ambiental e áreas indígenas são produzidos pelo cruzamento com os dados disponíveis pelo Terrabrasilis. Já informações como desmatamento da área podem ser obtidas por meio do DETER (INPE..., 2004), enquanto informações Áreas Públicas, Privadas, Desconhecidas e Áreas de Cadastro Ambiental Rural são obtidas a partir do Cadastro Ambiental Rural (BRASILEIRO, 2023).

O código de aquisição e cruzamento foi feito com base no modelo desenvolvido por Bruno Scholles (DIAS, 2023), por meio do qual foi-se possível adicionarmos as seguintes *features* ao *dataframe*:

- **areakm2\_UC**: Área em  $km^2$  da detecção pertencente à uma unidade de conservação;
- **perc\_in\_UC**: Porcentagem da área da detecção da detecção pertencente à uma unidade de conservação;

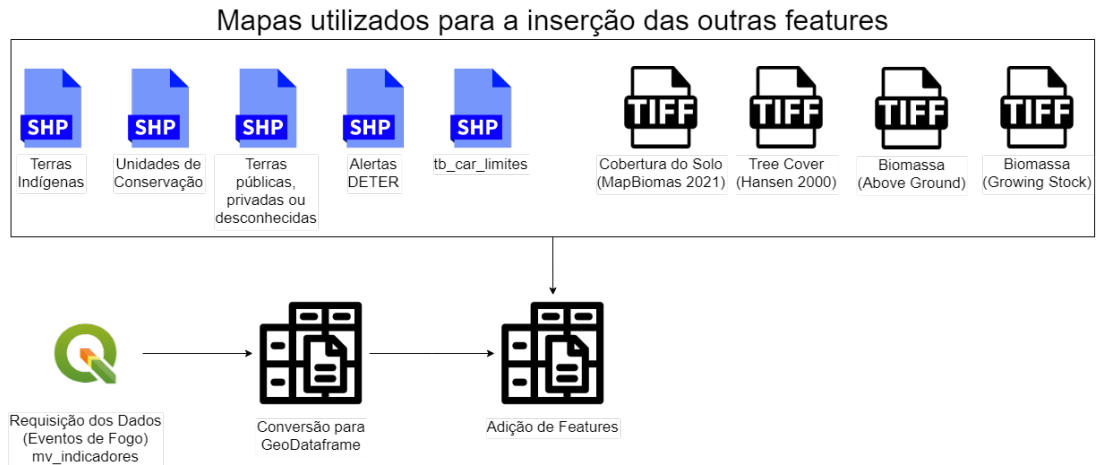


Figura 5.18 – Esquemático do processo de obtenção de *features*

- **areakm2\_IN**: Área em  $km^2$  da detecção pertencente à território indígena;
- **perc\_in\_IN**: Porcentagem da área da detecção da detecção pertencente à território indígena;
- **deter\_area**: Área em  $km^2$  da detecção pertencente à área em que ocorreu desmatamento;
- **deter\_days**: Quantidade de dias passados desde a demarcação do DETER como área em que ocorreu desmatamento;
- **deter\_perc**: Porcentagem da área da detecção da detecção pertencente à área em que ocorreu desmatamento;
- **tree\_cover**: Cobertura de árvore, métrica referente à porcentagem da área que é coberta por copas de árvores ou vegetação alta.
- **a\_cover\_X e p\_cover\_X (onde X pode ser 3, 4, 5, 9, 11, 12, 15, 20, 21, 23, 24, 25, 29, 30, 32, 33, 39, 40, 41, 48, 62)**: Cobertura do solo (tipo de vegetação) medida em área e porcentagem;
- **biomass\_AG**: biomassa *Above Ground* média dentro do perímetro
- **biomass\_GS**: biomassa *Growing Stock* média dentro do perímetro
- **a\_terr\_pri e p\_terr\_pri**: Perímetro da detecção pertencente à área privada;
- **a\_terr\_pub e p\_terr\_pub**: Perímetro da detecção pertencente à área pública;
- **a\_terr\_unk e p\_terr\_unk**: Perímetro da detecção pertencente à área não classificada como pública ou privada (*unknown*);
- **area\_CAR e per\_car**: Perímetro da detecção pertencente à área de Cadastro Ambiental Rural;



## 5.5 Dados nos Servidores do Censipam

A parceria com o Censipam se deu com o intuito de integrar o modelo de aprendizado como uma ferramenta classificadora em seu sistema. Para isto, foi-se fornecido acesso a uma máquina por meio de tunelamento SSH para que o projeto tenha acesso ao banco de dados em tempo real.

Dessa forma, o código foi implementado no sistema para que rode continuamente. Atualmente, o programa aguarda até as 23 horas de cada dia para iniciar seu processamento. Neste momento, são requisitados os dados novos, referentes ao dia atual, e os dados antigos de mesmo ID, com 29 *features*. A partir delas, o algoritmo de adição de *features* adiciona as demais. Em sequência, os dados passam pela etapa de pré-processamento, sendo normalizados e efetuado o *padding*. Por fim, o algoritmo de aprendizado de máquina classifica estes dados e monta arquivos dos tipos CSV e SHP com todos os dados classificados.

### 5.5.1 Dados de Treinamento e Dados em Produção

Os fornecidos para o treinamento do algoritmo e os dados obtidos em produção nos servidores são diferentes. Primeiramente, os dados de treinamento vão até o ano 2022 enquanto os dados disponíveis nos servidores se iniciam em 2023. Além disso, é interessante de se estudar a continuidade entre estes dados, de forma a garantir que sua qualidade é similar à dos dados utilizados no treinamento, ou se houve alguma mudança na forma como são produzidos, como por exemplo, mudança de métrica de metros para quilômetros.

### 5.5.2 Comparando dados do Treinamento com os dados em Produção

O treinamento de um modelo de aprendizado de máquina deve envolver o uso de um *dataset* de treino suficientemente parecido com dados de casos práticos, ou aumentam-se as chances de que a performance do modelo em operação não seja satisfatória. Dessa forma, foram-se gerados vários gráficos de caixa comparando as distribuições dos dados utilizados no treinamento com os dados obtidos diretamente dos servidores do Censipam, onde o modelo está em operação diária.

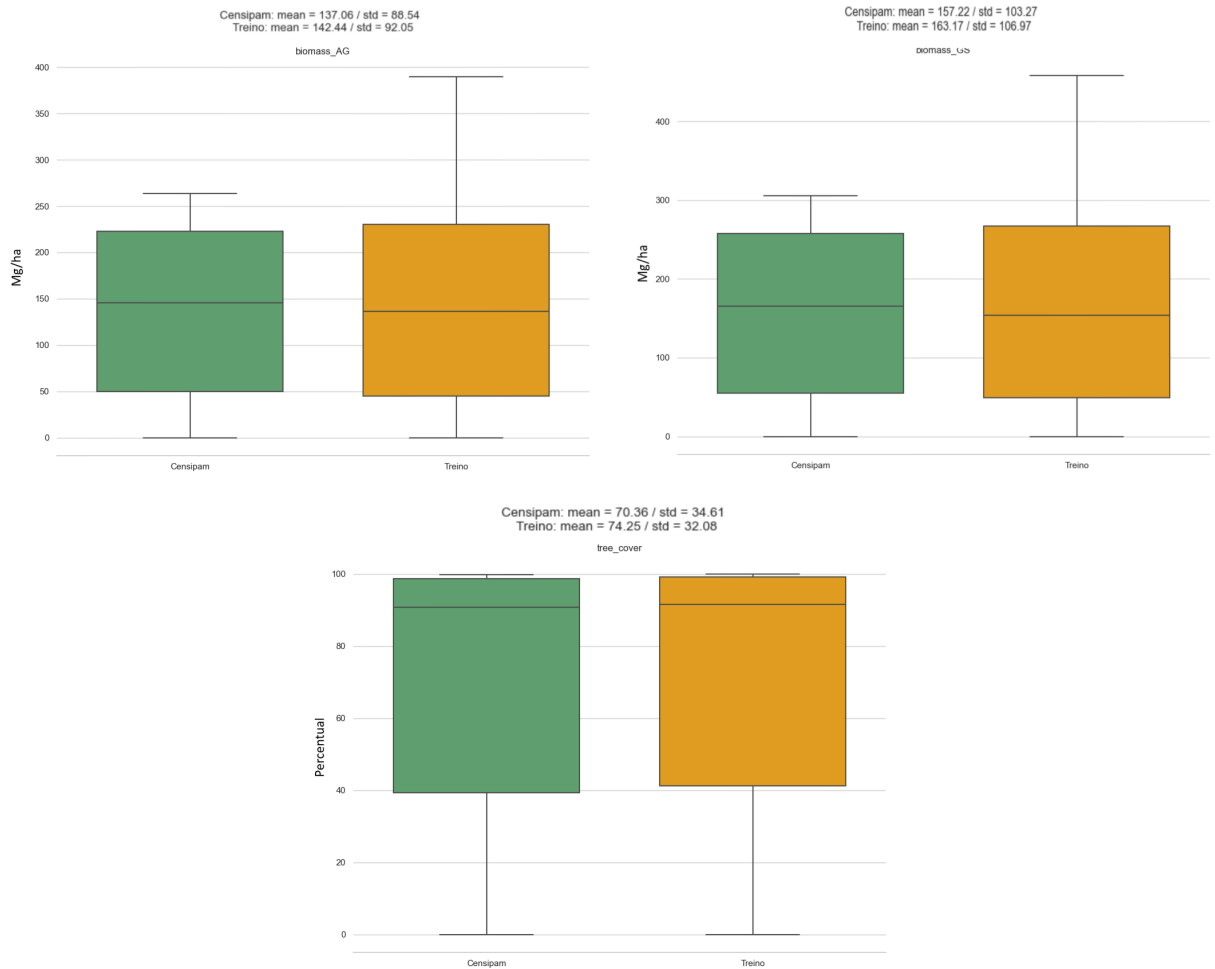


Figura 5.19 – Diagramas de Caixa de dados semelhantes e com valores de média e desvio padrão próximos

A imagem 5.19 contém gráficos de caixa além de valores de média e desvio padrão (representados como mean e std, respectivamente) de alguns atributos cujo as distribuições mostraram-se bastante similares entre os dois **datasets**. Estes dados possuem distribuições semelhantes, com valores de média e desvio padrão próximos indicando que ambos os **datasets** se assemelham em relação a estes dados.

No entanto, a diferença entre dados é significativa para vários outros dados. Na imagem 5.20 encontra-se o gráfico de caixa da *feature* relativa a variação da área do fogo. A diferença de escala entre as duas medidas é óbvia e indica que há algo de errado nestes dados. Este problema se revela em todos os dados relativos a área.

A escala sugere que talvez os dados estejam em unidades de medida diferentes, uma é metros quadrados e a outra em quilômetros quadrados. A imagem 5.21 contém o mesmo gráfico, porém com os dados do servidor ajustados de quilômetro quadrado para metro quadrado. Como a semelhança entre os dados é alta, a hipótese de que se trata de uma diferença de escala utilizada na geração dos dados é provavelmente verdadeira.

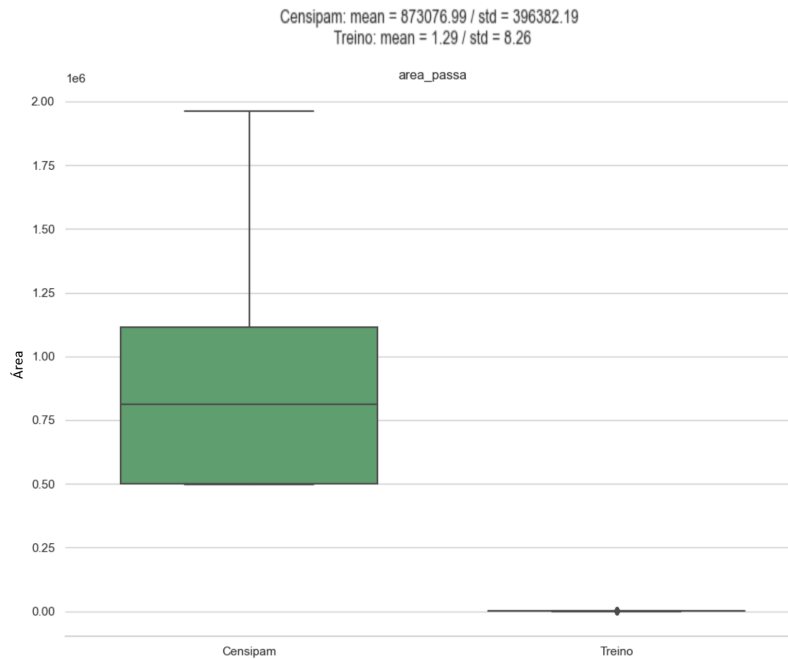


Figura 5.20 – Diagramas de Caixa de variação de área da queimada

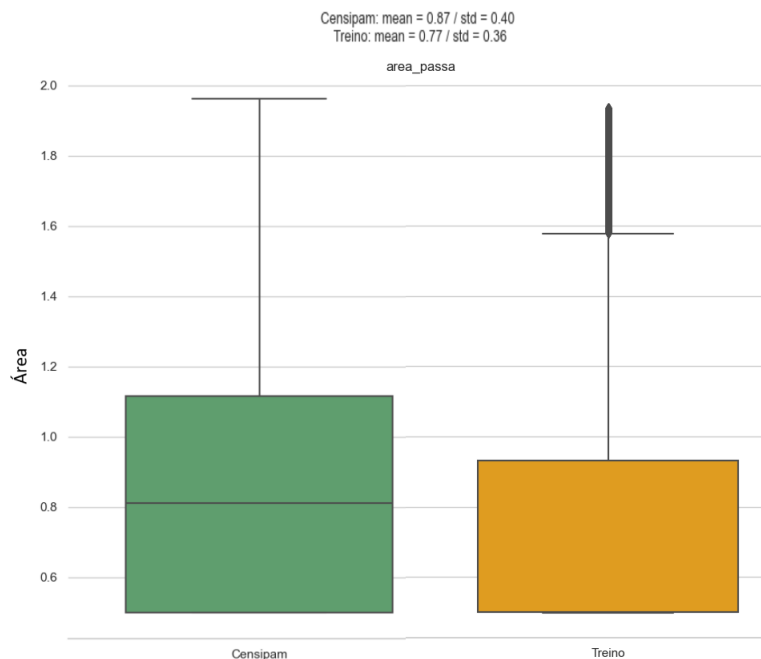


Figura 5.21 – Diagramas de Caixa de variação de área da queimada corrigido em 10e6

Além disto, a *features* geradas costumam apresentar problemas em geral, com todos os dados sendo aproximadamente zero, indicando algum problema com o algoritmo que as gera. Como este erro não surgiu durante o processo de gerar dados para o treinamento, nem ao se gerarem os dados em produção para o trabalho de (DIAS, 2023). A hipótese mais provável é de que há alguma diferença nestes dados disponíveis no servidor do Censipam

para o projeto atual que não foi levada em consideração durante o desenvolvimento do algoritmo. A inconsistência específica não foi encontrada.

Dessa forma, espera-se que o modelo não performe bem em produção, visto que não está recebendo dados adequados. Até o momento não tem como se verificar sua verdadeira performance visto que os dados novos do servidor não foram rotulados manualmente pelo pessoal do Censipam.

## 6 Resultados

### 6.1 LSTM

#### 6.1.1 Modelo

Encontram-se no apêndice A os códigos em Python da modelagem via Pytorch do modelo LSTM e do código de função de custo *Focal Loss*.

Os hiperparâmetros são:

Parâmetros	Significado	Valor
NUM_LAYERS	Número de camadas	2
input_dim	Dimensão da entrada (número de features)	68
HIDDEN_SIZE	Tamanho do estado oculto	60

Tabela 6.1 – Parâmetros do Modelo

#### 6.1.2 Lidando com *Overfitting*

O problema de *overfitting* foi encontrado durante alguns testes como consequência do mal balanceamento do *dataset*. A classe 3, referente às queimadas de sub-bosque, representa 1.8% do *dataset*, o que resulta em 0% de acurácia para esta classe para um modelo treinando sem qualquer algoritmo de balanceamento.

É comum na literatura encontrar exemplos de "data augmentation" (SHORTEN; KHOSHGOFTAAR, 2019). Trata-se de algoritmos capaz de expandir um *dataset* sem comprometer de forma significativa seu valor como dado de treinamento. Neste trabalho, optou-se por empregar uma redução de dados como estratégia de balanceamento. Esta estratégia envolve eliminar de forma aleatória dados de rótulos diferentes de forma a se aumentar a representação do tipo de dados de menor presença no *dataset*. Este tipo de estratégia acabou resultando em *overfitting*, conforme a imagem a seguir.

Na Figura 6.22 pode-se ver que embora a acurácia de treinamento esteja aumentando, a acurácia de validação não está melhorando. Dessa forma, pode-se dizer que a rede está se tornando "viciada" nos dados de treinamento, aprendendo a interpretá-los bem, mas não se tornando capaz de prever dados diferentes deste.

Tendo em mãos estes resultados, optou-se por utilizar um leve balanceamento, aumentando a representatividade de 1.8% para 5%. Esta proporção foi obtida de forma empírica como um valor que representa uma melhora significativa na acurácia para a classe 3, porém sem reduzir demais a acurácia para as demais classes.

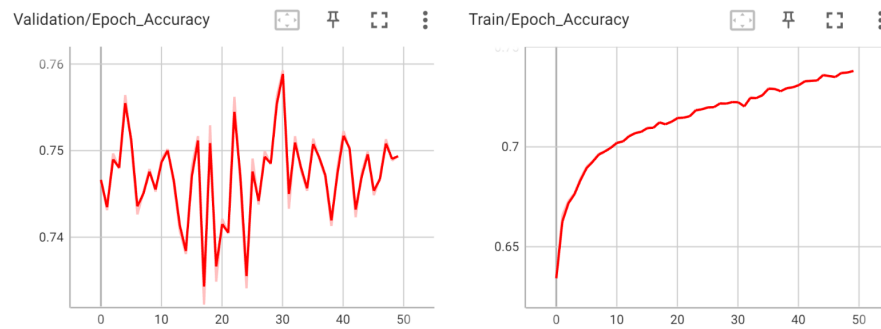


Figura 6.22 – *Overfitting* no treinamento

### 6.1.3 Organização em Série Temporal

Transformar os dados existentes em uma série temporal envolve a adição de uma nova dimensão, a temporal. Em termos práticos, isto é efetuado por meio de uma dimensão acrescentada aos tensores dos dados transformando eventos discretos em dados sequenciais.

Os dados disponibilizados pelo Censipam possuem identificadores rotulados como "id\_evento". Estes códigos agrupam diferentes detecções de fogo como pertencentes a um mesmo foco e é necessário para a transformação dos dados em uma série temporal. Em outras palavras, cada dados possui três dimensões: o ID representando o foco da queimada, a quantidade de detecções relativas a este mesmo fogo, e todos os dados de cada uma destas detecções.

No entanto, esta abordagem envolve o desafio de transformar estes dados em tensores. Todo o código utilizado é construído a partir do Pytorch (PASZKE et al., 2019), que utiliza tensores como tipo de dado. Tensores são matrizes multidimensionais, apenas uma forma de se organizar dados de forma lógica. Por exemplo, pode-se imaginar um tensor tridimensional como um cubo, tratando-se de um tensor cujo as três dimensões são iguais.

Em outras palavras, tensores são uma forma de se estruturar dados de forma regular, onde cada amostra de dados possui dimensões constantes e condizentes com as dimensões do tensor. No entanto, os dados em questão não são regulares pois são obtidos a partir de fenômenos aleatórios e caóticos, de queimadas sobre um território florestal de proporções continentais. Por isso é necessário o emprego de alguma técnica de tratamento de dados que possa regularizá-los, permitindo sua manipulação como tensores, como por exemplo, *Padding*, como explicado na seção 4.4.3.2.

### 6.1.4 Fluxograma do Desenvolvimento

Por fim, o funcionamento do código de treinamento pode ser então ilustrado por meio do seguinte fluxograma:

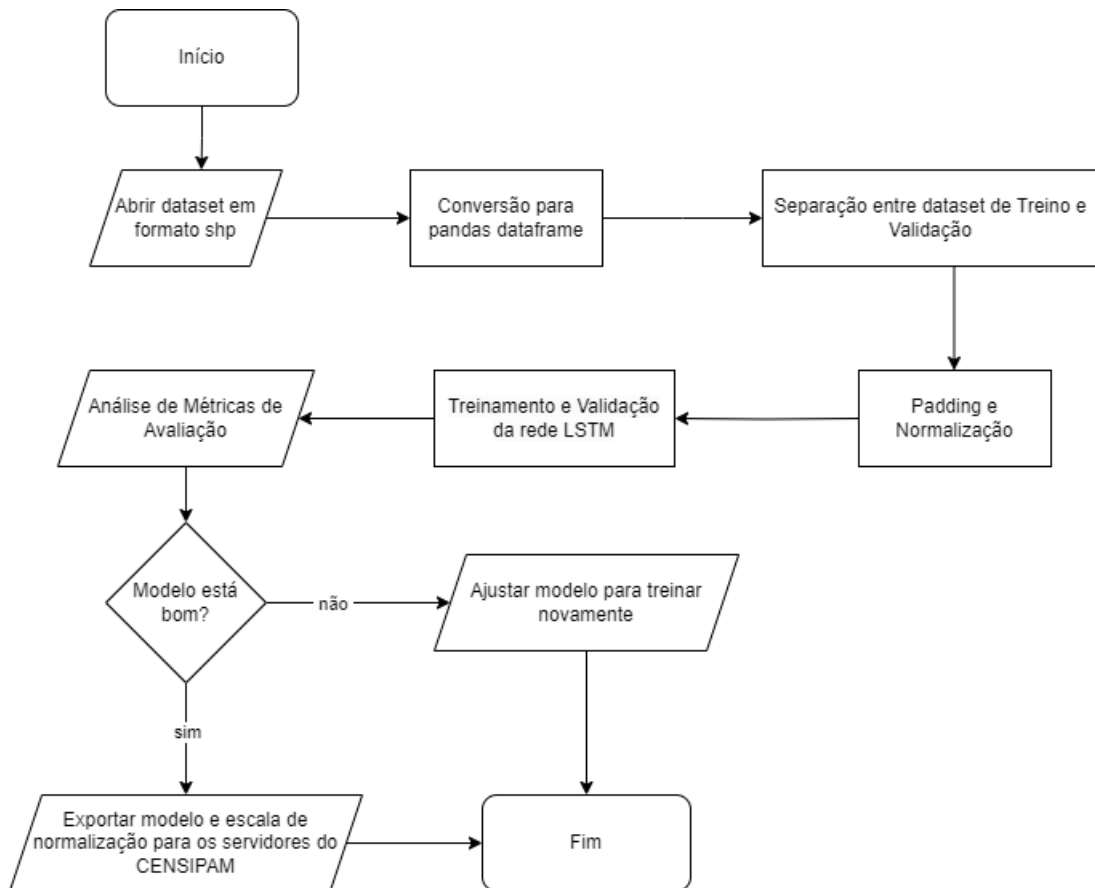


Figura 6.23 – Fluxograma do processo de treinamento, análise e produção do sistema em LSTM

### 6.1.5 Desempenho

No algoritmo de treinamento do modelo, separou-se o *dataset* entre validação e treinamento, sendo 80% destinado para o treinamento e 20% para a validação, separados cronologicamente. As métricas são obtidas de época em época e podem ser observadas por meio do módulo Tensorboard, disponível pelo Tensorflow (ABADI et al., 2015), ou pelo terminal do programa como uma tabela ao finalizar-se o treinamento, conforme visto na tabela 6.2.

As tabelas de métricas obtidas contém uma linha das métricas ponderadas, onde o peso é a quantidade de dados de cada classe.

### 6.1.6 Balanceamento

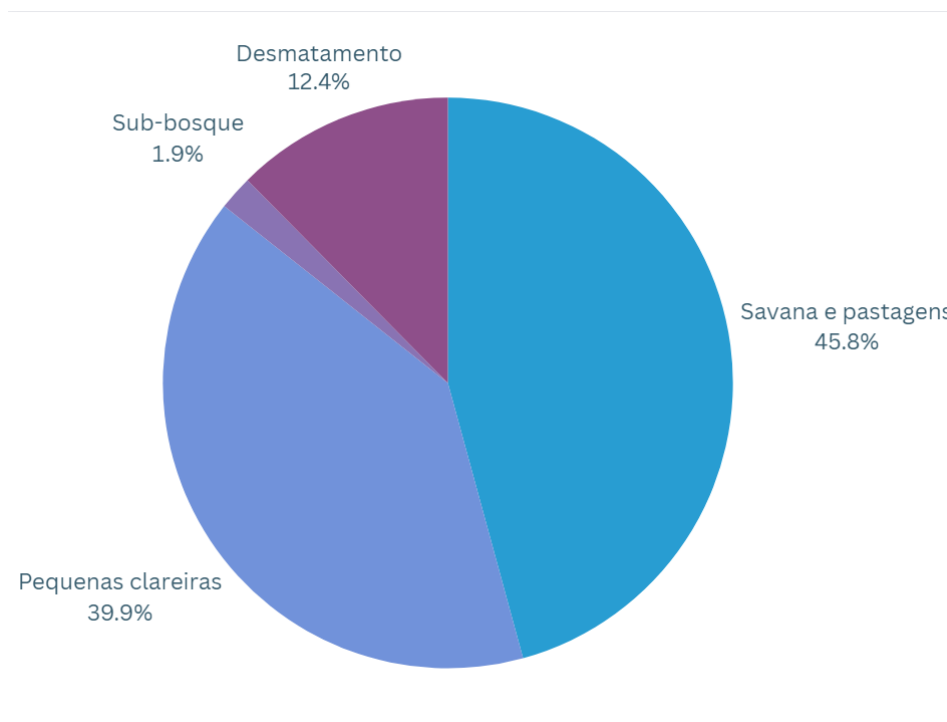
Como podemos ver na tabela 6.2, a rede em seu treinamento praticamente desconsiderou a presença da classe Sub-Bosque dentro do *dataset*, observando a imagem 6.24 podemos entender o porquê disso. A quantidade de eventos de sub-bosque é bastante inferior à das outras classes, em situações como essa, a rede aprende a não prever aquela classe, isso porque ela é tão inexpressiva que mesmo que ele a desconsidere, a acurácia da rede permanece alta. Como queremos que o trabalho seja capaz de fornecer todas os 4 tipos de

	Acurácia	Precisão	Recall	F1-Score
Savana e Pastagem	0.839	0.869	0.839	0.854
Pequenas Clareiras	0.821	0.763	0.821	0.791
Sub-Bosque	0	0	0	0
Desmatamento	0.536	0.523	0.534	0.529
Ponderada	0.779	0.768	0.779	0.773

Classe	Quantidade de eventos
Savana e Pastagem	73647
Pequenas Clareiras	64205
Sub-Bosque	3136
Desmatamento	19986

Tabela 6.2 – Resultados da Avaliação

fogo, é necessário fazer um balanceamento do *dataset* para aumentar a expressividade das classes menos presentes.

Figura 6.24 – Gráfico da presença de cada classe dentro do *Dataset*

O balanceamento pode ser feito utilizando de duas estratégias: remoção de dados e o chamado *data augmentation*, onde aumentamos a quantidade de um certo tipo de dado. Aqui, ambas foram utilizadas, como pode ser observado na tabela 6.3, diminuimos a quantidade de eventos para todas as classes exceto a 3, enquanto aumentamos a quantidade de eventos dela. Tal aumento foi feito repetindo os eventos dentro do banco de dados. A tabela também possui o desempenho do modelo para o treinamento com o banco de dados balanceado.



	Acurácia	Precisão	Recall	F1-Score
Savana e Pastagem	0.795	0.904	0.795	0.846
Pequenas Clareiras	0.804	0.772	0.804	0.788
Sub-Bosque	0.290	0.237	0.289	0.261
Desmatamento	0.620	0.485	0.619	0.544
Ponderada	0.767	0.786	0.767	0.774

Classe	Quantidade de eventos
Savana e Pastagem	25088
Pequenas Clareiras	23866
Sub-Bosque	9408
Desmatamento	10662

Tabela 6.3 – Resultados da Avaliação

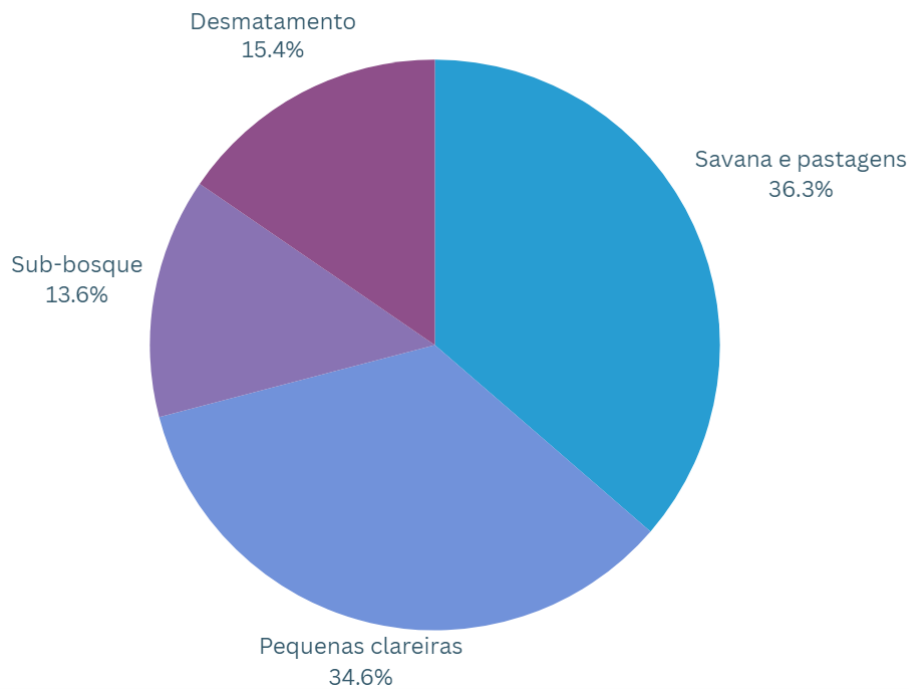


Figura 6.25 – Gráfico do Dataset após data augmentation e remoção aleatória de dados

O resultado do balanceamento é evidente nas métricas da classe Sub-Bosque, que saiu de 0% para todas as métricas para até 29% de acurácia e *Recall*. Apesar dessa melhora expressiva, a performance do modelo para esta classe permanece ruim. Além disso, não houve muita mudança nas métricas para as demais classes. É possível que não exista uma solução para a performance da classe Sub-Bosque, se tratando de uma limitação dos dados, porém isto não significa que novas possíveis soluções não possam ser testadas.

Podemos associar a baixa presença da classe de Sub-bosque devido a sua natureza, é um tipo de fogo que acontece no nível do solo abaixo da vegetação, tornando-o mais difícil de detectar do que os demais tipos. Além disso, ele é mais dificilmente diferenciável do que os outros, logo obter dados classificados como esse tipo é um desafio maior também.

## 7 Conclusão

O trabalho apresentado teve como objetivo o desenvolvimento de um sistema de classificação de queimadas no território da Amazônia Legal, por meio de um algoritmo de inteligência artificial para séries temporais. Ademais, também se foi desenvolvido com o propósito de eventualmente ser integrado ao sistema do Painel do Fogo do Censipam.

A revisão literária indicou que embora este tipo de abordagem não seja inovador no ramo do sensoriamento remoto em geral, é inovador no âmbito do sensoriamento dentro do território da Amazônia Legal.

O algoritmo desenvolvido é capaz de puxar dados oficiais e complementá-los, além de ser capaz de classificá-los com uma taxa de acurácia satisfatória.

Para trabalhos futuros, reserva-se a possibilidade de testar diferentes arquiteturas de redes neurais para séries temporais, como outras redes recorrentes ou redes convolucionais. Além disso, espera-se integrar o programa ao sistema oficial do Censipam como uma ferramenta classificadora de seus dados, podendo assim, serem utilizados por autoridades competentes no combate à incêndios florestais.

# Referências

A taxa consolidada de desmatamento para os nove estados da Amazônia Legal em 2022 foi de 11.594 km<sup>2</sup>. 2022. Disponível em: <[https://www.gov.br/inpe/pt-br/assuntos/ultimas-noticias/a-taxa-consolidada-de-desmatamento-para-os-nove-estados-da-amazonia-legal-em-2022-foi-de-11-594-km2-1/NotaTecnica\\_consolidado\\_PRODES\\_2022.pdf](https://www.gov.br/inpe/pt-br/assuntos/ultimas-noticias/a-taxa-consolidada-de-desmatamento-para-os-nove-estados-da-amazonia-legal-em-2022-foi-de-11-594-km2-1/NotaTecnica_consolidado_PRODES_2022.pdf)>. Citado nas pp. 9 e 13.

ABADI, M.; AGARWAL, A.; BARHAM, P.; BREVDO, E.; CHEN, Z.; CITRO, C.; CORRADO, G. S.; DAVIS, A.; DEAN, J.; DEVIN, M.; GHEMAWAT, S.; GOODFELLOW, I.; HARP, A.; IRVING, G.; ISARD, M.; JIA, Y.; JOZEFOWICZ, R.; KAISER, L.; KUDLUR, M.; LEVENBERG, J.; MANÉ, D.; MONGA, R.; MOORE, S.; MURRAY, D.; OLAH, C.; SCHUSTER, M.; SHLENS, J.; STEINER, B.; SUTSKEVER, I.; TALWAR, K.; TUCKER, P.; VANHOUCKE, V.; VASUDEVAN, V.; VIÉGAS, F.; VINYALS, O.; WARDEN, P.; WATTENBERG, M.; WICKE, M.; YU, Y.; ZHENG, X. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. Software available from tensorflow.org. Disponível em: <<https://www.tensorflow.org/>>. Citado na p. 46.

ABID, F. A survey of machine learning algorithms based forest fires prediction and detection systems. *Fire Technology*, v. 57, n. 2, p. 559–590, Mar 2021. ISSN 1572-8099. Disponível em: <<https://doi.org/10.1007/s10694-020-01056-z>>. Citado na p. 19.

AMAZONIA Legal - O que é. 2023. Disponível em: <<https://www.ibge.gov.br/geociencias/cartas-e-mapas/mapas-regionais/15819-amazonia-legal.html?=&t=o-que-e>>. Citado na p. 13.

ANDELA, N.; MORTON, D.; GIGLIO, L.; PAUGAM, R.; CHEN, Y.; HANTSON, S.; WERF, G.; RANDERSON, J. The global fire atlas of individual fire size, duration, speed, and direction. *Earth System Science Data Discussions*, p. 1–28, 09 2018. Citado na p. 15.

ARRUDA, V. L.; PIONTEKOWSKI, V. J.; ALENCAR, A.; PEREIRA, R. S.; MATRICARDI, E. A. An alternative approach for mapping burn scars using Landsat imagery, Google Earth Engine, and deep learning in the Brazilian savanna. *Remote Sensing Applications: Society and Environment*, v. 22, p. 100472, 2021. ISSN 2352-9385. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S2352938521000082>>. Citado na p. 19.

ASSIS, L. F. F. G.; FERREIRA, K. R.; VINHAS, L.; MAURANO, L.; ALMEIDA, C.; CARVALHO, A.; RODRIGUES, J.; MACIEL, A.; CAMARGO, C. Terrabrasilis: A spatial data analytics infrastructure for large-scale thematic mapping. *ISPRS International Journal of Geo-Information*, v. 8, n. 513, 2019. Citado na p. 38.

BARLOW, E. B. J.; CARMENTA, R.; FRANÇA, F. Clarifying Amazonia's burning crisis. *Glob Change Biol*, p. 319–321, 2020. Citado na p. 14.

BATISTA, A. C. Detecção de incêndios florestais por satélite. *Floresta* 34, Curitiba, Pr, 2004. Citado na p. 21.

BISHOP, C. *Neural networks for pattern recognition*. [S.l.]: Oxford University Press, USA, 1995. Citado na p. 24.

BOT, K.; BORGES, J. A systematic review of applications of machine learning techniques for wildfire management decision support. *Inventions*, 2022. Disponível em: <<https://www.mdpi.com/2411-5134/7/1/15>>. Citado na p. 16.

BRASILEIRO, S. F. *SICAR*. 2023. Acessado em: 23/11/2023. Disponível em: <<https://www.car.gov.br/publico/imoveis/index>>. Citado na p. 38.

BUITINCK, L.; LOUPPE, G.; BLONDEL, M.; PEDREGOSA, F.; MUELLER, A.; GRISEL, O.; NICULAE, V.; PRETTENHOFER, P.; GRAMFORT, A.; GROBLER, J.; LAYTON, R.; VANDERPLAS, J.; JOLY, A.; HOLT, B.; VAROQUAUX, G. API design for machine learning software: experiences from the scikit-learn project. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. [S.l.: s.n.], 2013. p. 108–122. Citado na p. 23.

CAUCHY, A. Methode generale pour la resolution des systemes d'equations simultanees. *C.R. Acad. Sci. Paris*, v. 25, p. 536–538, 1847. Disponível em: <<https://cir.nii.ac.jp/crid/1573387450834953216>>. Citado na p. 25.

CENSIPAM. *Painel do Fogo*. 2022. Acessado em: 26/10/2022. Disponível em: <<https://paineldofogo.sipam.gov.br>>. Citado nas pp. 14, 31 e 36.

CENTRO Gestor e Operacional do Sistema de Proteção da Amazônia - CENSIPAM. 2023. Disponível em: <<https://www.gov.br/censipam/pt-br>>. Citado na p. 31.

COORDENACAO-GERAL de Observação da Terra, INPE. 2019. Disponível em: <<http://www.obt.inpe.br/OBT/noticias-obt-inpe/memoria-historia-do-monitoramento-da-floresta-amazonica-projetos-prodes-deter-e-panamazonia>>. Citado nas pp. 9 e 17.

CÁRDENAS-MONTES, M. Sobreajuste - overfitting. 2015. Citado nas pp. 9 e 27.

DAMAGE, U.; BANDARANAYAKE, L.; WANASINGHE, R.; KOTTAHACHCHI, K.; JAYA-SANKA, B. Forest fire detection system using wireless sensor networks and machine learning. *Scientific Reports*, 2022. ISSN 2045-2322. Disponível em: <<https://www.nature.com/articles/s41598-021-03882-9>>. Citado na p. 19.

DIAS, B. S. S. Tipificação de incêndios florestais na amazônia legal através de aprendizado de máquina e aprendizado profundo. Trabalho de Conclusão de Curso não apresentado. 2023. Citado nas pp. 20, 38 e 42.

EU Horizon 2020 Work Programme 2018–2020 Cross-Cutting Activities. Luxemburg: Publications Office of the European Union, 2020. 70-73 p. Citado na p. 14.

FARIA, D. R. G. de; BERNINI, H.; ASSIS, J.; JUNIOR, H. V. Multicriteria severity indicator using remote sensing for forest firefighting dispatch in the brazilian amazon. 2022. Citado na p. 19.

GERS, F. A.; SCHMIDHUBER, J. Lstm recurrent networks learn simple context free and context sensitive languages. 2001. Citado na p. 29.

GLOBAL Fire Emission Database. 2022. Disponível em: <<https://globalfiredata.org/pages/amazon-dashboard/>>. Citado nas pp. 18 e 38.

GOODFELLOW, I. J.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. <<http://www.deeplearningbook.org>>. Citado na p. 24.

- GUIA do Usuário QGIS. [S.l.], 2023. Disponível em: <[https://docs.qgis.org/3.22/pt\\_BR/docs/user\\_manual/index.html](https://docs.qgis.org/3.22/pt_BR/docs/user_manual/index.html)>. Citado na p. 32.
- HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, v. 9, p. 1735–80, 12 1997. Citado na p. 29.
- INPE. 1992. Disponível em: <<http://www.dsr.inpe.br/laf/panamazonia/>>. Citado na p. 18.
- INPE. 2022. Disponível em: <[http://www.inpe.br/noticias/noticia.php?Cod\\_Noticia=5876](http://www.inpe.br/noticias/noticia.php?Cod_Noticia=5876)>. Citado na p. 14.
- INPE - Perguntas frequentes. 2023. Disponível em: <<http://www.inpe.br/faq/index.php?pai=6>>. Citado nas pp. 13 e 19.
- INPE. Banco de Dados de queimadas. 2004. Disponível em: <<http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/deter/deter>>. Citado nas pp. 18, 19 e 38.
- INPE. Banco de Dados de queimadas. 2022. Disponível em: <<http://www.inpe.br/queimadas/bdqueimadas>>. Citado na p. 14.
- JAIN, P.; COOGAN, S.; SUBRAMANIAN, S.; CROWLEY, M.; TAYLOR, S.; FLANNIGAN, M. A review of machine learning applications in wildfire science and management. p. 478–505, 2020. Citado na p. 21.
- LIMA, A.; SHIMABUKURO, Y.; ADAMI, M.; FREITAS, R.; LUIZ, E.; ARAGÃO, L.; FORMAGGIO, A.; RIVERA-LOMBARDI, R. Mapeamento de cicatrizes de queimadas na amazônia brasileira a partir da aplicação do modelo linear de mistura espectral em imagens do sensor MODIS. 01 2009. Citado na p. 19.
- LIN, T.-Y.; GOYAL, P.; GIRSHICK, R.; HE, K.; DOLLAR, P. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. Citado nas pp. 9 e 54.
- LSTM Cell. 2023. Disponível em: <<https://static.googleusercontent.com/media/research.google.com/pt-BR//pubs/archive/43905.pdf>>. Citado nas pp. 9, 30 e 31.
- MINSKY, M.; PAPERT, S. *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA, USA: MIT Press, 1969. Citado na p. 24.
- MODIS (or Moderate Resolution Imaging Spectroradiometer). NASA, 2023. Disponível em: <<https://modis.gsfc.nasa.gov/about/>>. Citado nas pp. 19 e 34.
- NAQA, I. E.; MURPHY, M. What is machine learning? *Machine Learning Techniques for Multimedia*, erlin/Heidelberg, Germany, p. 3–11, 2015. Citado na p. 21.
- PASZKE, A.; GROSS, S.; MASSA, F.; LERER, A.; BRADBURY, J.; CHANAN, G.; KILLEEN, T.; LIN, Z.; GIMELSHEIN, N.; ANTIGA, L.; DESMAISON, A.; KOPF, A.; YANG, E.; DEVITO, Z.; RAISON, M.; TEJANI, A.; CHILAMKURTHY, S.; STEINER, B.; FANG, L.; BAI, J.; CHINTALA, S. Pytorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019. p. 8024–8035. Disponível em: <<http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>>. Citado nas pp. 23 e 45.

---

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na p. 23.

PENHA, T. V. *Deteção de áreas queimadas na Amazônia utilizando imagens de média resolução espacial, técnicas de GEOBIA e mineração de dados*. Dissertação (Programa de Pós-Graduação do INPE em Sensoriamento Remoto) — Instituto Nacional de Pesquisas Espaciais (INPE), 2018. Disponível em: <<http://urlib.net/sid.inpe.br/mtc-m21c/2018/04.29.22.45>>. Citado na p. 13.

Projeto MapBiomas. *Mapeamento de cicatrizes de fogo no Brasil Coleção 1*. 2022. Disponível em: <<https://plataforma.brasil.mapbiomas.org>>. Citado nas pp. 19 e 38.

ROSSUM, G. V.; JR, F. L. D. *Python reference manual*. [S.l.]: Centrum voor Wiskunde en Informatica Amsterdam, 1995. Citado na p. 23.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data*, v. 6, n. 1, 2019. Citado na p. 44.

SINGH, D.; SINGH, B. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, v. 97, p. 105524, 2020. ISSN 1568-4946. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1568494619302947>>. Citado na p. 27.

TRAN, B.; HOA, N. Approach to detecting forest fire by image processing captured from IP cameras. *Advances in Image and Video Processing*, v. 5, 10 2017. Citado na p. 19.

VIEIRA, I.; SÁ, T.; PIEDADE, M. T. F. *Os incêndios criminosos na Amazônia devem ser contidos*. Oeco, 2021. Disponível em: <<https://oeco.org.br/analises/os-incendios-criminosos-na-amazonia-devem-ser-contidos/>>. Citado na p. 14.

# Apêndice A – Códigos

```

class Lstm(Module):
    def __init__(self, input_dim):
        super(Lstm, self).__init__()
        self.num_layers = NUM_LAYERS
        self.input_size = input_dim
        self.hidden_size = HIDDEN_SIZE
        self.lstm = LSTM(input_size=input_dim , hidden_size = HIDDEN_SIZE , num_layers= NUM_LAYERS, batch_first= True)
        self.fc = Linear(HIDDEN_SIZE, NUM_CLASSES)
        self.softmax = Softmax(dim=1)
        self.hn = torch.zeros(self.num_layers , BATCH_SIZE , self.hidden_size).to(DEVICE)
        self.cn = torch.zeros(self.num_layers , BATCH_SIZE , self.hidden_size).to(DEVICE)

    def forward(self,x, hn, cn):
        out , (hn,cn) = self.lstm(x , (hn,cn))
        final_out = self.fc(out)
        for i in range(len(final_out)):
            final_out[i] = self.softmax(final_out[i].clone())
        self.hn = hn
        self.cn = cn
        return final_out, hn, cn

    def init(self):
        h0 = torch.zeros(self.num_layers , BATCH_SIZE , self.hidden_size).to(DEVICE)
        c0 = torch.zeros(self.num_layers , BATCH_SIZE , self.hidden_size).to(DEVICE)
        return h0 , c0

```

Figura A.26 – Código do modelo LSTM

```

class FocalLoss(modules.loss._WeightedLoss):
    def __init__(self, weight=WEIGHTS, gamma=2,reduction='mean'):
        super(FocalLoss, self).__init__(weight,reduction=reduction)
        self.gamma = gamma
        self.weight = weight #weight parameter will act as the alpha parameter to balance class weights

    def forward(self, input, target):

        ce_loss = F.cross_entropy(input, target,reduction=self.reduction,weight=self.weight)
        pt = torch.exp(-ce_loss)
        focal_loss = ((1 - pt) ** self.gamma * ce_loss).mean()
        return focal_loss

```

Figura A.27 – Código do modelo de Loss Focal em pytorch (LIN et al., 2017)